ADVANCES IN GEOPHYSICS

Volume 8

H. E. Landsberg &
J. Van Mieghem

ADVANCES IN

GEOPHYSICS

VOLUME 8

This Page Intentionally Left Blank

Advances in
GEOPHYSICS

Edited by

H. E. LANDSBERG

U. S. Weather Bureau
Washington, D. C.

J. VAN MIEGHEM

Royal Belgian Meteorological Institute
Uccle, Belgium

VOLUME 8

1961

Academic Press • New York and London

ACADEMIC PRESS INC.
111 FIFTH AVENUE
NEW YORK 3, N. Y.

United Kingdom Edition
Published by
ACADEMIC PRESS INC. (LONDON) LTD.
17 OLD QUEEN STREET, LONDON S.W. 1

Library of Congress Catalog Card Number 52-12266

PRINTED IN THE UNITED STATES OF AMERICA

LIST OF CONTRIBUTORS

J. K. ANGELL, *U. S. Weather Bureau, Washington, D. C.*

R. GRANT ATHAY, *High Altitude Observatory, University of Colorado, Boulder, Colorado*

R. N. BRACEWELL, *Radioscience Laboratory, Stanford University, Stanford, California*

ALLAN COX, *U. S. Geological Survey, Department of the Interior, Menlo Park, California*

RICHARD R. DOELL, *U. S. Geological Survey, Department of the Interior, Menlo Park, California*

OWEN K. GARRIOTT, *Department of Electrical Engineering, Stanford University, Stanford, California*

CONSTANCE S. WARWICK, *National Bureau of Standards, Boulder, Colorado*

PIERRE WELANDER, *Swedish Research Council of the Natural Sciences and International Meteorological Institute, Stockholm, Sweden*[1]

[1] Present address: Meteorologic Institute, University of Stockholm, Stockholm, Sweden.

This Page Intentionally Left Blank

FOREWORD

The rapid flow of new literature has confronted scientists of all branches with a new problem: It is the difficulty of how to keep up with the new knowledge without becoming too narrowly specialized. In geophysics this has become a very acute dilemma. It is well recognized that advances in science are often based upon techniques and findings in neighboring or even distant fields. The flood of papers and reports in geophysics, while gratifying testimony to the rapid progress, has made it well-nigh impossible to read all the original literature.

Some see the solution to this problem in indexing and machine-recording of papers coupled with a rapid recall and reproduction system. This may perhaps serve adequately for encyclopedic and bibliographic work. If the indexing can be adequately done—and this is by no means assured—it may help in bringing together all items that bear on one subject. However, these systems will exercise no judgment on the quality of a contribution, nor will they separate lasting knowledge from the ephemeral. This can only be done by the discerning scientist. Machine selection will also be incapable of bringing to the fore relevant but not specifically so designated information. Again the connection will have to be noted by the intelligent mind.

Collections of review articles covering broad sectors of science, as we have attempted to present in these volumes, are still the best way of sifting new knowledge critically. They also serve as guides to the literature and as stimuli to thought. The present volume, we believe, reflects this tenet again.

Our plans for future volumes include, among other topics, articles on satellite geodesy, airborne dusts, avalanches, focal mechanism of earthquakes, tsunamis, atmospheric heat balance, atmospheric ozone.

We gratefully acknowledge the helpful suggestions received from our editorial advisory committee and the cooperative spirit of the authors of this volume.

H. E. LANDSBERG
J. VAN MIEGHEM

April, 1961

This Page Intentionally Left Blank

CONTENTS

Indices of Solar Activity

by R. Grant Athay and Constance S. Warwick

Satellite Studies of the Ionization in Space by Radio

by Owen K. Garriott and R. N. Bracewell

Use of Constant Level Balloons in Meteorology

by J. K. Angell

Paleomagnetism

BY RICHARD R. DOELL AND ALLAN COX

Numerical Prediction of Storm Surges

BY PIERRE WELANDER

INDICES OF SOLAR ACTIVITY

R. Grant Athay

High Altitude Observatory of the University of Colorado, Boulder, Colorado

and

Constance S. Warwick

National Bureau of Standards, Boulder, Colorado

1. Introduction

1.1. Scope of this Review

This review of solar physics and solar activity is intended as an aid to research workers in fields bordering astrophysics. More specifically, it is written as a guide through which geophysicists interested in solar-terrestrial research may be helped in their interpretation and use of the rather bewildering array of indices of solar activity. It is not intended, however, as a comprehensive treatise of either solar physics or solar activity in all its ramifications.

We have endeavored to describe the sun and its phenomena in such a way that the physical meaning and relative merits of the available indices will be clearly defined, insofar as this is possible. To the extent that we have succeeded, this review will make it possible for research workers in solar-terrestrial relationships to make an intelligent choice of the solar index appropriate to a given problem.

We have sharply distinguished between indirect indices derived from solar-induced phenomena of the terrestrial atmosphere and indices directly related to solar observations. Finally, we have endeavored to list the major published sources of solar indices of various types.

1.2. General Features of the Sun and Solar Radiation

The sun is rather typical of the multitudinous stars that make up our local Milky Way. Like all stars it is gaseous throughout, though in the interior the pressure is so great that the gases become degenerate and have high density. In stellar spectral class, the sun is a type G, main-sequence star, which marks it as a member of one of the most abundant classes.

The sun's relative proximity to earth is all that marks it as unusual among stars. The star nearest the earth, other than the sun, is more than 250,000 times more distant. As a consequence, the sun is the only star for which we are able to observe surface features. All other stars are simply points of light to even the largest telescopes.

TABLE I. Basic facts concerning the sun.

Distance at perihelion	$= 1.4710 \times 10^{13}$ cm
Distance at aphelion	$= 1.5210 \times 10^{13}$ cm
Radius	$= 6.96 \times 10^{10}$ cm
Apparent angular radius	$= 959.63''$.
Mean density	$= 1.410$ gm cm^{-3}
Surface gravity	$= 2.74 \times 10^{4}$ cm sec^{-2}
Total radiation	$= 3.76 \times 10^{33}$ erg sec^{-1}
Emittance at surface	$= 6.18 \times 10^{10}$ erg cm^{-2} sec^{-1}
Effective temperature	$= 5750$°K
Hydrogen content (mass)	$\approx 80\%$
Helium content (mass)	$\approx 19\%$
Other elements	$\approx 1\%$
Apparent rotation period at equator	$= 26.8$ days
Date of maximum northern solar latitude (7.25°) for sun-earth line	$\approx$ September 7
Date of maximum southern solar latitude (7.25°) for sun-earth line	$\approx$ March 6

The radiant energy from the sun, when considered *in toto*, is generally assumed to be practically constant. The basis for this assumption lies partly in the fact that the energy is generated by nuclear fusion in a small central core whose properties supposedly vary only over an extremely long time scale, and partly in the observation of the solar constant carried out over some 30 years by the Smithsonian Institution. These observations show no recognizable systematic variations as large as 0.17% [1], although C. G. Abbot [2] claims to find several periodicities in the data by rather dubious analytical methods. Johnson and Iriarte [3], at Lowell Observatory, have recently reported a large systematic increase in the solar constant amounting to 2% over the period January 1953 to July 1958. The solar constant is deduced in their work from the blue magnitudes of Uranus and Neptune, which are calibrated relative to background stars. This technique frees the observations from the effects of atmospheric absorption and is capable of high accuracy. However, it seems possible that the reported variations in planetary magnitudes could result from changes other than the solar constant. More data are needed before the reality of these suspected changes of the solar constant can be established. So far as the energy generation is concerned, calculations on the rate of fusion of hydrogen into helium suggest that the energy output will not change appreciably over a period of 10^9 years.

In view of the foregoing arguments, the constancy of the solar energy output must not be regarded as firmly established. There are, in the universe, a nontrivial percentage of variable stars whose energy output does vary considerably over short periods of time. Furthermore, the measurements of the solar constant represent only the energy in the electromagnetic spectrum that falls within the wavelengths that penetrate the atmosphere. The intensity of the ultraviolet and x-ray spectrum of the sun is known to vary with solar activity but by an unknown amount. Similarly, the sun is known to emit varying amounts of its energy in the form of corpuscular radiation, and, in this case, both the amount of energy radiated and the amount by which it varies are unknown. We can, however, with a reasonable degree of assurance, place limits on the energy radiated in these relatively unobserved regions of the spectrum.

The solar constant, which includes the electromagnetic radiations penetrating the atmosphere plus corrections for atmospheric absorption, is 2.00 cal cm^{-2} min^{-1} = 1.39×10^6 ergs cm^{-2} sec^{-1}. This value is uncertain to perhaps 2 or 3% [4]. The visible spectrum of the sun is, broadly speaking, consistent with blackbody radiation at a temperature of about 6000°C. In detail, however, the spectrum shows a great deal of structure in the form of the Fraunhofer absorption lines and in the form of systematic changes in the characteristic temperature with wavelength. In general, the near ultraviolet spectrum is characterized by somewhat lower temperatures than the red end of the spectrum.

While much can be said with regard to the Fraunhofer spectrum, most of the discussion would be irrelevant to our purpose in this chapter. Suffice it to say, at this point, that the simple interpretation of the Fraunhofer lines as being the absorption spectrum of a relatively cool "reversing layer" overlying the layers where the continuum is formed is a great oversimplification of the actual problem. We shall have more to say about this later in the chapter.

Although the energy in the far ultraviolet and x-ray spectrum is incompletely observed and is variable, existing observations of the line spectrum between 2000 and 84 A at a rocket altitude of just over 200 km give about 20 ergs cm^{-2} sec^{-1} [5], about one-third of which is in the Lyman-α line of hydrogen. The atmospheric absorption at 200 km is still appreciable for many of the lines observed, but its exact value is unknown. The most important correction is for the 304-A line of He II, whose observed intensity is considerably less than the Lyman-α intensity at 200 km, but whose intensity at the top of the earth's atmosphere may exceed the Lyman-α intensity. In any case, it seems unlikely that the total intensity in this spectral range at the top of the earth's atmosphere will normally exceed 100 ergs cm^{-2} sec^{-1}. X-rays at wavelengths between about 50 and 3 A have been observed with intensities of the order of 1 erg cm^{-2} sec^{-1} [6]. Thus, the energy radiated in the ultraviolet and x-ray regions of the spectrum normally amount to less than 10^{-4} of the solar constant. Relatively large variations are known to occur, but we are as yet unable to state with any degree of certainty the range of these variations. It seems unlikely, however, that this energy would exceed 10^{-3} of the solar constant even under extreme conditions, although this is mere speculation based upon a rather vague understanding of the physics of the problem.

The corpuscular radiation from the sun is even more uncertain than the x-ray and ultraviolet radiation. The "solar wind" proposed by Parker [7] seems, in the light of empirical evidence, to be an upper limit to the normal corpuscular radiation from the quiet sun. At the orbit of earth, in Parker's theory, there are about 10^2 protons cm^{-3} moving with a velocity of about 10^8 cm sec^{-1}. The flow of energy in this solar wind is therefore of the order of 100 ergs cm^{-2} sec^{-1} or about 10^{-4} of the solar constant. A more realistic evaluation of the corpuscular energy can perhaps be arrived at from spicule data (Section 2.3). At a height of 5000 km in the chromosphere, the total outward flow of energy represented by the upward motion of spicules is of the order of 10^{27} ergs sec^{-1}, which amounts to about 10^{-6} of the solar constant. Variations in the corpuscular radiation by relatively large factors are known to occur, but again we can only speculate about the extreme ranges. Satellite observations in the outer Van Allen radiation belts are, at the time of this writing, just beginning to indicate the true nature of the corpuscular radiation.

In summary, present evidence suggests that the combined energies radiated by the sun in the ultraviolet, x-ray, and corpuscular spectra probably never exceeds 10^{-2} of the solar constant and may well have an upper limit of 10^{-3} of the solar constant. It should be noted, however, that the energies relative to the solar constant may be an entirely misleading comparison. Most of the total solar energy incident upon the atmosphere is either reflected back into outer space or passes through the atmosphere with but little attenuation. On the other hand, the ultraviolet, x-ray, and corpuscular radiations referred to above are completely absorbed by the atmosphere itself. Furthermore, the corpuscular radiation is selectively channeled into the polar regions, and "locally," in these regions, may represent a major component of energy. The Van Allen radiation belts provide an additional factor to be considered. It is not inconceivable, for example, that solar particles are stored in the outer belts and later released to the atmosphere in large quantity through perturbation induced by new streams of solar particles. Future observations of the particle flux in outer space will soon answer many of these questions.

2. Features of the Undisturbed Sun

2.1. General Concepts

Observational solar astronomy had its serious origins in 1611, when Fabricius, Scheiner, and Galileo discovered sunspots, each independently of the other. From these earliest times it has been apparent that the sun's normal background state undergoes frequent perturbations localized in both time and space. The indices of solar activity in use today derive from direct or indirect measurements of this activity. However, it is historically apparent that an entirely unperturbed sun is rare, and that disturbances of one degree or another are almost always present. Nonetheless, it turns out to be highly convenient to discuss the ideally undisturbed sun, and to delineate its stratified layers, as a point of departure for study of solar activity.

In this section we review, briefly and in general terms, the observed features of the sun. The section is intended as a background to the discussion of indices of solar activity, since it is principally by observation of these features that the direct indices of solar activity are derived. Figure 1 illustrates the different layers of the solar atmosphere as distinguished by their physical properties. These layers provide a convenient framework for reference.

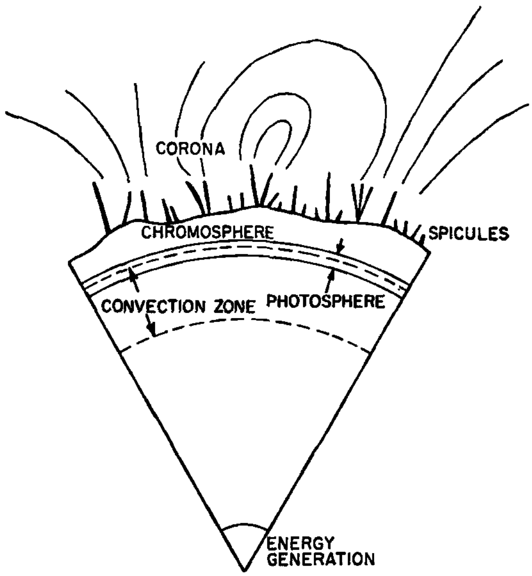


FIG. 1. A schematic representation of solar structure with an exaggerated photosphere and chromosphere. If drawn to scale, the thickness of the photosphere would be about 0.001 radii and of the chromosphere about 0.01 radii. Similarly, the spicule diameter would be about the same as the photospheric thickness.

Before discussing the different regions of the solar atmosphere, it is useful to define the astronomical usage of "optical depths." In an atmosphere with density ρ and absorption coefficient κ, an element of optical thickness $d\tau$ is defined by

$$d\tau = -\kappa\rho dx \tag{2.1}$$

where dx is an element of geometrical thickness. Usually, x is measured along a radial direction in the atmosphere. For our purposes, however, we shall define x as the coordinate in the line of sight, measured positively outward in the sun, and we shall designate the radial direction by subscript "r" and the tangential direction by subscript "t." Because of the negative sign in equation (2.1), the optical thickness increases inward. Inasmuch as we must discuss the properties of the solar atmosphere in terms of its radiation and opacities, the optical thickness is a more convenient tool than geometrical thickness.

Since the absorption coefficient is generally a function of wavelength, so also is τ. Thus, a given geometrical depth does not have a unique optical depth. Unless otherwise specified, our subsequent usage of τ will apply to a wavelength of 5000 A.

Radiation of intensity I at a point in the atmosphere where the optical depth is τ will emerge from the atmosphere with an intensity I_0 given by

(2.2) $$I_0 = Ie^{-\tau}$$

2.2. The Photosphere and Granulation

The *photosphere* of the sun is, by definition, the layer that produces the radiation in the visual region of the spectrum. It is, in other words, the visible layer of the solar atmosphere as observed by the eye without special optical devices.

The sun's edge (limb) when viewed in white light exhibits an apparently sharp boundary. From this we deduce that through a depth of some few hundred kilometers the gaseous layers of the sun change from being opaque to being virtually transparent, when viewed tangentially. The sun's limb, as viewed at a wavelength of 5000 A is, for purposes of an explicit definition, the top of the photosphere. More specifically, the top of the photosphere is defined by $\tau_t = 1$. By knowing the scale height for τ and the radius of the sun, we may compute the corresponding value of τ_r. The scale height is of the order of 60 km, which yields $\tau_r = 3 \times 10^{-3}$. This latter value is uncertain, however, because the scale height is not accurately known in this region of the solar atmosphere and because it depends upon the implicit assumption of spherical symmetry. As we shall note later, this assumption may be a rather poor one.

The base of the photosphere is usually rather loosely defined by $\tau_r \approx 10$, since only a small fraction of the radiation originating at greater τ escapes without being reabsorbed. The total thickness of the photosphere, in this definition, is of the order of 500 km. At the mean distance of the sun, this corresponds to less than 1 sec of arc (725 km).

The material density at the top of the photosphere is of the order of 10^{-8} gm cm^{-3}, and the gas pressure is about 6×10^{-3} atm. At the base of the photosphere, both these values are increased by about a factor of 50. By comparison with the lower terrestrial atmosphere, therefore, the solar photosphere is extremely tenuous. Its thickness, relative to the radius of the sun, is also extremely small.

Since the great bulk of energy in the solar constant is contained in the visible spectrum, the photosphere is the region of the sun that determines the solar constant. The energy, however, is not generated in these layers but in the much deeper central core of the sun. Its diffusion outward from the central core, where it is generated by thermonuclear fusion of hydrogen into helium, is by radiative processes. Each quantum of energy follows a random walk consisting of the short steps taken by photons as they are successively emitted and reabsorbed. The journey to the photosphere requires a great many years and traverses a wide range of physical conditions beginning at the core where the temperature and density are of the order of 10^7°K and 100 gm/cm^3 and ending in the tenuous, relatively cool photosphere. In its outward journey, the radiation is continually readjusting to its local environment following the laws of thermodynamic equilibrium. Thus, the photospheric radiation is largely characteristic of the thermodynamic state of the photosphere. However, it is equally true that the thermodynamic state of the photosphere, taken as a whole, must be such that it can dispose of the vast quantity of radiant energy fed to it from the solar interior.

Beginning in the lowest layer of the photosphere and extending downward some 10^5 km is a zone of convective instability induced primarily by the increased opacity resulting from the ionization of hydrogen. The total flux of energy transferred by convection is believed to be small relative to that transferred by radiation, but the problem has not been satisfactorily solved. It is probably this convection zone that gives rise, in one way or another, to the varied manifestations of solar activity.

Within the photosphere itself, temperature and density decrease outward. Since the tangential optical thickness of the photosphere exceeds the radial optical thickness, the effective depth at which the escaping radiation originates increases steadily from the limb to the center of the solar disc; i.e., we see deeper in the solar atmosphere at the center of the disc than at the limb. As a result, the radiation seen at the center of the disc is brighter and characterized by higher temperature than that seen at the limb. This so-called "limb darkening" provides a powerful tool for studying the temperature structure of the solar photosphere. The amount of limb darkening varies with wavelength, being more pronounced in the ultraviolet than in the red end of the spectrum. It also varies sharply depending upon whether it is observed in the continuum or in the Fraunhofer lines, and it varies from line to line. Some lines, in fact, show limb brightening rather than limb darkening. Our current knowledge of the temperature structure is derived almost entirely from the continuum limb darkening.

At the wavelengths of the Fraunhofer absorption lines, the absorption coefficient is greater than in the adjacent continuum. Thus, the radiation in the lines originates at a greater geometrical height in the solar atmosphere than the radiation in the continuum. The fainter lines, as observed near the center of the disc, are formed well within the bounds of the photosphere. The cores of the stronger lines, however, are formed above the photosphere in the still more tenuous layers of the *chromosphere* (see Section 2.3). The now outmoded concept of a reversing layer lying just above the photosphere was based on an attempt to account for the Fraunhofer lines in terms of Kirchhoff's radiation law, which states, in effect, that a cool gas lying in front of a relatively hot radiating surface will exhibit an absorption line spectrum. In thermodynamic equilibrium, the converse statement is also true, *viz.*, an absorption line spectrum requires a relatively cool layer of gas between the observer and the continuum source. In the absence of thermodynamic equilibrium, this latter statement is not necessarily true. In fact, for such strong lines as the Balmer-line of hydrogen and the H and K lines of Ca II (singly ionized calcium), where the radiation temperature at the line center drops below 4000°K, it is known to be grossly incorrect. In such cases, the radiation temperature cannot be identified with, say, the temperature characterizing the kinetic energies of the particles [8]. Attempts to extend the atmospheric model to lower optical depths through the use of the stronger Fraunhofer lines have been unsuccessful for this reason.

A telescopic view of the photosphere, in integrated light, through suitable eye-protecting devices, reveals a wealth of minute detail. Most of these features, such as the sunspots and faculae discussed below, are concentrated into reasonably well-defined regions of the sun, the active regions. The small photospheric granules, on the other hand, are found in all portions of the solar disc in roughly uniform distribution. Photographs taken by Schwarzschild [9] from high altitude balloons show the granules as fine, irregular bright surfaces of 300 to 1800 km diam., separated by a network of dark lanes that are often very narrow (Fig. 2). Schwarzschild [9] points out that these bright cells have irregular, but often polygonal shapes, and resemble in character the nonstationary convection cells that have been explored by Siedentopf [10] in laboratory studies. Such convection regimes are intermediate between random flow and the laminar flow that gives rise to regular polygonal Bénard cells of relatively fixed size. The granules are presumably the visual manifestation of the upper reaches of the convection zones lying mainly below the photosphere.

The root-mean-square brightness fluctuation between granules and dark lanes is about 4.6%, corresponding to a temperature fluctuation of about ± 60°K. It must be remembered, however, that this temperature fluctuation corresponds to a surface of constant optical depth and not constant geometrical depth. The opacity in the photosphere increases rapidly with increasing temperature. The granules, therefore, are more likely elevated above the dark lanes, and since the temperature decreases outward the temperature fluctuations on a surface of constant geometrical depth probably exceed ± 60°K by quite a large amount. This roughness of the temperature contours may greatly complicate the interpretation of limb-darkening data and has never been properly taken into account.

The granular structure exhibits no known systematic changes with time. It evidently characterizes the normal undisturbed solar surface at all latitudes and longitudes. There may be small perturbations in its average

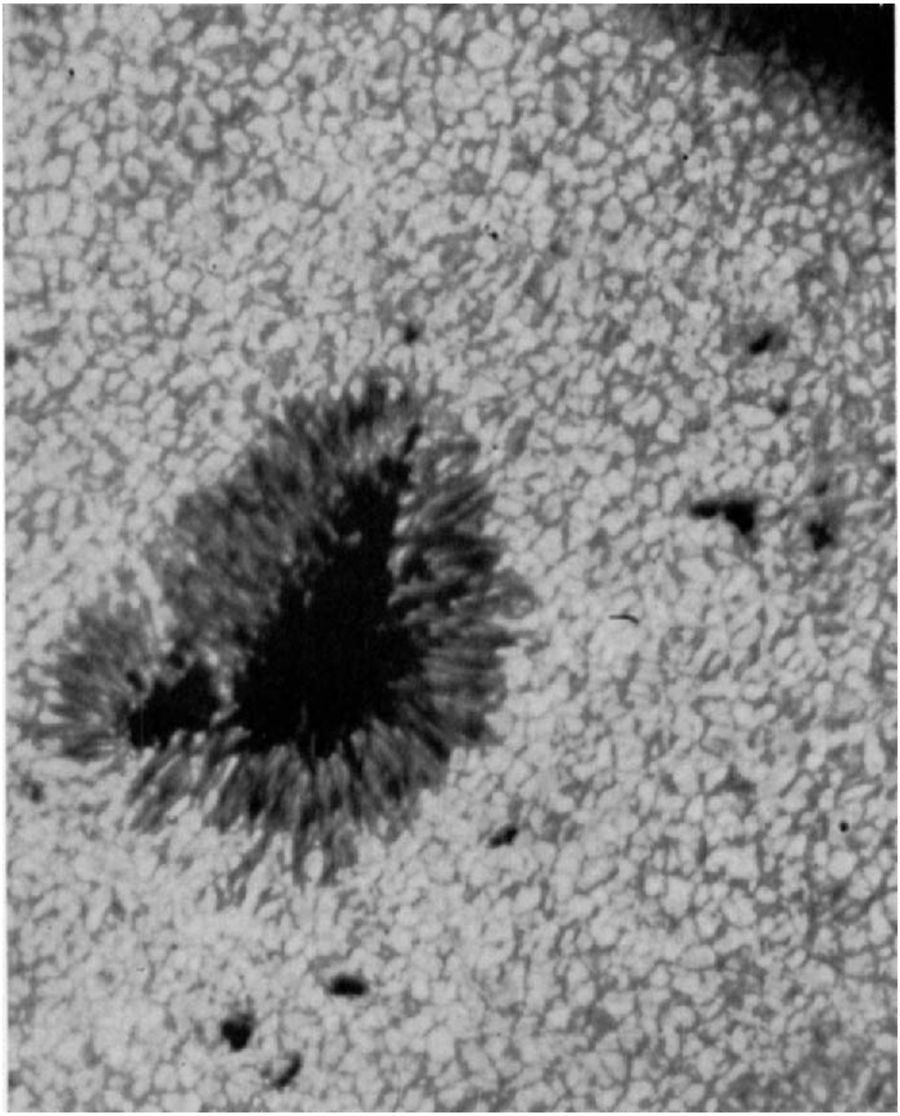


FIG. 2. A high altitude balloon photograph of solar granulation, including one moderate sized sunspot and several small spots (courtesy Martin Schwarzschild, Project Stratoscope of Princeton University, sponsored by the Office of Naval Research and the National Science Foundation).

scale very near highly disturbed regions of the solar disc, but observations of adequate quality to reveal such differences are just now becoming possible as a result of the development of large balloon-borne telescopes.

2.3. The Chromosphere and Spicules

2.3.1. Observed Features. The photospheric layers are defined simply in terms of their opacity rather than any distinguishing thermodynamic or spectral properties. There is, however, near the upper boundary of the photosphere, a marked change in both the thermodynamic and spectral properties of the solar atmosphere. During times of total solar eclipses, just as the moon occults the last remaining portion of the solar disc the continuum brightness drops suddenly and bright emission lines flash into view in the so-called "flash spectrum." In a rough sort of way, this change in the spectrum can be illustrated as shown schematically in Fig. 3. The point at which the brightness in the continuum equals the brightness at the center of the line marks the height in the solar atmosphere where the character of the spectrum changes from absorption lines to emission lines. The actual height where this occurs varies from line to line but by relatively small amounts.

In detail, the lines of the chromospheric spectrum differ considerably from the lines of the photospheric spectrum. Relatively faint photospheric lines may be bright in the chromospheric spectrum, and the stronger photospheric lines may weaken in the chromosphere. In particular, several lines of neutral and singly ionized helium are prominent in the chromospheric spectrum but absent from the photospheric spectrum. The general trend is toward higher states of excitation in the chromosphere.

The flash spectrum was first observed during the total eclipse of 1870.

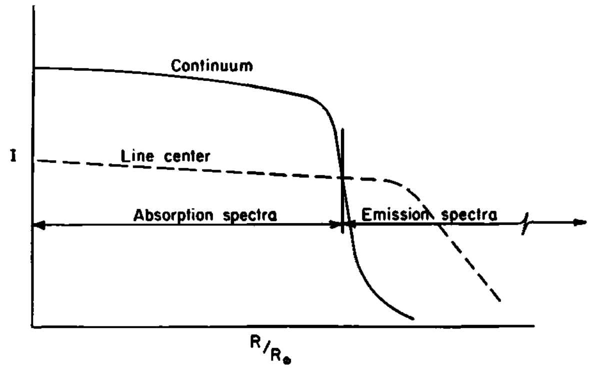


FIG. 3. A schematic illustration of the change from an absorption line spectrum on the solar disc to an emission line spectrum at the limb.

The strongest line in the visual region of the spectrum is the red Balmer-α line of hydrogen, which gives a slight reddish color to the integrated light and lends meaning to the name *chromosphere* used to denote this region of the solar atmosphere. Originally, the name chromosphere was applied, all inclusively, to those regions of the solar atmosphere showing the Balmer-α line in emission. In this definition, the base of the chromosphere does not necessarily coincide with the top of the photosphere. On the other hand, the great solar prominences suspended in the corona high above the photosphere fall into the category of chromospheric phenomena regardless of their origin.

As our understanding of the sun advanced it became apparent that a finer distinction of these phenomena was desirable. Thus, the definition of the chromosphere has evolved considerably. We now define the base of the chromosphere as being synonymous with the top of the photosphere; i.e., we define it in terms of the opacity through the condition $\tau_t = 1$. The top of the chromosphere is defined somewhat vaguely by the upper boundary of the Balmer-α emission exclusive of prominences. Although this definition is not entirely unambiguous, it is found to be useful in a practical sense.

In terms of geometrical height, the top of the chromosphere is highly irregular and transitory. Early observers described it as being similar in appearance to a "burning prairie." Over an area something in excess of 99% of the total surface area, the top of the chromosphere lies between 3000–5000 km above its base. This range in heights represents partly an uncertainty in the actual values and partly a real variation from point to point on the sun. Over the remaining area of 1% or less the chromosphere extends to greater heights in the form of fine, bristling jets of matter ejecting upward into the corona (see Fig. 4). These fine jets, known as spicules [11], are more or less randomly distributed over the sun and undergo rapid evolution. Individual spicules have lifetimes of the order of 3–5 min, during which they elongate at a mean velocity of 25–30 km/sec to heights often in excess of 12,000 km above the base of the chromosphere. Their diameters

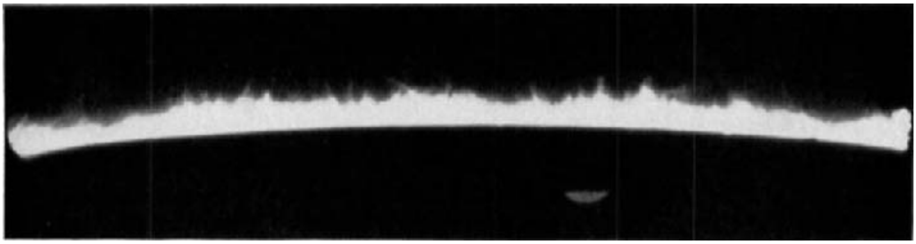


FIG. 4. Chromospheric spicules photographed in the Hα line (courtesy R. B. Dunn, Sacramento Peak Observatory, Geophysics Research Directorate, AFCRC, Sunspot, New Mexico).

are typically less than 1 sec of arc ($\approx$ 725 km), although their exact size is still uncertain.

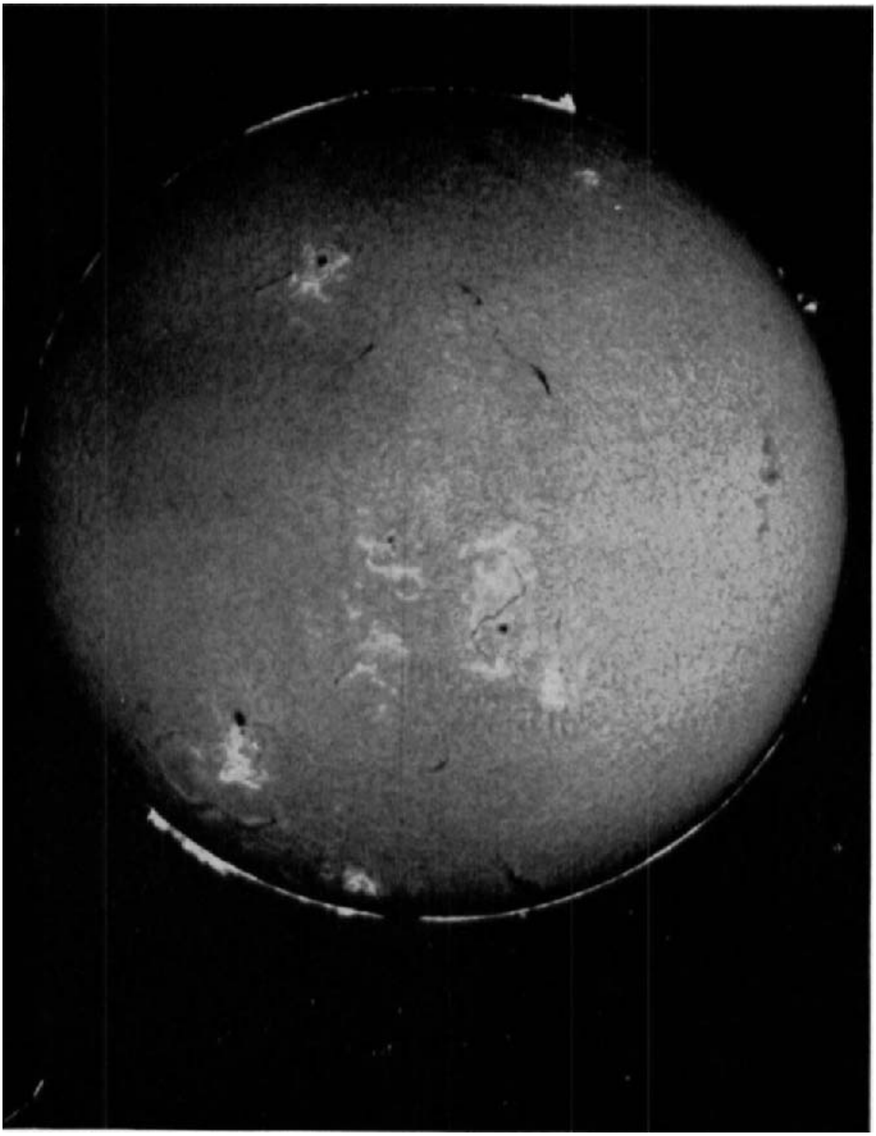
When viewed at the limb, spicules appear to cover a large portion of the surface, but this is an illusion created by the long path length over which spicules may be observed. For example, all spicules extending 10,000 km or more above the normal upper boundary of the chromosphere and lying within 100,000 km of the true limb show in projection at the limb. Although they are often classed as a type of prominence, spicules are an integral part of the chromosphere. In a 24-hr period, more than half of the chromospheric surface will have been disrupted by spicules, assuming that they are randomly located.

The chromosphere at the limb is best observed at eclipse. However, Lyot's invention of the coronagraph for producing artificial eclipses greatly improved the extra-eclipse observations. Much valuable information is coming from coronagraph observations, particularly in the study of spicules [11–13].

The chromosphere may also be observed on the disc of the sun by isolating the light in the cones of the strongest Fraunhofer lines. This is commonly accomplished with a spectroheliograph, which employs a spectrograph to isolate the line, or with a birefringent filter designed to pass only the wavelengths of the line in question. The Balmer-α line of hydrogen and the H and K lines of Ca II are especially suitable for such observations. Figure 5 shows the chromosphere as observed in the Balmer-α and K lines. The patchy nature of the surface is further evidence of the irregular structure. Differences in the structure seen in the two lines reflect differences in the heights at which the lines are formed and differences in the physical processes involved in line formation. In spite of the great wealth of information in such observations, little attempt has been made to exploit them.

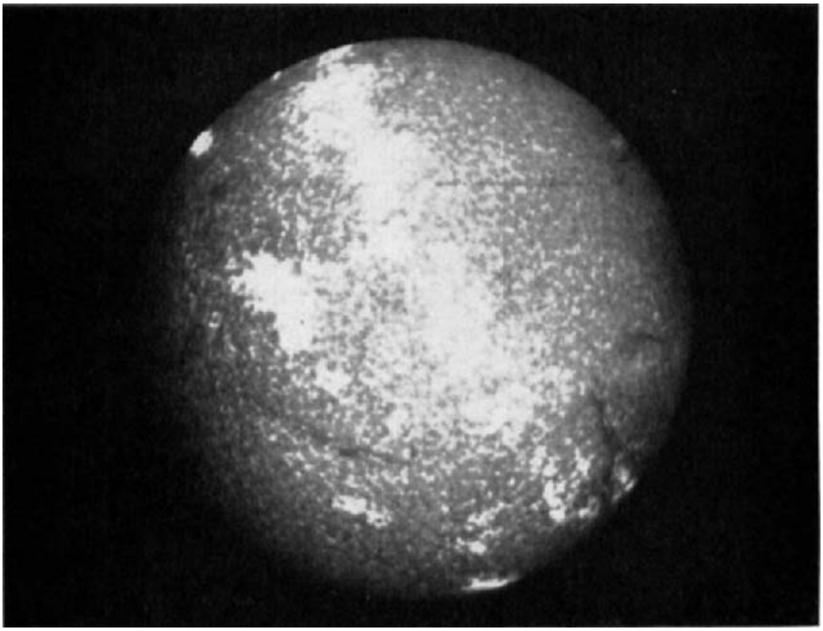
In addition to the general mottling of the background, a variety of large-scale features are observed in the chromosphere on the disc. Sunspots show only inconspicuously, but sunspot areas are surrounded by large bright features known as plages, less commonly called "chromospheric faculae." Flares (Section 3.6) occur principally in plage areas and are properly classed as chromospheric phenomena. The larger dark ribbonlike features are prominences elevated above the chromosphere. Each of these features will be discussed in more detail later in this chapter.

2.3.2. Physical Processes. In order to emphasize the important role of the chromosphere in the over-all phenomenon of the solar atmosphere, we return to a few comments on the photosphere and the anticipated structure of the higher layers. To a very good degree of approximation, the photospheric layers are in hydrostatic equilibrium. So far as the continuous spectrum is concerned, the photosphere is also in thermodynamic and



(a)

Fig. 5. The solar disc photographed in (a) the Hα line (courtesy Sacramento Peak Observatory, Geophysics Research Directorate, AFCRC, Sunspot, New Mexico), (b) the K line (Mount Wilson Observatory). The bright features are plages and the long dark features are dark filaments. Sunspots are also visible in the Hα photograph. [See facing page for part (b) of figure.]



(b)

radiative equilibrium. These conditions (assuming that they continue to be valid when combined with the known temperature, density, and chemical composition) should determine the structure of the overlying layers. However, in such a picture, the chromosphere would be replaced by a shallow layer less than 1000 km in thickness, and no such thing as spicules would exist. Furthermore, there would be no corona.

It is quite evident then that a fundamental difference from the photospheric structure sets in near the base of the chromosphere. The nature of this difference is only now coming to light after many years of investigation. The first widely accepted attempts to account for the anomalous extent of the chromosphere retained the assumptions of thermodynamic and radiative equilibrium and abandoned hydrostatic equilibrium in favor of a system of "turbulent" cells, which exerted a "pressure" and distended the atmosphere. Such a picture is *still* widely accepted in principle [14, 15], though not in detail. The attraction offered by this picture is simply that an alternative picture, which abandons thermodynamic and radiative equilibrium, greatly complicates the problems of interpreting observational data.

The difficulty with the above picture of the chromosphere, which has not been properly recognized by those who advocate it, is that the turbulent motions must be supersonic in order to produce the required support and must, therefore, dissipate energy to their surroundings. Once this is realized, then one cannot simultaneously drop the assumption of hydrostatic equilibrium and retain the assumption of radiative equilibrium with a support mechanism that is coupled to the thermodynamic properties of the atmosphere.

Recently, it has been shown quite conclusively from both theoretical and empirical models that the lowest layers of the chromosphere are, in fact, in hydrostatic equilibrium but depart rather sharply from both thermodynamic and radiative equilibrium [8]. At some point very near the upper boundary of the photosphere, the temperature passes through a minimum and thereafter increases outward through a series of sharp transitions. The thermal structure is determined by a nonradiative energy supply, as yet unspecified, but presumably related to the convective transport of energy in the deeper layers of the photosphere, and a radiative loss of energy to outer space.

The quantity of energy that a given atomic species is capable of radiating rises through a series of maxima as the degree of ionization is increased as a result of increased temperature, as illustrated in Fig. 6.

A maximum is reached for each stage of ionization, and, in general, each successive maximum is higher than the preceding one. A sharp rise in temperature occurs whenever the particular ion providing the principal radiation loss reaches its maximum radiative efficiency. At this point, the temperature field becomes unstable against small perturbations unless the principal energy loss shifts to a new ion. In the lowest layers of the chromosphere, hydrogen provides the energy sink while the temperature climbs to about 11,000–13,000°K. At about 1500 km above the base of the chromosphere, the energy sink shifts to He II at a temperature of 40,000–50,000°K. The top of the chromosphere occurs when the temperature reaches ~70,-

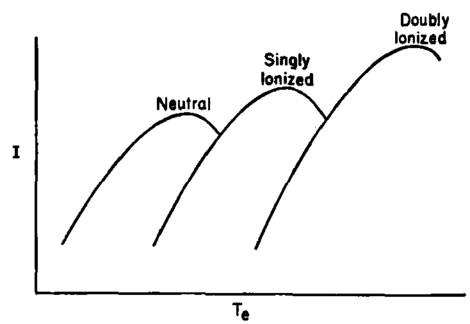


FIG. 6. Radiant energy as a function of temperature from an idealized, monatomic, optically thin gas at constant density.

000°K and the energy sink shifts from He II to multiply ionized metals at a temperature of about 2×10^{6}°K.

The temperature structure is directly related to the chemical composition of the solar gases. He II, and not He I, follows hydrogen as the dominant source of radiative energy loss because of the lower abundance of helium relative to hydrogen. Similarly, the low relative abundance of the heavy elements requires that they be multiply ionized and reach temperatures of the order of 2×10^{6}°K before their radiation can compete with the radiation of more abundant hydrogen and helium.

The height range near 1000–1500 km in the chromosphere is of particular interest. At these heights and above, the chromosphere departs strongly from spherical symmetry, and over most of the chromosphere the principal energy sink shifts from hydrogen to He II. It is apparently near these heights where the spicules are formed and accelerated outward [8]. The more or less random distribution of spicules over the surface of the sun suggests a basic instability in the chromospheric plasma. This region of the solar atmosphere may indeed hold the clue to a wide variety of solar phenomena including the corona itself, as well as many of the transient features of solar activity. As an illustration of this point, the mass flux flowing outward in spicules at the 5000-km level is sufficient to replace the entire corona in about 3 hr time and is an order of magnitude larger than that required to produce the so-called "solar wind" [7]. The question of a return flow of spicule matter to the chromosphere, or even an eventual escape from the sun, is completely unsettled.

A further point of interest in these lower levels of the chromosphere is that they apparently require a much larger supply of nonradiative energy than the much hotter and, in some ways, more spectacular corona, although reliable quantitive estimates of the energy requirements are still lacking. In any case, our attention has shifted somewhat from the corona to the chromosphere as the focal point of solar activity and the anomalous phenomena of the "quiet" solar atmosphere.

As a consequence of the strong departures from thermodynamic equilibrium in the chromosphere, its radiation intensity departs by several orders of magnitude from what it would be in thermodynamic equilibrium. For example, the equivalent blackbody temperatures of the radiation intensity observed at the centers of the resonance lines of hydrogen (1216 A) and He II (304 A) are about 7000 and 20,000°K, respectively. These lines are both formed in the high chromosphere or in the transition region from the chromosphere to the corona. For the resonance line of hydrogen, the kinetic temperature of the free electrons is in excess of 70,000°K where the center of the line is formed, and for He II it is in excess of 100,000°K [16]. The difference between the observed radiation intensity and that predicted from thermodynamic equilibrium amounts to about eight orders of magnitude in the case of the He II line. Similar, though less drastic, differences occur in such lines as the Balmer-α line of hydrogen and the H and K lines of Ca II. As a consequence, the radiation intensity in any of these lines is not a good measure of the kinetic temperature, as is so often implied in the literature.

2.4. The Corona

By definition, the corona of the sun begins at the top of the chromosphere and extends outward to the ultimate limit of observed solar radiation. At total solar eclipse it presents a magnificent spectacle, extending millions of miles into interplanetary space as an irregular, faint, white halo. Its total output of visible light is approximately that of the full moon, or less than 10^{-5} times that of the photosphere, which makes it, in spite of its large dimensions, difficult to study outside of solar eclipses.

In common usage, the term "solar corona" embraces two sharply different physical phenomena. One of these is truly a solar phenomenon, and includes the so-called "electron" of "*K*" corona and the "emission line" or "*E*" corona. The other is not a physical property of the sun, but a halo produced by interplanetary dust particles, the so-called "*F*" corona. In subsequent usage, we shall refer to the "true" solar corona as the solar corona or, for convenience, simply as the corona, and to the interplanetary corona as the dust corona. This differs from the usual *E*, *F*, and *K* coronal designations used among solar physicists. However, these designations are based largely upon phenomenological features rather than physical features, and for this reason they are not convenient for our discussion. Our interest in this review is primarily in the solar corona.

2.4.1. The Solar Corona. The solar corona is in a complex physical state. Its gases are highly tenuous at pressures comparable with a good laboratory vacuum. Its kinetic temperature is measured in millions of degrees, though the grossness of departure from thermodynamic equilibrium makes simple assignment of temperatures questionable. Its atomic particles are very highly ionized, as, for example, Fe X, Fe XIV, Ca XIII, etc. The fundamental spectral lines from these ions lie in the far ultraviolet and soft x-ray region and, with one or two possible exceptions [5], have not yet been observed as discrete lines. Its visible spectrum contains a continuum component similar to photospheric light, but with significant differences, and an emission line component corresponding to "forbidden" transitions in the highly ionized atoms. The total number of these lines so far discovered is about 30, and their total energy is about 1% of the continuum energy. It is probable that throughout its large dimensions weak magnetic fields play a major role in the physical processes operative in the corona.

The continuum in the spectrum of the solar corona is a product of Thompson scattering of photospheric light by free electrons. Thermal Doppler motions of these electrons, because of their high kinetic temperature, broaden the Fraunhofer absorption lines to such an extent that they are no longer resolvable. The spectrum, therefore, resembles the photospheric spectrum with the Fraunhofer lines "smeared out."

Electron densities in the corona are of the order 10^9 cm^{-3} at its base and decrease outward roughly exponentially with a scale height of about 10^5 km. This is only an average picture, however, and should not be taken literally. Eclipse observations of the coronal continuum show a great deal of structure in both radial and tangential directions, as shown in Fig. 7. As such observations are made with the line-of-sight tangential to the

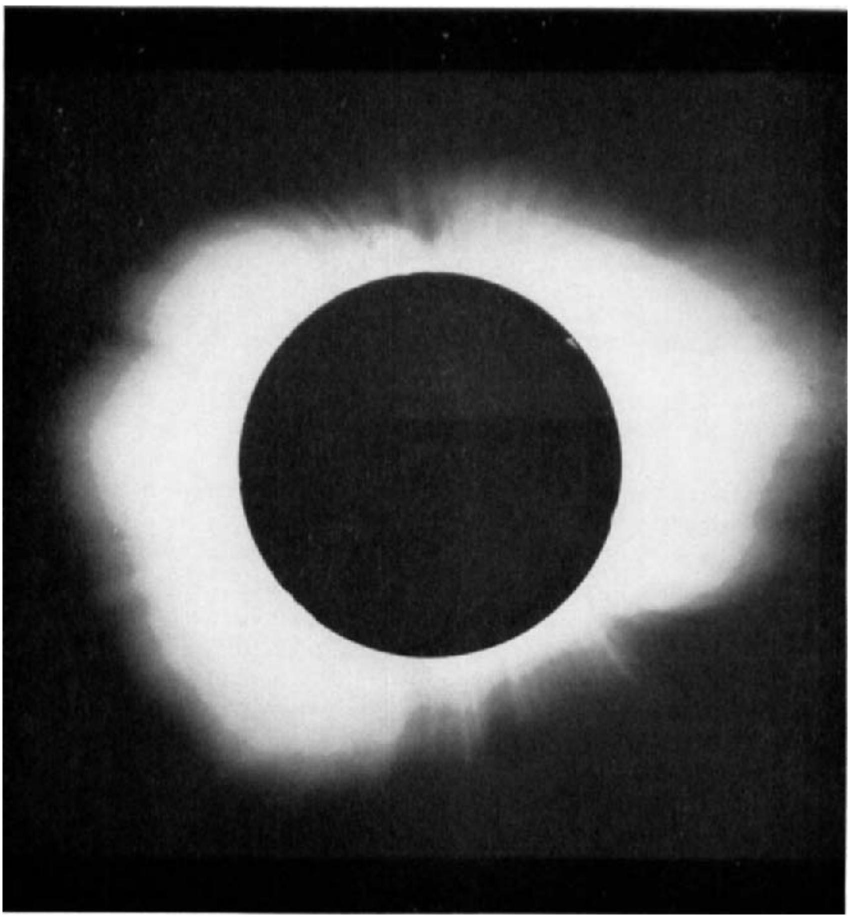


FIG. 7. The white light corona photographed at the total eclipse of 1918 (Mount Wilson Observatory).

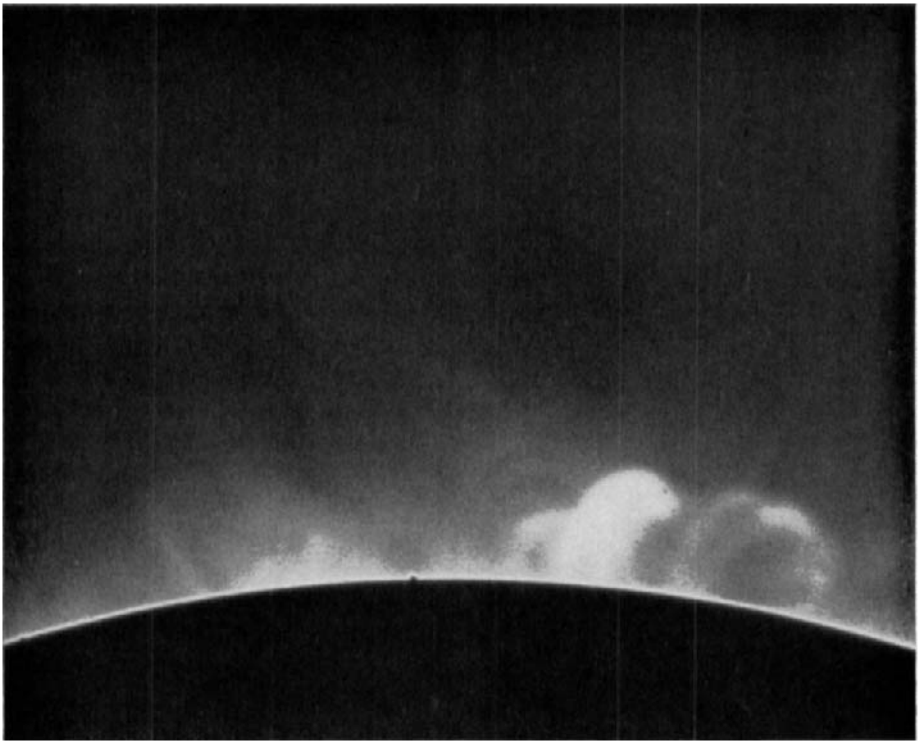


FIG. 8. The coronal green line, λ5303 A, photographed outside of eclipse (courtesy R. B. Dunn, Sacramento Peak Observatory, Geophysics Research Directorate, AFCRC, Sunspot, New Mexico).

solar surface, and as the scale height is so large, the true structure is partially obscured by the integration over the long effective path length. The line spectrum of the corona exhibits the structure more clearly than the continuum because the line intensity in the low corona is proportional to the square of the electron density rather than to the first power as in the case of the continuum. Figure 8 is a photograph of the green coronal line of Fe XIV at 5303 A obtained by R. B. Dunn with the coronagraph at Sacramento Peak observatory. Figure 9 shows the same line observed with a slitless spectrograph during the total eclipse of 1952 by the High Altitude Observatory expedition. We shall have more to say about these coronal structures later in this chapter.

As a given volume element in the corona receives photospheric light from one hemisphere only, the scattered coronal light is partially plane polarized with the magnetic vector parallel to the radial direction. The degree of polarization is consistent with the electron density distribution obtained

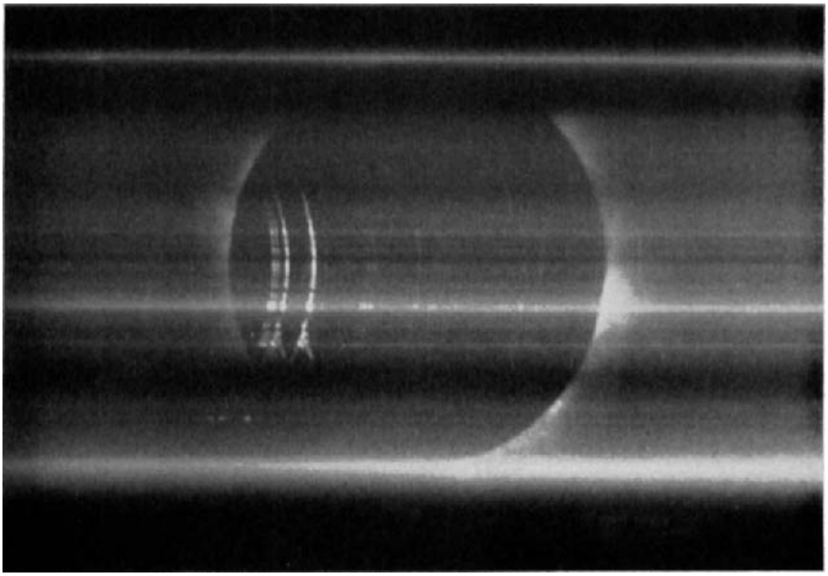


FIG. 9. The coronal green line, λ5303 A, photographed at the total eclipse of 1952 using a slitless spectrograph. The group of four lines includes the green Mg triplet of the chromospheric spectrum and a small prominence.

from the intensity of the scattered radiation. There have been claims by some observers that in certain transitory features, the plane of polarization is rotated with respect to its normal direction. These claims have been interpreted by some to indicate the presence of synchrotron radiation from relativistic electrons spiraling in a magnetic field. However, in all cases where detailed observations are available, no significant departures from radial polarization are observed, and there is no good reason to postulate sources other than electron scattering for the optical continuum. Later, when we discuss the transitory features of the corona and the continuum at radio wavelengths, the need for sources other than electron scattering will become abundantly clear.

The coronal emission lines in the visual spectrum, from the time of their discovery in 1869 by Harkness and Young to their identification in 1942 by Edlén, presented one of the most challenging problems of astrophysics. Edlén's identification of these lines with forbidden transitions in common metals in advanced stages of ionization, such as Fe XIV, Ca XIII, Ni XV, requiring a kinetic temperature of about 10^6°K for excitation cleared up this long-standing mystery. Lyot's invention of the coronagraph and his subsequent careful observations of the coronal lines was one of the strong contributing factors in Edlén's success.

Subsequent studies of the profiles of coronal lines and blackbody emission from the corona at radio wavelengths have reinforced the conclusion that the corona has a kinetic temperature of the order of 10^6°K. In some regions of the enhanced line emission such as those shown in Figs. 8 and 9, both the excitation stages of the ions and the line profiles indicate considerably higher temperatures of the order of $3–5 \times 10^6$°K. There are, to be sure, still some unexplained discrepancies between the kinetic temperature indicated by line profiles, the excitation stages observed, and the radiation in the radio spectrum. Thus, as in the case with the chromosphere and photosphere, we still have much to learn about the corona and its physical processes.

The processes of excitation and ionization in the corona are largely by electron collisions, whereas the balancing de-excitations and recombinations are by radiative transitions. This represents a complete departure from thermodynamic equilibrium where every process is balanced by its own inverse process. That this is not a trivial change is seen by noting that the ratio of Fe X to Fe XIV in the corona is greater by a factor of about 10^{54} than it would be in thermodynamic equilibrium at a kinetic temperature of 10^6°K.

Forbidden lines appear in the coronal spectrum, in opposition to normal laboratory spectra, because of the combination of high kinetic temperature and low electron density. In these lines, the corona is highly transparent, which contributes to their value as probes of coronal activity. The same ions which produce these lines must also radiate in the far ultraviolet and x-ray spectrum with much greater intensity than in the faint forbidden lines.

The corona is not transparent in many of the stronger permitted lines. However, because of the strong departures from thermodynamic equilibrium, the energy in these lines falls many orders of magnitude below that expected in the same wavelength interval from a blackbody at 2×10^6°K [17].

Two problems of long-standing interest in the corona are: (1) the source of energy for producing the high kinetic temperature, and (2) the explanation for the highly irregular geometry. Neither of these questions can be answered with any degree of confidence at this time. Both the geometry and the strength of the line emission in the corona vary systematically through the sunspot cycle. At sunspot maximum the outer corona is more nearly spherical in outline, although the inner corona appears to be highly irregular. The line emission is at its highest average level, both as a result of higher densities and higher temperatures. At sunspot minimum the outer corona becomes grossly aspherical with long equatorial streamers extending several solar diameters into interplanetary space, and shorter tufts of polar plumes reminiscent of lines of force about a magnetic dipole.

Coronal line emission, electron density, and temperature are at their lowest average values.

In the average picture, the density gradient within the corona is roughly consistent with a hydrostatic equilibrium model. The variety of complex spatial structures of loops and streamers observed in some regions in both the continuum and the coronal lines and the over-all irregular geometry both argue strongly against such a simple picture. Magnetic fields undoubtedly play a role in coronal phenomena; whether it is a minor or major role we cannot yet say. A magnetic field of one gauss, for example, would exert a pressure comparable to the gas pressure. Photospheric observations of magnetic fields suggest that coronal fields in excess of one gauss are highly probable in active regions. The coronal structures and prominence motions suggest a major influence from magnetic fields, but direct observational evidence is lacking. Both the magnetic field phenomena and the coronal energy source are of great interest in solar-terrestrial research.

2.4.4. The Dust Corona. The dust corona is distinguished from the true solar corona in a variety of ways. In the first place it is apparently spherically symmetric. Secondly, it is unpolarized, and finally it exhibits the photospheric spectrum complete with the Fraunhofer absorption lines at their ordinary intensity and breadth. Its explanation as a diffraction phenomenon due to interplanetary dust particles was successfully demonstrated in 1946 and 1947 independently by Allen and van de Hulst. The dust particles are of the order 10^{-3} cm in diameter, and their number density is about 10^{-13} cm^{-3}. In other words, there is about one such particle in an earth to sun column of 1 cm^2 cross section.

Although the observations are still incomplete, it is quite apparent that the dust corona and the zodiacal light will merge smoothly together and that they represent the same basic phenomenon. Recent observations of zodiacal light by Blackwell [18] from a high altitude observing site in South America have done much to close this gap. Figure 10, adapted from van de Hulst [14] and Newkirk and Roach [19], illustrates the comparative brightness of the solar corona and dust corona with more commonly observed objects.

3. Phenomena of the Disturbed Sun

3.1. General Features

The variable activity of the sun derives from outbreaks that have foci in localized regions of the sun's surface, the "active regions" or "active centers." Each of these regions undergoes a history of activity that rises to a maximum and then declines. Sometimes regions exhibit several separate

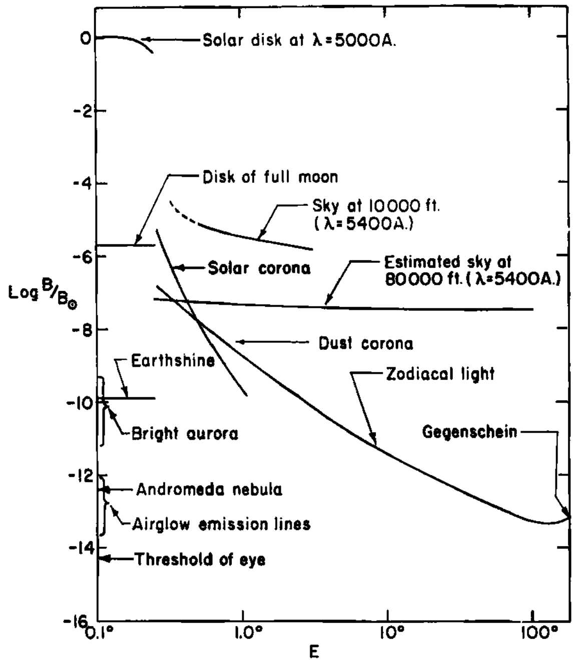


FIG. 10. Relative brightness of several comparison objects. When applicable, the brightness is given as a function of elongation angle from the sun.

peaks of disturbance before declining to a quiet state. Periods of high solar activity are characterized by the simultaneous occurrence of numerous active regions in different solar positions, with different degrees of disturbance.

The locations of the active centers on the sun display a systematic trend through the 11-year solar activity cycle. For example, the major active centers display a progressive equator-ward trend in their average latitudinal position, which begins near 30° in the early phases of the 11-year solar activity cycle and ends at about 5° at the end of the cycle. Centers of a declining cycle are differentiated from those of a new cycle, when the two overlap, by their location in latitude, though the reversed magnetic polarities of the associated sunspots or photospheric magnetic fields can also be used, in many instances, to make the distinction.

Generally speaking, the commonly used indices of solar activity combine together all active centers of the sun or of given defined areas of the sun. Some published indices, as described in Section 5, measure the activity of individual regions based on evaluation of the level of disturbance exhibited by the individual phenomena discussed.

3.2. Sunspots

Sunspots are the most readily apparent features of solar activity. They are relatively dark areas of the solar photosphere, when seen in integrated light. They consist of fluted "penumbral" areas of intermediate darkness, and darker "umbral" areas that also reveal brightness fluctuations when studied under ideal conditions. Figure 11 shows a typical large sunspot, with umbral and penumbral areas labeled.

Historical records of positions, areas, or numbers of sunspots reach back to 1611 and are relatively reliable from about 1750. These are generally results derived from hand drawings of the disc of the sun, carefully made for this purpose.

Sunspots tend to group together into compact associations numbering from a single or, more frequently, a few disconnected spots to dozens of individual spots. Areas of the individual spots, including penumbral areas, range up to about 10^{10} km^2, although those with areas in excess of 10^9 km^2, which are visible to the unaided eye, are rare. A large spot group may cover 2×10^{10} km^2. Areas are often given in millionths of the area of the visible solar hemisphere. One such unit equals 3.04×10^6 km^2. Thus, a spot group of substantial size may be 5000 to 6000 millionths of the solar hemisphere. Published sunspot areas have generally been corrected for foreshortening as they approach the solar limb.

The umbras of sunspots have effective temperatures about 1500°K cooler than the adjacent photosphere, whence it is obvious that they are, by terrestrial standards, very bright luminous surfaces and that they are dark only by comparison with the still more luminous photosphere. As a result of the lower temperature in sunspots, the opacity of the solar gases is decreased. Thus, a sunspot probably represents a depression in the photosphere, although the amount of the depression is unknown. As the effective temperature refers to the radiation from the visible layers, and as the temperature increases inward in the photosphere, the temperature difference between photosphere and sunspot on a constant height surface probably exceeds 1500°K.

Averages for sunspots are sometimes difficult to evaluate because of the difficulty of properly including the smaller spots near the threshold of observation. For this reason, values reported by different authors for average lifetimes, sizes, etc., vary greatly. Thus, Kiepenheuer [20] states that the lifetime of most spot groups is less than a day, whereas Allen [21] gives the lifetime of the average group as six days and Abetti [22] states that most individual spots live several days. It is clear, however, that spots and spot groups that develop to where they are easily observed have average lifetimes of at least a few days. The larger spot groups may persist much

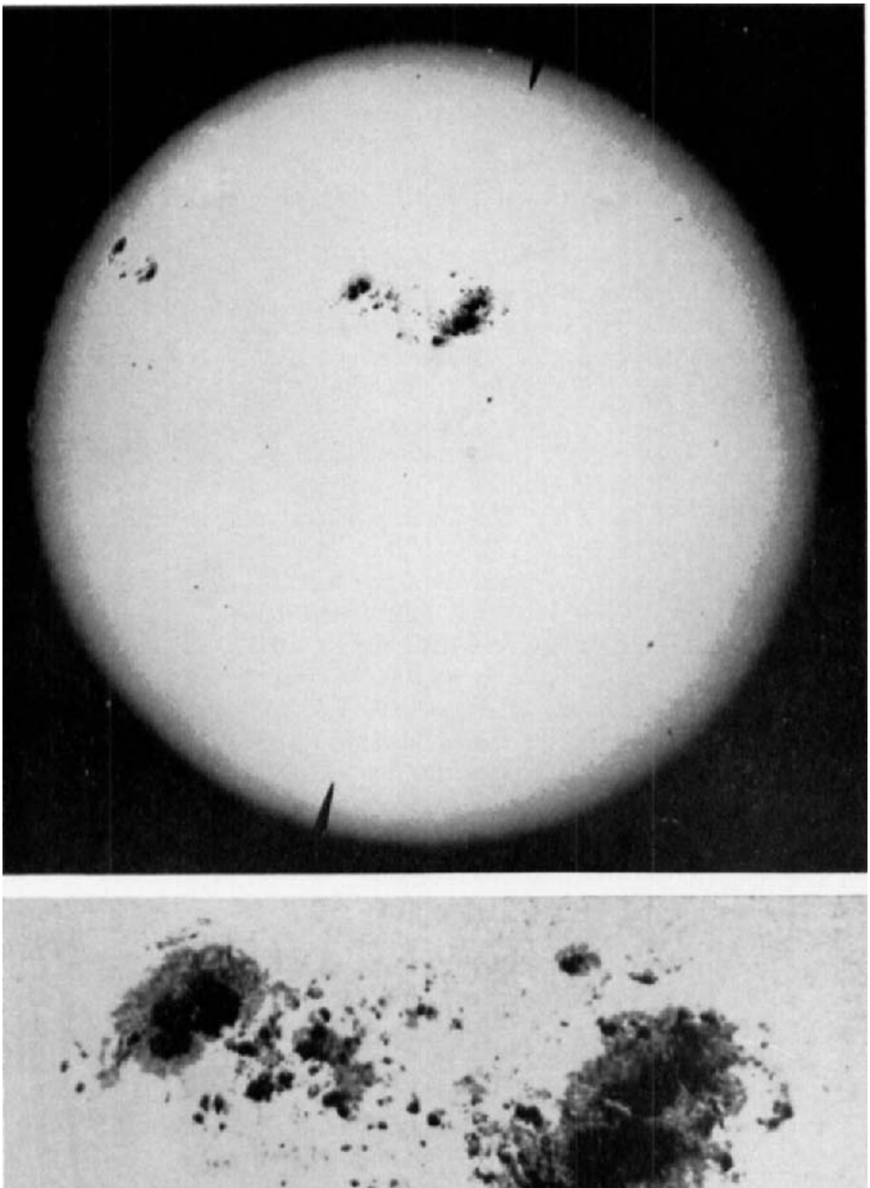


FIG. 11. The large sunspot group of April 7, 1947 (courtesy Mount Wilson Observatory).

longer, often being recognizable for six to eight months. On the other hand, even a large spot group may change significantly from one day to the next, particularly in its early development.

One of the more interesting features of sunspots is their unusually strong magnetic fields. All spots of sufficient size to permit a study of the Zeeman splitting of their spectral lines reveal strong magnetic fields. The strength of the field increases statistically with the size of the umbral area in the individual spot, approaching 4000 gauss for the larger spots. Within individual spots and within spot groups, there are marked changes in the magnetic field strength. Most often spot groups are bipolar in character with a dominant spot in the forward (westward) edge of the group with one polarity and the following less developed spots of opposite polarity. However, a fairly large percentage of spot groups show a unipolar character with all spots of the same polarity.

Sunspot polarities tend to be systematic throughout a given hemisphere with the dominant spots exhibiting the same polarity. In the opposite hemisphere and in alternate sunspot cycles, the polarity of the dominant spots show a strong tendency to be reversed from the reference hemisphere and reference cycle. This curious reversal of polarities has led to much speculation concerning 22 year cycles of solar activity versus the normal 11 year cycle. However, it is of more than passing interest to note that the magnetic polarity laws are not hard and fast rules and are not infrequently violated.

Spot groups are classified both according to their visible features and their magnetic features. The details of these classifications along with the great wealth of statistical data on the morphology, evolution, and interdependent features of sunspots are beyond the scope of this review. For a discussion of these features the reader is referred to the various textbooks listed in the bibliography.

The role of sunspots in the over-all picture of solar activity is difficult to assess at our present state of knowledge. Sunspot groups mark the locations of all important centers of solar activity, and the spots are among the first features to be observed in the region. However, long after the sunspots have disappeared, there are a variety of visual manifestations of continued activity in most active regions. Because of the ease with which sunspots are observed, they are generally considered to be the primary indicator of solar activity. This should not be construed to mean, however, that sunspots are the primary physical phenomenon of an active region. Like other manifestations of solar activity, they should be interpreted as a symptom of a more general disturbance in the solar atmosphere, the cause and extent of which are unknown.

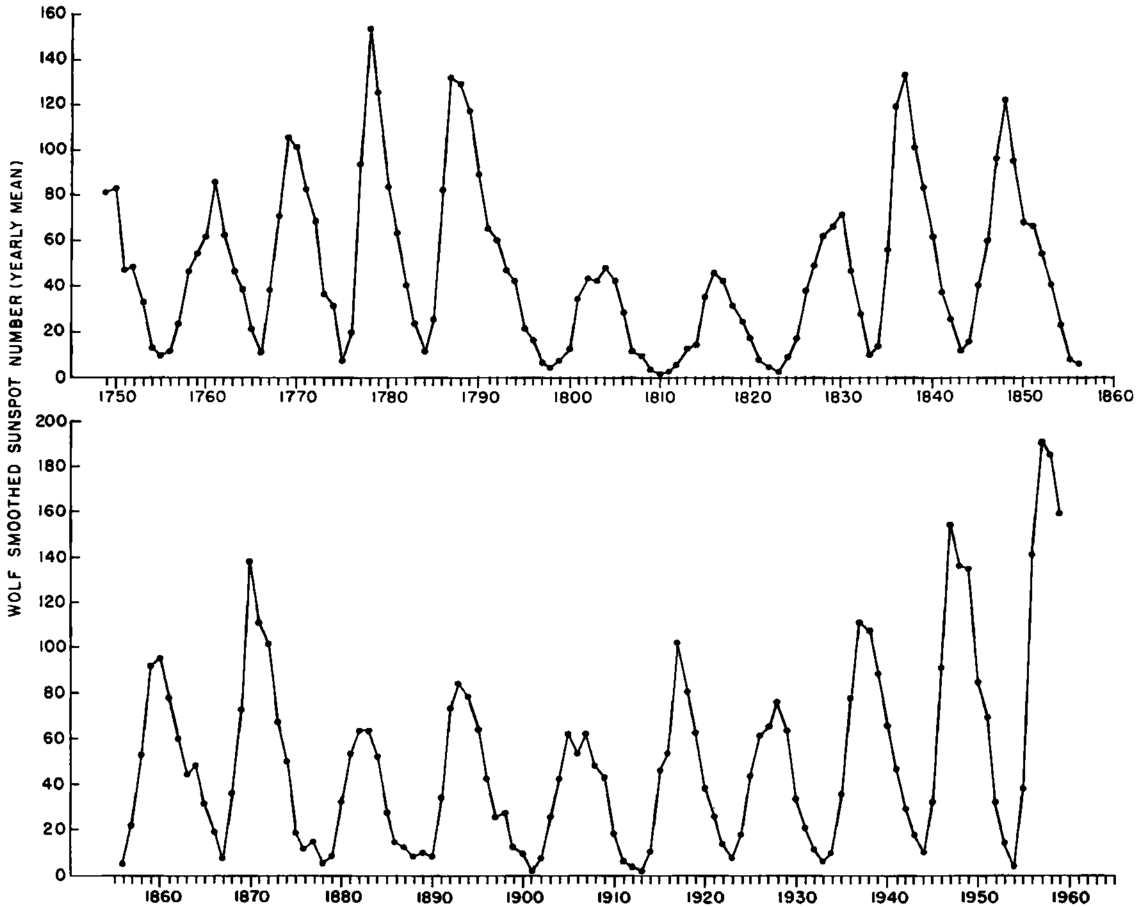


Fig. 12. Annual mean sunspot numbers as a function of time.

The most common index of sunspot activity is the Wolf number R defined by

$$R = 10g + s \tag{3.1}$$

where g is the number of groups and s is the number of individually countable spots. While this number may be criticized because of the rather heavy weighting of groups versus individual spots (for example, a large well-developed spot with 9 moderate-to-small size spots in the same group has $R = 20$, whereas two small isolated spots have $R = 22$), and because of the lack of consideration of area or magnetic properties, its long-standing usage and systematic variation in time make it a valuable parameter. Sunspot numbers are available on a daily basis or as monthly and yearly averages.

Yearly sunspot numbers show a general rise to maximum and a subsequent decline to minimum extending over an average period of 11.2 years. Once again, however, irregularity is almost as much the rule as is regularity. Actual observed sunspot cycles vary in length depending on the particular cycle and whether the period is measured between minima or between maxima. The periods between minima vary from about 8.5 years to about 14 years, and the periods between maxima from about 7.3 years to about 17 years. Similarly, the annual mean of $R_{\min}$ varies from about 0 to 10, and of $R_{\max}$ from about 50 to 190. There is some evidence in sunspot data for longer periods superposed on the average 11.2 year period. A long-period cycle of about 90 years is suggested by visual inspection of a plot of R versus time as in Fig. 12. However, the observing history of sunspots is not sufficiently long to verify this suggestion. During years of relatively low values of $R_{\max}$ there is a suggestion of alternate relatively high and low cycles. For example, the eight cycles between 1858 and 1942 each show a successive alternation in amplitude. This effect, if real, disappears during years of relatively large $R_{\max}$.

Many attempts have been made to represent sunspot data analytically with the thought of predicting future sunspot activity. Successful representations have been achieved for periods of time in excess of 100 years, but all attempts to predict subsequent cycles from these analytical representations have been marked by failure relative to simple statistical predictions.

3.3. Faculae

Solar photospheric faculae, like sunspots, are seen in integrated light. Unlike spots, they are visible only near the solar limb, where they appear as patchy, veinlike regions somewhat brighter than the adjacent "normal" photosphere. They surround all sunspot groups and are most pronounced near large, active spot groups. On the other hand, they also occur in high latitudes well removed from sunspot activity. These high latitude faculae are fainter and less extensive than those directly associated with sunspots.

As the sun rotates, and faculae seen at the east solar edge are carried toward the center of the apparent solar disc, the faculae fade from view. Three days of solar rotation (about 40°) suffice, generally, to erase the brightness difference between faculae and the photosphere.

The increased contrast between faculae and photosphere near the solar limb means, of course, that faculae show less limb darkening than the normal photosphere and that they rise above the normal photospheric layers seen at the center of the disc. Their abnormal brightness requires that they have a somewhat higher temperature than their surroundings. Since the opacity of the photosphere will increase with increasing temperature, the requirement that faculae be elevated in the photosphere is consistent with their brightness.

Faculae are often the first sign of the birth of an active region, appearing somewhat before the first signs of sunspots. And they can often be seen long after other signs of an active region have died out. Their average lifetime usually exceeds that of the attendant spot group by at least threefold.

Relatively little is known about faculae other than the brief statements made here. Nevertheless they may well represent one of the more important manifestations of solar activity.

3.4. Plages

Photospheric faculae have a counterpart in the bright *plages* (or *plages faculaires*) seen in solar images formed in the monochromatic light at the cores of strong Fraunhofer lines (Hα, H, and K of Ca II^{+}, etc.) with birefringent filters or spectroheliographs. Because they are observed in lines that are formed within the chromosphere, they are a chromospheric phenomenon. Historically, they have also been called "bright flocculi" and "chromospheric faculae," but these terms are undesirably ambiguous and to be avoided.

Even though plages show as bright features against a darker background as illustrated in Fig. 5, they are bright only in comparison with the cores of the strong Fraunhofer lines. A spectrum of the disc in the plage region still exhibits absorption lines, though not quite as deep as is normally observed. Quite often, in the literature, one finds reference to the "emission lines" of a plage. These so-called "emission lines" are obtained by subtracting the normal disc spectrum from the plage spectrum. The lines that are less deep in the plage spectrum thereby show up as emission lines after the subtraction. There is no physical justification for assuming that the spectrum of the gases underneath the plage is the same as that of the undisturbed disc. In fact, there is every reason to believe from the theory of radiative transfer that this is not the case. One should, therefore, regard these emission lines as simply the difference between the spectrum of a plage region and a normal region. They should not be regarded as the spectrum of the plage itself.

Plages are visible in all parts of the solar disc with no apparent change in contrast from center to limb. Otherwise, they resemble the faculae in distribution, association with sunspots, and in the fact that they often precede the formation of spot groups and generally outlive other signs of activity in active centers.

A large plage is 6×10^{10} km^2 or 20,000 millionths of the solar hemisphere. Tabulated plage and facular areas are generally corrected for foreshortening near the solar limb. Thus areas based on data taken near the limb can be highly misleading because of the large importance of the correction. Although plages are considerably more extensive in area than faculae, in all other respects they appear to be the chromospheric counterpart of faculae. The physical conditions within plages are even less well understood than those within faculae. Observations of plages, however, usually receive more attention than observations of faculae, and more data are available. The reason is that once a monochromator of sufficiently narrow band-pass is obtained, plages are easier to observe than faculae and they can be followed across the entire solar disc.

Plages are of great interest in solar-terrestrial research because of the relative ease with which they are observed. Although there is nothing about the plage observations to suggest direct terrestrial influence of any consequence, the strong association between plages and other solar phenomena more difficult to observe, but with direct terrestrial influences, makes plage data extremely valuable as an index of solar activity.

3.5. Prominences and Filaments

Solar prominences are perhaps the most striking and most varied of the changeable features of the sun, when viewed appropriately. At total eclipse they can be seen in white light as bright clouds imbedded in the corona. Their spectra, however, resemble that of the chromosphere and are composed of a faint continuum with bright emission lines. With a coronagraph or spectroheliograph, prominences can be seen only by isolating the light in their stronger emission lines. Again, the Hα line of hydrogen and the H and K lines of Ca II are well suited for this purpose.

Prominences show large, readily discernible differences in both their spectra and their morphology and evolution. These differences have been used to classify prominences according to several alternative schemes. For our purposes, two main classes are of interest: active region prominences and quiescent prominences. The dark filaments observed in monochromatic light against the disc of the sun as seen in Fig. 5 are prominences belonging to this latter category. When seen at the limb, they are elevated well above the chromosphere with the upper boundaries typically extending to heights of 40,000–50,000 km. Active region prominences can also be seen against the disc on occasion, but they are much more transitory and much smaller in lateral dimensions than the dark filaments.

Generally speaking, the active region prominences display (a) broad line profiles suggesting kinetic temperatures $\gtrsim$30,000°K, (b) relative intensification of spectrum lines of high excitation atomic states such as He II compared to low excitation lines of such atoms as Fe I, (c) rapid motions and evolutions, (d) association with active centers, (e) frequent large changes of brightness, and (f) lifetimes of the order of hours.

There are many distinctive forms of active region prominences, ranging from pronounced closed "loops" lying over sunspots to violently ejective "surges." Speeds of motion for active region prominences generally range from 50 km/sec to 100 km/sec, but often reach 500 km/sec and occasionally exceed 1000 km/sec.

Quiescent prominences are often of enormous size and brightness. They cannot be accurately called a phenomenon of the undisturbed sun, since they occur in isolated places and because they appear quite generally to have an association with solar active regions. Typically, they form on the poleward boundaries of active regions. Many drift slowly poleward as the region evolves and may reach large distances from their parent region.

The great average longitudinal extent of dark filaments, or quiescent prominences, is evident in observations at the limb as well as on the disc. It is not unusual, for example, for a large quiescent prominence to be seen at the limb for 7 successive days, during which the sun rotates approximately 90°. Lifetimes may run for as long as a few months.

The quiescent prominences, when observed at the limb, appear to be of great optical depth in their stronger spectral lines. The fainter lines, however, where there is little self-absorption, are intrinsically narrow, testimony to their relatively low kinetic temperatures $\sim$10,000°K.

A particularly characteristic form of quiescent prominence is the "hedgerow" prominence (see Fig. 13) which stretches out ribbon-like, sometimes for hundreds of thousands of kilometers.

The internal structure of prominences consists of a fine network of filamentation, generally radial in orientation. Motion within prominences is generally sunward along the filamentary structure. In quiescent prominences, the motions have moderate speeds of about 10–20 km/sec, much slower than the motions within active region prominences. Certain types of active region prominences, notably the surges, puffs, and sprays have a predominant outward motion. The majority of these prominences decelerate and eventually fade from view or fall back into lower layers. Many, however, show no deceleration and apparently escape from the sun.

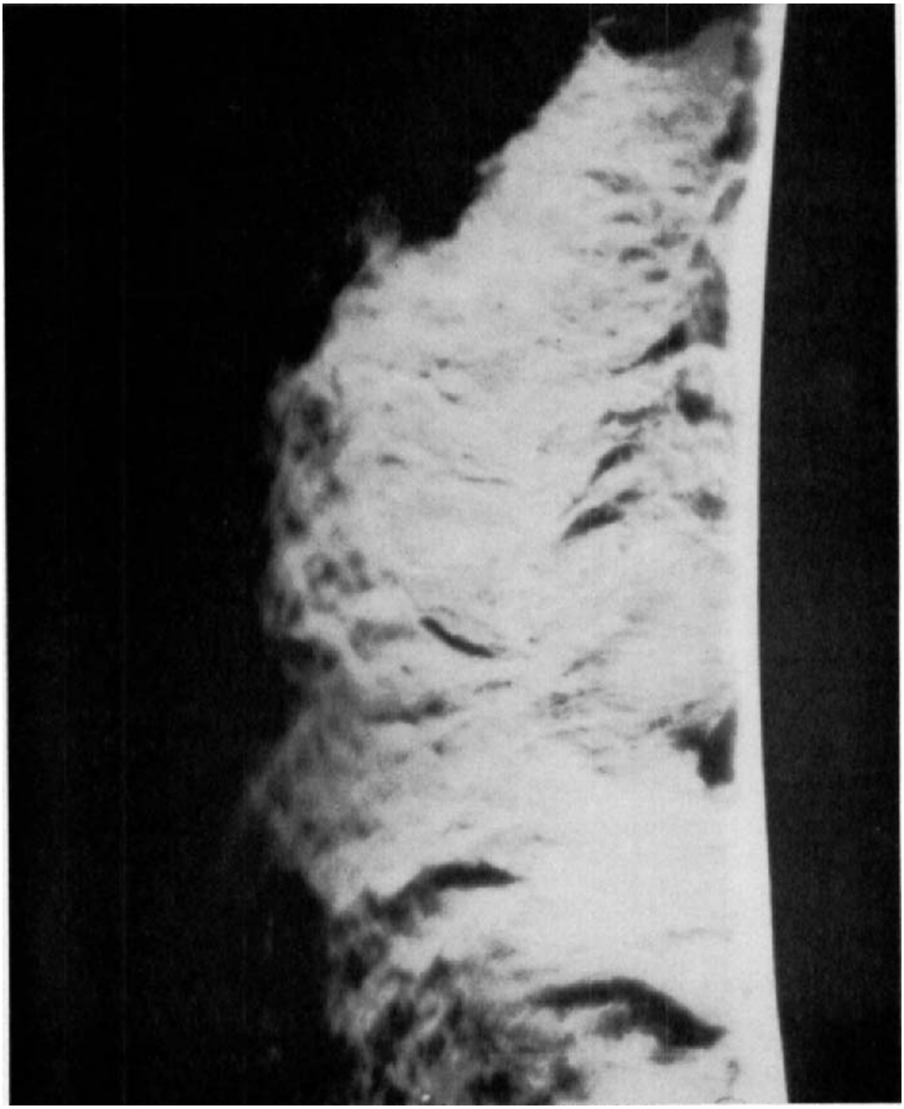
On occasion, the dark filaments, or quiescent prominences, become active and erupt bodily outward from the sun at high speeds. These "sudden disappearances," as they are called, undoubtedly reflect changes in the parent active region with which the prominence is associated. Prominences are of much higher density than the surrounding coronal gases, and they must be supported against the strong gravitational attraction of the sun. The support mechanisms are presumably the magnetic fields of the active regions, and the sudden eruptions are presumably due to disturbances initiated in the active region and propagated along the magnetic lines of force. Quite commonly, following the eruption, which may require a few hours, a new quiescent prominence forms within a relatively short time in the pattern of the old quiescent before its eruption, suggesting that the eruption was only a temporary disturbance.

Recently, Moreton [23] has discovered disappearances and reappearances of dark filaments that occur much more rapidly than even the "sudden disappearances." These "abrupt disappearances" often last but a minute or so and are directly associated with the outbreak of a flare in an active region. The disappearance of the filament, in this case, is more likely caused by bombardment by high speed particles ejected by the flare [24] (see Section 3.6).

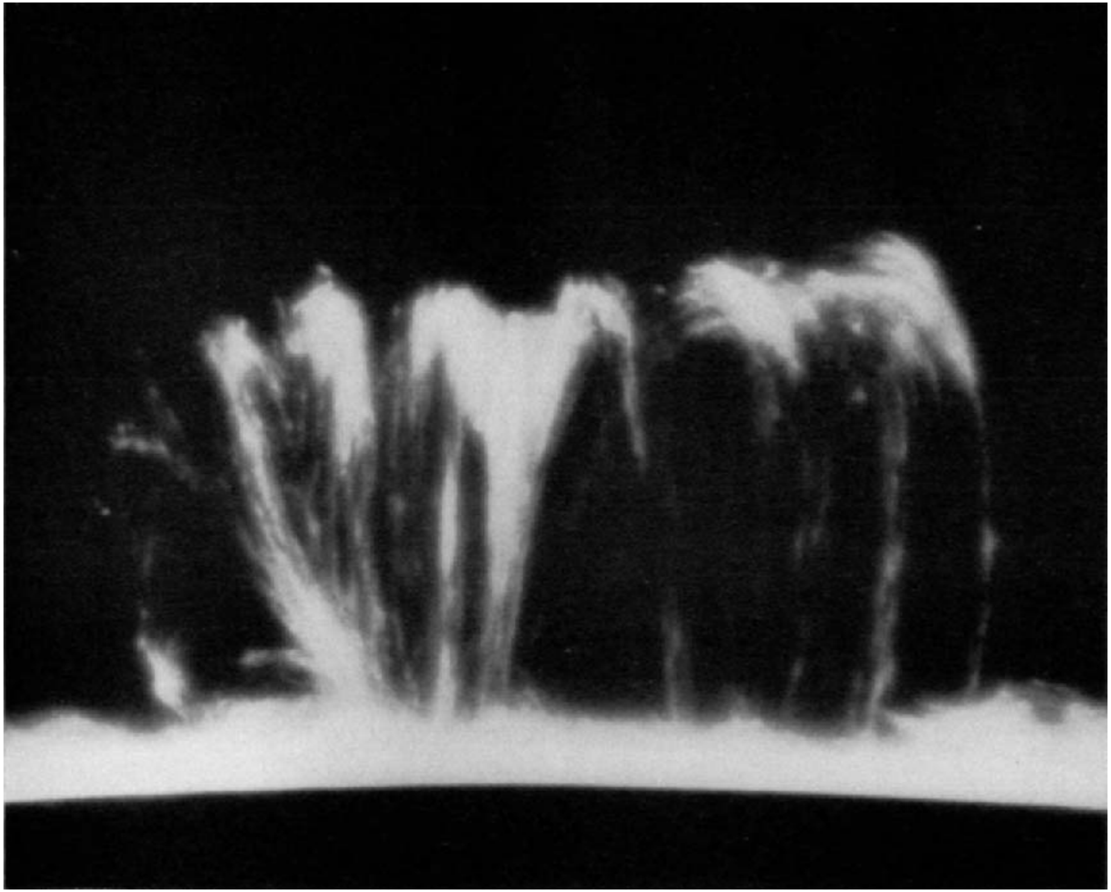
Prominences, like most other phenomena of the active sun, are very poorly understood. Their spectra are formed under conditions departing grossly from thermodynamic equilibrium, and their geometry is extremely complex. For these reasons, their spectra are difficult to interpret in terms of the physical conditions and processes operative within the prominences. Without this knowledge, we cannot properly assess from a solar point of view their value as indices of solar activity for use in solar-terrestrial research (see Section 4).

3.6. Flares

Solar flares are in many ways the most spectacular of all known solar events. Like plages and prominences, they are visible mainly at the wavelengths of the strong Fraunhofer lines, where they show up as emission features of unusual brightness. Their most distinguishing feature, however, is the explosive suddenness of their appearance. It is not unusual for a large flare to rise to maximum brightness in less than 5 min with a subsequent decline lasting typically from 20 min to 1 hr. Figure 14 exhibits a time sequence of photographs showing the development of a large flare.



(a)



(b)

Fig. 13. The Hα prominences (a) large quiescent and (b) active region loops (courtesy Sacramento Peak Observatory, Geophysics Research Directorate, AFCRC, Sunspot, New Mexico). [See facing page for part (a) of figure.]

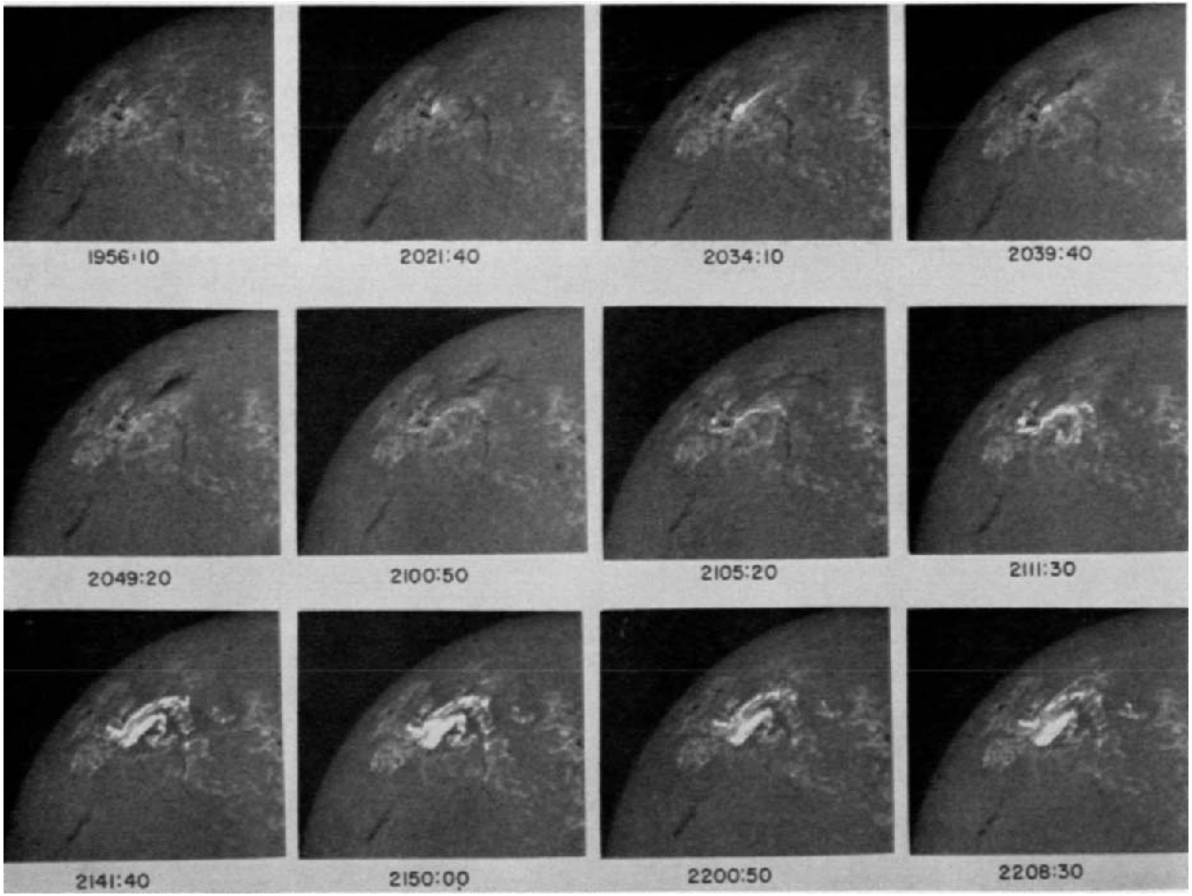


FIG. 14(a). A sequence of Hα pictures showing the development of a high-speed bright which evolves into a high-speed dark followed by a 3+ flare, 10 May 1959.

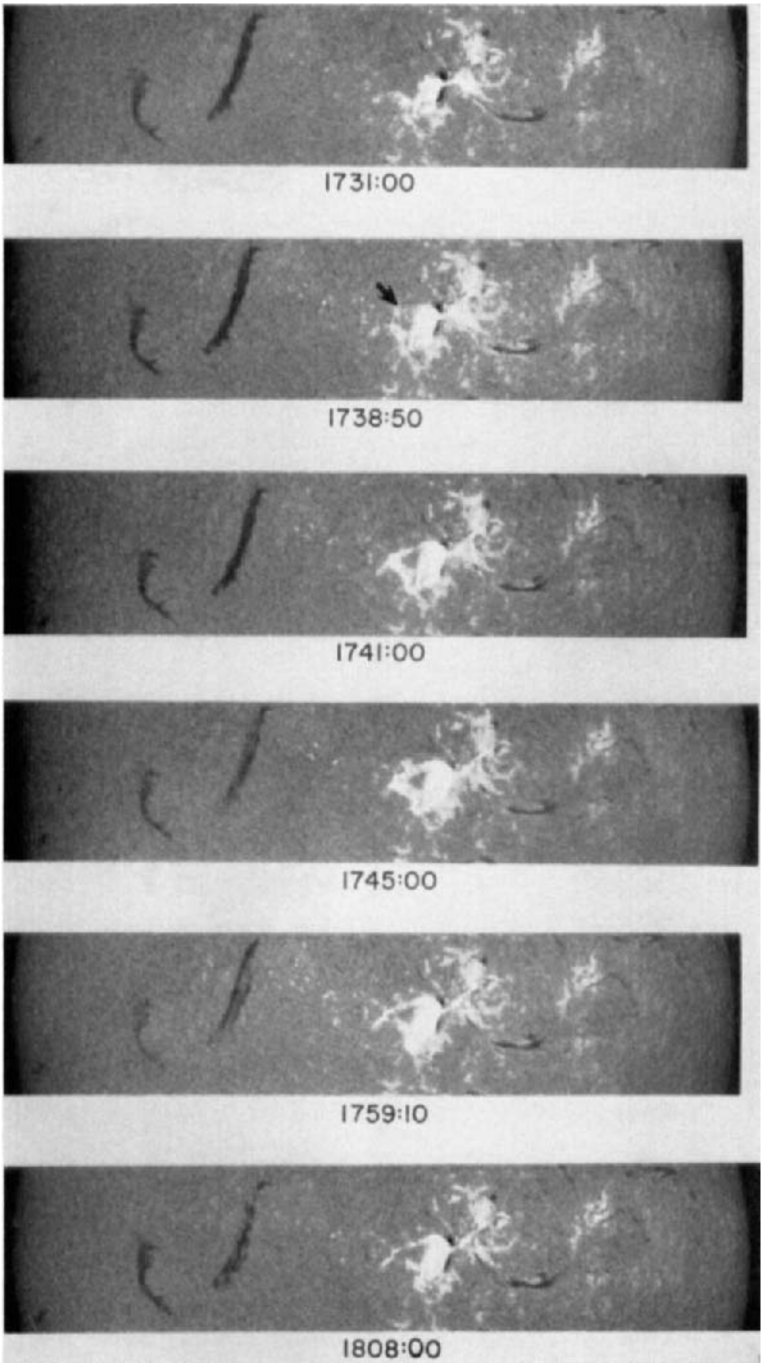


FIG. 14(b). A sequence of Hα pictures showing the development of a 3+ flare followed by activation of a distant dark filament, 30 November 1959. (Courtesy Gail Moreton, Lockheed Solar Observatory).

When a flare is observed against the disc in the Hα line, the central intensity of the line may exceed that of the adjacent continuum so that the normal Fraunhofer absorption line becomes an emission line. At the wavelengths of the stronger helium lines, not normally present in the Fraunhofer spectrum, flares may produce either absorption or emission features. Sometimes both occur in adjacent regions of the same flare.

Flare spectra on the disc have been much discussed in the literature, but with rather fruitless results. Great difficulties in the interpretation of the spectra again arise because of gross departures from thermodynamic equilibrium and an unknown geometry. As is the case with plages, most observers assume that the spectrum of the flare itself can be obtained simply by subtracting from the observed spectrum the normal disc spectrum. For some lines such an assumption may be valid, but for such lines as Hα it is certainly in gross error.

Somewhat less than 10% of all observed flares are seen at the limb above the chromosphere and in projection against the sky and corona. In these cases, there is little ambiguity in the true spectrum of the flare, and one can hope to infer the physical conditions within the flare. As yet, however, adequate data are lacking, and we can do little more than guess that the temperature conditions within flares lie somewhat intermediate to the chromosphere and the corona. Their material density is comparable to that of the low chromosphere.

The relatively small percentage of limb flares argues that in general flares are not elevated much beyond the upper boundary of the chromosphere. Some are observed well above the chromosphere, however, and they should perhaps be regarded as phenomena of both the upper chromosphere and low corona. Since their density is high relative to these atmospheric regions, at least part of the flare phenomenon seems to be a sudden increase in density resulting from some such mechanism as compression, condensation, or injection.

For the most part, flare spectra show little evidence of systematic motions. In some flares, on the other hand, rapid motions are indicated both in the morphology of the flare and in the Doppler shifts in the spectral lines.

Flares occur only near active centers, usually over a pre-existent plage. They are most frequent during periods of rapid evolution of complex sunspot groups, particularly during periods of growth. However, some occur in regions where the spots have disappeared. In size and brightness they vary greatly. Flares exceeding 3000 millionths of the visible hemisphere are rare, and the smallest extend down to the limit of present resolution. Their brightness in Hα ranges from about a fourfold increase over the normal disc brightness (about twice plage brightness) to about a fifteenfold increase (three times the brightness of the continuum adjacent to Hα).

The shapes of solar flares are usually highly irregular, and frequently areas that are well removed from each other and not otherwise visually connected brighten simultaneously. They often follow the outlines of the minor filamentary markings in the vicinity of the sunspots. Also, quite commonly they seem to border the larger dark filaments near active regions but, just as commonly, the outbreak of a flare results in the disruption of a distant dark filament. In still other cases, loop prominences exhibit flarelike characteristics and may themselves become flares. The reasons for these varied associations are not at all clearly understood, but it seems very probable that magnetic fields play a major role in flare phenomena. Different active regions vary enormously in their flare productivity, which may result from peculiarities in the magnetic fields.

A comparative wealth of data are available for flares and related phenomena showing complex interrelationships. Although some outstanding effects will be discussed in Section 4, the reader should consult the more detailed discussions available in the literature [25].

Of all directly observed solar phenomena, flares are perhaps of the most interest in solar-terrestrial research and exhibit the closest correlations with terrestrial phenomena. Their terrestrial responses range from sudden increased ionization (i.e., SID's) in the ionosphere, particularly the *D* layer, resulting from increased ultraviolet or x-ray radiation, to the spectacular "polar cap absorption events" resulting from proton bombardment of the atmosphere overlying the poles. Additional evidence of strong corpuscular radiation associated with flares comes from their correlation with auroral and geomagnetic activity. More directly, several large increases in cosmic ray intensity and changes of particle density in the outer Van Allen belts have been observed following flares.

3.7. Active Coronal Regions

The corona overlying centers of activity in the photosphere and chromosphere reveals much of the same complexity of phenomena as the lower layers. Generally speaking, both the line and continuum emission increases markedly in the low corona above active sunspot regions. In the case of the lines, the increase may exceed a factor of 100. The line profiles broaden considerably, and the intensities of the lines from the ions with higher ionization potentials increase relative to those from ions with lower ionization potentials. It is only in such regions that the rare yellow line of Ca XV (ionization potential $\sim$820 volts) at 5694 A is observed.

The dimensions of the active regions in the corona are commonly in excess of 10^5 km. The general increase in intensities of the line and continuum emission is due primarily to a density increase, which may be as much as a factor of ten in the low corona. In general, the relative enhancement of the emission from ions with high ionization potential and the widening of the line profile both suggest higher temperatures in these active regions than in normal coronal regions. The increased density and emission diminish in both the outward radial and tangential directions, and in a rough sort of way the coronal region can be visualized as a localized condensation of increased density and temperature in the lower corona. Indications are that the temperature may reach 3–5 $\times$ 10^6°K.

The structure of coronal regions (see Figs. 8 and 9) is highly individual and varies from tightly closed loops to open diverging rays. Observations of the structure are still severely limited by lack of data. In the limited observations available, however, there is clear evidence of rapid changes in structure over short periods of time. Loops of increased brightness appear to move outward, or in some cases, to oscillate in the vertical direction. Horizontal motions are also suggested. Rays of increased brightness, which may be generally radial but somewhat curved, have on a few observed occasions shown signs of violent motion. On two occasions, distended loops have been observed to break apart in a violent "whiplike" motion suggesting an outward surge of matter.

It is not known to what extent these apparent motions are real motions of matter or simply changing thermodynamic conditions. Doppler shifts in the spectral lines show clearly that real motions of the type observed are not uncommon, and the nature of the more violent motions suggests a real motion of matter rather than a "wave of excitation."

Since the corona is the outermost region of the solar atmosphere, any radiations reaching the earth must either traverse it or originate within it. Solar x-rays and part of the ultraviolet radiation originate within the corona, and their variations are best described by coronal parameters. Much, if not all, of the solar corpuscular radiation probably originates in the corona. That which originates at lower levels, say in the chromosphere, can hardly traverse the corona without exhibiting some such major effect as the whiplike phenomena observed. Since the mean free paths for nonrelativistic particles are small compared to the thickness of the corona, corpuscular clouds ejected from the chromosphere and escaping from the sun must push aside or carry along the overlying coronal gases. For these reasons, the observations of coronal structure such as those initiated for the coronal lines at Sacramento Peak by Dunn and those initiated for the coronal continuum at the High Altitude Observatory by Newkirk [26] will, in the future, provide valuable new data for use in solar-terrestrial research.

Observations in the radio spectrum at wavelengths above about 10 cm, for which the corona becomes either totally or partially opaque, are adding a great wealth of information about the corona and its physical processes, particularly in active regions. These observations are discussed in Section 3.9.

3.8. Magnetic Fields

The discovery by G. E. Hale in 1908 of the strong magnetic fields in active regions of the sun opened the way for a new class of theories of solar activity in which magnetic field energy played its rightfully major role. Since that time it has been found that the sun also possesses a general magnetic field oriented in a direction parallel to the rotation axis, suspected but never properly established by Hale. It has also been discovered that there are extensive regions of the sun's disc, not characterized by any pronounced visual features of activity, which exhibit measurable magnetic field anomalies. Some of these are unipolar, in the sense that over a substantial area there is a consistent magnetic polarity which is not equalized by the existence of concentrated poles of opposite sign nearby. Others may be of a bipolar or multipolar nature. All direct observations of solar magnetic fields apply to the photospheric layers only. In the external layers there is strong, though indirect, evidence of magnetic field phenomena, such as that suggested by prominence motions and coronal structure. The behavior of solar magnetic fields has taken on greatly increased importance in recent years because of the apparent importance of these fields in the corona and interplanetary space. It is likely that changes in the behavior of the associated magnetic fields play an important role in all phenomena of varying solar activity.

For many years it had been suspected, from the shape of coronal polar plumes, that the sun possesses a general magnetic field similar to that of a bar magnet lying parallel to the solar rotation axis. If the shape of the rays of the solar corona, seen at total eclipse, is taken to define the direction of the mangetic field, it can be deduced that the character of the field is like that of a simple dipole two-thirds the diameter of the sun in length, symmetrically located along the rotation axis of the sun. This approximate dipole character can be clearly seen only in the polar regions of the sun, to a distance of approximately 30° from the rotation axis. Outside this angular range the behavior of the corona suggests that the magnetic fields are considerably more complex. In the photosphere, the magnetic fields show evidence of this same behavior, with a general dipolar character near the poles and an irregular complex structure in lower latitudes. Also, there is evidence from cosmic ray data [27] that in equatorial regions the solar magnetic field departs from that of a dipole, even at great distances from the sun, and perhaps more nearly resembles a radial field.

Observational establishment of the existence of the general or polar field awaited the development of a very precise electronic magnetometer by H. W. and H. D. Babcock [28]. With this instrument the Babcocks established that there are magnetic fields of consistent sign in the sun's polar caps.

Their measurements were confined to the longitudinal component of the field as shown by the Zeeman splitting of certain iron lines in the solar photosphere. But they conclusively established the existence of a field of reasonably uniform character with a strength of about one gauss at the poles.

From the first observations with the solar magnetograph in 1953 until May 1957, the general magnetic field of the sun was oppositely oriented to that of the earth; i.e., the north rotation pole of the sun was of opposite polarity to the north terrestrial pole. In May 1957, the southern rotation pole underwent a reversal in magnetic polarity that was followed in November 1958 by a similar reversal in the Northern Hemisphere [29]. This change of polarity has persisted to the present and will probably continue for at least several years. The reason for this conjecture is that the reversal of polarity occurred just during the period of sunspot maximum, and is probably related directly to the sunspot cycle.

Whether or not a similar reversal of the solar field will occur at every sunspot maximum is not known, of course. It is possible that the reversals, if truly related to the sunspot cycle, would show an alternate cycle effect with reversals at alternate maxima. On the other hand, it is not implausible to suppose that the decline of solar activity following sunspot maximum is itself due to a reversal of the general field.

For more than a year during the reversal of the field, the two rotation poles of the sun had the same magnetic polarity, which further emphasizes the departure from a true dipolar character.

The magnetic field structure of the lower solar latitudes, from about 50° to the equator, is highly complex. In all active regions there are magnetic fields whose longitudinal field components measure from a few tens of gauss throughout most of the active region to thousands of gauss in the centers of the larger sunspots. These magnetic fields often show extremely large gradients over short distances. The active centers that show the most complex sunspot groups, and that produce the most flares, appear to be those whose magnetic field structure is the most complex. It is not unusual for a single active region to have several different centers of strong magnetic field strength, with complex arrangements of polarity. Electronic magnetographs similar to those developed for weaker fields by the Babcocks are now being applied to detailed study of the structure of active region magnetic fields within the photosphere. Promising new techniques are also being developed for detection of photospheric magnetic fields as well as for chromospheric and coronal magnetic fields [30–32]. Figure 15 exhibits the magnetic field pattern around a sunspot group and plage area as observed by Leighton [32], using a photographic technique.

Speculation regarding the nature of the magnetic fields in the atmosphere

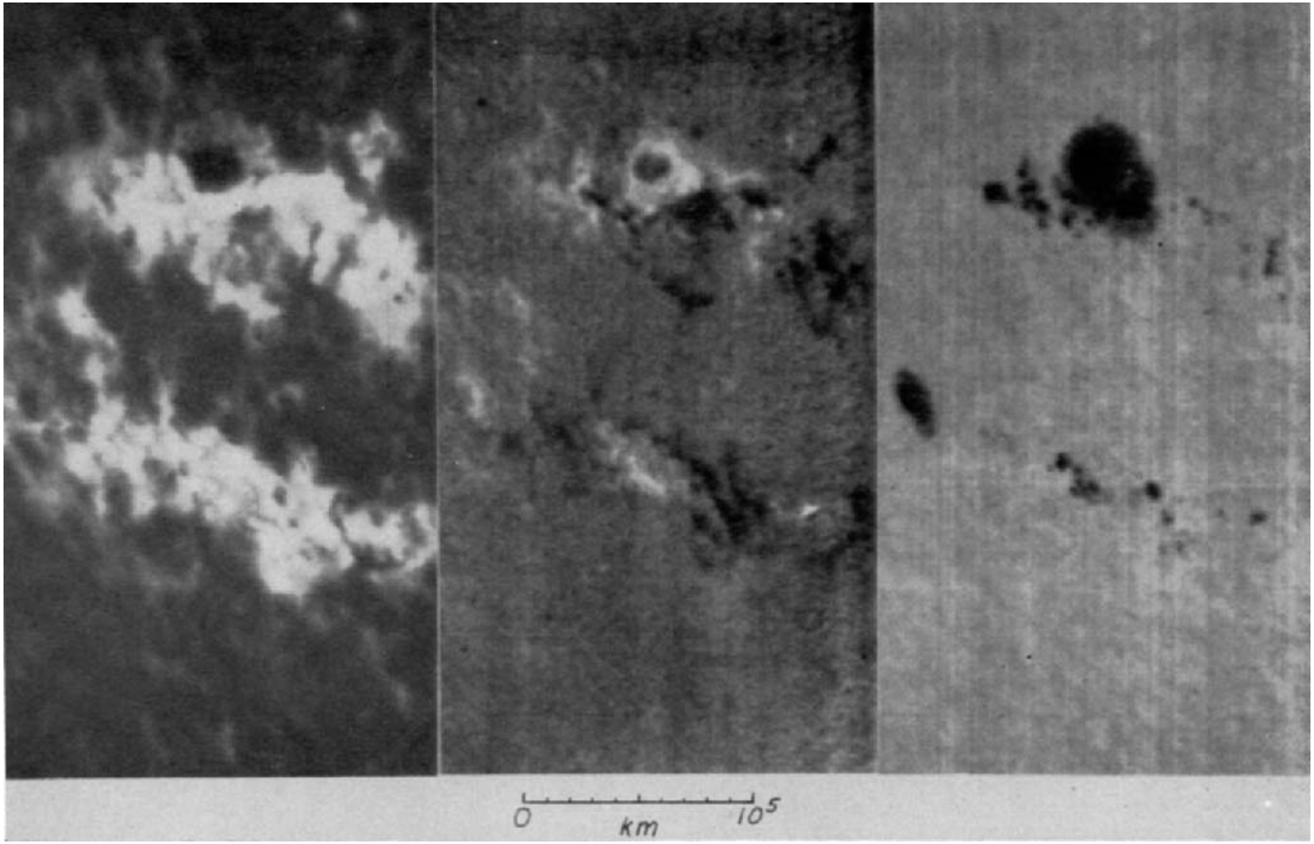


FIG. 15. The magnetic field pattern of an active region observed by a photographic technique [32]. The center picture shows the field pattern with bright and dark features indicating opposite polarity. The right-hand picture shows the spot group as observed in one wing of a Zeeman-sensitive line (λ6102.8 of Ca I), and the left-hand picture shows the plage structure in the K line of Ca II. Note the remarkable similarity between plage and magnetic field patterns (courtesy R. B. Leighton, California Institute of Technology and Mount Wilson Observatory).

above active centers is of great interest. Work by some authors [32] suggests that, at least in some instances, the atmospheric magnetic fields over active regions resemble dipoles with vertically oriented axes, and that they possess a large-scale continuity. Solar radio noise polarization phenomena also suggest that the atmospheric magnetic fields may be considerably simpler than the photospheric fields and that they possess a single polarity over large regions. This of course would not be entirely unexpected if one polarity predominated in the center of a large sunspot group with the lines of force closing to the opposite pole only at some distance from the center itself. Though all interpretive work on solar magnetic fields above the photosphere is as yet speculative, it is a subject of very great importance. In it are likely contained the nature of solar corposcular emission; the acceleration of solar cosmic rays; problems of conductivity and possibly heating of the coronal and chromospheric plasma; the production of spicules, flares, and plages; and the motions, formation, and support of prominences.

The unipolar magnetic field regions on the solar disc appear to be anomalies in the magnetic field structure which are not usually associated with any visible features of the active sun. They sometimes lie away from active centers and cover areas up to some 250,000 km in diameter with field intensities of a few gauss. The Babcocks have suggested that these unipolar magnetic field regions may be the long sought solar "*M* regions" postulated by Bartels in 1934 on the basis of the 27-day recurrent storms in the earth's magnetic field. However, the data on which this suggestion is based are not adequate to firmly establish such a conclusion.

3.9. Variable Emission in the Radio Spectrum

The recent development of observations of solar radio emission is adding tremendously to our knowledge of the outer layers of the sun's atmosphere. This is because the corona, highly ionized as it is, is practically transparent and hence difficult to observe at most optical frequencies. However, the coronal electron gas is highly efficient in radiating, absorbing, and refracting electromagnetic waves at radio frequencies. A study of solar emission at these frequencies brings out information about the density, temperature, and structure of the corona and chromosphere and provides a semi-independent check on models derived from optical data. Of especial importance is the fact that emission in the radio spectrum is sensitive to the same sort of changes in coronal and chromospheric conditions that bring about variations in short-wave and corpuscular emission, changes that may produce very little effect on visible features. Thus solar radio flux, observed with relative ease at the earth, gives quite reliable clues as to the behavior of other solar emissions which, while important in their terrestrial effects, are observed only with great difficulty.

3.9.1. Long-lived Radio Emission. At wavelengths of 1 cm to 1 meter, the total solar flux changes by as much as a factor of 2 over a solar rotation period, showing pronounced 27-day periodicity, and also changing in step with solar activity over the sunspot cycle [33, 34]. This is the slowly varying component. With a radio interferometer, the size of the regions of enhanced emission and their position on the sun can be determined. The emission centers are found to have diameters of the order of 5 to 10′ (the sun's apparent diameter being 30′) and to occur over chromospheric plages at a height (deduced from the rotation rate) of about 40,000 km [35, 36]. The lifetime of one of these centers is several months.

This type of enhancement can be explained as thermal emission—the result of the acceleration of free electrons at encounters with positive ions. At these low frequencies, the Planck function reduces to the Rayleigh-Jeans approximation, in which the intensity of emission at a given frequency varies directly as the temperature. The optical depth changes with the square of the electron density, and inversely as the temperature to the $3/2$ power [37]. For small optical thickness of the source we find, surprisingly, that the intensity varies inversely as the square root of the temperature, while in the case of large optical depth, the intensity varies directly with the temperature. In either case, the intensity of thermal emission increases directly with the square of the electron density. Newkirk [38] has given a fairly detailed explanation of the characteristics of the slowly varying component by assuming thermal emission from a coronal region of high electron density and from the underlying plage. Electron densities need to be about double those of the normal corona through an extended region, which may be identified with a coronal streamer. Temperatures are comparable to normal coronal temperatures.

There are also smaller, more transient features observed in the wavelength range 1 to 10 cm. These have diameters of the order of 1′ and lifetimes of hours or days [36]. These features may also be due to thermal emission, but from smaller regions with higher values of density and temperature.

At metric wavelengths, from 1 to 5 meters, there are also localized regions of enhanced emission, but these seem to have an essentially different character from the slowly varying component of the decimetric region. They have diameters of 1 to 10′, and lifetimes of one to several days. The emission from these "*R* regions" (also called "noise storms") [39] shows many narrow-band bursts, superimposed on a slowly changing continuum, that extends over a broad frequency band. These regions show a general correspondence to optically active regions and appear to be at heights of $1/3$ to 1 solar radius above the photosphere, but they may wander over the solar disc from day to day, making somewhat difficult their identification with a given optical region.

At the longest wavelengths observed—15 to 20 meters—is seen still another type of emission region, with lifetime of a day or two. These regions occur less frequently than the *R* regions, usually show a bursty character, and have been observed as far as 3 solar radii above the sun, sometimes showing rapid and irregular motions [40].

The enhanced emission at wavelengths of about a meter or longer can hardly be explained as thermal emission, since temperatures of the order of 10^9 and more would be required. Even higher temperatures are suggested by some of the bursts to be described below. For these events, a mechanism like the thermal one, in which radiation is continuously emitted, absorbed, and re-emitted, until it finally escapes from the solar atmosphere seems hopelessly inadequate. What is needed is some special region, or particular group of particles, in which the electrons are moving in an organized fashion, so that their separate radiations not only add together in a limited range of wavelengths and directions but can pass through the surrounding atmosphere of normal electrons without being heavily absorbed. Four such non-thermal mechanisms have been suggested to explain various high-energy radio emissions; plasma, gyro, synchrotron, and Čerenkov radiation. The first, plasma radiation, comes about when electrons are caused to oscillate in phase. The frequency of the oscillation is directly proportional to the square root of the electron density, so that in the corona different frequencies are emitted and absorbed at different heights. The decrease in density with height in the corona allows radiation emitted at a given point to pass outward through regions of lower density without being heavily absorbed. However, to explain by plasma radiation the high intensities observed, one needs to assume some especially efficient process for channeling the energy of the electron motion into the form of electromagnetic radiation.

In gyro radiation, electrons spiral around lines of magnetic force with a frequency that depends directly on the field strength. Again, to attain intensities higher than that of thermal radiation, the emitting electrons must move together and differently from the electrons of the surrounding medium. Unless the surrounding medium is so low in density that its absorption can be ignored, some further special condition—a sharp change in density, strength, or direction of the magnetic field—is necessary in order for this radiation to escape from the solar atmosphere.

In synchrotron radiation, electrons spiral around lines of magnetic force but with velocities comparable to that of light. Then many harmonics of the gyro frequency are radiated, so that an essentially continuous spectrum is emitted. Propagation of this emission is easy to achieve, once the problem of accelerating the electrons to such high velocities has been met.

Čerenkov radiation also requires high-velocity electrons, exceeding the velocity of light in the medium. This is possible when the refractive index is greater than 1. With this mechanism, too, the problem lies in the high electron velocities required, though also problems in propagation arise because of the high refractive index required for emission. It is quite possible that each of these four mechanisms is operative in one or another type of radio emission.

3.9.2. Solar Radio Bursts. Several schemes of classification have been developed for bursts at fixed frequencies. These are generally described in the data sources, and we will review here only the types recognized on radio spectrum records, where the time variation of intensity of radio emission is displayed over a broad range of frequencies.

We have already mentioned the noise storm. The bursts that occur during this event are called type I or storm bursts. They appear most frequently at wavelengths longer than 1 meter. Storm bursts extend over only a narrow frequency bandwidth, perhaps 5 to 30 Mc, while their lifetimes are 0.2 to 0.3 sec up to nearly a minute [41, 42]. During the lifetime of a single burst, the frequency of emission may drift with time either to higher or to lower frequencies, the direction of drift depending on the position of the source [42].

A schematic representation of other types of solar burst is shown in Fig. 16, in the sequence that tends to follow certain large solar flares. Type III bursts are very frequent, and many occur outside of such sequences, while type IV is relatively rare, and will often be lacking.

The chief characteristic of type II (slow drift) and type III (fast drift) bursts is the drift of the emission toward lower frequency with time. This drift can be explained as due to a source moving outward through the corona, emitting at the plasma frequency, which decreases as the electron density decreases. Interferometer observations show that the sources actually do move outward, with velocities that fit with the density-height

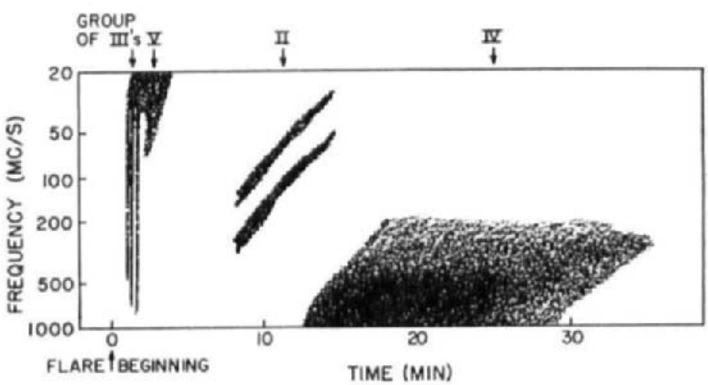


FIG. 16. Schematic representation of radio bursts following a large solar flare.

relation in a coronal streamer, and that the emission at a given height is essentially monochromatic. Velocities for type II bursts are about 1000 km/sec, while those for type III bursts are $\frac{1}{3}$ or $\frac{1}{2}$ the velocity of light [43, 44]. Both these types often show harmonic structure—two similar and simultaneous bursts, one at double the frequency of the other. The intensity of the second harmonic may be as great as that of the fundamental, but no sure case of a third harmonic has been observed.

The type IV event is particularly important in solar-terrestrial relations, as it shows a strong association with solar corpuscular radiation, evidenced by polar cap absorption (PCA) events. First recognized on interferometer records, these long-lived emissions come from a source of large extent that moves in the solar corona to heights of the order of a solar radius [45]. Radiation occurs over a broad band of frequencies at each height [43]. The identification of these events with the long-lasting continuum that sometimes follows a type II burst and with high intensity, long duration bursts in the centimeter range seems quite certain [46]. Probably the noise sources at low frequencies are also related to type IV events. A particularly strong example of type IV emission occurred on August 22, 1958, accompanied by a cosmic-ray increase at high altitude. In this case, intense continuum emission occurred over the whole range of radio frequencies, drifting slowly toward lower frequencies with time [47]. The characteristics of type IV radiation can be explained by synchrotron emission of electrons, trapped in a magnetic cloud that is close to the sun but able to move about in the corona [48].

Type V emission is continuum emission of shorter duration, following a type III burst. Its characteristics outside the frequency range 40 to 200 Mc have not been described, but it seems to be similar to type IV [43].

Still another type of burst is the U-burst, which starts out like a type III burst but turns over at some frequency (often around 150 Mc) and then drifts back toward higher frequency [41].

Detailed studies of solar radio bursts and their relation to optical events is rapidly increasing our understanding of the solar corona and of the high-energy particles that must occasionally speed through it, while their association with terrestrial effects gives us new leverage on the problem of solar outbursts of short-wave and particle radiation.

4. Indirect Indices of Solar Activity

4.1. Introduction

Reactions of the earth's magnetic field and ionosphere to changes in solar radiation provide a number of indirect indices of solar activity. Some observable change in physical conditions on the earth is taken as a measure of an unobservable solar emission. Often the physical connection between terrestrial change and solar radiation is uncertain, and we can only guess as to what type of radiation is being measured. This unsatisfactory state of affairs is a natural consequence of the fact that indirect indices are most useful for the parts of the solar spectrum that cannot be measured directly from the earth's surface—the ultraviolet and x-ray regions and particle radiation.

Derivation of an indirect solar index can fall into three parts: (*1*) demonstration of a correlation between the suspected indirect index and some direct solar index; (*2*) theoretical arguments as to what solar radiation produced the variations, and why the indirect index may be expected to be a valid measure of this radiation; (*3*) arguments that the indirect index is a better measure of the indicated radiation than is the direct solar index with which it was shown to be related.

With such a necessarily indirect and tenuous approach it is not surprising that recent direct observations from balloons, rockets, and satellites have occasionally upset classical ideas as to the meaning of indirect indices more than they have immediately clarified them. The importance of such direct observations to basic interpretation is obvious, and the problem of fitting together necessarily fragmentary direct observations with plentiful but difficult to interpret indirect measures forms a fascinating puzzle in contemporary solar and terrestrial physics.

4.2. Indirect Measures of Solar Ultraviolet and X-Radiation

4.2.1. Long-Term Changes. Many studies have been made of the variation of geophysical quantities over the cycle of solar activity. In general, monthly mean values are used in these studies, or even annual or 13-month running means. Thus the general rise and fall of solar activity is shown without the details that are most useful in differentiating the behavior of one type of solar activity from another.

Measures of ionospheric densities in the form of critical or penetration frequencies have long been used as an indication of the state of the ionosphere and therefore of the radiation incident on it. A plot of depth of penetration against wavelength of solar ultraviolet and x-radiation shows plateaus where the height changes only slightly with wavelength [49]. These occur at heights corresponding to those of ionospheric layers or maxima in electron density. These maxima can be detected routinely by sounding the ionosphere with radio waves. A typical vertical incidence ionogram presents the effective height of reflection of a radio wave as a function of its frequency. A transmitter beamed directly overhead emits a signal that sweeps toward higher frequencies with time. The signal is reflected down again to an adjacent receiver that records the reflected signal at each frequency along with the delay time from the transmitted signal. This delay time is proportional to the "virtual height" of reflection, which is the quantity appearing on the vertical scale of the ionogram.

On the ionogram one sees a curve that shows flat portions, with sudden increases of virtual height at the critical frequencies of the E, $F1$, and $F2$ layers. The square of the frequency of the reflected wave is a direct measure of the electron density at the point of reflection, which is a point of maximum electron density for a critical frequency.

Interpretation of the virtual height h' is not so simple. The group or signal velocity of the wave decreases sharply near the point of reflection, so the time lag is related in a complex way to the distance traversed by the wave. Derivation of true heights of reflection requires numerical solution of an integral equation relating virtual height to electron density at each height. Use of electronic computing machines has made this solution possible on a fairly large scale, but true height data are still much scarcer than data on critical frequencies and virtual heights. The critical frequency of the F_2 layer, $foF2$, shows a very definite and regular change with solar activity over longer periods, and, in fact, is more sensitive than foE and $foF1$ to these long-term changes in solar radiation. Allen's [50] study of the relation of ionospheric critical frequencies to a number of solar indices is perhaps the most illuminating single study of this type. Working with monthly mean indices, he determined the time lag that gave the highest correlation with sunspot number for each of the solar and ionospheric indices. Comparison of these time lags showed that the E layer varied in phase with plages and faculae, while the $F1$ and $F2$ regions were more closely related to solar features that persist even longer in the life of the active regions.

Various studies have been made relating $foF2$ to sunspot number. Over these long periods, the relation is clear and approximately linear for moderate sunspot numbers. For very high sunspot numbers, $foF2$ levels off at a nearly constant value [51]. Minnis [52] has formalized the long-term quasi-linear relation by expressing the mean value of $foF2$ on the same scale as sunspot numbers. However, the relation is different for different data samples [53], showing that sunspot number is probably not a completely adequate measure of the ionizing radiation. The relation between 10-cm flux and $foF2$ shows a similar nonlinearity but less scatter than when sunspot number is used as the direct solar index [34]. The dependence of foE and of foF on sunspot number disappears in the auroral zone and reappears in the polar caps [54].

The earth's atmosphere moves under the action of the sun and moon. The effect can be seen in a daily cycle in barometric pressure, which differs from ocean tides in that the variation tends to stay in phase with the sun rather than with the moon. The importance of the solar influence may result from the thermal action of the sun on the atmosphere, or from the temperature structure of the atmosphere, which gives it a natural oscillation period close to the 12-hr period of the solar gravitational effect. As the charged particles of the ionosphere are pulled in this tidal motion across the lines of force of the earth's magnetic field, an electric dynamo current is set up, and this can be measured at the surface of the earth as a change in intensity and direction of the magnetic field. This change, the S_q variation, is especially marked near the geomagnetic equator. The diurnal and semi-diurnal changes show solar and lunar components much like the tidal pressure variations though complicated by the additional dependence on geomagnetic coordinates. Because the geomagnetic field is subject to major disturbances, the diurnal variation, even at the equator, can be measured only on geomagnetically quiet days. The amplitude of the daily variation might be expected to vary with the number of ions involved in the dynamo action, and so with the electron density, as well as with velocity of the electron motion.

The daily amplitude of S_q shows a pronounced and regular variation with sunspot number over the solar cycle [55], and monthly means of *foE* also closely follow the variations of sunspot number and 10-cm flux [34]. Again, a plot of *foE* against 10-cm flux shows smaller scatter but larger departure from linearity than a plot of *foE* against sunspot number.

Sunspot number (assumed to measure the ionizing radiation) varies roughly as the fourth power of the critical frequency, or the square of the electron density, in the *E* and *F*1 regions, but as the square of *foF*2, or directly as the electron density, in this highest region. This does not necessarily mean that the radiation effective in ionizing the *F*2 region is of a different type, whose intensity varies in a different manner from the radiation ionizing the *F*1 region. The effect can arise through different recombination processes. An ionized atom of oxygen does not combine readily with an electron—much more probable is charge exchange between the atom and a molecule, followed by dissociative recombination of the resulting molecule-ion. If the first of these reactions is rate-limiting, the electron density varies directly as the ionizing radiation; whereas if the dissociative recombination is rate-limiting, the density varies as the square root of the radiation [101].

In the region of interest (from about 100 to about 910 A), line emission occurs down to 912 A from the hydrogen Lyman series, in the range 500–600 A from neutral helium, and 200–300 A from ionized helium. Observations of the emission line of He II at 304 A [56] suggest that this may be an important component of the radiation ionizing the *F* region.

In the *E* layer, rocket observations of radiation intensity and calculated models are in agreement if the ionizing radiation falls in the soft x-ray region.

Although these studies using long-term averages are useful in demonstrating solar control of ionospheric processes and also in predicting ionospheric conditions on the basis of solar activity, they are not in general particularly helpful in clarifying details of solar-ionospheric relations, since all solar indices vary so nearly together. As can be seen from the above discussion, the principle has usually been to assume a direct solar index such as sunspot number as a measure of the solar ionizing radiation, rather than to use the ionospheric index as a measure of the unobservable solar emission.

4.2.2. Variations over a Day to a Month. Many terrestrial phenomena show a well-defined change over the 11-year solar activity cycle, and such short-term phenomena as ionospheric fadeouts show an almost perfect correspondence with solar flares, but day-to-day relations are fewer and less definite. A 27-day periodicity in geophysical data, corresponding to the rotation period of the sun, in itself demonstrates an influence of solar active regions. Examples of such variations are found in the critical frequencies of the ionospheric E and F layers, in the amplitude of diurnal variation of the S_q ionospheric current, and in rate of deceleration of earth satellites.

The latter relation, the most recently discovered, shows 27-day waves in the rate of change of period of the satellite's orbital motion, superimposed on the general decrease in period as the satellite falls toward the earth. Although there is a possible "charge drag" on the satellite, the period changes have been generally interpreted as the effect of changing atmospheric density on the frictional force tending to decrease the satellite velocity. The force is such as to decrease the velocity, but the dimensions of the orbit decrease so that the period also decreases. Because atmospheric density decreases very rapidly with height, the major effect of atmospheric drag on the satellite's motion occurs close to the perigee point. The rate of decrease of period is approximately proportional to the atmospheric density at perigee and to the square root of the logarithmic decrement of the density. (It depends also on characteristics of the satellite and its orbit.) The periodic variation tends to appear most strongly when satellite perigee occurs on the sunlit side of the earth [57], a fact that suggests that wave radiation is being measured rather than particle radiation. The variation is detected for satellites with perigee heights from 200 to 600 km and is stronger at greater heights. It indicates a higher atmospheric density at a given height when solar activity is high. This is interpreted as a general expansion of the atmosphere due to heating as a larger amount of radiation is absorbed [57–60].

According to an analysis by Johnson [61], the solar radiation most effective in heating the atmosphere lies in the wavelength range between 200 and 900 A. A likely source of variable emission in this region is the line of He II at λ304 A [56].

Simultaneous analysis of the motions of several satellites should make it possible to disentangle temporal variations from the variations due to changes in position of perigee point, and to derive an accurate description of atmospheric density as a function of height and latitude, with its diurnal variation and dependence on solar emission.

Ionospheric critical frequencies again are assumed to measure the ionizing radiation for the corresponding layer, and we may speculate that the amplitude of S_q varies with conditions in the E layer. The $F2$ layer shows large and erratic changes from day to day, partly because of effects of particle emission. In spite of its close relation to solar activity over the sunspot cycle, $foF2$ does not provide a good index of shorter term changes in solar short-wave radiation. Nevertheless, a small (10%) variation with 27-day periodicity has been detected in the F layer [62]. An index devised by Van Zandt [58], based on ionospheric true height data through the F regions, appears more promising. This index shows considerably more stability than $foF2$ and appears to vary similarly to satellite deceleration, as it should if both are measures of atmospheric heating.

Daily values of foE, when reduced to the subsolar point (removing seasonal and geographical variations dependent on the angle of incidence of the solar radiation), show a clear 27-day variation in phase with satellite deceleration and with 10-cm solar flux. Noontime absorption in the D layer varies with a 27-day period but also shows geographic and seasonal changes that have not been well determined as yet [63].

Day-to-day variations in the amplitude of the S_q current are more complex. Changes in geographical location together with amplitude of the S_q current system have been related to solar activity only after careful consideration of seasonal effects. The relations are interpreted in terms of variations in the velocity of the ionospheric wind as well as in electron density [64].

4.2.3. Variations over Minutes to Hours. Simultaneously (within minutes) with certain solar flares, electron density below the E region increases sharply. Evidently there is a significant increase in the amount of the short wavelength, high-energy solar radiation that can penetrate the normal ionosphere. Various observational aspects of this phenomenon have received separate names, and these are listed below.

Sudden ionospheric disturbance (SID) or *perturbation ionosphérique à début brusque* (PIDB) includes the whole complex of observed ionospheric changes associated with a solar flare.

Short wave fadeout (SWF) describes the effect on long distance radio communication. Radio transmissions at short and medium wavelengths (frequencies down to about 2 Mc/s) that are normally reflected from the ionosphere are absorbed, and on long distance radio communication circuits the signal suddenly becomes weaker or disappears. The fadeout is characterized by the intensity or amount of fading of the signal, by the geographic extent of the effect (which is confined to the sunlit hemisphere), and by the suddenness of its onset.

Sudden phase anomaly (SPA) shows the change in phase of a received very long wave due to a decrease in height of reflection. For these waves, with frequencies of the order of 15 to 100 kc/s, and for oblique incidence, the signal strength is enhanced. The height of reflection decreases by as much as 15 km (usually about 4 km) from its normal value of around 75 km.

Sudden enhancement of atmospherics (SEA) describes the increase in amplitude of received very low frequency radio signals (around 30 kc/s) originating in atmospheric disturbances such as thunderstorms, as the ionosphere becomes a better reflector for these very long wave emissions.

Sudden cosmic noise absorption (SCNA) measures the decrease in intensity of galactic radio noise of about 20 Mc/s frequency when absorption increases in the lower ionosphere. Receivers designed to monitor the level of galactic radio noise are called indirect flare detectors, because essentially all SID's are associated with solar flares and so offer a means of detecting the occurrence of a flare in the absence of visual observations.

Solar flare effect (SFE), or magnetic crochet, shows the effect on the geomagnetic field of ionospheric currents set up when the ionization suddenly increases during an SID.

In an SID, the ionospheric changes take place so suddenly that the usually assumed ionospheric model of a "Chapman layer" becomes inapplicable, and special approximations and models must be developed. Using an exhaustion model, Bracewell [65] estimated that an increase of incident ionizing radiation by a factor of 15 would account for an observed decrease in height of the reflecting layer by 15 km. Friedman *et al.* [66] estimate the total Lyman-α flux would need to increase by a factor of 10^4 during a flare to account for the height decrease.

Until recently, Lyman-α radiation was considered to be the principal ionizing radiation and nitric oxide the principal source of electrons in the D region during an SID. Close and detailed relations between SID's and the behavior of Hα emission in flares led to this conclusion, which was strengthened by agreement between the observed height of enhanced ionization with the depth of penetration of Lα radiation. But rocket observations present quite a different picture. Because an SID occurs suddenly and to some extent unpredictably, direct measurement of the ionizing radiation with the help of rockets presents a formidable problem. Nevertheless, workers at the U. S. Naval Research Laboratory have on several occasions succeeded in making such measures during solar flares and have consistently detected x-rays [66]. Fitting this observation to solar and ionospheric theory may drastically change ideas about flare emission and ionospheric reaction.

Analysis of the change of absorption with time during SCNA's suggests that the ionizing radiation occurs in a "burst" near the beginning of the Hα flare [67, 68]. The sharpness of the onset of absorption and the shape of the recovery indicate an abrupt increase in intensity of ionizing radiation, followed by a rapid decay. According to this picture, the "burst" of radiation would have been concluded before any of the existent rocket observations had begun. Attempts to measure solar ultraviolet and x-radiation from satellites have been hindered by radiation from the Van Allen belts, which saturates the instruments. Such observations during the initial phase of a flare and SID would be of great importance and, it is hoped, will be forthcoming with the use of spectrographic techniques in low-altitude satellites.

Recent observations suggest that more than one type of radiation plays a role in the SID phenomenon. Continued enhancement of E layer ionization after the end of the D layer absorption of SWF's [69], a difference in recovery rate of ionospheric sounding echoes at different heights [70], and somewhat different flare relations for SWF's classified as sudden or gradual [71] all indicate that two types of enhanced short-wave radiation may occur at the time of a solar flare and that these differ in the amount of intensity increase over normal, in their decay rate, and in the height of absorption in the ionosphere and height of emission in the solar atmosphere.

Changes in the E region during a solar eclipse cannot be explained completely by simple models of the ionosphere and of the sun. The observations have been explained either by a time or height (temperature) dependency of the recombination coefficient [72] or by assuming that the ionizing radiation does not correspond exactly to the visible solar radiation but shows different center-limb variation or is preferentially emitted from active regions.

The latter explanation seems the most satisfactory in most cases. The picture of the sun in ionizing radiation that has been built up from eclipse observations of foE has about 10% of the total radiation coming from regions beyond the visible limb and up to 25% from limited areas corresponding to plages [73]. This picture is remarkably like a direct photograph of the solar disc in the light of 50 A x-rays that was recently obtained from a rocket [74]. A model in which emission of ionizing radiation is enhanced in the neighborhood of solar active regions is consistent also with the 27-day and sunspot cycle variations of E region densities.

4.3. Indirect Measures of Solar Particle Emission

Variations in the earth's magnetic field have been measured for more than a century, and the larger disturbances have been almost universally interpreted as the result of an encounter between the earth and a stream or cloud of charged particles from the sun. More recently, measurement of the velocity of auroral protons and the direct observation of such particles from balloons have shown that energetic charged particles actually do penetrate the earth's atmosphere at times of geomagnetic disturbance. However, the discovery of the Van Allen radiation belts, which provide possible storage and acceleration of particles, has led to a review of the evidence for direct influx of solar particles.

Perhaps the most certain proof that particles sometimes come directly from the sun to the earth is provided by a class of event that was described and extensively observed only recently—the polar cap absorption event [75–79]. The effects of solar particles and, from balloons, the particles themselves are observed in striking association with large solar flares, less than an hour after the beginning of the flare. This very close time association leaves little doubt that solar particles are emitted from the sun at the time of the flare and travel to the earth with little delay.

Another quite direct observation of solar particles has been made recently by Blackwell and Ingham [80]. Measuring the intensity of the zodiacal light, he found that the intensity increased and the position of maximum intensity changed at the time of a major geomagnetic storm. This, too, provides quite direct evidence of particles moving between sun and earth on at least one occasion.

Other direct observations of energetic particles in the earth's atmosphere include the Doppler displaced hydrogen emission in the aurora and balloon observations of changes of flux of protons and of x-rays in detailed relation with geomagnetic and auroral variations. The possible role of Van Allen radiation in these cases has not yet been fully assessed, so one cannot be completely sure that these particles represent directly the time dependency of solar emission.

4.3.1. Particle Emission under Undisturbed Conditions. Even when solar and geomagnetic activity are both very low, an influx of particles at the earth is recorded by the presence of aurora and irregular variations in the earth's magnetic field, especially in the auroral zones, the regions near 68° geomagnetic latitude. One may speculate that solar particles stored in the outer Van Allen belt tend to escape slowly from the horns, or mirror points, of the trapping region. Reaching the atmosphere, they excite the visible aurora. Variations in the intensity, location, and direction of electric currents in the ionosphere should then explain the magnetic fluctuations. In this case, the relation of these "quiet" phenomena to solar particle emission would not be very precise over short periods of time. A difficulty with this model is that the latitude of the mirror points of the Van Allen zone, as determined from satellites, is at a lower latitude than the auroral zones.

The realization that the space between sun and earth is not a vacuum has led to several theoretical studies of the particle emission one might expect from the quiet sun. Chapman has attacked the problem from the point of view of heat conduction, assuming hydrostatic equilibrium and negligible heat loss by radiation from the outer corona [81]. He concludes that the sun is surrounded by a hot, tenuous envelope, an extension of the solar corona well past the orbit of the earth. Parker has considered the "solar wind" that results if one considers the corona as a gas in hydrodynamic equilibrium [7, 82, 83]; he finds a continuous radial outflow of solar material with considerable density and velocity. Chamberlain [84], approaching the problem as one of evaporation or escape of the highest energy particles from a gaseous sphere, finds an outflow of material from the sun with more modest velocity values.

The observational evidence for such a steady corpuscular emission lies mainly in the tendency of comet tails to stream away from the sun. The effect of pressure of electromagnetic radiation has been calculated to be insufficient to explain this tendency, and a rather large flux of particle emission would be required [85]. Observations of comets are of course infrequent, and individual cases are difficult to interpret, but there is some evidence for changes in the form or intensity of comets at the time of solar or geomagnetic activity, as would be expected if their form is largely controlled by solar particle flux.

A doubly indirect measure of corpuscular emission may be provided by the intensity of cosmic rays received at the earth. Assuming that a uniform flux of cosmic rays is present, one may interpret the variation at the earth in terms of a magnetic shield formed by field-containing clouds ejected by the sun. A recent observation of such a Forbush decrease at a distance of eight earth's radii shows that the shield probably has dimensions like the inner solar system [86]. When solar activity is high, more of these solar clouds are emitted, the shield becomes generally more effective, and fewer cosmic rays are observed. Although only the magnetic field and not the particles themselves could affect the cosmic rays, the particle clouds must be present if the fields are there, so we can conclude that an isotropic cosmic-ray flux from outside the solar system varies at the earth in some way inverse to solar particle emission.

Geomagnetic micropulsations are regular, wavelike oscillations of the earth's field, of shorter period and smaller amplitude than disturbance variations. Recent theory has linked this phenomenon to the effect of relatively low-energy particles impinging on the interface between the earth's atmosphere and the interplanetary medium. Observations suggest that there are two types of pulsations. One, *Pc*, has periods in the 1–30 sec range, amplitudes of about 0.1γ, and may last for hours. These occur primarily in the daytime. The *Pt* type has a longer period and larger amplitude and is heavily damped, lasting only 10 min or so, and tends to occur at night. Giant micropulsations, with amplitudes as great as 40γ occur in the auroral zone. (The amplitude of geomagnetic disturbance is about 150γ or more.)

The characteristics of these oscillations, especially the *Pc* type, and their dependence on latitude and local time have been the subject of much recent research. Twenty-seven day periodicity in frequency of occurrence and a general association with geomagnetic disturbance suggest a solar dependence. A close relation with fluctuations of auroral intensity [87] has also been observed on occasion. With additional observations and further interpretation, micropulsations may turn out to be a valuable index of low-energy corpuscular radiation from the sun.

Another phenomenon that shows promise of measuring the effect of solar particles at distances of several earth's radii is that of very low frequency (VLF) emission. These are bursts, or broader band emission ("hiss") in the 2–8-kc/s range of frequencies. Very low frequency emission shows a general correspondence with geomagnetic disturbance, and on some occasions, a minute-to-minute relation with airglow or auroral emission at $\lambda 6300$ [88, 89].

These emissions may be due to solar particles at several earth's radii. The hiss may be related to a steady stream of solar particles, while the narrow frequency range of the discrete bursts suggests small clouds of particles, only about 100 km in diameter. Direction finding techniques applied to VLF emission have measured a source size of the order of 500 km [90]. Interpreted as the effect of interaction between solar corpuscular clouds of ionized particles and electromagnetic radiation with a phase velocity similar to the particle velocity, these emissions permit the derivation of velocity for the charged particles, which is of the order of 10^4 km/s [88]. This is like the velocities deduced for auroral protons, and an order of magnitude higher than velocities corresponding to the time delay between solar flares and geomagnetic disturbance.

Studies of airglow, micropulsations, and VLF emission during the coming period of low solar activity may be expected to contribute greatly to our knowledge about particle emission of the quiet sun.

4.3.2. Effects of Enhanced Corpuscular Emission at Time of a Solar Flare. We will describe here an idealized event, comprising the outstanding features of the terrestrial effects of a sudden enhancement in solar particle emission. Such an event can usually be related unambiguously to a solar flare, impressive for its emission in Hα and at radio frequencies (see interrelations). However, the reader should be aware of the fact that none of the suggested relations is without exception.

Less than an hour after the initiating flare, an effect called "polar cap absorption" may commence. This recently distinguished effect provides striking evidence for particles coming directly from the sun to the earth. This event, of which about 24 examples were detected during IGY [77], is characterized by strong absorption of radio noise from the galaxy, and of very high frequency (VHF) forward scatter signals, with sometimes an early enhancement of the latter. In contrast to auroral absorption, this type shows no sudden temporal changes or evidence of patchy structure and no simultaneous auroral or geomagnetic changes. It occurs within the polar cap and cuts off sharply at the auroral zones, though on some occasions it has been observed at lower latitudes during magnetic disturbance. Absorption sets in within a few hours or less of certain major solar flares and continues for several days, although usually auroral absorption, accompanying geomagnetic disturbance, confuses the observation after a day or two. Polar cap absorption, which has been called type III absorption, shows a strong diurnal variation; but this need not require any diurnal variation in ionizing flux. It can be explained by the nighttime loss of electrons by attachment, a process which is countered by photo-detachment during the day [91]. Consideration of the time of transition between daytime and nighttime absorption values, compared to the time of sunrise and sunset, shows that the absorption occurs at a height as low as 50 km [76].

The depth of penetration (height of absorption) and the amount of absorption, the time relation to the flare, and the geographical distribution were shown to be consistent with the effect of small numbers of protons of high energy (100 Mev). This result was beautifully confirmed by balloon observations, which detected directly the presence of such protons. Their energy spectrum has been measured; the number of protons is inversely proportional to about the fifth power of the energy, varying somewhat from event to event. The balloon observations also showed that protons continue to arrive at the earth for several days. In one case, protons were still detected 10 days after the beginning of the event, although the active solar region in which the initiating flare occurred had by then rotated to the opposite side of the sun [79]. The proton flux decayed regularly with time during the whole period. The close time relation with flares suggests that such high-energy particles can only be produced at the time of a flare. The smooth decay, and the presence of absorption over the polar cap are evidence against storage at the earth, at least in the Van Allen regions. It seems likely that in this event particles were stored in the vicinity of the sun. If this region rotated with the sun, it must have been large enough and far enough from the sun that the fact that the sun lay between the emitting region and the earth had little effect on the particle flux.

Extreme examples of the emission of high-energy particles from the sun are the solar cosmic-ray events, of which only 10 have been observed at

TABLE II. Solar cosmic-ray events.

Date, UT	Associated flare or SID	
	Beginning	Position
1942 Feb. 28 1000–1200	Before 1100	N06 E05
1942 Mar. 07 0400–0600	About 0440	N06 W90?
1946 July 25 about 1700	1504 or 1616	N21 E16
1949 Nov. 19 1045	Before 1029	S03 W72
1956 Feb. 23 0345	Before 0334	N23 W80
1959 July 17 before 0200	July 16 2114	N15 W30
1960 May 04 1031	1015	N12 W90
1960 Nov. 12 1342	Before 1323	N26 W04
1960 Nov. 15 0242	0207	N26 W33
1960 Nov. 20 2100	2018	N25 W90

the ground (see Table II in Section 5). In these cases, the energy of the particles is so great that they reach the surface of the earth and are detected as a sharp increase in cosmic-ray intensity. The most thoroughly studied solar cosmic-ray event is that of February 23, 1956. Particle energies are calculated to be as high as 2 Bev [75], and the densities of these most energetic particles are of the order of 10^{-11} cm^{-3}. The very low density explains the lack of geomagnetic effect—these particles move in the earth's field without affecting the field significantly and without interacting with each other.

Data on satellite motions showed two sudden discontinuities in the rate of period decrease in 1958. These effects were associated with major geomagnetic disturbance and appeared only at high latitude, so probably are due to the presence of solar particles [92].

On three occasions in August 1958, changes were observed in the structure of the Van Allen radiation at the time of PCA's. The particle flux increased in the region beneath and toward the pole from the outer zone. All the measures are consistent with the characteristics of PCA protons as deduced from absorption and balloon observations. The relative number of low-energy particles increased with time during these events; this could happen either if the energy spectrum of the incident radiation changes with time, the less energetic particles becoming relatively more numerous late in the event, or it could be the result of magnetic disturbance distorting the earth's field so that particles of lower energy than usual were able to penetrate the magnetic barrier.

During magnetic disturbance itself, changes in the intensity of the Van Allen radiation are rather complex. Basing our description sometimes on evidence from a single case, we may say that the radiation intensity seems to increase at low altitudes beneath the outer zone but seems to decrease in the high latitude part of the outer zone [93, 94]. On the other hand, the average energy of the particles increases in the outer zone. The maximum effect of the decrease seems to occur about a day and a half after the beginning of the storm, while the low altitude increase begins very near the beginning of the storm [95]. In the later phase of the storm, radiation increases in the outer zone [94, 96].

Geomagnetic storms are more frequent than polar cap absorption events, but almost all PCA's are followed after a day or two by particularly strong geomagnetic disturbance. A typical "sudden commencement" geomagnetic disturbance, the type of major storm that shows definite association with solar activity, can be divided into three portions: the sudden commencement (SC), the main phase, and the recovery.

The SC is a sharp increase (or occasionally a decrease) in the horizontal component of the geomagnetic field. The essential property is the discontinuous character of the variation. The form or type of SC depends to some extent on the latitude and local time. Sudden commencements can occur without a subsequent storm: at Greenwich, 35% of SC's are followed by no disturbance or only weak disturbance [97].

The form of the main phase of a geomagnetic disturbance, besides varying from one case to another, depends on latitude and local time in a regular way. The time dependence of magnetic variation can be divided into two parts, one dependent on the local time and one on the time since the commencement of the storm. At subauroral latitudes, the local-time dependent part of the variation, called S_D, is similar in character to the quiet day magnetic variation, S_q, but with a larger amplitude. At higher latitudes, there are strong effects that can be described in terms of intense electric currents flowing across the poles and around the auroral zones. The part of the variation that depends on the phase of the storm, called D_{st}, can be described even more simply as a current flowing from east to west around the whole earth with an intensity maximum at the auroral zone.

The occurrence of visible aurora depends strongly on latitude. In the auroral zones, about 22° from the geomagnetic poles, some aurora may be observed practically every clear night. Aurora is most likely to occur around local midnight, and radar studies have confirmed that the frequency of occurrence is higher during the night than during the day. During a magnetic storm, auroral displays intensify and move to lower latitudes, so that at temperate latitudes the occurrence of aurora is closely related to geomagnetic and solar activity. The IGY observations in the Antarctic show that aurora occurs simultaneously in both auroral zones in 99% of the cases [98] but that the form of the aurora may be different in the two hemispheres.

Ionospheric disturbance accompanies a geomagnetic storm, often interrupting long distance radio communication for periods of hours. Effects are strongest in the F layer; typically, critical frequency decreases and height increases, but dependence on latitude, season, local time, and phase of the disturbance complicates the picture. The main effects are consistent with heating and expansion of the upper atmosphere.

Detailed relations between changes in the magnetic field and the form and spectrum of auroral emission have recently been described [99]. In rough outline an increase in H, the horizontal component of the earth's field, is followed by the appearance of a red arc, with intense hydrogen emission, which stretches longitudinally across the sky. As the night progresses, the arc breaks up to form rays—long, thin formations aligned with the geomagnetic lines of force. This is accompanied by a decrease in the relative intensity of hydrogen emission and by the occurrence of a negative bay in H.

The hydrogen emission in aurora provides an important piece of evidence for incoming solar particles, in that the Doppler displacement of the emission line shows that the hydrogen atoms are moving along the magnetic field lines with velocities up to 3000 km/sec. Since a fast-moving proton entering the atmosphere can capture an electron and so emit as a hydrogen atom only after it is slowed down to this velocity or less, this value represents a lower limit for the velocities of incoming protons. Thus velocities are deduced that are considerably greater than the velocity of around 1000 to 2000 km/s derived from the delay between a solar flare and the beginning of magnetic disturbance and aurora on the earth. Such high velocities are also needed to explain the depth of penetration or height of occurrence of auroral emission. The difference in velocity values could be explained if the energetic auroral particles were trapped by magnetic fields in a slower moving cloud during their passage from sun to earth or if the particles were accelerated near the earth.

During aurora, in the auroral zones and polar caps ionospheric absorption often occurs. This is known as polar blackout, type II absorption, or auroral absorption. Associated with active aurora and magnetic bays, this type of nighttime absorption tends to recur for two or three nights.

On one occasion when an intense auroral arc was observed at subauroral latitudes, closely corresponding changes were observed in the position of maximum intensity of the outer Van Allen zone, where the "horns" of the zone come closest to the earth [100]. This can be interpreted as the effect of atmospheric heating by an influx of Van Allen particles [101].

In 1958, the artificial introduction of additional high-energy particles into the region between the inner and outer trapping regions by a nuclear explosion resulted in local aurora, magnetic, and ionospheric disturbance closely paralleling the natural phenomena except for the low latitude of the artificial effects.

According to present observations, it seems unlikely that the Van Allen regions could provide enough energetic particles to explain a major magnetic disturbance, and we are probably safe in continuing to assume that geomagnetic disturbance essentially measures the flux of solar particles at the earth. In any case, it seems that we must rely on solar particles, at least as a trigger for the process, to distort the geomagnetic field to the extent that the Van Allen particles can escape into the atmosphere.

Balloon flights have shown bursts of x-rays, with energies of the order of 100 kev, that correspond in detail to the formation or brightening of auroral rays, and, on another occasion, with magnetic variations in the early hours of a magnetic disturbance [102, 103, 104]. Here again, we are faced with the paradox that highly energetic particles that appear to be closely associated with geomagnetic variations and with aurora are observed at the earth 43 hr after the solar flare that seems most certainly associated with the phenomenon. We need to assume either that these fast particles ride a slow magnetic cloud from the sun, that they are accelerated by some special process near the earth, the accelerating process being related closely to the magnetic and auroral variation, or that the flare particles act only as a trigger, letting the energetic particles escape from the trapping regions at the earth.

4.3.3. Magnetic Disturbance at Times of Low Solar Activity. Before the differences between sudden commencement and gradual commencement storms had been described, Bartels introduced the term "*M* region" to designate the region on the sun associated with any geomagnetic disturbance. Since then, SC storms have been shown to be closely related to certain solar flares, and the ambiguous term "*M* region" is now used mainly in relation to GC storms or recurrent storms, whose solar connection is still not well-defined; hence, these are sometimes referred to as "*M*-region storms."

The main differences between these storms and the SC storm described above are: (*1*) the change in the geomagnetic field occurs more gradually, without the sudden jump of the SC; (*2*) in general, these storms are less intense; (*3*) a tendency for disturbance to recur after 27 days may be clearly seen during low solar activity, especially during the period before sunspot minimum; (*4*) these disturbances show somewhat different relations to solar activity, which will be described in the section on interrelations.

4.3.4. Theories of Geomagnetic Disturbance and Aurora. Theories of geomagnetic disturbance and aurora are numerous and sometimes slanted toward explanation of a single observational detail. The main lines of current theory were laid down by Chapman [105] and Ferraro. A stream or cloud of solar particles, composed of approximately equal numbers of protons and electrons, is emitted from the sun and encounters the earth a day or two later. The earth's magnetic field forms a barrier for these charged particles, and the effect of the initial interaction between moving charged particles and fields is seen in the SC. The stream subsequently tends to flow past and around the earth, and at its closing in around the earth a ring current forms, encircling the earth and producing the main phase of the disturbance. The gradual dissipation of the ring current, either by the moving away of the solar cloud or by a charge exchange process, that leaves the charged particles with low velocity, corresponds to the recovery phase of the disturbance.

Chapman and Ferraro also showed that particles from a solar cloud can penetrate closest to the earth at the auroral zones. The intense auroral displays accompanying geomagnetic disturbance would then be due to the direct influx of solar protons and electrons at these latitudes.

The main differences between this model and more recent theories arise from the concept that the space between earth and sun contains both matter and magnetic fields. More sophisticated ideas are introduced to describe the interaction of particles and field, and the excitation and propagation of waves in the interplanetary and atmospheric media. Some models consider a more or less stable interplanetary magnetic field, while others suppose that the magnetic fields of solar active regions extend as far as, and beyond, the earth. A very recent observation of magnetic variations at several earth's radii, closely paralleling simultaneous variations at the earth's surface during magnetic disturbance [106], suggests that the magnetic field of the solar clouds may be more important than the interaction of charged particles with the earth's field, at least on one occasion.

The change from a static to a dynamic view of the sun-earth medium, with consideration of the interaction between the motions and magnetic fields of the medium and the cloud of moving charged particles, perhaps itself carrying a magnetic field, promises to be most fruitful in explaining the puzzling details of geomagnetic disturbance.

4.3.5. Long-Term Changes in Solar Particle Emission. During the sunspot cycle, geomagnetic disturbance changes character more than it changes in average intensity. The typical major disturbance occurring at times of high solar activity, intense, with sudden commencement, tends to disappear as solar activity dies away. Then the less intense gradual commencement storms can be discerned. The over-all level of disturbance does decrease as solar activity decreases, but moderately large disturbance is frequent even when solar activity is very low. As already mentioned, the range of S_q (the daily range in magnetic intensity during a quiet day) changes in a regular fashion with sunspot activity. We may say that quiet days are quieter at solar activity minimum, and the frequency of occurrence of moderately disturbed days does not change greatly or very regularly.

The change of the frequency of occurrence of aurora during the sunspot cycle is obscured by the southward shift in geographical position of maximum auroral frequency with increasing geomagnetic activity. At subauroral latitudes, the frequency of auroral displays is definitely greater during high solar activity, but in the normal auroral zone the two effects—increasing auroral activity plus a shifting out of the zone of the most intense displays—partially compensate, and the absolute frequency may not show a very definite change over the sunspot cycle.

4.3.6. Indices of Geomagnetic Disturbance. From the brief discussion of theories of magnetic disturbance, it should be clear that we cannot expect that an index of magnetic disturbance provides a reliable measure of solar corpuscular emission, or even of the flux of solar particles at the earth. However, since such indices are the best measures we have of this emission, some widely used indices of disturbance are described here.

The amplitude of actual variation of the earth's field, dependent as it is on latitude and local time, is not a convenient index for statistical studies. Several indices have been devised that combine data from various stations to give a world-wide description of magnetic conditions.

The international character figure C_i is a mean of character figures from different stations. These are based on the daily range (the difference between the highest and lowest values of a given magnetic component during a Greenwich day) and on the general character of the variability during the day. Being somewhat subjective, this estimate may change systematically over the sunspot cycle. An attempt to attain greater long-term uniformity is represented by the K index, which measures the range of variation over each 3-hr period. The K values for each station are scaled according to the usual values at that station and their average represented by K_p, the planetary index. The sum of the eight values for a given day is used as a daily index. The indices a_p (3 hr) are related to K_p in an approximately exponential way, and the daily index A_p is the sum of the eight a_p values. The A_p changes more drastically for large disturbances than does K_p.

Another index of magnetic disturbance is the U index. This is the change from the preceding day in the component of the earth's field in the direction of the magnetic axis, averaged over various stations. Since it measures a day-to-day change in the field, it is sensitive to the occurrence of disturbance. Its use is appropriate only for monthly or annual means. The index U_1 is the index U reduced to a scale similar to that of sunspot numbers.

5. Interrelations

5.1. Introduction

In a solar active region at maximum development so many optical and radio phenomena occur almost simultaneously that the separation of their terrestrial effects becomes difficult. Flares tend to occur in regions of large, bright plages, with rapidly developing, complex sunspot groups. If the region reaches this stage of development near the limb of the sun, surges and active loop prominences and yellow coronal line emission will appear; and in the light of the green coronal line, unusually intense and intricate structure may be seen, with occasional evidence of rapid motions in the corona. On the disc, the prominence activity can be detected as "high speed dark" features, dark loops showing unusually large radial velocities, sudden disappearances, and "winking" filaments. This type of intense activity usually lasts only a few days for a given active region, but the plage and enhanced green coronal emission may continue for months, although when solar activity is high, they will be obscured by new intense activity (see Table III).

TABLE III. Indices of Solar Radiation.

Period	Solar index	Measured directly by	Closely related phenomena	Measures indirectly
Sec to min	Flare ejections, sprays, surges, high speed dark features	*Hα*, λ6563 A Velocities of ~500 km/sec	Flares, yellow line, etc.	
	"Winking" filaments	Hα, velocities of ~1500 km/sec	Flares, high speed dark features	Particles, 5 kev
	Type III radio bursts	Entire radio range, velocities ~ 10^5km/sec	Flares, high speed dark features	
Min to hr	Flare	*Hα*, λ6563	Yellow line, loop and surge prominences, bursts A and B, type IV, SID, SC storm, PCA, cosmic ray event	*Lα*, x-ray, particle
	Yellow coronal line emission	λ5694 A	Loop prominences, flare	x-ray
	Radio burst A	Centimeter, decimeter	Flare, SID	*Lα*, x-ray
	Radio burst B	Meter, decameter	Flare, prominence ejections, SC storm	Particles, 10 kev
	Type IV emission	Entire radio range	Flare, PCA, cosmic ray event, SC storm, *R* region	Particles, 50 Mev
	SID	Reflection and absorption of radio waves	Flares, burst A	X-ray, *Lα*
	Geomagnetic micropulsation	Variation of earth's field	Aurora?	Particles, 3–10 kev?
	VLF emission	2–10 kc/sec	Aurora, airglow?	Particles, 3–10 kev?
	SC storm	Variation of earth's field	Flare, type IV, burst B, *R* region	Particles, 10 kev
	Aurora	Visible and infared	Flare, type IV, SC storm	Particles, 3–10 kev
	Polar cap absorption	Absorption of cosmic noise, 30 Mc/sec; absorption of VHF forward scatter transmission, 50 Mc/sec	Flare, type IV, SC storm	Particles, 50 Mev
	Cosmic ray event	Ground-level cosmic ray counts	Flare, type IV, SC storm	Particles, Bev

TABLE III—*Continued*

Period	Solar index	Measured directly by	Closely related phenomena	Measures indirectly
Days	Sunspot number, area	Sunspots seen in white light	Plage, corona, centimeter flux, flare frequency, R region, SC storm, K_p	uv, x-ray, particles 10 kev
	Sunspot, Zürich type	Appearance, stage of development of spot group	Flare frequency, coronal yellow line, R region	uv, x-ray, particle
	Sunspot, Mt. Wilson type	Magnetic characteristics of spot groups	Flare, frequency, R region	uv, x-ray, particle
	Plage	Calcium H and K, Hα	Coronal emission, cm flux	uv, x-ray, particle 400 ev?
	Green coronal emission	λ5303 A	Plage, cm flux, foE	uv, x-ray
	Slowly varying component	Centimeter, decimeter	Plage, corona, foE, satellite drag	uv, x-ray
	R region, noise storm	Metric flux	Sunspots, yellow line, K_p	Particle
	Noise source	Decametric flux	Active region, SC storm, Jupiter decametric emission	Particle
	E-layer critical frequency	Reflection of transmission at 3 Mc/sec	Centimeter flux, satellite drag, sunspot number	uv ($L\beta$); x-ray
	F layer, scale height	Reflection of transmission at 10 Mc/sec	Centimeter flux, satellite drag, sunspot number	uv
	Satellite drag	Change of period of orbital motion	Decimeter flux, foE, $H(F_2)$	uv
	Daily magnetic indices C_i, K_p, A_p	Range of variation of earth's field	Flare, type IV, R region, foF	Particle, 10 kev
	SC storm	Range of variation of earth's field	Flare, type IV, R region, burst B	Particle, 10 kev
	Recurrent storm	Range of variation of earth's field	Sunspot, corona, plage, (negative?)	Particle, 400 ev?
Months to years	*Monthly or annual means of:*			
	Sunspot number, area	Sunspots in white light	cm flux, plage, corona, flare frequency, range in S_q	uv, x-ray, particle
	Slowly varying component	Centimeter, decimeter flux	Sunspot number, foE, foF$_2$	uv, x-ray
	E-layer critical frequency	Reflection of 3 Mc/sec	Sunspot number, corona, centimeter flux	x-ray
	F-layer critical frequency	Reflection of 10 Mc/sec	Sunspot number, cm flux	uv, He II, λ304?
	Magnetic index C_i, K_p, A_p, U or frequency of magnetic disturbance	Range in variability of earth's field	Sunspot number	Particles
	S_q, range in H or D	Diurnal range in earth's field for quiet days	Sunspot number	uv, x-ray

5.2. Relations of Slowly Varying Features

Of the long-lasting effects, the slowly varying component of radio flux in the centimeter and decimeter range has been quite positively identified with plages, and coronal streamers detected with the *K* coronameter. Ultraviolet and x-ray photographs of the sun show that these emissions are intensified in large regions centered on plage regions, although the time scale, especially for the latter, is still uncertain. Ionization of the *E* layer closely follows radio flux, as does atmospheric density, derived from satellite deceleration.

The relation of corpuscular radiation to these long-lived active centers is rather confused. The existence of long series of geomagnetic disturbances, recurring with a period of 27 days, shows that such a relation exists. Statistical treatment of geomagnetic data as related to optical solar data such as plages, green coronal emission, or sunspots suggests a negative relation—geomagnetic activity tends to be unusually low two days after such a region passes central meridian, and to be high after solar activity is low [107–109]. This effect is illustrated by the superposed epoch diagrams in Fig. 17. The

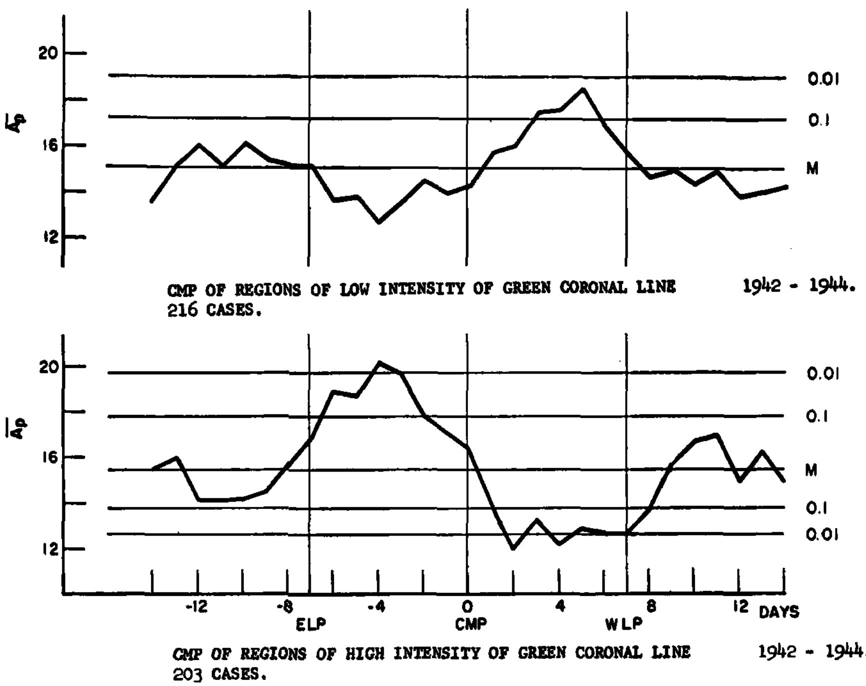


Fig. 17. Relation of geomagnetic disturbance to solar activity at sunspot minimum (originally published in *J. Geophys. Research.*)

mean value of a daily geomagnetic index is plotted against n for the days which fall n days after central meridian passage of a region of unusually low intensity of the green coronal line (unusually high intensity in the second case). The minimum at $n = 2$ for high intensity regions and the maximum for low intensity regions are the features that lead to the conclusion that a negative relation exists between solar activity and geomagnetic activity. The maxima on either side of the minimum are typical, and have led to the concept of a "cone of avoidance" [110] over an active region. In this model, particles emitted from the active region are systematically deflected to either side, away from the radial direction, so that particularly intense particle emission occurs before and after the passage of the region, while just at meridian passage, particle radiation is at a minimum. The particles are supposed to take approximately 2 days to reach the earth, so the earth's response lags two days behind the solar emission.

Still another interpretation is that the particles are emitted radially from such regions of moderate activity, but they are of much lower energy than those emitted at the time of flares and require 6 to 10 days to reach the earth [111, 112]. This explains only one feature of the statistical results shown above, the second maximum in the lower figure, but has the logical advantage that a positive relation applies at all phases of the sunspot cycle. Very limited data show a positive relation between meridian passage of regions of enhanced flux in the decimeter range and geomagnetic activity 6 days later [113], a result in accord with this second interpretation, although no relation was found between daily flux at metric wavelengths and geomagnetic disturbance. Clearly, further analysis is necessary to even determine the nature of the relation between solar activity and recurrent geomagnetic storms.

For more intense activity, relations are more clearly defined. The noise centers in the metric range, or R regions, are located fairly close to optically active regions, usually those with a complex sunspot group near the height of its development. Before interferometer observations allowed the location of the radio noise centers, noise storms had been identified with regions of yellow line emission [114] and of active loop prominences [115] and the likelihood of geomagnetic disturbance following central meridian passage of such regions had been demonstrated [114, 116].

5.3. Relations of Specific Events

Even closer relations appear when individual flares are brought into the picture. The fact that some large central flares are sources of particle emission was shown as early as 1945 [117]. Type II (slow drift) bursts and type IV (continuum) emission occur at the time of large flares that appear in R regions. The type IV event with a large flare is almost always followed by geomagnetic disturbance [118] and sometimes is accompanied by polar cap absorption [119, 120]. Polar cap absorption events tend to be followed by particularly intense magnetic disturbance but show no stronger relation to disturbance than would be expected from the fact that they follow large flares with type IV, and usually occur in *R* regions [121].

Surprisingly, the type II burst, whose velocity corresponds to that of the particles associated with geomagnetic disturbance and which is associated with optically observed ejections [122], seems to show little relation to such disturbance, unless type IV is also present [123]. Major bursts at metric wavelengths, occurring early in the life of a flare, also show a strong statistical relation to geomagnetic disturbance [124], a fact that is somewhat difficult to interpret in terms of spectral burst classification, since these early bursts probably correspond to type III. Major bursts with second part (major + or type 9) may correspond to type IV, and these do show a particularly strong relation to geomagnetic disturbance. However, since the second part tends to occur only with the optically more important flares, the interpretation of this relation is not clear-cut.

The U-bursts and the numerous type III bursts may occur with even small flares and also with such optical events as high speed dark features and flare ejections [125].

The attenuation of cosmic radiation known as a Forbush decrease tends to occur just before or near the beginning of a magnetic storm. This effect, too, shows a relation to type IV events [126].

An examination of the position on the solar disc of the solar flares associated with SC magnetic storms and with PCA and cosmic ray events suggests that the particles responsible for both these events must be emitted over wide angles, for while central flares are particularly likely to be followed by geomagnetic disturbance, flares close to the solar limb often are the only events that can be reasonably associated with the terrestrial event. The polar cap and cosmic ray events show a peculiar preference for flares on the western hemisphere of the sun (see Table II, and [120]). Statistics for even large numbers of flares must be interpreted with great caution, because flares, grouped in time and position as they are, will give mean values that show random deviations far greater than those occurring in a sample of independent events. However this effect for PCA and cosmic ray events, coupled with a systematic tendency for Forbush decreases to begin earliest at stations where the local time is between 4 and 10 hr, [127] strongly suggests that both charged particles and the solar magnetic clouds supposed to cause the Forbush effect are constrained to move in an antisymmetric way. An interplanetary magnetic field that spirals about the sun, the lines of force following the locus of radially emitted particles [128], would fit with both results. Both results also suggest a high curvature for the lines of force, which would result if they follow the locus of very slow particles, with velocities less than 100 km/sec. It is difficult to see how the interplanetary lines of force could be constrained to follow such slow particles and yet could, in turn, guide the motion of very energetic particles, unless the latter have very low density (which is the case for the PCA protons). A greater difficulty for the PCA protons is that the uniform absorption over the polar cap shows that they enter the atmosphere from all directions, which would be unlikely if their motion were strongly guided.

Another series of events associated with flares shows the effects of enhanced ultraviolet and x-radiation. These are the phenomena included in the sudden ionospheric disturbance described in Section 4. The relation of fadeout to flares is so strong that we can say that practically no case is known of a SWF in the absence of unusual Hα-brightening [129]. An interesting exception to this rule occurred on June 9, 1959 [130]. Very severe SID effects occurred with radio emission in the centimeter and decameter ranges but *not* at the intermediate metric wavelengths. Any flare during the first half-hour of this SID must have been behind the east limb of the sun. The area, height, and especially the Hα-intensity of the flare are strongly related to the occurrence of accompanying fadeout [71]. An even closer connection appears between the occurrence of fadeout and a radio burst in the centimeter-decimeter range, while in individual cases, the ionospheric absorption and centimeter flux show similar time development [131, 132]. The intensity variation and the location in the solar atmosphere must be similar for emission in these widely separated spectral regions: at decimeter wavelengths, in the visible region at Hα, and in the region of the short wave ionizing radiation.

In a case of unusual emission of γ-rays at the time of a solar flare [133], a radio burst occurred that was extremely abrupt and intense at 3 cm, decreased in intensity toward longer wavelengths, and was unobservable at 60 cm [134]. In a second case of unusually hard flare emission [135] a radio burst occurred in the centimeter range that was more unusual for its long duration than for its intensity, but which showed the same type of spectrum, decreasing sharply in intensity toward longer wavelengths. The intensity decreased by a factor of one-half from 3.2 to 8 cm, and had disappeared at 30 cm [136]. In the first case, the short wave radiation is estimated to have considerably shorter wavelength than in the second, so some differences might be expected. In both these cases, the short wave emission, as well as the radio emission, requires some nonthermal mechanism of radiation.

Although magnetic crochets are an ionospheric effect and occur along with fadeouts, the two phenomena show somewhat different relations to solar flares. Each is most likely to accompany a flare that is large and bright in Hα, but crochets occur particularly with flares near the limb [137], while fadeouts show a slight tendency to accompany central flares.

The enhanced ionization in the lower ionosphere that takes place during the usual fadeout is not sufficient, it seems, to produce the ionospheric currents responsible for the crochet. Along with the usual ionizing radiation, there must also be present some other radiation that has the peculiar property of being emitted most intensely in nonradial directions. The radiation (bremsstrahlung) that results from the sudden deceleration of a stream of charged particles would have this property of directivity [138], and there are other emission mechanisms, such as synchrotron and Čerenkov, that are nonisotropic. All require highly energetic particles moving at a velocity close to that light.

Rocket observations of x-rays close to the times of solar flares suggest that a hardening of the x-ray spectrum, rather than an intensification, accompanies the most severe fadeouts and crochets [66].

Apart from the relation of each to large solar flares, there is little connection between SID phenomena and geomagnetic activity; it seems that solar particle emission is only incidentally related to radiation in the short wave end of the spectrum. As a general rule, radiation in the centimeter and decimeter range of the radio spectrum is associated with ultraviolet and x-ray solar emission, and radiation at the longer wavelengths, metric and decametric with particle emission [139].

Interesting new evidence of solar particle emission is the decametric radio noise from the planet Jupiter. While Jupiter emission in the decimeter region is rather steady, at the longer wavelengths it appears only when the planet's rotation brings a certain region on to the side facing the earth. Even then, emission occurs only sporadically, apparently in close association with solar radio emission at these wavelengths, and, less certainly, with geomagnetic activity [140]. As our understanding of these relations increases, we may be able to detect solar outbursts of particle emission at sun, earth, and other planets and to derive a clearer picture of the form and velocity of these solar particle clouds.

5.4. Uses of Solar Indices

In searching for a solar relation, the researcher must not only decide which phenomenon of solar activity is most likely to reflect the conditions in which he is interested but must also choose a numerical expression of the occurrence or intensity or size of the phenomenon. Practically all of this article has been devoted to guiding the first choice, and we will add only a few words about the second decision.

For some types of activity, various measures may be available. General sunspot activity, for example, may be represented by the sunspot number, the total area of spots on the disc, or the total area within the central zone of the disc. In most flare reports are given the measured flare area, the area corrected for foreshortening when the flare is not at the center of the disc, and the importance, a rough index that is based on corrected area (sometimes measured area) but also takes into account the intensity of the flare. Sometimes a measure of the intensity and the linewidth of Hα is also given.

Both sunspot and flare data serve as examples of an inhomogeneity, often present in solar data, that results from the relative invisibility of features near the limb. The number of flares observed drops by about a factor of two from center to limb, and surprisingly this is true for flares rated as most "important" as well as for the smaller ones. Since the largest or most important flares should be easily visible even at the limb, the applied correction for foreshortening must be, in general, insufficient. A more puzzling systematic effect shows up in sunspot data and sometimes for flares, filaments, etc. Consistently greater numbers of spots are seen on the eastern hemisphere of the sun than on the western hemisphere. The effect is very small, but it is consistent in different data samples. White light faculae, of course, are seen only near the limb, thus showing a center-limb effect opposite that of sunspots and flares.

Another type of inhomogeneity arises when data from different observatories are combined in a single study. Absolute values of intensity are notoriously difficult to define, and when to this difficulty is added the fact that different methods of observation are used, large systematic differences may arise between different observatories. An amusing example of this sort of difference is seen in the fact that during IGY more than twice as many flares were reported to begin between the hours of 8 and 10 UT as between 20 and 22 UT [141].

Most terrestrial indices show a pronounced seasonal change; for example, the semiannual variation in number of geomagnetic disturbances. Ionospheric and atmospheric effects are presumably due to the changing inclination of the sun's rays through the year, and the geomagnetic effect is probably caused by one or another change in sun-earth orientation. A solar index such as flare frequency may also show a seasonal variation, because the majority of observing stations are located in the Northern Hemisphere, and many flares can be observed during long summer days. These variations can lead to a high correlation between quantities that are unrelated except that both show such a seasonal change. Thus, caution is necessary in analyzing and interpreting data that extend over periods of the order of a year or more.

Coronal observations are particularly sensitive to differences in observational technique: whether the slit is oriented tangentially or radially, and how far from the limb the slit is placed. In some cases, the research worker may prefer to avoid this difficulty by using data from a single observatory, in others, he may determine the systematic differences in order to combine data from different observatories, or he may choose to use an index that already represents combined data such as the Zürich sunspot number or the planetary geomagnetic index.

Finally, we present a list, far from exhaustive, of some sources of solar data, which we hope will be useful in introducing the reader to this extensive literature. We thank Richard T. Hansen, World Data Center A for Solar Activity, for his major part in the compilation of Table III and the following selected references for solar indices. We are indebted to Bernhard Haurwitz, Thomas Van Zandt, and Walter Roberts for many helpful suggestions concerning the manuscript.

Selected References for Solar Indices

Sunspots, Relative Numbers (Wolf)

Since 1749: Waldmeier, M., *Astronomische Mitteilungen Zürich* **14**, No. 140, 551, 1941.

International Astronomical Union, *Quarterly Bulletin on Solar Activity*, Eidgenössische Sternwarte Zürich.

Journal of Geophysical Research.

Sunspots, Group Classification

Since 1947: Waldmeier, M., *Publikationen Zürich Obs.* **9**, 1, 1947. *Publikationen der Eidgenössischen Sternwarte Zürich.*

Sunspots, Magnetic Properties

Since 1917: Hale, G. E. and S. B. Nicholson, "Magnetic Observations of Sunspots," 1917–1924 (Washington: Carnegie Institution of Washington), Part I, p. 56, 1938. *Publications of the Astronomical Society of the Pacific.*

Since 1956: Central Astronomical Observatory of Academy of Sciences, *Solar Data*, Moscow.

Sunspots, Positions, Areas and Counts

Since 1874: Greenwich Royal Observatory, *Greenwich Photoheliographic Results*, London, Her Majesty's Stationery Office.

Since 1948: U. S. Naval Observatory, *Circulars*, Washington, D. C.

Faculae, Areas, and Positions

Since 1874: Greenwich Royal Observatory, *Greenwich Photoheliographic Results*, London, Her Majesty's Stationery Office.

Faculae, Outlines, and Positions

Publikationen der Eidgenössischen Sternwarte Zürich.

Filaments, Dimensions and Positions

Since 1919: Observatoire de Meudon, *Cartes Synoptiques de la Chromosphère Solaire et Catalogue de la Couche Superieure*, Paris.

Filaments, Sudden Disappearances

Since 1919: Observatoire de Meudon, *Cartes Synoptiques de la Chromosphère Solaire et Catalogue de la Couche Superieure*, Paris.

Filaments, Character Figures

1917–1944: International Astronomical Union, *Quarterly Bulletin on Solar Activity*, Eidgenössische Sternwarte Zürich.

Plages, Outlines, and Positions

Since 1919: Observatoire de Meudon, *Cartes Synoptiques de la Chromosphère Solaire et Catalogue de la Couche Superieure*, Paris.

Plages, Positions, and Areas
Since 1955: National Bureau of Standards, CRPL, *Solar Geophysical Data*, Boulder, Colorado (Observations by McMath-Hulbert Obs.).

Prominences, Surges, and Active Regions
July 1957–1959: High Altitude Observatory, *IGY Solar Activity Report Series*, Boulder, Colorado.

Flares
Since 1936: International Astronomical Union, *Quarterly Bulletin on Solar Activity*, Eidgenössische Sternwarte Zürich (Complete, final listings from world wide observatories after a lag of 1–2 years).
National Bureau of Standards, CRPL, *Solar Geophysical Data*, Boulder, Colorado (fast schedule publication of available data after 1–2 months).

Radio Noise, Flux, and Outstanding Occurrences
Since 1947: International Astronomical Union, *Quarterly Bulletin on Solar Activity*, Eidgenössische Sternwarte Zürich (complete, final listings from world wide observatories after a lag of 1–2 years).
National Bureau of Standards, CRPL, *Solar Geophysical Data*, Boulder, Colorado (fast schedule publication of available data after 1–2 months).

Corona, Emission Line Intensities
Since 1947: International Astronomical Union, *Quarterly Bulletin on Solar Activity*, Eidgenössische Sternwarte Zürich.

Corona, Daily Diagrams
1939–1950: Waldmeier, M., *Die Sonnenkorona*, Vol. 1 (Basel: Verlag Birkhäuser), 1951.

Corona, Isophotes
1941–1949: Waldmeier, M., *Die Sonnenkorona*, Vol. 2 (Basel: Verlag Birkhäuser), 1957.
Since 1951: High Altitude Observatory, *Solar Activity Summaries*, Boulder, Colorado.

Corona, Electron
1956–1958: Newkirk, G., G. W. Curtis and D. K. Watson, "Observations of the Solar Electron Corona, Sept. 1956–Jan. 1958," *IGY Solar Activity Report Series*, No. 4, 1958, High Altitude Observatory.

Solar Maps, Daily
Since 1955: Fraunhofer Institute, *Daily Maps of the Sun*, Freiburg.
Since 1956: Central Astronomical Observatory of Academy of Sciences, *Solar Data*, Moscow.

Solar Maps, Synoptic
Since 1919: Observatoire de Meudon, *Cartes Synoptiques de la Chromosphère Solaire et Catalogue de la Couche Superieure*, Paris.
Since 1946: Fraunhofer Institute, *Sonnenzirkular*, Freiburg.
Since 1951: High Altitude Observatory, *Preliminary Reports of Solar Activity*, Boulder, Colorado (issued each week).
Since 1951: High Altitude Observatory, *Solar Activity Summary*, Boulder, Colorado (issued each quarter-year).
Since 1956: Central Astronomical Observatory of Academy of Sciences, *Solar Data*, Moscow.
Since 1949: Tokyo Astronomical Observatory, *Bulletin of Solar Phenomena*, Tokyo.

Sudden Ionospheric Disturbance (SWF, SEA, SCNA)
National Bureau of Standards, CRPL, *Solar-Geophysical Data*.

Solar Flare Effect

IAGA (IATME) Bulletins, "Geomagnetic Indices K and C."

Journal of Geophysical Research.

Polar Cap Absorption

Reid, G. C., and Leinbach, H. (1959). "Low Energy Cosmic-Ray Events Associated with Solar Flares," *Journal of Geophysical Research*, **64**, 1802.

Bailey, Dana, National Bureau of Standards, Boulder, Colorado. Data to be published.

Ionospheric Critical Frequencies and Virtual Heights; hourly values for E and F_2 layers, fE_s, f-plots for a number of stations.

National Bureau of Standards, CRPL, *Ionospheric Data.*

Ionospheric Electron Densities and True Heights

Schmerling, E. R., and Thomas, J. O. (1955). "Tables of F_2-layer Electron Density on International Quiet Days," Cavendish Laboratory Report Series A no. 1.

Thomas, J. O., and Robbins, A. R. (1955). "Tables of F_2-layer Electron Density on International Quiet Days," Cavendish Laboratory Report Series A no. 2.

Thomas, J. O., Haselgrove, J., and Robbins, A. R. (1957). "Tables of Ionospheric Electron Density, Slough (England)." Cavendish Laboratory Report Series B, no. 1.

Croom, S. A., Robbins, A. R., and Thomas, J. O. (1959). "Tables of Ionospheric Electron Density, Watheroo, Maui, Huancayo," Cavendish Laboratory Report Series B, nos. 2, 3, 4.

Schmerling, E. R. (1958–1959). "Ionospheric Electron Density for Washington, D.C., Panama, Talara and Huancayo," Pennsylvania State University, Ionosphere Research Laboratory, Scientific Reports.

National Bureau of Standards, CRPL, *Ionospheric Data.*

Geomagnetic Daily Character Figures, C, K_p, A_p

IAGA (IATME) Bulletins, "Geomagnetic Indices K and C."

National Bureau of Standards, CRPL, *Solar-Geophysical Data* (daily values, monthly means, musical diagrams).

Lists of Magnetic Storms

"Sunspot and Geomagnetic Storm Data Derived from Greenwich Observations 1874–1954," London: Her Majesty's Stationery Office 1955.

IAGA (IATME) Bulletins, "Geomagnetic Indices K and C."

Journal of Geophysical Research.

Occurrence of SC

IAGA (IATME) Bulletins, "Geomagnetic Indices K and C."

Journal of Geophysical Research.

Selected Quiet and Disturbed Days

IAGA (IATME) Bulletins, "Geomagnetic Indices K and C."

Journal of Geophysical Research.

Compilations of Solar, Ionospheric and Geomagnetic Data for Selected Periods, *Reports of Ionospheric Research in Japan.*

References

1. Sterne, R. E., and Dieter, N. (1958). The constancy of the solar constant. *Smithsonian Contrib. Astrophys.* **3**, 9.
2. Abbot, C. G. (1958). On Sterne and Dieter's paper, "The constancy of the solar constant." *Smithsonian Contrib. Astrophys.* **3**, 13.

3. Johnson, H. L., and Iriarte, G. (1959). The sun as a variable star. *Lowell Obs. Bull.* **4,** No. 96, 99.
4. Minnaert, M. (1953). The photosphere. *In* "The Sun" (G. P. Kuiper, ed.), p. 91. Univ. of Chicago Press, Chicago, Illinois.
5. Violett, T., and Rense, W. A. (1959). Solar emission lines in the extreme ultraviolet. *Astrophys. J.* **130,** 954.
6. Friedman, H. (1957). Photoelectric measurements of solar x-rays and ultraviolet radiation. 9th report of the Commission for the Study of Solar-Terrestrial Relations, p. 41.
7. Parker, E. N. (1959). Extension of the solar corona into interplanetary space. *J. Geophys. Res.* **64,** 1675.
8. Thomas, R. N., and Athay, R. G. (1960). "Physics of the Chromosphere." Interscience, New York, New York.
9. Schwarzschild, M. (1959). Photographs of the solar granulations taken from the stratosphere. *Astrophys. J.* **130,** 345.
10. Siedentopf, H. (1948). Zellförmige Strömingen im Laboratorium und in der Natur. *Naturw. Rundschau* **1,** 107–115.
11. Roberts, W. O. (1945). A preliminary report on chromospheric spicules of extremely short lifetime. *Astrophys. J.* **101,** 136.
12. Dunn, R. B. (1960). Photometry of the solar chromosphere. Thesis, Harvard College Observatory.
13. Athay, R. G. (1959). The number of spicules in the middle chromosphere. *Astrophys. J.* **129,** 164; (1958). Hα and D_3 profiles in chromospheric spicules. *Ann. Astrophys.* **21,** 98; (1960). Hβ, Hγ and K profiles in chromospheric spicules. *ibid.* **23,** 250.
14. Hulst, H. C. van de (1953). The chromosphere and the corona. *In* "The Sun" (G. P. Kuiper, ed.), Chapter 5. Univ. Chicago Press, Chicago, Illinois.
15. Jager, C. de (1959). "Handbuch der Astrophysik," Ed. S. Flügge, p. 112. Springer, Berlin.
16. Athay, R. G. (1960). The equilibria and ultraviolet spectra of H, He I and He II in the solar atmosphere. *Astrophys. J.* **131,** 705.
17. Elwert, G. (1954). Die weiche Röntgenstrahlung der ungestörten Sonnenkorona. *Z. Naturforschung* **9a,** 637.
18. Blackwell, D. E. (1960). Interplanetary electrons. Paper presented at Symposium on Solar Emissions and the Interplanetary Medium, 97th Annual Meeting of National Academy of Sciences, Washington, D.C.
19. Newkirk, G. and Roach, F. (1959). Private communication.
20. Kiepenheuer, K. O. (1953). Solar activity. *In* "The Sun" (G. P. Kuiper, ed.), p. 340. Univ. Chicago Press, Chicago, Illinois.
21. Allen, C. W. (1955). "Astrophysical Quantities," p. 150. Athlone Press, London, England.
22. Abetti, G. (1957). "The Sun," p. 81. Macmillan, New York.
23. Moreton, G. (1960). Private communication.
24. Athay, R. G., and Moreton, G. (1961). Impulsive phenomena of the solar atmosphere. *Astrophys. J.* **133,** No. 3.
25. Dodson, H. W., Hedeman, E. R., and McMath, R. R. (1956). Photometry of solar flares. *Astrophys. J. Suppl.* **2,** 241.
26. Newkirk, G. (1961). The solar corona in active regions and the thermal origin of the slowly varying component of solar radio radiation. *Astrophys. J.* **133,** No. 3.
27. Sakurai, Kunitomo (1959). Motions of solar cosmic rays and the properties of the general magnetic field of the sun. *J. Geomagn. Geoelectricity* **11,** 21.

28. Babcock, H. W., and Babcock, H. D. (1955). The sun's magnetic field, 1952–1954. *Astrophys. J.* **121**, 349.
29. Babcock, H. D. (1959). The sun's polar magnetic field. *Astrophys. J.* **130**, 366.
30. Nikulin, N. S., Severny, A. B., and Stepanov, V. E. (1958). "The Solar Magnetograph of the Crimean Astrophysical Observatory," Vol. 19, p. 30 Publications of Crimean Astrophysical Observatory.
31. Roberts, W. O., and Correll, M. (1958). Atmospheric magnetic fields above active solar region of April 13, 1950. *Astrophys. J.* **127**, 726.
32. Leighton, R. B. (1959). Observations of solar magnetic fields in plage regions. *Astrophys. J.* **130**, 366.
33. Covington, A. E., and Broten, N. W. (1954). Brightness of the solar disk at a wave length of 10.3 cm. *Astrophys. J.* **119**, 569–589.
34. Kundu, M. R., and Denisse, J. F. (1958). Solar radiation on decimeter waves as an index for ionospheric studies. *J. Atmos. Terr. Phys.* **13**, 176–178.
35. Christiansen, W. N., and Matthewson, D. S. (1959). The origin of the slowly varying component. *In* "Paris Symposium on Radio Astronomy" (R. N. Bracewell, ed.), pp. 108–116. Stanford Univ. Press, Stanford, California.
36. Kundu, M. R. (1959). Étude interférométrique des sources d'activité solaire sur 3 cm de longeur d'onde. *In* "Paris Symposium on Radio Astronomy" (R. N. Bracewell, ed.). Stanford Univ. Press, Stanford, California.
37. Pawsey, J. L., and Bracewell, R. N. (1955). "Radio Astronomy," pp. 82–83 and 94–100. Clarendon Press, London and New York.
38. Newkirk, G. N. (1959). A model of the electron corona with reference to radio observations. *In* "Paris Symposium on Radio Astronomy" (R. N. Bracewell, ed.). Stanford Univ. Press, Stanford, California.
39. Avignon, Y., Boischot, A., and Simon, P. (1959). Observation interférométrique à 169 Mc/s des centres R, sources des orages de bruit. *In* "Paris Symposium on Radio Astronomy," (R. N. Bracewell, ed.), pp. 240–244. Stanford Univ. Press, Stanford, California.
40. Boischot, A., Lee, R. H., and Warwick, J. W. (1960). Low frequency solar bursts and noise storms. *Astrophys. J.* **131**, 61–67.
41. Maxwell, A., Swarup, G., and Thompson, A. R. (1958). The radio spectrum of solar activity. *Proc. I.R.E.* **46**, 142–148.
42. Elgarøy, ø., (1959). Frequency drift and fine structure of 200 Mc/s solar bursts. *In* "Paris Symposium on Radio Astronomy" (R. N. Bracewell, ed.), pp. 248–250 Stanford Univ. Press, Stanford, California.
43. Wild, J. P., Sheridan, K. V., and Trent, G. H. (1959). The transverse motions of the sources of solar radio bursts. *In* "Paris Symposium on Radio Astronomy" (R. N. Bracewell, ed.), pp. 176–185. Stanford Univ. Press, Stanford, California.
44. Wild, J. P., Sheridan, K. V., and Neylan, A. A. (1959). An investigation of the speed of the solar disturbances responsible for type III radio bursts. *Australian J. Phys.* **12**, 369–398.
45. Boischot, A. (1957). Caractères d'un type d'émission hertzienne associé à certaines éruptions chromosphériques. *Compt. rend. Acad. Sci.* **244**, 1326–1329.
46. Avignon, Y., and Pick-Gutmann, M. (1959). Relation entre les émissions de type IV et d'autres formes d'activité solaire. *Compt. rend. Acad. Sci.* **248**, 368–371.
47. Boischot, A., and Warwick, J. W. (1959). Radio emission following the flare of August 22, 1958. *J. Geophys. Res.* **64**, 683–684.
48. Boischot, A., and Denisse, J. F. (1957). Les émissions de type IV et l'origine des rayons cosmiques associées aux éruptions chromosphériques. *Compt. rend. Acad. Sci.* **245**, 2194–2197.

49. Watanabe, K. (1958). Ultraviolet absorption processes in the upper atmosphere. *Advances in Geophys.* **5,** 203.
50. Allen, C. W. (1948). Critical frequencies, sunspots, and the sun's ultra-violet radiation. *Terr. Magn.* **53,** 433–448.
51. Huang, Chun-Ming (1960). Some abnormalities in the variations of F_2-layer critical frequency during the period of high solar activity of solar cycle 8–19. *J. Geophys. Res.* **65,** 897–906.
52. Minnis, C. M., and Bazzard, G. H. (1959). Some indices of solar activity based on ionospheric and radio noise measurements. *J. Atmos. Terr. Phys.* **14,** 213–228.
53. Ostrow, S. M., and Po Kempner, M. (1952). The differences in the relationship between ionospheric critical frequencies and sunspot number for different sunspot cycles. *J. Geophys. Res.* **57,** 473–480.
54. Scott, J. C. W. (1952). The solar control of the E and F_1 layers at high latitudes. *J. Geophys. Res.* **57,** 369–386.
55. Chapman, S., and Bartels, J. (1940). "Geomagnetism," pp. 220–225, 410–412. Oxford Univ. Press, London and New York.
56. Rees, M. H., and Rense, W. A. (1959). Note on the cause of ionization in the F-region. *J. Geophys. Res.* **64,** 1251–1255.
57. Jacchia, L. G. (1959). Solar effects on the acceleration of artificial satellites. *Smithsonian Inst. Astrophys. Obs., Spec. Rept. No.* **29.**
58. Van Zandt, T. (1960). Temperature in the upper atmosphere and fluctuations with solar activity. Presented at conference in Boulder, Colorado, May 27, 1960.
59. Martin, H. A., and Priester, W. (1960). Measurement of solar and diurnal effects in the high atmosphere by artificial satellites. *Nature* **185,** 600–601.
60. Nonweiler, T. (1958). Effect of solar flares on earth satellite 1957 β. *Nature* **182,** 468–469.
61. Johnson, F. S. (1958). Temperature in the high atmosphere. *Ann. Geophys.* **14,** 94–108.
62. Bartels, J. (1950). 27-day variations in F_2 layer critical frequencies at Huancayo. *J. Atmos. Terr. Phys.* **1,** 2–12.
63. Lange-Hesse, G. (1953). 27-tägige Variationen in der D-Schicht-Absorption der Ionosphäre über Singapore und Slough. *J. Atmos. Terr. Phys.* **3,** 153–162.
64. Matsushita, S. (1960). Day-to-day and seasonal changes of the S_q overhead current system. Paper presented at conference in Boulder, Colorado, May 27, 1960.
65. Bracewell, R. N. (1952). Theory of formation of an ionospheric layer below E-layer based on eclipse and solar-flare effects at 16 kc/sec. *J. Atmos. Terr. Phys.* **2,** 226–235.
66. Friedman, H., Chubb, T. A., Kupperian, J. E. Jr., Kreplin, R. W., and Lindsay, J. C. (1958). X-ray and ultra-violet emission of solar flares (IGY Project 10.3). *Ann. Géophys.* **14,** 232–235.
67. Warwick, J. W., and Zirin, H. (1957). Diurnal absorption in the D-region. *J. Atmos. Terr. Phys.* **11,** 187–191.
68. Warwick, J. W., and Zirin, H. (1957). Rocket observations of x-ray emission in a solar flare. *Nature* **180,** 500–501.
69. Taubenheim, J. (1957). Über den Einfluss von Sonneneruptionen auf die ionosphärische E-Schicht. *J. Atmos. Terr. Phys.* **11,** 14–22.
70. Gardner, F. F. (1959). The effects of SID's on 2.28 Mc/s pulse reflections from the lower ionosphere. *Australian J. Phys.* **12,** 42–53.
71. Warwick, C., and Wood, M. (1959). A study of limb flares and associated events. *Astrophys. J.* **129,** 801–811.
72. Burkard, O. (1956). Temperature-controlled variations of the ionosphere during

an eclipse. *In* "Solar Eclipses and the Ionosphere" (W. J. G. Beynon and G. M. Brown, eds.), pp. 69–73. Pergamon Press, New York.
73. Minnis, C. M. (1956). The origin of the E- and F_1-layer ionizing radiation. *In* "Solar Eclipses and the Ionosphere" (W. J. G. Beynon and G. M. Brown, eds.), pp. 158–163. Pergamon Press, New York.
74. Purcell, J. D., Packer, D. M., and Tousey, R. (1959). Lyman-alpha photographs of the sun. *Nature* **184**, 8–10.
75. Bailey, D. K. (1959). Abnormal ionization in the lower ionosphere associated with cosmic ray flux enhancements. *Proc. I.R.E.* **47**, 255–266.
76. Hultquist, B., and Ortner, J. (1959). Strongly absorbing layers below 50 km. *Plan. and Space Sci.* **1**, 193–204.
77. Reid, G. C., and Leinbach, H. (1959). Low-energy cosmic-ray events associated with solar flares. *J. Geophys. Res.* **64**, 1801–1805.
78. Obayashi, T., and Hakura, Y. (1960). Enhanced ionization in the polar ionosphere caused by solar corpuscular emissions. *J. Radio Res. Labs.* **7**, 27–66.
79. Anderson, K. A., and Enemark, D. C. (1960). Observations of cosmic rays near the north magnetic pole. *J. Geophys. Res.* **65**, 2657–2671.
80. Blackwell, D. E., and Ingham, M. F. (1961). Observations of the zodiacal light from a very high altitude station. III The disturbed zodiacal light and corpuscular radiation. *Monthly Notices Roy. Astron. Soc.* **122**, 143–155.
81. Chapman, S. (1957). Notes on the solar corona and the terrestrial ionosphere. *Smithsonian Contrib. Astrophys.* **2**, 1–14.
82. Parker, E. N. (1958). Suprathermal particle generation in the solar corona. *Astrophys. J.* **128**, 677–685.
83. Parker, E. N. (1958). Dynamics of the interplanetary gas and magnetic fields. *Astrophys. J.* **128**, 664–676.
84. Chamberlain, J. (1960). Interplanetary gas II. Expansion of a model solar corona. *Astrophys. J.* **131**, 47–56.
85. Biermann, L. (1957). Solar corpuscular radiation and the interplanetary gas. *Observatory* **77**, 109–110.
86. Fan, C. Y., Meyer, P., and Simpson, J. A. (1960). Preliminary results from the space probe Pioneer V. *J. Geophys. Res.* **65**, 1862–1863.
87. Campbell, W. H. (1960). Micropulsations in the earth's magnetic field simultaneous with pulsating aurora. *Nature* **185**, 677.
88. Gallet, R. M. (1959). The very low frequency emissions generated in the earth's exosphere. *Proc. I.R.E.* **47**, 211–231.
89. Duncan, R. A., and Ellis, G. R. (1959). Simultaneous occurrence of sub-visual aurora and radio bursts on 4.6 kc/s. *Nature* **183**, 1618–1619.
90. Ellis, G. R. A., and Cartwright, D. G. (1959). Directional observations of radio noise from the outer atmosphere. *Nature* **184**, 1307–1308.
91. Chapman, S., and Little, C. G. (1957). The nondeviative absorption of high-frequency radio waves in auroral latitudes. *J. Atmos. Terr. Phys.* **10**, 20–31.
92. Jacchia, L. G. (1959). Corpuscular radiation and the acceleration of artificial satellites. *Nature* **183**, 1662–1663.
93. Rothwell, P., and McIlwain, C. (1959). Satellite observations of solar cosmic rays. *Nature* **184**, 138–140.
94. Arnoldy, R. L., Hoffman, R. A., and Winckler, J. R. (1960). Observations of the Van Allen radiation regions during August and September 1959, Part I. *J. Geophys. Res.* **65**, 1361–1376.
95. Rothwell, P., and McIlwain, C. (1960). Magnetic storms and the Van Allen radiation belts—observations from satellite 1958ϵ (Explorer IV). *J. Geophys. Res.* **65**, 799–806.

96. Van Allen, J. A., and Frank, L. A. (1959). Radiation measurements to 658,300 km with Pioneer IV. *Nature* **184,** 219–224.
97. Newton, H. W. (1947). Note on "Sudden Commencements" and other small characteristic impulses. *Terr. Magn.* **52,** 441–447.
98. Fillius, R. W., Gartlein, C. W., and Sprague, G. (1959). Ellsworth auroral data—1957. Cornell Univ. Rept. W.D.C.A. IV (a).
99. Fan, C. Y. (1958). Time variations of the intensity of auroral hydrogen emission and the magnetic disturbance. *Astrophys. J.* **128,** 420–427.
100. O'Brien, B. J., Van Allen, J. A., Roach, F. E., and Gartlein, C. W. (1960). Correlation of an auroral arc and a sub-visible monochromatic 6300A arc with outer-zone radiation on 28 November 1959. *J. Geophys. Res.* **65,** 2759–2766.
101. King, G. A. M., and Roach, F. E. (1961). Relationship between oblique ionospheric echoes and red auroral arcs. *J. Res. Natl. Bur. Stand.* **65D** (*Radio Prop.*), 129–135.
102. Winckler, J. R., Peterson, L., Arnoldy, R., and Hoffman, R. (1958). X-rays from visible aurorae at Minneapolis. *Phys. Rev.* **110,** 1221–1231.
103. Anderson, K. A. (1958). Soft radiation events at high altitude during the magnetic storm of August 29–30, 1957. *Phys. Rev.* **111,** 1397–1405.
104. Winckler, J. R., Peterson, L., Hoffman, R., and Arnoldy, R. (1959). Auroral x-rays, cosmic rays and related phenomena during the storm of February 10–11, 1958. *J. Geophys. Res.* **64,** 597–610.
105. Chapman, S., and Bartels, J. (1940). "Geomagnetism," pp. 850–897. Oxford Univ. Press, London and New York.
106. Coleman, P. J., Jr., Sonett, C. P., Judge, D. L., Smith, E. J. (1960). Some preliminary results of the Pioneer V magnetometer experiment. *J. Geophys. Res.* **65,** 1856–1857.
107. Allen, C. W. (1944). Relation between magnetic storms and solar activity. *Monthly Not. Roy. Ast. Soc.* **104,** 13–21.
108. Bell, B., and Glazer, H. (1957). Geomagnetism and the emission-line corona, 1950–1953. *Smithsonian Contrib. Astrophys.* **2,** 51–107.
109. Sinno, K. (1956). On the origin of the long-lived solar corpuscular streams which appeared last solar cycle, 1950 to 53. *Rept. Ionosphere Res. Japan* **10,** 250–260.
110. Pecker, J. C., and Roberts, W. O. (1955). Solar corpuscles responsible for geomagnetic disturbances. *J. Geophys. Res.* **60,** 33–44.
111. Mustel, E. R. (1959). Statistical effects, due to the connection between flocculi and geomagnetic disturbances. Transl. Sov. Ast., *Astron. J.* **3,** 221–227.
112. Mitropolskaya, O. N. (1959). Some results of statistical treatment of geomagnetic disturbances of the declining branch of solar activity *Ast. Zh.* **36,** 224–232. Transl. Sov. Ast., *Astron. J.* **3,** 228–236.
113. Hansen, R. T. (1960). Bright 21-cm solar regions and geomagnetic storms in 1952–1953. *J. Geophys. Res.* **65,** 3827–3829.
114. Simon, P. (1956). Centres solaires radioémissifs et non radioémissifs. I. Propriétés optiques et géomagnétiques. *Ann. Astrophys.* **19,** 122–141.
115. Wild, J. P., and Zirin, H. (1956). On the association of solar radio emission and solar prominences. *Australian J. Phys.* **9,** 315–323.
116. Denisse, J. F., and Simon, P. (1954). Relation entre l'apparition de la raie jaune coronale et l'activité géomagnétique. *Compt. rend. Acad. Sci.* **238,** 1775–1778.
117. Newton, H. W. (1943). Solar flares and magnetic storms. *Monthly Not. Roy. Ast. Soc.* **103,** 244–257.
118. Simon, P. (1960). Les éruptions chromosphériques associées aux sursauts radio-

électriques de type IV: effets ionosphériques et géomagnetiques. *Ann. Astrophys.* **23,** 102–110.
119. Obayashi, T., and Hakura, Y. (1960). Enhanced ionization in the polar ionosphere caused by solar corpuscular emissions. *J. Radio Res. Labs.* **7,** 27–66.
120. Avignon, Y., and Pick-Gutmann, M. (1959). Relation entre les émissions solaires de rayons cosmiques et les sursauts de type IV. *Compt. rend. Acad. Sci.* **249,** 2276–2278.
121. Warwick, C. S., and Wood, M. B. (1959). Geomagnetic disturbance following flares associated with type IV radio bursts and with radio noisy regions. Paper presented at Iowa Cosmic Ray Conference, October 31, 1959.
122. Giovanelli, R. G., and Roberts, J. A. (1958). Optical observations of the solar disturbances causing type II radio bursts. *Australian J. Phys.* **11,** 353–359.
123. McLean, D. J. (1959). Solar radio emission of spectral type IV and its association with geomagnetic storms. *Australian J. Phys.* **12,** 404–417.
124. Dodson, H. W., and Hedeman, E. R. (1958). Geomagnetic disturbances associated with solar flares with major premaximum bursts at radio frequencies ≤ 200 Mc/s. *J. Geophys. Res.* **63,** 77–96.
125. Giovanelli, R. G. (1958). Flare puffs as a cause of type III radio bursts. *Australian J. Phys.* **11,** 350–352.
126. Legrand, J. P. (1960). Les prébaisses de rayons cosmiques en périodes de maximum de l'activité solaire (avril 1957–décembre 1958). *Ann. Astrophys.* **16,** 140–142.
127. Fenton, A. G., McCracken, K. G., Rose, D. C., and Wilson, B. G. (1960). The onset times of Forbush-type cosmic ray intensity decreases. *In* "Proceedings of the Symposium on Physical Processes in the Sun-Earth Environment" (C. Collins, ed.), pp. 287–302. Defence Research Telecommunications Establishment Publ. No. 1025, Ottawa.
128. Chapman, S., and Bartels, J. (1940). "Geomagnetism," pp. 804–808. Oxford Univ. Press, London and New York.
129. DeMastus, H., and Wood, M. (1960). Short-wave fadeouts without reported flares. *J. Geophys. Res.* **65,** 609–611.
130. Coates, R. J., Covington, A. E., and Edelson, S. (1959). The 4.3 mm and 10.7 cm outbursts of June 9, 1959. *Rept. NRL Prog.* July 1959, pp. 23–24.
131. Hachenberg, O., and Volland, H. (1959). Vergleich zwischen 3.2 cm—Radiostrahlung der Sonne und ionosphärischer Dämpfung im D-Gebiet während einer Sonneneruption. *Z. Astrophys.* **47,** 69–80.
132. Hachenberg, O., and Krüger, A. (1959). The correlation of bursts of solar radio emission in the centimetre range with flares and sudden ionospheric disturbances. *J. Atmos. Terr. Phys.* **17,** 20–33.
133. Peterson, L. E., and Winckler, J. R. (1959). Gamma-ray burst from a solar flare. *J. Geophys. Res.* **64,** 697–707.
134. Denisse, J. F. (1959). Relation entre les émissions de rayons cosmiques solaires et certains sursauts radioélectriques. *In* "Paris Symposium on Radio Astronomy" (R. N. Bracewell, ed.), pp. 237–239. Stanford Univ. Press, Stanford, California.
135. Chubb, T. A., Kreplin, R. W., McNutt, D. P., and Friedman, H. (1959). Project Sunflare II. *Rept NRL Prog.* Oct. 1959, pp. 1–7.
136. "Solar Radio Observations, Toyokawa, Japan," IGY Report on outstanding occurrences, August 1959.
137. Dodson, H. W., and Hedeman, E. R. (1958). Crochet-associated flares. *Astrophys. J.* **128,** 636–645.

138. Parthasarathy, R. (1959). On the origin of the flare-associated x-ray emission. *Observatory* **79**, 65–67.
139. Sinno, K., and Hakura, Y. (1958). On the relation of solar eruptions to geomagnetic and ionospheric disturbances. I. On the power spectrum of solar radio outbursts, II. On the types of solar radio outbursts. *Rept. Ionosphere Res. in Japan* **12**, 285–300.
140. Warwick, J. W. (1960). Observations of the spectrum of Jupiter's decametric radio emission, January and February 1960. Talk presented at meeting of American Geophysical Union, April 30, 1960.
141. Dodson, Helen W., and Hedeman, E. Ruth (1960). Survey of number of solar flares observed during the International Geophysical Year. *J. Geophys. Res.* **65**, 123–131.

This Page Intentionally Left Blank

SATELLITE STUDIES OF THE IONIZATION IN SPACE BY RADIO

Owen K. Garriott and R. N. Bracewell

Department of Electrical Engineering and Radioscience Laboratory, Stanford University, Stanford, California

Page

I. Introduction

Many types of investigation are possible with earth satellites. To begin with, measurements can be made of physical conditions at the location of the satellite, and this is a very important aspect, for it is often very difficult indeed with ground-based apparatus to measure conditions at the great heights attainable by satellites. Thus one can attempt direct observations of local chemical composition, pressure, temperature and electron density, and the local intensity of the electromagnetic fields through the whole spectrum from gamma radiation down to the static magnetic field.

In another class of experiment the unique situation of the satellite is used as a vantage point for observing things that are remote from the satellite. Meteorological and astronomical observations are examples.

In this chapter we shall restrict ourselves to satellite studies of the ionization in space by radio methods. We find that we are not limited to investigating the electron density at the satellite, but that the whole space below the satellite is opened to our study, if we equip the satellite with sources of radio waves. We receive these waves at the ground and then by noting the effects impressed during transit, we make deductions about the ionization of the intervening medium.

The most important observable effects of ionization are to change (1) the angle of arrival of the ray, (2) the Doppler shift, and (3) the polarization at arrival. In this Introduction these effects are discussed in an elementary way to lay a basis for later sections in which various necessary complications have to be introduced.

1.1. Frequency Shift

We know that, if a source of electromagnetic waves of frequency f_S is moving with velocity v relative to an observer, then the observed frequency f_O may not be the same as the emitted frequency. This is the Doppler effect[1] for electromagnetic waves and it is usually expressed quantitatively by the relation

$$f_O = f_S \frac{1 - (v_r/c)}{[1 - (v^2/c^2)]^{1/2}} \tag{1.1}$$

where v_r is the velocity of recession of the source resolved along the straight line from observer to source, and c is the velocity of light in free space. If $v \ll c$, which is so in our case,

$$f_O = f_S - f_S \frac{v_r}{c} \tag{1.2}$$

In this form of the relation it is implied that the source and observer are embedded in free space and that there is nothing but free space intervening between them.

The medium, however, introduces important effects which are the basis of the present work. The Doppler effect relation for a stationary, isotropic, inhomogeneous medium, having dielectric constants different from unity, becomes

$$f_O = f_S - f_S \frac{v \cos \phi}{c/\mu_S} \tag{1.3}$$

where μ_S is the refractive index of the medium at the source, v is the velocity of the source, and ϕ is the supplement of the angle between the source

[1] There is a second frequency shift associated with the gravitational potential difference between source and observer, of an amount such that quanta at the two frequencies differ in energy by an amount equal to the potential difference. The possibility of detecting this effect by radio methods has been discussed, but it is very much smaller than the frequency shifts of concern here.

trajectory and the ray. Thus

$$\frac{\mu_S}{c} v \cos \phi$$

is the ratio of the velocity of recession resolved along the ray ($v \cos \phi$) to the phase velocity at the source (c/μ_S). If the source is surrounded by a medium of dielectric constant less than unity, which is the usual ionospheric situation, then this velocity ratio would be diminished and the Doppler shift would tend to be less than in the free space case. This is not the full story, however, because along with the changed refractive index will go a changed ray path; the source velocity component along the ray will depend on the new direction of departure of the ray.

We see, then, that there are two effects: one associated with the refractive index of the medium at the source and another associated with the whole refractive index structure of the intervening medium. In terms of the wavelength λ in the medium at the source,

$$f_O = f_S - \frac{v \cos \phi}{\lambda} \tag{1.4}$$

In this expression, both λ and ϕ vary as the satellite moves, making it difficult to visualize their joint influence on the time variation of observed frequency. The following physical picture is offered as a means of describing the Doppler effect in inhomogeneous media. We limit ourselves, for the time being, to an isotropic case.

Suppose that there is a source of radio waves located at the ground in the position of the observer O. At some instant of time, describe a set of isophase surfaces on which the field is retarded by an integral number of cycles of phase relative to O (see Fig. 1). We shall call these surfaces "wavecrests." They are separated by exactly one wavelength as determined by the local refractive index. This is a stationary picture, drawn for the location of the observer, through which a satellite source S carrying its own radio transmitter is now imagined to move. As it crosses each wavecrest, e.g., in going from S to S_1 or from S_1 to S_2, the observer will see a phase retardation of one cycle. He will interpret this as a frequency drop in cycles per second equal to the number of wavecrests per second crossed by the satellite.

Let the phase path from the observer to the satellite be defined by

$$P = \int_O^S \mu \, ds \tag{1.5}$$

where μ is the refractive index and the integral is taken over the elements ds of the ray. Then the total number of cycles of phase between the observer

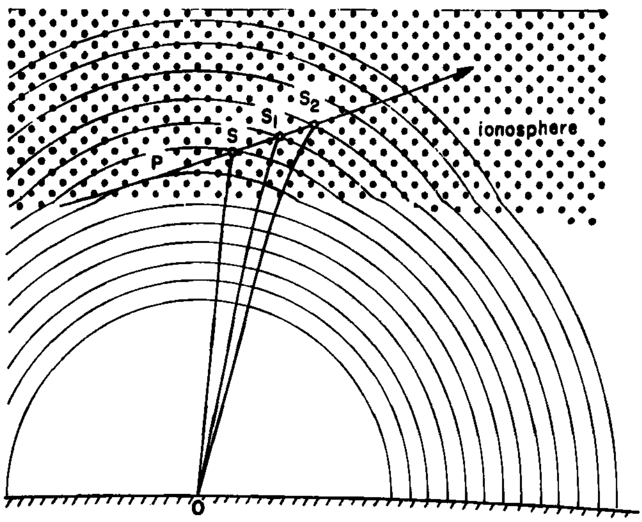


FIG. 1. Wavecrests, supposed to emanate from O, shown penetrating an inhomogeneous, but isotropic, ionosphere.

and the satellite will be

$$\int_{\mathrm{O}}^{\mathrm{S}} \frac{ds}{\lambda} = \frac{1}{\lambda_0} \int_{\mathrm{O}}^{\mathrm{S}} \mu \, ds = \frac{P}{\lambda_0} \tag{1.6}$$

where

$$\lambda = \frac{\lambda_0}{\mu} \tag{1.7}$$

is the wavelength at a point in the medium and λ_0 is the wavelength in free space. The rate at which the total number of cycles is changing gives the observed frequency[2] in the form

$$f_{\mathrm{O}} = f_{\mathrm{S}} - \frac{\dot{P}}{\lambda_0} \tag{1.8}$$

where $\dot{P}$ is the time derivative of P. The Doppler shift is $-\dot{P}/\lambda_0$.

The time variation of phase path is shown in Fig. 2 for the situation of Fig. 1. At the point P the phase path will be stationary and the observed frequency will be equal to the source frequency; before P is reached the observer will see a progressively advancing phase, interpretable as a frequency higher than that of the satellite transmitter. The graph of observed fre-

[2] Even if $v = 0$ there may be a frequency change if the medium is nonstationary, as is normally the case, for example, with the ionosphere. Equation (1.8) includes this possibility, but equation (1.3) would require a supplementary term.

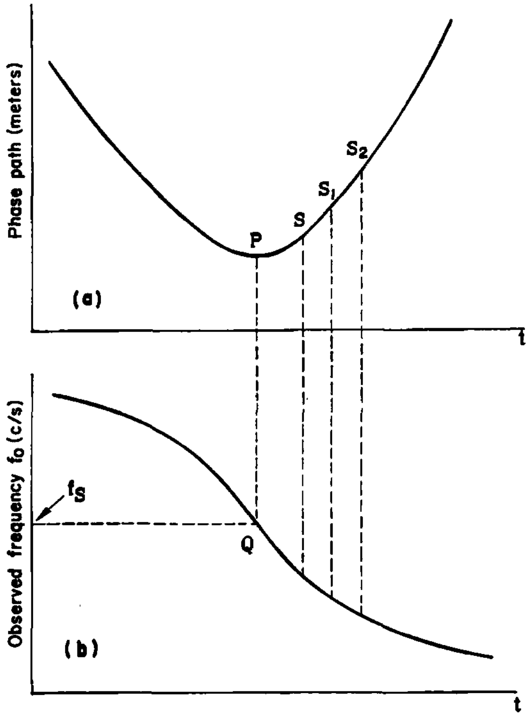


FIG. 2. The phase path and associated observed frequency $f_S - \dot{P}/\lambda_0$.

quency is also shown in Fig. 2. The point P in Fig. 1 will be referred to as the proximal point of the satellite orbit. It is usually near but not exactly the same as the point of closest approach; its precise location depends on the source frequency and on the electron density distribution.

The rate of change of observed frequency at P (i.e., the slope in Fig. 2 at Q) turns out later to be important for the theory. It assumes greater values, the greater the curvature of the wavecrests in the plane of departure (the plane defined by the orbit and the departing ray at the satellite).

1.2. Angle of Arrival

In an inhomogeneous but spherically symmetrical refracting medium in which the refractive index $\mu(\rho)$ is a function only of distance ρ from the center of the earth, Snell's law takes the form

$$\mu(\rho)\rho \sin \chi = \text{const.} \tag{1.9}$$

where χ is the angle of incidence of the ray on the spherical strata. Hence,

$$\mu_O\rho_O \sin \chi_O = \mu_S\rho_S \sin \chi_S \tag{1.10}$$

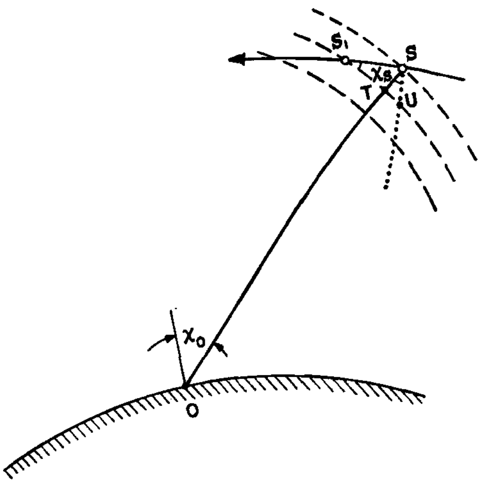


FIG. 3. Satellite in circular orbit cutting the stationary pattern of wavecrests supposed to have emanated from the observer.

where the subscripts O refer to the observer and S to the satellite. Since to a good approximation $\mu_O = 1$, for an observer on the ground, we have for the angle of arrival at the ground

$$\sin \chi_O = \frac{\mu_S \rho_S \sin \chi_S}{\rho_O} \tag{1.11}$$

As a special case we shall consider a satellite moving in a circular orbit through a stationary medium as shown in Fig. 3, the observer being in the plane of the orbit. The Doppler shift in cycles per second will, as we have seen, be equal to the rate of crossing the fixed wavecrests shown in broken outline. The spacing ST between wavecrests is equal to λ, the local wavelength in the medium. Hence, the Doppler shift will be proportional to the reciprocal of SS_1, which, from the triangle SS_1T in Fig. 3 is equal to

$$\lambda/\sin \chi_S$$

Now, in terms of the local refractive index μ_S, the local wavelength is given by $\lambda = \lambda_o/\mu_S$. Hence, we have

$$\text{Doppler shift} \propto \frac{1}{SS_1} = \frac{\mu_S \sin \chi_S}{\lambda_o} \tag{1.12}$$

Thus the two important quantities, local wavelength and direction of departure of the ray from the satellite, produce their joint influence on Doppler shift through the product $\mu_S \sin \chi_S$, which is practically a constant along any given ray. To be precise, from equation (1.10)

$$\mu_S \sin \chi_S = \frac{\rho_O}{\rho_S} \sin \chi_O \tag{1.13}$$

For a satellite describing an orbit at a constant height (which is known, albeit only with moderate precision) the Doppler shift is thus uniquely connected with the angle of arrival of the ray at the receiver. This also proves to be true when the observer is not situated in the plane of the orbit, but it is not necessarily accurate for noncircular orbits or nonspherical stratification. When the satellite has a velocity component perpendicular to the surface of constant index of refraction, there is an extra Doppler shift proportional to the local electron density. Since the angle of arrival is related only to the ionospheric structure below the satellite, the two measurements no longer provide precisely the same information.

1.3 Faraday Effect

An entirely different method of obtaining information about the medium is contributed by the Faraday effect, the phenomenon of birefringence induced in suitable media by a magnetic field. In the ionosphere, the presence of a magnetic field alters the motion of the electrons under the influence of the electric field of a radiowave. For example, consider a *linearly* polarized wave launched in the direction of the magnetic field. In the absence of a magnetic field an electron would be forced to oscillate in a straight line in the direction of the electric field of the wave, but when a weak magnetic field is present its trajectory is slightly curved. Consequently there is a small velocity component perpendicular to the electric field. In the absence of a magnetic field, the reradiated field of the oscillating electrons is parallel to the incident field but in time quadrature, and combines with it to advance the phase in a way that we describe, on a macroscopic basis, as a drop in refractive index. But when the magnetic field is present the radiated field contains a perpendicular component that causes a shift of the plane of polarization of the radio wave. The plane of polarization is twisted into a gentle helicoid.

In another way of looking at the phenomenon, we launch a *circularly* polarized wave in the direction of the magnetic field. Then under the influence of the rotating electric field of the wave an electron is forced to rotate in a circle. When the magnetic field is present, it will either increase the curvature of the electron trajectory, or, if the chosen polarization is in the opposite sense, it will decrease it. In the first case the amplitude of electron motion is reduced and in the second it is enhanced. In the first case the effect on refractive index due to the presence of electrons in the medium is diminished by the magnetic field, and the wave propagates with unchanged polarization but with a phase velocity nearer to the free space value than would have resulted from the given electron density with no magnetic field. We thus find two different refractive indices for the medium, one for each sense of polarization. Except when the direction of propagation is very close indeed to being perpendicular to the magnetic field, these two refractive indices are, in the absence of collisions,

$$\left[1 - \frac{Ne^2}{\epsilon_0 m\omega(\omega \pm \omega_L)}\right]^{1/2} \tag{1.14}$$

where e and m are the charge and mass of the electron, the "longitudinal gyrofrequency" ω_L is given by $\omega_L = e\mu_0 H_L/m$, ϵ_0 and μ_0 are the permittivity and permeability of free space, and H_L is the component of magnetic field in the direction of propagation.

When we launch a wave of any other polarization we can split it into circular components, handle each separately, and recombine them subsequently. For example, a linearly polarized wave splits into two equal and opposite circular components, which, after traveling with different phase velocities, recombine into a wave polarized linearly in a new direction.

Observation shows that the plane of polarization of radio waves received from satellites is indeed in a state of slow rotation. Even if no provision is made to measure the polarization, and one receives with a simple antenna, the Faraday effect makes itself quite obtrusive (Fig. 4) in the form of deep fading as the polarization at arrival periodically becomes unsuitable for reception by the antenna.

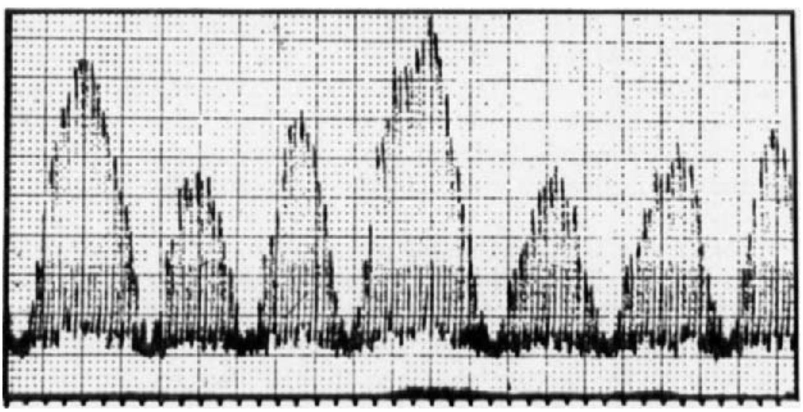


FIG. 4. A field strength record of the signal from Sputnik III received at Stanford with a simple dipole, showing deep fading caused by the Faraday effect. The time marks are at 1-sec intervals.

The pitch of the helicoid into which the plane of polarization is twisted is equal to the distance which has to be traveled in order for the extraordinary wave to get ahead of the ordinary wave by two full cycles. Two cycles are necessary because the total field vector bisects the angle between the field vectors of the two constituent modes; consequently, the angle of polarization rotation is only one-half the angular phase shift between the two modes. The propagation constants for the two modes are

$$\frac{2\pi}{\lambda_o}\mu_1 \quad \text{and} \quad \frac{2\pi}{\lambda_o}\mu_2 \qquad \text{radians/meter} \tag{1.15}$$

and therefore the difference in the two propagation constants is

$$\frac{2\pi}{\lambda_o}(\mu_1 - \mu_2) = \frac{2\pi}{\lambda_o}\Delta\mu \qquad \text{radians/meter} \tag{1.16}$$

where $\Delta\mu$ is the difference of the two refractive indices μ_1 and μ_2. Hence, the number of meters of path required for two cycles of phase difference to set in is

$$\frac{2\lambda_o}{\Delta\mu} \tag{1.17}$$

As when treating the Doppler effect, we simplify the situation by conceiving that radio waves are emitted from the receiving point, this time of both circular polarizations. Figure 5 shows the spherical wavecrests im-

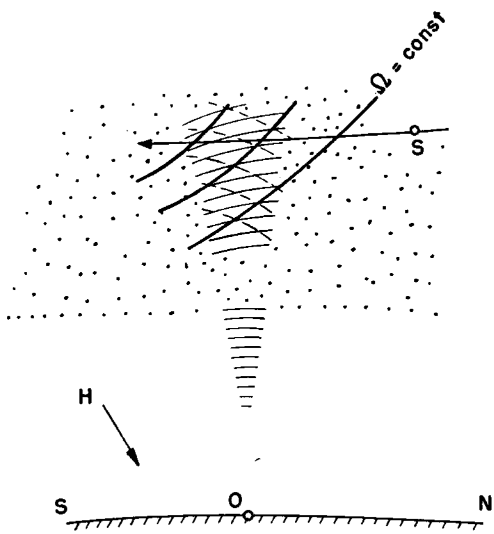


FIG. 5. The broken lines are wavecrests for the extraordinary mode and the fine solid lines for the ordinary mode. The heavy curves are the loci of constant path difference and are spaced at intervals of two cycles.

pinging on the bottom of the ionosphere after emission from O. Since both circular polarizations travel with the same velocity below the ionosphere their wavecrests coincide. Inside the ionosphere the two sets of wavecrests diverge in the way shown because of the different propagation characteristics of the two modes, and for any point, one can read off the phase path back to O, by either of the two modes. The heavy lines are the loci of points such that the difference in the two phase paths is a constant. Thus, if a radio source were to move down one of the heavy curves, emitting both modes at once, they would follow their independent paths to an observer at O and recombine without relative phase change. But the signal from a satellite moving more or less horizontally as shown, would exhibit a rotation in polarization as the difference in the phase paths changed.

The heavy curves are drawn at intervals of two cycles of relative phase change, which corresponds to one complete rotation of the plane of polarization. We call the total number of rotations between satellite and observer Ω and it will be seen that $\Omega = 0$ at the bottom of the ionosphere. One may conclude by looking at this diagram that the plane of polarization rotates at a more or less steady rate as the satellite passes by and that there is no symmetry about the position of closest approach. In fact it will be shown later that the rotation rate is connected with the total number of electrons contained in a column reaching from the ground to the satellite.

1.4. The Integrated Electron Density

We have seen that the presence of ionospheric electrons changes the Doppler shift and angle of arrival of a ray received from a satellite in a way that involves the refractive index distribution along the ray. The Doppler shift was found [equations (1.8) and (1.5)] to involve $\int \mu ds$. From (1.14) quoted above for the magneto-ionic theory we know that, in the absence of a steady magnetic field,

$$\mu = \left[1 - \frac{Ne^2}{\epsilon_0 m\omega^2}\right]^{1/2} \approx 1 - \frac{e^2}{2\epsilon_0 m\omega^2} N \tag{1.18}$$

Thus we have an approximately linear dependence of μ on N for moderate ionization densities which do not lower μ much below unity.

When we come to integrate μ along the ray, and to compare the result with what it would have been had there been no electrons, we find that the *change* in Doppler shift due to the ionosphere involves the integral of N along the ray. Consequently, the integrated electron density $\int Nds$ is the quantity about which we can expect to obtain information from a careful measurement of the Doppler shift. Observing the change in angle of arrival caused by the ionosphere must lead to this same quantity since, as we have already seen, the angle of arrival is closely related to the Doppler shift.

If a satellite moves in a circular orbit in a spherically stratified ionosphere, the ray path integral of N changes. The principal cause of this change is the varying obliquity of the ray relative to the ionized strata: the integral is essentially proportional to the secant of the ray inclination to the vertical. Hence any ray path integral, to a first order, is equivalent in information content to

$$\int N\,dh \tag{1.19}$$

the number of electrons per unit cross-section area in a vertical column reaching to the height of the satellite. The detailed distribution of N and in particular the value of N at the satellite are not determined by either angle of arrival or Doppler shift measurements.

One does, however, obtain information about the value of N at the vehicle if it has a vertical component of motion (more generally, a component perpendicular to the stratification). This is most clearly seen in the case of a vertically moving rocket where the change in the vertical ray-path integral from one moment to the next is determined by N at the rocket.

Although the Faraday effect is a quite distinct phenomenon, we note that it depends on the *difference* of two refractive indices, each of which, for moderate ionization densities, departs from unity in linear dependence on N. Thus from equation (1.14)

$$\mu_{1,2} \approx 1 - \frac{e^2}{2\epsilon_0 m\omega(\omega \pm \omega_L)} N \tag{1.20}$$

and when we compare the polarization imposed by the ionosphere with what it would have been in the absence of electrons, we find again that our measurement involves the ray-path integral of $\Delta\mu$, which is proportional to the integral of N.

As before, the total number of electrons in a vertical column of unit cross section will be revealed, again with the proviso that sources with vertical components of motion can contribute local electron density information.

2. Satellites as a Tool in Ionospheric Research

2.1. *Doppler Studies*

In the absence of an ionosphere, and neglecting relativistic effects, the received Doppler shift would be directly proportional to the component of satellite velocity in the direction of the observer. However, the presence of the ionosphere results in refraction of the ray, a reduction in the index of refraction along some portions of the ray path, and less Doppler shift than would otherwise be expected. The difference between the observed shift and that which would be observed with no ionosphere may be called the *Doppler shift offset*. An accurate measurement of this quantity provides a considerable amount of information about the ionosphere, as we shall see below.

Of course, in order to measure the Doppler shift offset it is necessary to know what would have been observed had there been no ionosphere. This problem is elegantly solved by arranging for the transmitter to radiate a higher harmonic whose propagation is relatively unaffected by the ionosphere.

In discussing Doppler studies we shall neglect the birefringence caused by the earth's magnetic field and assume that the source frequency is high compared with the critical frequency of the ionosphere.

2.1.1. Free-Space Doppler Shift. First we deduce the Doppler shift and its rate of change in the absence of the ionosphere. Then the ionosphere will be included and the departure from free space conditions will be found.

In Fig. 6, C is the center of the earth, O is the observer, S the satellite, and P the proximal point, which, for free space conditions, is the foot of the perpendicular from O onto the orbit SP. The subsatellite point G is at the intersection of the earth's surface with CS, and the subproximal point M is the intersection with CP. The great circle GM is thus the ground track of the satellite. From the triangle OCS

$$R^2 = \rho_O^2 + \rho_S^2 - 2\rho_O\rho_S \cos\beta \tag{2.1}$$

where R is the distance of the satellite from the observer, ρ_S is its distance from the center of the earth, and β is the angular distance from the observer to the subsatellite point G.

From the right-angled spherical triangle GOM, we have

$$\cos\beta = \cos\theta \cos\beta_m \tag{2.2}$$

where β_m is the angular distance between the observer and the subproximal point M, and θ is the angular distance between the subsatellite and subproximal points. Hence

$$R^2 = \rho_O^2 + \rho_S^2 - 2\rho_O\rho_S \cos\theta \cos\beta_m \tag{2.3}$$

Differentiating with respect to time, and denoting the derivatives by dots, we obtain

$$R\dot{R} = \rho_S\dot{\rho}_S - \rho_O\dot{\rho}_S \cos\theta \cos\beta_m + \rho_O\rho_S\dot{\theta} \sin\theta \cos\beta_m \tag{2.4}$$

From equation (1.8) the Doppler shift is

$$-\frac{\dot{P}}{\lambda_o} = -\frac{\dot{R}}{\lambda_o} = -\frac{\rho_S\dot{\rho}_S}{R\lambda_o} + \frac{\rho_O\dot{\rho}_S}{R\lambda_o} \cos\theta \cos\beta_m - \frac{\rho_O\rho_S}{R\lambda_o}\dot{\theta} \sin\theta \cos\beta_m \tag{2.5}$$

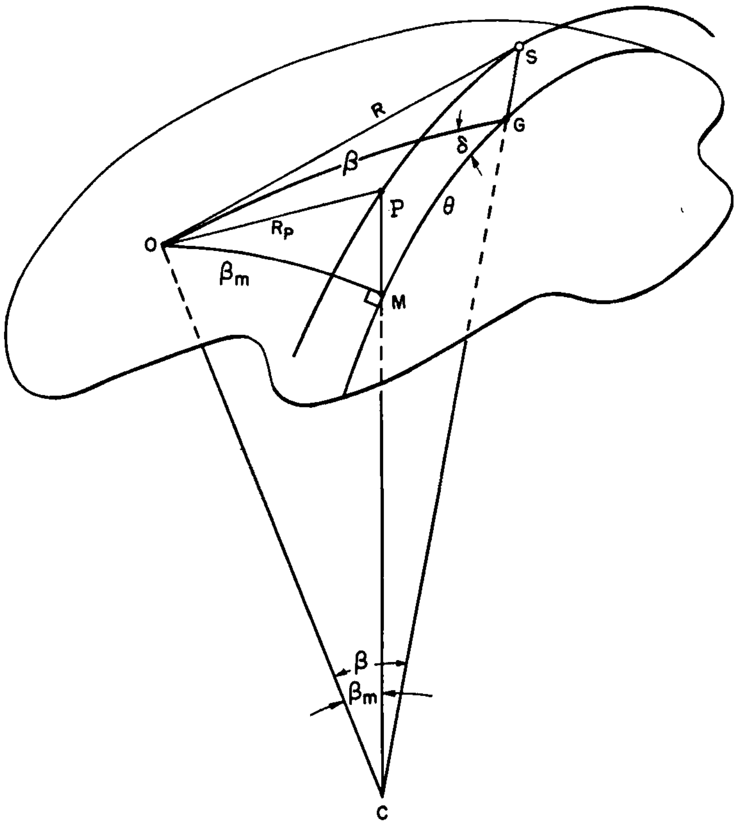


FIG. 6. Geometry necessary for the analysis of Doppler shift.

We can obtain a simpler result for the special case of circular orbits by putting the vertical velocity component $\dot{\rho}_S$ equal to zero. Then

$$-\frac{\dot{P}}{\lambda_0} = -\frac{\dot{R}}{\lambda_0} = -\frac{\rho_O v}{\lambda_0 R} \sin\theta \cos\beta_m \quad \text{cycles/second} \tag{2.6}$$

where the satellite velocity v is given by

$$v = \rho_S \dot{\theta} \tag{2.7}$$

To obtain the rate of change of the Doppler shift as time elapses, we differentiate equation (2.6) to find

$$-\frac{\ddot{R}}{\lambda_0} = \frac{-\rho_O v \cos\beta_m}{\lambda_0} \frac{d}{dt}\left(\frac{\sin\theta}{R}\right) \tag{2.8}$$

$$= \frac{-\rho_O v \cos\beta_m}{\lambda_0}\left(\frac{\dot{\theta}\cos\theta}{R} - \frac{\dot{R}\sin\theta}{R^2}\right) \tag{2.9}$$

Under our conditions the maximum (negative) rate of change of Doppler shift occurs when the satellite is at the minimum distance R_P from the observer. Then $\theta = 0$, and hence

$$\max\left(\frac{-\ddot{R}}{\lambda_o}\right) = \frac{-\rho_O v\dot{\theta}\cos\beta_m}{\lambda_o R_P} \tag{2.10}$$

$$= \frac{-\rho_O v^2 \cos\beta_m}{\rho_S \lambda_o R_P} \tag{2.11}$$

which is the equation obtained by Brito [1]. The corresponding result for a source moving in a straight line with velocity v and minimum distance R_P is

$$\frac{v^2}{\lambda_o R_P} \tag{2.12}$$

This "flat earth approximation," which has been rather widely used, is seen to be subject to serious error.

No specific mention has yet been made of the contribution to Doppler shift owing to the observer's motion on the surface of the rotating earth. The observer's motion is implicitly included when the velocity v is specified as the relative satellite velocity. However, the value of $\dot{\theta}$ in (2.7) will not quite be constant in this case, even for a circular orbit, since the angle between the satellite and observer velocities will change slowly with time. For passages near the observer these effects are quite small, but for more remote passages, a noticeable asymmetry in the Doppler curve may be found.

2.1.2. Ionospheric Effect with a Circular Orbit. Now the effects of the ionosphere will be considered. We assume that the satellite is moving in a circular orbit for the sake of concentrating on the peculiarly ionospheric phenomena. The increase in phase velocity caused by the ionosphere results in two things; first, the ray is bent by refraction and hence is longer than the straight free space ray; secondly, the number of cycles in the path is diminished because of an increased phase velocity. Of these two effects, the second is usually much the larger and so a first-order approximation can be obtained by calculating with the changed phase velocity in conjunction with the straight ray.

The phase path between the satellite and the observer is

$$\int \mu \, ds \tag{2.13}$$

In a medium with electron density N, in which collisions and the earth's magnetic field are not important, we have from equation (1.14)

$$\mu^2 = 1 - \frac{80.6N}{f^2} \tag{2.14}$$

where the frequency f is measured in cycles per second and N is the number of electrons per cubic meter. At sufficiently high frequencies; i.e., when the refractive index does not depart too far from unity

$$\mu \approx 1 - \frac{40.3N}{f^2} \tag{2.15}$$

The ionosphere therefore causes a reduction ΔP in phase path, from a value R in free space to a value P in the presence of the ionosphere. The difference between these quantities may be defined as the *phase path defect*, and it is given by

$$\Delta P = R - P = \int_O^S ds - \int_O^S \mu \, ds \tag{2.16}$$

$$= \int_O^S (1 - \mu) \, ds \tag{2.17}$$

$$\approx \frac{40.3}{f^2} \int_O^S N \, ds \tag{2.18}$$

where f is the frequency emitted by the satellite.[3]

In Fig. 7 we show the straight ray from the satellite to the observer, making an angle $\chi(h)$ with the vertical at height h. The angle χ varies slowly along the ray because of the curvature of the earth, and at the level of maximum ionization density, which will be in the F region, assumes a value χ_F. Thus

$$ds = \sec \chi dh \tag{2.19}$$

and

$$\Delta P = \frac{40.3}{f^2} \int_O^{h_S} N \sec \chi \, dh \tag{2.20}$$

$$\approx \frac{40.3 \sec \chi_F}{f^2} \int_O^{h_S} N \, dh = \mathfrak{N} \sec \chi_F \tag{2.21}$$

The quantity $\mathfrak{N}$ is equal to the phase path difference at vertical incidence. As an approximation, sec χ has been taken out of the integral and given the value sec χ_F which corresponds to the value of sec χ in the region where the bulk of the electrons is located. Similarly

[3] Strictly speaking one should use the Doppler-shifted frequency to evaluate the refractive index.

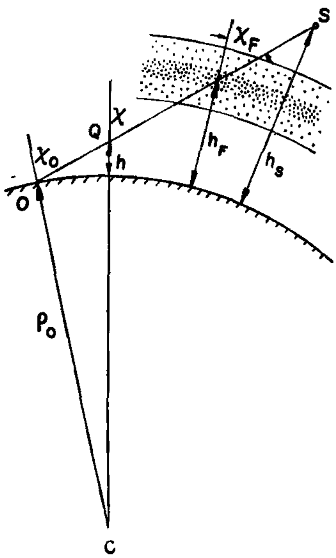


FIG. 7. Additional geometry, used when the ionospheric effect is included in the Doppler shift equations.

$$R = \int ds = h_S \sec \hat{\chi} \tag{2.22}$$

where $\hat{\chi}$ is an average value close to $\chi(h_S/2)$, it follows that

$$\frac{\Delta P}{R} = \frac{\mathfrak{N} \sec \chi_F}{h_S \sec \hat{\chi}} = \mathcal{R} \tag{2.23}$$

This ratio, which we have denoted by $\mathcal{R}$, is practically constant as time elapses, since there is little difference between χ_F and $\hat{\chi}$ and $\mathfrak{N}$ is a constant when the ionosphere is spherically stratified. Hence

$$\frac{\Delta \dot{P}}{\dot{R}} = \mathcal{R} = \frac{\mathfrak{N} \sec \chi_F}{R} \approx \frac{\mathfrak{N}}{h_S} \tag{2.24}$$

Furthermore, $\Delta\ddot{P}/\ddot{R} = \mathcal{R}$ and in particular

$$\frac{\max \Delta \ddot{P}}{\max \ddot{R}} = \mathcal{R} \tag{2.25}$$

Thus the ratio of the Doppler shift offset to the Doppler shift (or the slopes of these two quantities) may be used to determine the integrated electron density from

$$\mathfrak{N} = \frac{40.3}{f^2} \int_0^{h_S} N \, dh = R \cos \chi_F \frac{\Delta \dot{P}}{\dot{R}} \tag{2.26}$$

It is necessary to make these frequency measurements with such great precision that special techniques must be employed, if the required accuracy is to be achieved.

One method, which has proven satisfactory with the transmissions from Sputnik III, utilizes a phase-locked receiver at one of the satellite transmitter frequencies. Sputnik III radiated relatively strong signals at about 20 Mc and the second harmonic near 40 Mc was frequently detectable. During most of the satellite's lifetime, the transmissions were short pulses which effectively prevented the acquisition of a phase-lock other than intermittently. However, in the last few months of its life, the transmissions were usually CW and Doppler shift analysis was greatly facilitated. In this period an oscillator could be phase-locked with the incoming 40-Mc second harmonic. At Stanford we counted the phase-locked oscillations over a standard interval and printed the result automatically, thus obtaining an accurate tabulation of the observed high frequency.

In addition, the phase-locked signal was divided by two and compared with the incoming 20-Mc satellite frequency in another receiver. Although the two signals were radiated as harmonics, the effect of the ionosphere resulted in a nonharmonic relationship at the ground. This simply implies that the Doppler shifts imparted to the two signals were not precisely proportional to their frequencies. The frequency of the resultant beats between the generated 20-Mc signal (one-half of the received 40-Mc signal) and the received 20-Mc signal is proportional to the Doppler shift offset $\Delta\dot{P}/\lambda_0$.

Figure 8 is a plot of both the observed high frequency and Doppler shift offset obtained from a passage of Sputnik III near the end of its lifetime. At the time of the observations, the satellite was at a height of about 175 km and almost exactly at perigee where the vertical velocity component was zero. It has been determined that the plasma frequency never exceeded 6 Mc along the ray path and this implies that the refractive index was between 0.95 and unity at all points. The high frequency approximation of equation (2.15) and the neglect of refraction are therefore satisfactory assumptions in this case.

In Fig. 8, $\Delta\dot{P}$ may be obtained directly from the lower curve, but before $\dot{R}$ can be obtained from the upper curve it is necessary to know the satellite frequency f_S. A time separation between the zeros of Doppler shift offset and Doppler shift may result from ionospheric irregularities, horizontal gradients, or a vertical component in the satellite velocity. These factors were apparently not too serious in the passage under consideration, because the Doppler shift curve appears to be symmetrical about the time at which the Doppler shift offset passes through zero. Hence at this instant the observed frequency was set equal to the satellite oscillator frequency f_S. After marking the value of f_S as shown in Fig. 8 one can read $\dot{R}$. Values of the ratio $\Delta\dot{P}/\dot{R}$ can be averaged so as to minimize the fluctuation of

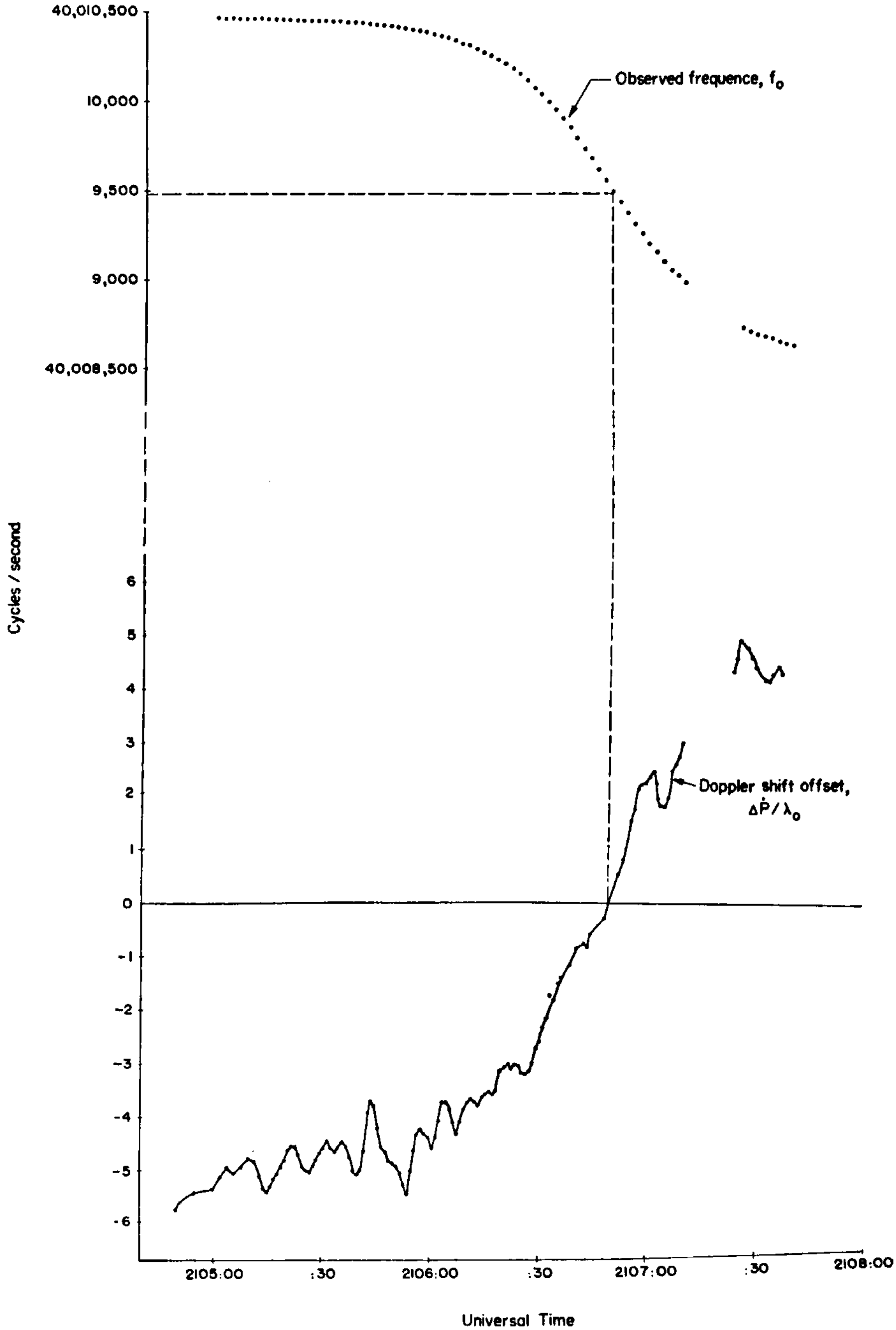


FIG. 8. The observed frequency f_O of the second harmonic of the 20-Mc transmitter aboard Sputnik III. The Doppler shift offset was obtained as four-thirds of the difference in frequency between the received 20-Mc signal and one-half the second harmonic. The data were obtained on March 21, 1960 when the satellite was very near perigee at approximately 175 km, heading SE.

$\Delta \dot{P}$ caused by inhomogeneities in the E region, and the value so found can be substituted in equation (2.26) to obtain the integrated electron density. From the observations of Fig. 8, a value of 1.7×10^{16} electrons/meter^2 was obtained. It is of interest to note that neither an accurate ephemeris nor even the distance of closest approach is needed for this deduction, since $R \cos \chi_F$ is closely equal to the satellite height.

In this description we have spoken as if 40 Mc is a frequency high enough to be unaffected by the ionosphere. But, in fact, the quantity ΔP is proportional to $1/f^2$ and its value at 40 Mc should be approximately ¼ of that at 20 Mc. Thus the correction factor to be applied to the observed beat frequency to obtain the Doppler shift offset used in equation (2.26) is

$$\frac{f_{\text{high}}^2}{f_{\text{high}}^2 - f_{\text{low}}^2} = \frac{4}{3} \tag{2.27}$$

Hibberd and Thomas [2] have suggested a procedure for estimating the electron density profile above the F-layer maximum density which is also based on the observation of the Doppler shift offset curve. They suggest that the satellite orbital elements, the true height profile $N(h)$ for the lower ionosphere, and various model profiles above the F peak be combined in a computer program. This program would then select the model that agreed best with the observed offset.[4]

2.1.3. Noncircular Orbits. We have thus far restricted the discussion to circular orbits; now let us consider the situation in which the satellite velocity may be in any general direction, although we retain the restriction of a spherically stratified ionosphere. The horizontal component of the velocity is $\rho_S \dot{\theta}_S$, and the vertical component is $\dot{\rho}_S$. It is required to find the Doppler shift.

From Fig. 1 we saw that the Doppler shift is equal to the number of wavecrests crossed per second as the satellite approaches. We now split the satellite velocity into its horizontal and vertical components and note that the net rate of crossing wavecrests is the sum of the rates at which they are being crossed horizontally and vertically.

[4] In the last two sections we have introduced two new quantities which, owing to the recent development of satellite studies, have thus far gone without definition. Perhaps the most basic effect of the ionization in space on radio waves is to reduce the phase path between the satellite and the receiver. This reduction, $\Delta P \equiv R - P$, has been termed the *phase path defect*. This in turn slightly alters the received Doppler shift by an amount $\Delta \dot{P}/\lambda_0$, defined as the *Doppler shift offset*. This quantity is closely related to the beat frequency obtained between two received radio frequencies which were originally radiated by a satellite as harmonics, as we have described in the last few pages. The frequency of this beat has been called "differential Doppler" or "dispersive Doppler" by several authors, and when corrected by the fraction given in 2.27 results in (to a first order) the Doppler shift offset.

From Fig. 3 we saw that the horizontal spacing in the plane of propagation (the vertical plane containing the ray) was $\lambda/\sin \chi_S$. But since in general the orbital plane will be inclined at an angle δ to the plane of propagation (see Fig. 6), the horizontal spacing in the plane of the orbit will be greater by a factor sec δ, viz.

$$\text{horizontal spacing} = \lambda \operatorname{cosec} \chi_S \sec \delta \tag{2.28}$$

The horizontal rate of crossing wavecrests is thus

$$\frac{\rho_S \dot{\theta}}{\lambda \operatorname{cosec} \chi_S \sec \delta} \tag{2.29}$$

where $\rho_S \dot{\theta}$ is the horizontal component of velocity.

Also from Fig. 3 the vertical spacing SU is given by

$$\text{vertical spacing} = \lambda \sec \chi_S \tag{2.30}$$

hence the vertical rate of crossing is

$$\frac{\dot{\rho}_S}{\lambda \sec \chi_S}$$

where $\dot{\rho}_S$ is the vertical component of velocity. Thus from the net number of wavecrests crossed per second we have

$$\text{Doppler shift} = -\frac{\rho_S \dot{\theta}}{\lambda \operatorname{cosec} \chi_S \sec \delta} - \frac{\dot{\rho}_S}{\lambda \sec \chi_S} \tag{2.31}$$

When the wavelength and angle of incidence are converted to ground-level values, using equation (1.11) as was done previously in equation (1.13), the Doppler shift becomes

$$-\frac{\rho_O \sin \chi_O}{\lambda_o \sec \delta} \dot{\theta} - \frac{\{\mu_S^2 - [(\rho_O/\rho_S) \sin \chi_O]^2\}^{1/2}}{\lambda_o} \dot{\rho}_S \tag{2.32}$$

From the spherical triangle OMG (Fig. 6), we may express sec δ in terms of θ, β_m, or β as required.

For horizontal overhead motion with velocity v, $\delta = 0$, $v = \rho_S \dot{\theta}$ and equation (2.32) reduces to

$$-\frac{\rho_O v \sin \chi_O}{\rho_S \lambda_o} \tag{2.33}$$

as derived by other authors [3, 4]. This special case does not involve the local refractive index μ_S. For vertical motion, as in the case of sounding rockets, equation (2.32) becomes

$$-\frac{\mu_S \dot{\rho}_S}{\lambda_o} \tag{2.34}$$

and we see that in this special case the Doppler shift gives information about the local refractive index irrespective of the structure of the intervening medium. This type of result will hold wherever the source trajectory is along a ray.

In the most general situation, a noncircular orbit and a nonspherically stratified medium, the expression for Doppler shift can be split into two terms, one that does and one that does not involve the local refractive index μ_S. Suppose that the satellite velocity is the sum of two nonorthogonal components, one along the ray and one along the stratification. Then if the satellite moves an infinitesimal distance along the ray, the change in phase path will depend on μ_S but not on the refractive index profile at lower heights because there will have been no change in the ray. Now suppose that the refractive index of a thin stratum containing the satellite is changed to a new value. Then if the satellite moves an infinitesimal distance within the stratum, the rays to the observer, and hence the Doppler shift, will approach their original conditions as the thickness of the stratum approaches zero. Thus the Doppler shift produced by motion along the stratification is independent of μ_S.

Thus we can write

$$\text{Doppler shift} = \frac{\sin i_S}{\lambda} v_{\text{strat}} + \frac{v_{\text{ray}}}{\lambda} \tag{2.35}$$

where i_S is the angle between the ray and the normal to the plane of stratification and is measured at the satellite, v_{ray} is the velocity component *down* the ray, and v_{strat} is the component along the stratification. The second term is proportional to μ_S while the first, as we have shown, is independent of μ_S. In the particular case of equation (2.32), $i_S = \chi_S$, $v_{\text{ray}} = -\dot{\rho}_S \sec \chi_S$ and $v_{\text{strat}} = \rho_S \dot{\theta} + \dot{\rho}_S \tan \chi_S$.

2.2. Faraday Rotation

The second kind of information furnished by a satellite equipped with a transmitter is the state of polarization of a wave that has been propagated down through the ionosphere to the ground. If the wave is linearly polarized the pitch of the helicoid into which the plane of polarization is twisted was shown to be $2\,\lambda_0/\Delta\mu$ in equation (1.17), where $\Delta\mu$ is the difference between the refractive indices of the two magneto-ionic modes; hence, the rotation $d\Omega$ (measured in full turns) in a distance ds is given by

$$d\Omega = \frac{\Delta\mu}{2\lambda_0}\, ds \tag{2.36}$$

The difference in refractive index can be obtained from equation (1.14). For sufficiently high frequencies we have

$$\Delta\mu = \frac{Ne^2\omega_L}{\epsilon_0 m\omega^3} = \frac{N}{N_c}\frac{\omega_L}{\omega} \tag{2.37}$$

where $\omega_L = e\mu_0 H_L/m$ and N_c is the critical electron density at the frequency ω.

In Section 2.1.2 we calculated the change in phase path produced by electrons by integrating along the unrefracted ray with the changed phase velocity. Now we estimate the total number of rotations Ω in the same way.

Although the rotation of polarization is due to the fact that the two magneto-ionic modes encounter different refractive indices, nevertheless we assume that they follow the same ray path. Let the ray make an angle χ with the vertical at each height h. Then the number of full rotations of polarization is

$$\Omega = \int_0^{h_s} \frac{Ne^2\omega_L \sec\chi \, dh}{2\lambda_0\epsilon_0 m\omega^3} \tag{2.38}$$

$$= \frac{e^3\mu_0}{16\pi^3 m^2\epsilon_0 c f^2}\int_0^{h_s} H_L N \sec\chi \, dh \tag{2.39}$$

The collection of constants may be reduced to

$$K = \frac{e^3\mu_0}{16\pi^3 m^2\epsilon_0 c} = 4.72 \times 10^{-3} \qquad \text{(mks)} \tag{2.40}$$

As a satellite travels from the south to the north of a northern hemisphere observer, the value of Ω is usually found to decrease. This implies that the difference in the phase paths between the two magneto-ionic modes is becoming smaller. The variations in the total rotation angle Ω can be observed in a very simple manner. When the satellite transmits a linearly polarized wave the received polarization will also be linear. When the signals are received on a dipole, the rotation of the plane of polarization results in periodic fading and each fade corresponds to an increase or decrease of π radians in the total rotation angle. This last ambiguity can be resolved by noting the phase of the fading on orthogonal dipoles.

If equation (2.39) is to be used to obtain a value of the integrated electron density, a method must be found to evaluate the magnitude of Ω at some time during the satellite passage. The value of Ω at any other time could be found by counting fades on the received signal strength record.

One possibility is provided by the geometry of the earth's magnetic field. In equation (2.39) the factor H_L indicates that Ω goes to zero when propagation becomes transverse to the earth's field. Although the above analysis has excluded propagation too close to perpendicular to the earth's field, in fact the difference in the phase paths does approach a small value at the time of transverse propagation. If a time of transverse propagation can be

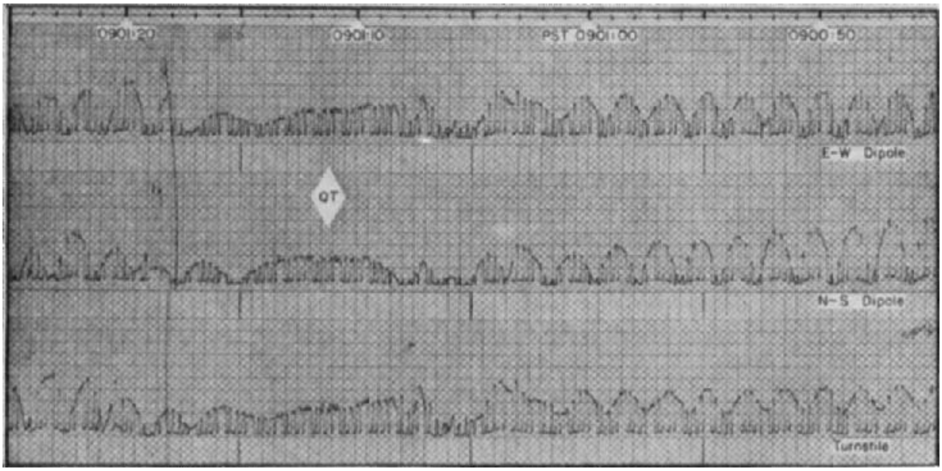


FIG. 9. Record of the fading due to Faraday rotation of the 20-Mc signals from 1958 δ_2 on May 24, 1958. The direction of rotation of the plane of polarization may be determined by comparing the relative phase of the fading on the upper two channels.

identified on the signal strength record, it can be assumed as a first approximation that Ω is zero at this time. Then the absolute value of Ω at the time the satellite passes nearest the observer's zenith can be determined from the fading record, and from (2.39) the value of the integrated electron density, $\int_0^{h_S} N\,dh$, calculated. Figure 9 is a recording of the signal strength of Sputnik III when received using several antennas. The region of "quasi-transverse" propagation is observed near the center of the record. The received polarization at this time was elliptical, rather than linear, and is an indication that the "quasi-longitudinal" approximation, used in the analysis above, is not valid in this region.[5]

The number of fades between the time marked QT in Fig. 9 and the time of the closest approach of the satellite is approximately equal to the number of half rotations of polarization which existed at the latter time. Knowing the position of the closest approach permits the value of $H_L \sec \chi$ in (2.39) to be estimated, and then the integrated electron density can be computed. On many occasions the time of transverse propagation cannot be identified; in these cases the absolute value of Ω at some time must be established by other means.

If refraction is neglected and a flat earth with horizontal ionospheric stratification assumed, it is quite simple to relate the rate of Faraday rota-

[5] A detailed discussion of the quasi-longitudinal and quasi-transverse conditions is given by Ratcliffe [5].

tion to integrated electron density [6]. Taking the time derivative of (2.39) leads to

$$\frac{d\Omega}{dt} = \frac{K}{f^2}\frac{vH_x}{h_S}\int_0^{h_S} N\,dh \text{ rotations/sec} \tag{2.41}$$

where v is the satellite velocity and H_x is the component of the earth's field in the direction of the satellite motion.

The Faraday rotation rate may be thought of as one-half the difference in the Doppler shifts imparted to the two magneto-ionic modes [7]. This concept is most important when computers are available for the calculation of each individual phase path length.

It is necessary to keep in mind the various approximations which have been included in equations (2.39) and (2.41). In both equations it has been assumed that the wave frequency is much higher than the plasma frequency, which permits the approximate form of $\Delta\mu$ given in (2.37) to be used. At 20 Mc this approximation is frequently in error by 10 % to 20 % in the daytime. A rather accurate correction can be based on the F-layer critical frequency at the time of the observations. Nevertheless, several authors have used equations similar to (2.41) without such a correction. It has been assumed that the satellite velocity had no vertical component and that there existed no horizontal ionospheric gradients. The assumption that both modes follow the same ray to the observer is analyzed in detail in reference [8].

It has also been assumed that the rotation of the satellite about its center of gravity has no effect. But in fact satellite rotation can produce deep fading, as nulls of the transmitting antenna pattern pass near the direction of the departing ray. This may completely mask the fading due to Faraday effect, as in examples given by Bracewell and Garriott [9]. The situation can be rectified by controlling the attitude of the satellite, or more simply by using a transmitting antenna pattern with rotational symmetry about the axis of rotation of the satellite. It is possible to insure that the free motion will settle down to rotation about a known body axis, under the influence of internal damping, by making the satellite disclike, and details of the dynamics have been given by Bracewell [10].

A turnstile antenna consisting of dipoles excited in quadrature is simple, virtually omnidirectional, and has good rotational symmetry. The electromagnetic field pattern rotates about the turnstile axis at the radio frequency. If now the satellite carrying the antenna rotates about the same axis, making F revolutions per second, the wave frequency is merely increased or decreased by F, with only minor effects on amplitude. Should the

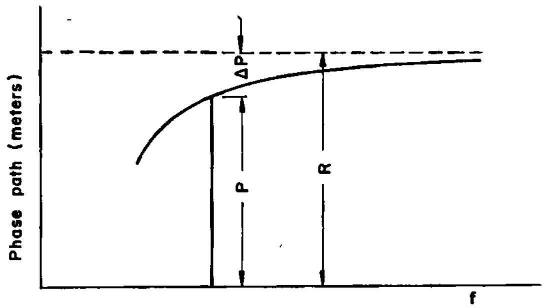


FIG. 10. The length of the phase path P vs. frequency from an observer to a geostationary satellite. At the highest frequencies, P approaches the true range R. The phase path defect is given by $(R - P)$.

satellite be emitting one fundamental frequency and its nth harmonic, the precise harmonic relationship would be destroyed. Consequently, there would be an apparent Doppler shift offset $\pm (1 - n^{-1}) F$, which should be applied before determining the zero of Doppler shift offset as explained in connection with Fig. 8. Even more simple, of course, would be a dipole aligned along the axis of rotation. In this case none of the effects of satellite rotation would be discernible on the ground.

2.3. Geostationary Satellite

Satellite orbits within a few thousand kilometers of the earth's surface have one characteristic which is frequently a serious disadvantage for ionospheric studies. This is the high relative velocity between the observer and the satellite. As a result of this rapid motion, observations can be made only for brief periods and at varying positions in the sky. These limitations can be avoided with a geostationary satellite, that is, a satellite in a west to east, circular equatorial orbit with a period of 24 hr. Such an orbit is possible at a geocentric range of about 6.5 earth radii.

From the radio transmissions of a geostationary satellite, the electron content of the ionosphere and its diurnal variation may be easily determined in theory. One method could be based on the dispersion of radio waves. Figure 10 is a sketch of the phase path P from the satellite to the observer as a function of frequency. We have

$$P = \int \mu \, ds = \int \left(1 - \frac{80.6N}{f^2}\right)^{1/2} ds = R - \Delta P \tag{2.42}$$

At the higher frequencies where the refractive index approaches unity the phase path approaches the range R, but at the lower frequencies the phase path is less than the range by an amount ΔP. Provided sufficiently high frequencies are considered, we have the approximation

$$P \approx \int \left(1 - \frac{40.3N}{f^2}\right) ds = R - \frac{40.3}{f^2} \int N\, ds \tag{2.43}$$

Hence if one could measure ΔP, the reduction in phase path relative to the range, which we have defined as the phase path defect, the electron content could be deduced.

Even if the range were not known, any *changes* of electron content would be revealed by changes in the phase path. This is the basis of a proposal by Garriott and Little [11] in which absolute phase changes of a 40-Mc signal from a satellite would be determined by reference to a signal radiated from the satellite on a harmonic frequency, say, 120 Mc. On reception, the 120-Mc signal, which would have suffered a small but accountable phase advance, would be divided by three and used to beat against the 40-Mc signal. The beat frequency

$$\frac{\Delta \dot{P}}{\lambda_0} = \frac{40.3}{fc} \frac{d}{dt} \int N\, ds \tag{2.44}$$

would then measure the rate of change of electron content.

The beat frequency given by equation (2.44) is analogous to the Doppler shift offset observed with satellites at lower heights. In the present case the beats occur because the medium is nonstationary [equation (1.8)]; but in the Doppler shift offset experiment with satellites at lower heights the phase path variation was predominantly due to satellite motion, any variation in electron density of the medium being small during the brief passage of the satellite.

In order to measure the total electron content, rather than merely its time changes, it is necessary to know ΔP. Since ΔP proves to be of the order of a few kilometers, the wavelength of 7.5 meters is a rather inconveniently small yardstick for the measurement. A yardstick of several kilometers would be more appropriate. A way of obtaining this is as follows. Let a 40-Mc signal be amplitude modulated at a frequency of 10 kc. Then the distance between peaks of the modulation envelope is 30,000 meters in free space and very little different in the ionosphere. Now the 40-Mc signal arrives at the ground advanced in phase by an amount depending on the phase velocity. The modulation envelope on the other hand moves at the group velocity and is retarded. Because the group velocity is as much below the velocity in free space as the phase velocity is above,[6] the modulation envelope slips back a distance ΔP. Hence, it is retarded in phase by $\Delta P/30{,}000$

cycles at the modulation frequency. In order to know the amount of this lag it is sufficient to impose the identical modulation on a carrier of higher frequency, say 120 Mc, and to receive this signal also. The modulation of the latter is little affected by the ionosphere and furnishes a reference waveform from which to measure the phase of the 10-kc modulation carried by the 40-Mc signal. The phase difference then gives ΔP and hence the electron content.

A separate calculation may be made by observing the received polarization on two closely spaced frequencies. By measuring the change $d\Omega$ in the number of Faraday rotations over a frequency difference df, the total rotation at either frequency is determined. Since $\Omega \propto f^{-2}$ it follows that

$$\Omega = -\frac{f}{2}\frac{d\Omega}{df} \tag{2.45}$$

and so from equation (2.39)

$$\int_0^{h_S} H_L N \sec \chi \, dh = \frac{-f^3}{9.44 \times 10^{-3}} \frac{d\Omega}{df} \tag{2.46}$$

This same technique could be used with lower satellites, as suggested by Daniels [12], but then only a few observations each day would be possible at times when the satellite passed nearby.

Since the Faraday rotation angle is weighted by the earth's magnetic field, electrons at heights above a few thousand kilometers are almost overlooked, as they fail to contribute to the polarization rotation. However, the dispersion experiment is equally sensitive to electrons at all heights. Therefore, a comparison of the two values of integrated electron density may permit the average electron density between 2 and 6 earth radii to be deduced.

It would seem even more desirable for the orbital period to be slightly nonsynchronous, perhaps either 23 or 25 hr. Then the satellite would very slowly move in longitude, and every month the entire world would have an opportunity to make measurements. Since the observations at any location would still be continuous for more than a week, the effects of solar flares, solar "winds," and geomagnetic storms could readily be studied.

2.4. Refraction Studies

The refraction of radio waves has presented some difficulty to the accurate positioning of radio stars. The radio position of a radio star tends to be closer to the zenith than is the true position of the source. However, the apparent shift in position of a radio emitter provides still another method

[6] In an ionized isotropic medium, the group velocity $c(1\text{-}80.6\ N/f^2)^{1/2}$ is such that the geometric mean of the phase and group velocities is c.

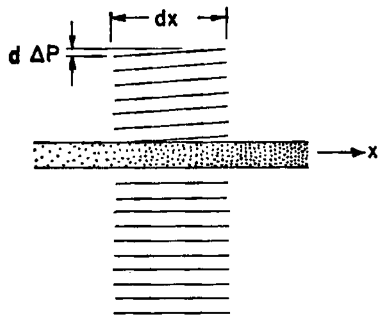


FIG. 11. The angular deviation is equal to the gradient of the phase path defect ΔP.

for estimating the total ionospheric electron content. Consider the time of transit of a radio star, which can be observed with precision by radio interferometry. Let the zenith distance of the star be χ; then the phase path defect due to passage through the ionosphere is $\Delta P = \int (1 - \mu) \sec \chi \, dh$. Let x and y be coordinates in the wavefront measured respectively to the west and to the north from the intersection of the ray with the ionosphere. If the value of ΔP for neighboring parallel rays is different, there will be a displacement in apparent position given in magnitude and direction by grad ΔP. Figure 11 shows the one-dimensional situation. Thus the westerly displacement in radians is given by

$$\frac{d\,\Delta P}{dx} = \frac{40.3}{f^2} \sec \chi \frac{d}{dx} \int N \, dh \tag{2.47}$$

and the northerly displacement by

$$\frac{d\,\Delta P}{dy} = \frac{40.3}{f^2} \sec \chi \frac{d}{dy} \int N \, dh \tag{2.48}$$

$$= \frac{40.3}{f^2} \sec^2 \chi \frac{d}{dy'} \int N \, dh \tag{2.49}$$

where y' is measured *horizontally*.

This situation was described rather early in the history of radio astronomy by F. G. Smith [13, 14] who, by measuring east-west displacements of four radio stars, estimated the total electron content by fitting the observations to the computed seasonal variation of the east-west gradient at the time of transit. All these displacements were less than one minute of arc at 81.5 Mc. More recent measurements have been made at 19.7 Mc, showing the refraction errors in both right ascension and declination. Komesaroff and Shain [15] find the refraction error in declination to be consistent with estimates of the total electron content at the time of the measurements.

Two inherent difficulties with the radio star measurements may be eliminated when a satellite-borne transmitter is used for the refraction studies. These are the poor signal to noise (S/N) ratios attainable and the finite angular dimensions of the source. With the coherent radiation from a satellite, filtering and cross-correlation techniques are easily employed which result in excellent S/N ratios, even with very low-powered transmitters. Although the early earth satellites do have the disadvantage of high relative velocity between the receiver and transmitter, even this difficulty would be removed with an approximately stationary satellite, as we have discussed previously.

Consider the situation shown in Fig. 12 where OS is the ray from the observer to the satellite. Due to refraction the satellite at S appears to be in position S′, where OS′ is the tangent to OS at O and S′ is, let us say, at the same height as S. Let the tangent S′O pass within a distance a of the center of the earth. From the small triangle we have

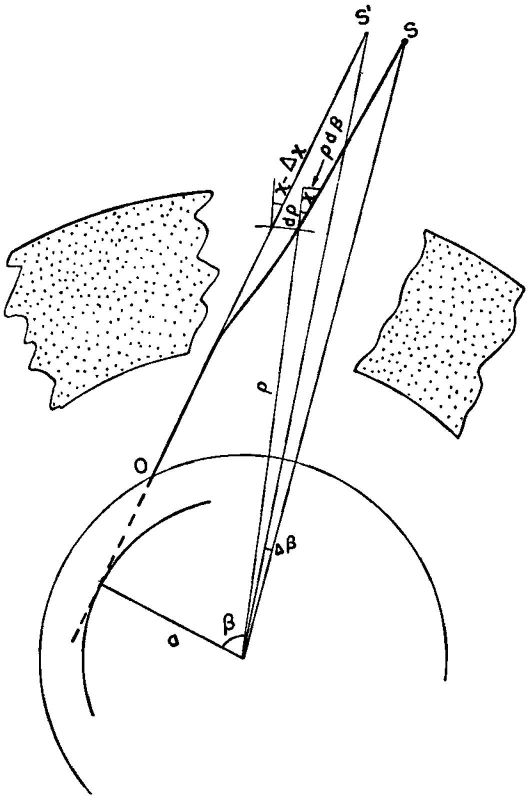


Fig. 12. Geometry necessary for the analysis of the ionospheric refraction of satellite radio transmissions.

(2.50) $$\tan \chi = \frac{\rho \, d\beta}{d\rho}$$

Hence, we can write the equation of the ray in polar coordinates (ρ,β) as

(2.51) $$\beta = \int_a^\rho \frac{\tan \chi}{\rho} \, d\rho$$

At any radius ρ, let the tangent OS′ make an angle $\chi - \Delta\chi$ with the vertical, where χ is the angle that the ray OS makes at the same height. Then the equation of the tangent is

(2.52) $$\beta = \int_a^\rho \frac{\tan (\chi - \Delta\chi)}{\rho} \, d\rho$$

Let $\Delta\beta$ be the difference in the values of β for S and S′ as given in the last two equations. Thus, provided $\Delta\chi$ is small

(2.53) $$\Delta\beta = \int^{\rho_S} \frac{\Delta\chi \sec^2 \chi}{\rho} \, d\rho$$

We obtain an expression for $\Delta\chi$ as follows. From Snell's law

(2.54) $$\mu\rho \sin \chi = a$$

at all points on the ray. Likewise, at all points on the tangent

(2.55) $$\rho \sin (\chi - \Delta\chi) = a$$

Hence,

(2.56) $$\sin (\chi - \Delta\chi) = \mu \sin \chi$$

and

(2.57) $$\Delta\chi = (1 - \mu) \tan \chi$$

(2.58) $$\approx \frac{40.3N}{f^2} \tan \chi$$

Thus

(2.59) $$\Delta\beta = \int^{\rho_S} \frac{40.3N \tan \chi \sec^2 \chi}{f^2\rho} \, d\rho$$

As the major contribution to the integral occurs near the height of maximum electron density in the F layer,

(2.60) $$\Delta\beta \approx \frac{40.3 \tan \chi_F \sec^2 \chi_F}{f^2\rho_F} \int^{\rho_S} N \, d\rho$$

which is equivalent to an approximate equation obtained by D. K. Bailey

[16] as early as 1947. Woyk [17] and Bremmer have also obtained a much more sophisticated solution which does not involve the approximations made above, although inhomogeneities and horizontal gradients in the ionosphere may prevent the achievement of any greater accuracy than that obtained with equation (2.60).

In practical use $\Delta\beta$ will probably not be directly measurable; more likely ϵ, defined as the angle subtended by SS′ at the observer O, will be obtained. In this case, $\Delta\beta$ can be computed trigonometrically. A simple relation may be shown to be

$$\Delta\beta = \frac{R}{\rho_S}\left(1 - \frac{a^2}{\rho_S{}^2}\right)^{-1/2} \sin\epsilon \tag{2.61}$$

Low altitude satellites have thus far not permitted use of the refraction techniques described above principally because of their high relative velocity to the observer. However, there are some advantages to angle of arrival measurements as compared with the more easily obtained Doppler shift and Faraday rotation. For example, the angular position of the source is not affected by the vertical velocity component of the satellite, as are the Doppler shift and Faraday rotation rate experiments.

2.5. Local Electron Density Measurements

2.5.1. High-Frequency Techniques. Perhaps the most useful single piece of ionospheric information would be the electron density as a function of height. The lower ionosphere below the height of maximum density has been studied for several decades with vertical incidence sounders [18], and the general features of the lower layers are now well established. In addition, rocket measurements have provided data up to a height of over 200 km [19–21]. In the method developed by Seddon, two harmonically related frequencies are radiated by a rocket in a nearly vertical trajectory. The received frequencies are not precise harmonics, since the refractive index at the rocket is smaller for the lower of the two frequencies and thereby affects the Doppler shifts differently. After dividing the higher frequency by the integer ratio of the two transmitted frequencies, the two low frequencies are compared. The beat between the two is a measure of the electron density at the satellite. This result was noted before when we considered the Doppler shift imparted to satellite signals.

In the last few years, the altitudes achieved by rockets have continued to increase and reliable F-layer measurements are beginning to be obtained [22, 23]. Still, the electron density distribution above the F-layer maximum is in reasonable doubt and a number of high-frequency (HF) techniques involving satellites have been proposed to measure the ionization density of the plasma.

A very interesting prototype experiment has been flown in a rocket by Bordeau, Jackson, Kane, and Serbu in which measurements were made directly upon the atmospheric environment [24]. Measurements of electron density were made by noting the capacitive detuning of a HF antenna, and these results were compared with those determined from the propagation-type experiment developed by Seddon. In addition, estimates of the sheath thickness around the rocket and values of the electron temperature and positive ion concentration were obtained at altitudes of 130 to 210 km. Properly calibrated experiments of this nature will certainly be of great value when carried out in satellite orbits above the F layer.

A more complicated but extremely valuable experiment involves placing a vertical incidence sounder in a satellite orbit well above the height of the F-layer maximum [25]. This "topside sounder" would transmit short pulses toward the ground and measure the delay time of the returned echo as a function of the transmitted frequency. When the frequency is less than the F-layer critical frequency, the pulses will be reflected from the top portion of the layer. The $h'(f)$ curve of pulse delay time vs. frequency which will be obtained can then be related to the electron density profile above the F-layer peak.

Another experiment which may be performed in the near future is primarily designed for the measurement of the cosmic noise level but also determines the local electron density as a secondary goal [26]. A receiver tuned to, say, 3 Mc is carried in a satellite whose orbit reaches above and below the level at which 3 Mc is the critical frequency. This level is indicated by the broken line in Figure 13. When the satellite is at S it can receive radiation only from sources lying in a volume bounded by the surface Σ. We refer to Σ as the limit-of-escape surface for a source at S; it is the envelope of rays leaving S in all possible directions. There are no known terrestrial sources within the limit-of-escape surface; on occasion it may include the sun, but

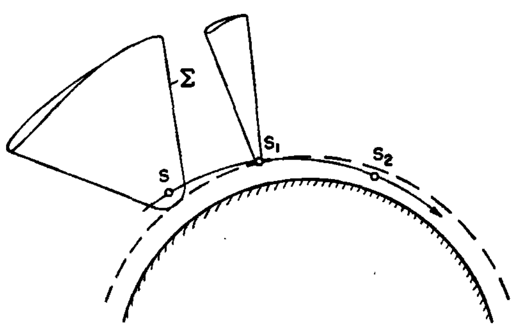


FIG. 13. Sketch of the variations in the limit-of-escape surface for a particular wave frequency as the satellite passes into a nonpropagating region.

the predominant source by far is that cone of the galaxy centered on the vertical.

By telemetering the received noise level back to the ground one could survey the galactic noise distribution over the sky, in a rough way. Very little is known of the galactic noise at 3 Mc, because hitherto the ionosphere has hindered the measurements, but it is considered an important matter in radio astronomy.

Now as the satellite drops in height to S_1, the cone of reception narrows, thus tending to limit the portion of the sky contributing to the received noise, until finally the cone collapses completely as the satellite crosses the critical level. At S_2 the satellite is surrounded by a medium whose plasma frequency is greater than 3 Mc (electron density greater than about 10^5 electrons/cm^3) and through which 3-Mc radiation cannot propagate. It is expected that the sharp drop in received power will reveal the location of the 3-Mc level.

In an extension of the experiment the receiver frequency would be swept, and the satellite would be placed in a highly eccentric orbit. Thus the local plasma frequency will be determined at numerous points along the satellite orbit.

Many effects may contribute to mar the sharpness of the expected cut-off. Thus the 3-Mc level will be split by the earth's magnetic field, and any inhomogeneities will upset the interpretation. There may also be sources of noise that we are unaware of.

All in all, very rich results may be expected from the topside sounder and galactic noise recorder.

2.5.2. A Very Low Frequency Experiment. An entirely different approach has been taken by Storey [27]. He has proposed that the local plasma frequency be determined by a measurement of the "local wave admittance of the medium." This measurement is to be accomplished with a ground-based very low frequency (VLF) transmitter and a satellite-borne receiver. A portion of the transmitted energy would propagate through the ionosphere in the "whistler" mode. At very low frequencies the index of refraction for the propagating mode is quite large in the ionosphere, perhaps 10 to 100, and it is this fact which makes the experiment so useful. The wave admittance is proportional to the index of refraction and thereby to the square root of the local electron density. The large variations in the magnitude of the refractive index imply a sensitive measurement of the plasma density.

As the satellite passes in the vicinity of the transmitter, the amplitude of the magnetic and electric fields of the propagating wave would be measured. This would be accomplished by connecting a loop antenna and a dipole to separate receivers. These antennas would be responsive to the magnetic and electric fields, respectively. The ratio of the amplitudes of the two signals is the desired wave admittance. The experiment is somewhat complicated in that the magnitude of the earth's magnetic field at the satellite must be known as well as its direction with respect to the wave normal of the propagating wave.

3. Some Results of the Satellite Experiments

Of the first 15 satellites launched by the United States (1958 Alpha to 1959 Lambda) only one, 1959 Iota, has carried a transmitter in the high-frequency range. The other satellites have used 108 Mc or higher, and the ionospheric effects at very high frequencies (VHF) are quite small. On the other hand, the Soviet satellites have continued to use 20 and 40 Mc for identification and some telemetry. These frequencies are much better suited for at least the initial ionospheric studies; thus, most of the ionospheric satellite measurements to date have been made with the Sputniks.

Of the three first Sputniks, the third (1958 δ_2) has provided the most information, principally due to the fact that its transmitter operated from solar power with storage provision when in the earth's shadow. The storage capability lasted for over a year, after which transmissions continued when the satellite was sunlit until very near the ultimate demise of the satellite in April 1960, a total of nearly two years operation. Because of rotation of the perigee position in the orbital plane, the height of the passages at any latitude varied slowly with date. This provided the opportunity for several different kinds of observations during the lifetime of Sputnik III. Figure 14 shows the variation in height of this satellite as a function of time for a latitude of 37°N.

In region A to C the NW to SE passages were considerably above the major portion of the F-layer ionization and measurements of the total ionospheric electron content were made. In this same period, a number of antipodal signals were detected at times midway between NW to SE passages. In the region B to C the height of the SW to NE passages was increasing up through the F layer. From the change of electron content with height, the ionospheric electron density profile was estimated. Similar opportunities were available in the period from C to D. The perigee rotation also permitted studies of the height of the region responsible for scintillation or the fast amplitude fading of the satellite signals. As the local time of the passages advanced an average of about 15 min/day (owing to precession of the orbit and the rotation of the earth about the sun), the diurnal variation of all these quantities could be studied. In the final months of the lifetime of Sputnik III, the transmissions were frequently CW, and Doppler studies at 20 Mc (and the second harmonic at 40 Mc) were greatly facilitated.

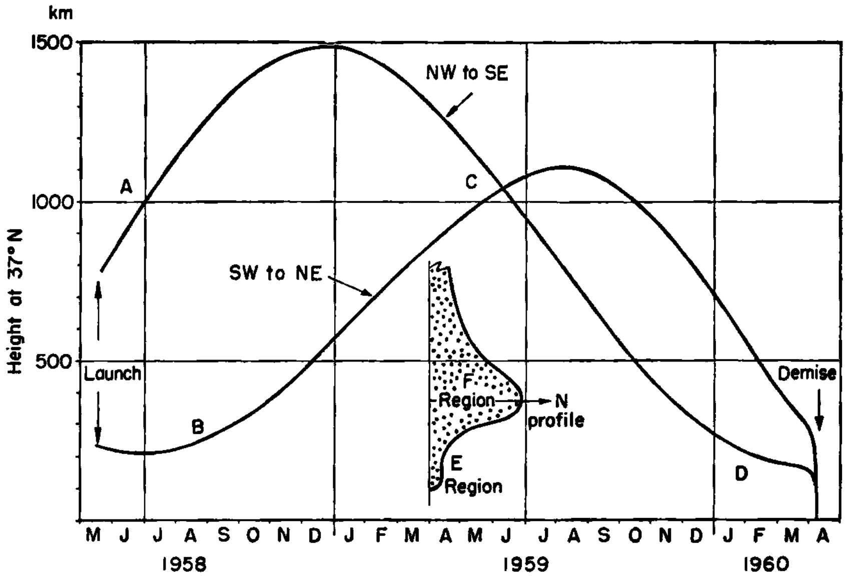


FIG. 14. Height variation of 1958 δ_2 (Sputnik III) at a latitude of 37°N. Several different types of measurements were appropriate in separate regions of the curves. An approximate daytime electron density profile is shown in the center of the figure.

3.1. Electron Content

3.1.1. Faraday Rotation. Measurements of the total ionospheric electron content and its diurnal variation have been made at Stanford University [28]. The results shown in Fig. 15 were obtained both by estimating the total Faraday rotation angle and by relating the rotation rate to the integrated electron density, as outlined in Section 2.2. It is of interest to compare these values of total content, defined as N_S in Fig. 15, with the content of electrons in the lower ionosphere below the height of maximum density, shown as N_b. The ratio N_S/N_b is about 4:1, as found by Evans [29] from moon echo data, but in addition a considerable diurnal variation is observed here. The ratio tends toward 3:1 in the daytime and 5:1 at night in the Stanford data.

Other measurements of electron content have been made by Blackband [30] and his co-workers from locations in England and Malta. They estimated the total rotation angle Ω in two ways; the first, as discussed in Section 2.2, depends upon locating the time at which propagation becomes transverse to the magnetic field direction. The second is usable whenever a portion of the satellite orbit is observed to fall beneath the ionosphere. The value of Ω is zero at this time, and the fading record permits the determina-

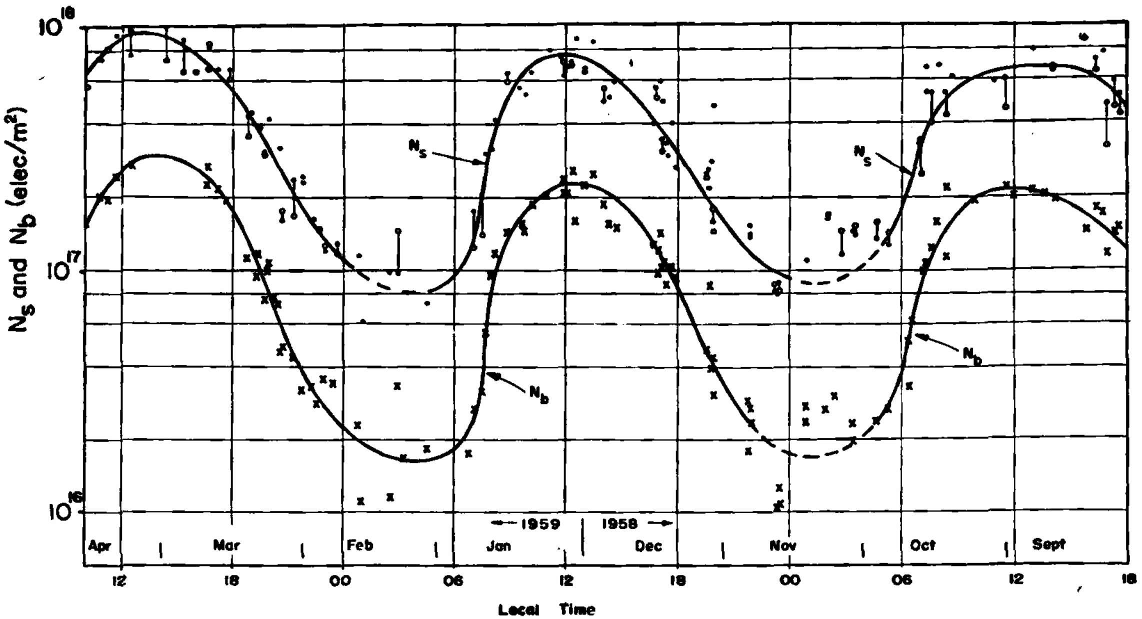


FIG. 15. Ionospheric electron content N_S as a function of local time. The content of the lower ionosphere below the height of maximum density is shown as N_b . The circles refer to calculations based upon estimates of the total rotation angle, and dots correspond to calculations based upon the rate of the Faraday rotation. Dots and circles which are connected were obtained from the same satellite passage.

tion of Ω and therefore the integrated electron density at other times. The calculated values of electron content are in reasonable agreement with extrapolations based on a true height analysis of simultaneous vertical incidence ionograms. Rather larger horizontal ionospheric gradients were found, and a preliminary model ionosphere was suggested in which the electron density had dropped to 45 % of its peak value at a distance of 250 km above the height of maximum density.

3.1.2. Doppler Effect. Measurements of electron content have also been made at Pennsylvania State University by W. J. Ross from a comparison of the 20-Mc and 40-Mc Doppler shifts of Sputnik III [46]. Since the satellite was above the *F* layer at the time of his observations, an equation similar to (2.26) was used, but with additional terms to include the effects of a spherical earth and ray refraction and to compensate for error in the high frequency approximation to the equation for the index of refraction. The results obtained in a sixteen month period are shown in Fig. 16.

The values shown in this figure were obtained between 1200 and 1600 local time from both south-bound and north-bound passages. When these values are compared with those shown in Fig. 15 for the same local time

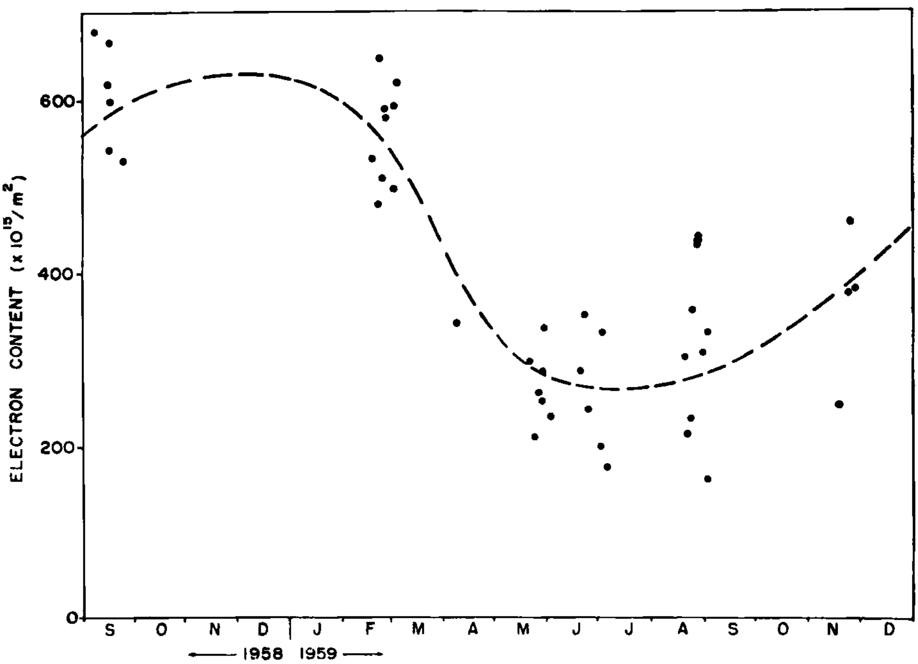


Fig. 16. Afternoon values of the electron content up to the height of Sputnik III, as observed at Pennsylvania State University (after Ross [46].

and season, it is found that the data of Ross are appreciably lower. It seems likely that this is due to the higher geomagnetic latitude of Pennsylvania State University (about 10° greater than Stanford). In addition, Fig. 16 shows a pronounced annual variation in electron content, although part of the apparent trend may be due to the decreasing solar activity.

Such an annual cycle is certainly not unexpected, since at mid-latitudes the value of the F-layer peak electron density is typically two or three times higher in the winter than in the summer. Continued observations of electron content at various latitudes should be of great assistance in understanding this major F-region anomaly.

3.2. *Scintillation*

"Spread-F" ionospheric conditions have long been associated with the scintillation of radio stars (see Briggs [31] for a more complete list of references). Spread-F echoes are obtained on vertical incidence ionograms, such as shown in Figs. 17(a) and 17(b), and are presumably caused by irregularities in the F-layer electron distribution. Although this phenomenon has been observed for many years, the reasons for its existence are still largely unknown. A recent attempt has been made by Martyn [32] to explain the presence of the inhomogeneities on the basis of a vertically drifting F layer. Several of the known characteristics of spread-F, such as the predominant nighttime occurrence, are then predicted.

Satellite studies have provided an important new tool in the investigation of scintillation. Some of the first information was obtained by O. B. Slee [33] using the 108-Mc transmissions of United States satellites 1958 Alpha and Gamma. His observations revealed scintillation when the latter satellite was as low as 200 km. The scintillation also correlated reasonably well with simultaneous radio star observations.

More recent work has been reported by Kent [34] and by Yeh and Swenson [35] from observations of the Sputniks. Figure 18 is a record obtained by Kent illustrating the "high-frequency fading" on the 40-Mc transmissions of Sputnik I. These rapid, irregular fluctuations in signal amplitude are defined here as scintillation. The autocorrelation of the fluctuations becomes small in approximately 0.1 sec for satellites in low orbits. From one nighttime autocorrelogram, the probable size of the ionospheric irregularities responsible for the scintillation has been estimated to be 0.5 km, but they are expected to be further elongated along the magnetic field. The irregularities appear to cover a vast area, perhaps several hundred kilometers in the N-S direction and as much as 1000-km E-W. The observations often indicate that the scintillation region has a sharp boundary. The occurrence of scintillation seems to be restricted to latitudes greater than 50° (48° magnetic) in England. The height of the irregularities is estimated to lie between 270 and 325 km. The data of Yeh and Swenson are largely con-

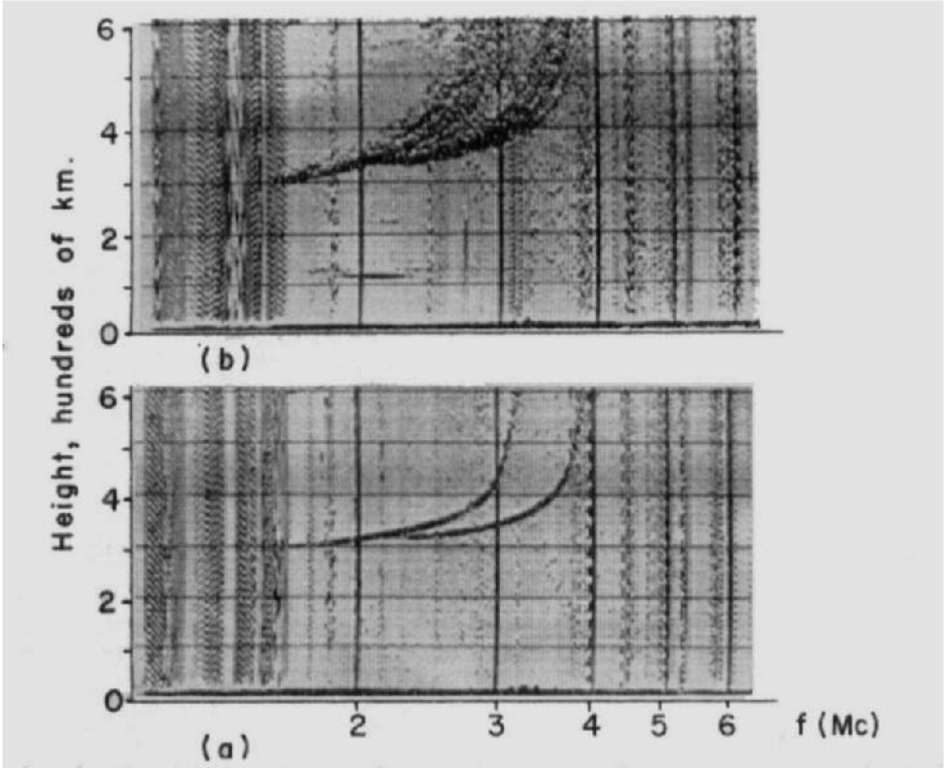


FIG. 17. Vertical incidence ionograms. Record (a) taken at Stanford on January 4, 1960, at 0001 local time, shows the well-defined traces of the two magneto-ionic modes and is indicative of a relatively smooth electron distribution. Record (b) obtained on January 7, 1960, at 0400 local time, shows the much more diffuse echoes, known as spread-*F*, which reveal an inhomogeneous structure in the *F*-layer.

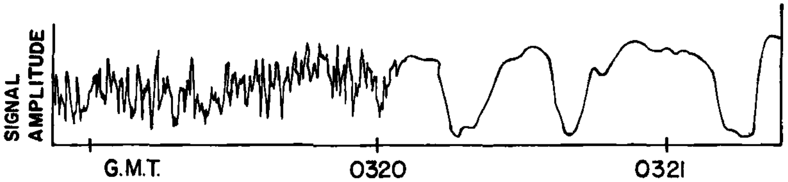


FIG. 18. Record of the 40-Mc signal received from Sputnik I, showing scintillation and its sudden cessation (after Kent [34]).

sistent with the foregoing, although heights between 220 and 300 are preferred for the irregularities. Also, the region of the irregularities extends as far south as 35°N (50° magnetic) in Illinois, indicating a rather strong longitudinal dependence for the southern boundary, which tends to confirm the conclusion of Kent that the region is controlled more by magnetic than by geographic latitude.

3.3. Large-Scale Irregularities in the Ionosphere

Munro has detected the presence of moving irregularities in the F region whose dimensions are on the order of hundreds of kilometers [36]. The same phenomenon has been studied more recently by Valverde using HF backscatter techniques [37].

The study of the polarization fading on satellite transmissions has revealed the presence of similar large-scale irregularities in the ionosphere and these may be of the same type as observed by Munro and Valverde. These irregularities were detected by Little and Lawrence by noting the deviation of the Faraday rotation rate from the smoothed rotation rate [38]. Their identification with the traveling disturbances is in doubt because it has not been possible to determine from the satellite observations whether or not the irregularities are in motion. It has been possible to conclude from the observations that the irregularities are on the order of 300 km in extent, that they have an appreciable vertical dimension in the F layer, and that the variations of integrated electron density may have a magnitude of about ± 1 %. Figure 19 shows the analysis of one passage in which deviations about the smoothed integrated density I_L are clearly demonstrated.

It should be understood that these large-scale irregularities do not correspond to the large geographical areas discussed in the previous section relating to the sources of radio star and satellite scintillation. The large-scale irregularities studied by Little and Lawrence are apparently quite homogeneous on a small scale, because no scintillation is observed at these times. In fact, when scintillation is present, the Faraday nulls are frequently so obscured that the precise measurements of Faraday rotation rate cannot be made.

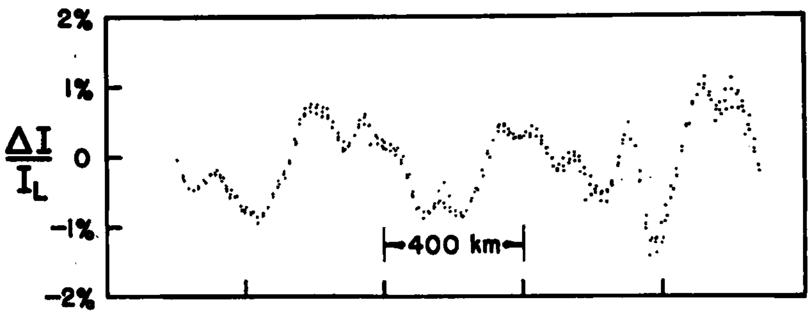


FIG. 19. Deviations of the integrated electron density about the smoothed electron density integral. Large-scale irregularities are in evidence (after Little and Lawrence [38]).

3.4. Propagation Effects

3.4.1. Antipodal Reception. A variety of interesting propagation effects has been observed with satellite transmissions. One of these is antipodal reception, first reported by H. W. Wells [39]. He reported the reception of the 40-Mc transmissions from Sputnik I (1957 α_2) at times when the satellite was in the vicinity of the antipodal point.

Numerous antipodal signals have been reported in work conducted at Stanford University, and these have served to confirm the phenomenon as well as shed further light on the propagation mode [40]. These results were obtained at 20 Mc with the signals from Sputnik III (1958 δ_2). It was found that the afternoon or evening hours at the receiving location were the most favorable for antipodal reception. The signals were received almost invariably from the south to southeast with a Doppler shift which was consistent with their direction of arrival. The signals therefore appear to have penetrated the F layer in a region prior to local sunrise in most cases and then traveled around the earth by internal ionospheric reflection. Horizontal gradients or "tilts" in the ionosphere are necessary to explain both the initial trapping of the energy and then again to provide sufficient refraction to bend the rays down to the earth's surface.

The propagation mode has also been investigated by Woyk [41]. She suggests that the ray paths follow reasonably close to the level at which the ray curvature just equals the curvature of the ionosphere. Ray directions on either side of this still propagate around the earth by the normal reflection process. A gross study of the world-wide variation of F-layer critical frequency reveals that the required tilts may be present at about the times found to be most favorable in the Stanford observations.

At least one antipodal report has been made at 108 Mc, although this frequency is so far above the maximum usable frequency that a much lower probability of detection should be expected. Trapping of the VHF energy at angles very close to the horizontal is to be expected, but it appears more difficult to reflect this energy down to the ground level.

3.4.2. Long Distance Propagation. Although the power radiated by earth satellites has been limited to a fraction of a watt, these signals have been detected without interruption for remarkably long periods of time. Continuous observation for the major portion of one orbital period appears to be fairly common at 20 or 40 Mc, and there are some reports of continuous reception for a number of hours [25]. Paetzold [42] has plotted the position of Sputniks I and III at the times when the transmissions were audible in Germany. For Sputnik I the region of audibility was quite large but appeared to be bounded by the geomagnetic equator as shown in Fig. 20. For Sputnik III the observations were made in the summer of 1958 and indicated a greatly reduced coverage as compared with those of the previ-

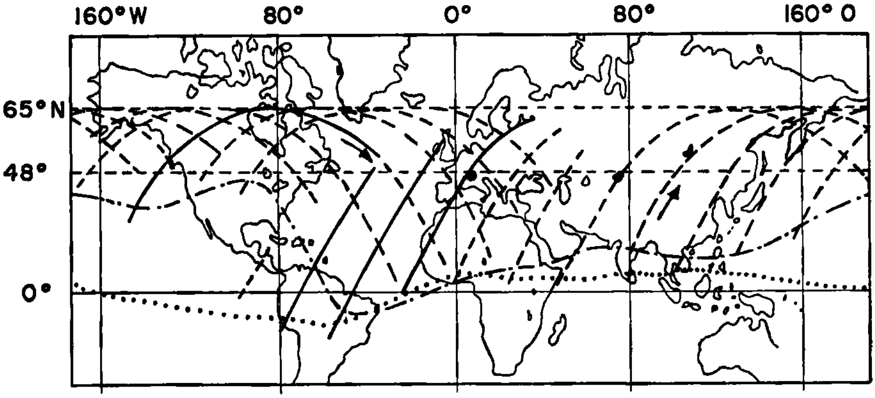


FIG. 20. Regions in which the signals from Sputnik I were detected on 20 Mc in West Germany on October 6 and 7, 1957. The solid and broken lines show the sub-satellite position at times when the signals were observed. The dotted line approximates the geomagnetic equator (after Paetzold [42]).

ous autumn. Paetzold attributes these effects to first, the "valley of electron density" which is known to exist above the geomagnetic equator, and second, the more disturbed ionospheric conditions during the summer observations. Although much of the time the satellite was far below the optical horizon, the observed azimuth of arrival corresponded very closely to the computed great circle path.

3.4.3. Skip Distance Phenomenon. Among those who listen to the radio transmissions from earth satellites, there must be very few who have not been surprised by the remarkably strong signals which are frequently received when the satellite is known to be near, or even far below, the horizon. These signals are usually observed to slowly increase in amplitude then suddenly disappear (or in the reverse sequence, if the satellite is receding), and they have been correctly explained as due to skip distance focusing [43]. The presence of several simultaneous ray paths, and the fact that the Doppler shift is, in general, different for each ray, cause the received signals to flutter rapidly. Some care must be taken to avoid confusing such records with those of scintillation due to ionospheric irregularities [35].

On many of the records of relative signal strength vs. time, there are not one, but two, apparent skip distances spaced in time by 2 to 20 sec [44]. Figure 21 is such a record which shows the slow build-ups and sudden decay which occur as the satellite passes inside the skip zone. This effect appears to be a consequence of path splitting between the ordinary and extraordinary wave components. As the extraordinary wave component has a somewhat greater "critical" frequency than the ordinary wave, it will also have a smaller skip distance. Estimates of the difference in the skip dis-

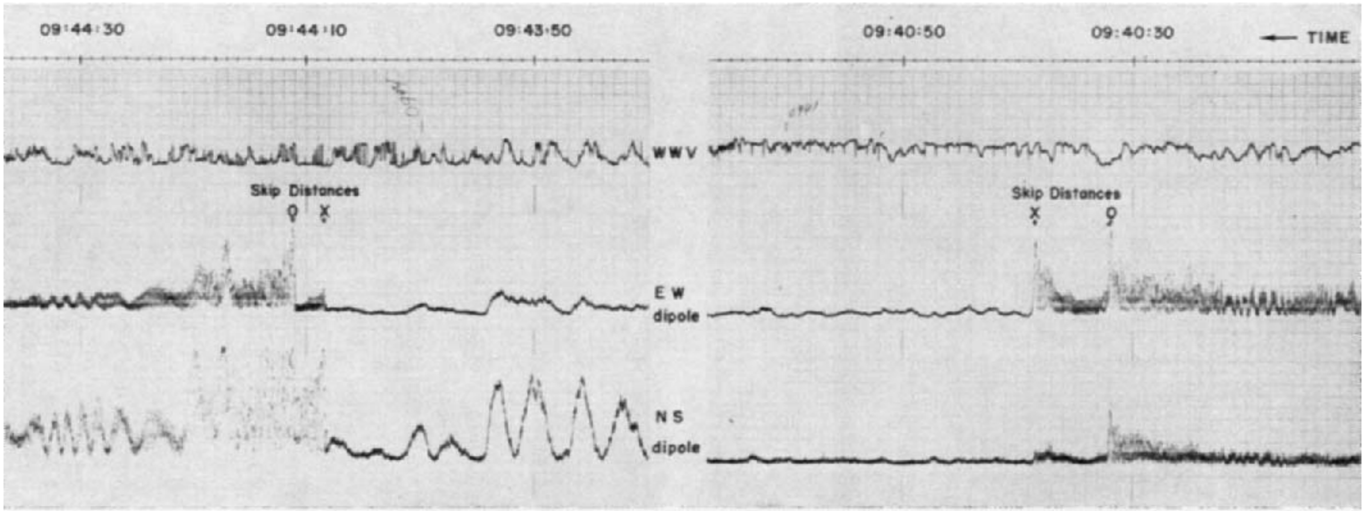


FIG. 21. The received signal amplitude from Sputnik III as the satellite entered and receded from the skip regions. As the satellite approached, the signal increased in amplitude, then suddenly disappeared at the ordinary wave skip distance. A few seconds later, a similar event occurred with the extraordinary wave. A reverse sequence is observed as the satellite departed the skip regions (after de Mendonça *et al.* [44]).

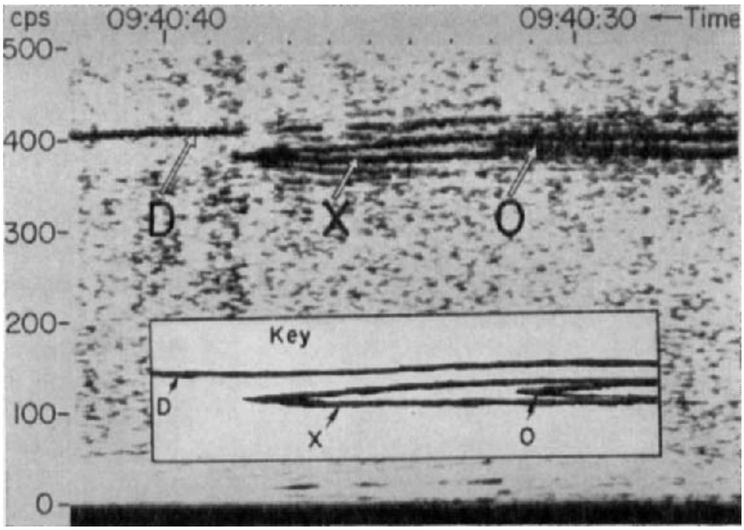


FIG. 22. Spectrogram of the signal received at the time the satellite passed into the skip regions as shown in Fig. 21. Early in the record five frequency components are received: one from the upper and one from the lower ionospheric ray for each of the two magneto-ionic modes, and the fifth component, identified as *D*, is a direct ray which does not involve ionospheric reflection. As the upper and lower rays for each of the two modes coalesce at their respective skip distances, the Doppler shifts are also observed to come together (after de Mendonça *et al.* [44]).

tances for the two waves, using transmission curves and vertical incidence ionograms (see N. Smith [45]), agree quite well with the 2- to 20-sec time difference found in the satellite observations.

Figure 22 is a spectrum analysis of the signal recorded on magnetic tape at the time the satellite passed into the skip region. As the satellite approached the ordinary wave skip distance, the Doppler shifts imparted to the upper and lower rays are observed to coalesce. A few seconds later, the same thing happens to the extraordinary mode.

3.5. Very Low Frequency Measurements in Explorer VI

Explorer VI was launched in August 1959 in a highly elliptical orbit about the earth. At apogee the satellite soared over six earth radii into space, returning about five hours later to a height of only 260 km. A VLF receiver tuned to 15.5 kc, the frequency of *NSS*, was included in the instrument package. A principal goal of the experiment was the determination of the VLF noise level in the outer regions of the ionosphere and exosphere. If possible, man-made signals were to be detected above the ionosphere, where propagation occurs in the whistler mode.

The preliminary results of the VLF experiment have been reported by R. A. Helliwell and his associates to the American Physical Society in November 1959. The *NSS* and atmospherics were received in the satellite during the launch phase and telemetered to the ground. These signals correlated precisely with the VLF signals received directly at the ground station. The received VLF signals in the satellite began to drop in strength at a height as low as 40 km and had completely disappeared at a height of 70 km. Neither *NSS* nor atmospherics was detected after that. The background noise level above the *F* region appears to be quite low and below the sensitivity of the satellite receiver. There are several most interesting theoretical questions raised by the unexpectedly low noise levels. For example, if the transmission loss through the ionosphere is sufficient to greatly attenuate the atmospheric noise, how is the relatively slight whistler attenuation to be explained?

4. Conclusion

The very first artificial earth satellites, launched withour prior experience in ionospheric studies using orbiting radio transmitters, have proved extraordinarily fruitful. Electron densities in the earth's atmosphere, out beyond the *F*-layer peak, have been broadly established, and the succeeding, more detailed, phase of investigation has already begun.

In the first experiments the Doppler and Faraday effects have proved to be basic. Each yields roughly the same information: if the velocity is split into two (nonorthogonal) components, one tangential to the stratification and one in the direction of the ray, the first produces frequency shifts and polarization rotations bearing on the electron content of a column reaching up from the receiver to the height of the transmitter, the second tells about the electron density at the transmitter.

There is a difference between the two effects. The electron density is weighted in proportion to the magnetic field in the case of Faraday effect; hence, in experiments where the magnetic field is not the same over the whole of the path, an extra piece of information bearing on the distribution of ionization over the path may be expected from measurements of both effects.

Refraction due to the ionosphere, a third measurable effect, proves to be intimately related to the Doppler effect. Thus in satellite experiments of the type so far conducted, where the velocity is essentially tangential to the stratification, it has not added worthwhile additional information. However, there is this difference between an observation of refraction and one of Doppler shift. Any departure of the velocity from the tangential direction contributes to the Doppler shift, without altering the refraction. Doppler data from rocket-borne transmitters with large vertical velocities may thus be supplemented by refraction measurements. But again the integrated electron density and the local density would be the two quantities deduced. Refraction of radio stars, whose broad-band emissions do not exhibit Doppler shift, yields geographic gradients of the integrated electron density.

In addition to the three basic effects, Doppler, Faraday, and refraction, there are a variety of phenomena connected with the structure of the ionosphere. Some of these are: scintillation due to inhomogeneities, propagation from the antipodal point, and interference effects as the skip-zone caustics of the two characteristic modes pass through the observer.

Future development of radio beacon experiments should be planned to exploit the ability of measuring integrated electron density. Suitably high frequencies should be selected to minimize path splitting and to facilitate the interpretation of results. For example, harmonically related 40 and 120 Mc frequencies would be suitable for measuring the Doppler shift offset due to the differential effect of the ionosphere on the two frequencies. A knowledge of the diurnal variation in electron content of the ionosphere and its dependence on latitude would contribute profoundly to the understanding of its production and behavior. In addition, peculiar phenonena at the geomagnetic equator where the horizontal magnetic field impedes the rise and fall of ionization would become accessible to study by suitably placed stations. The behavior of the F region during magnetic storms has been one of the more difficult ionospheric problems to attack theoretically (Maeda [47]). Measurements of electron content as a function of time throughout a magnetic storm at various latitudes would be of great help in understanding these events.

All these things have been studied before. The important extra feature contributed by satellite-borne transmitters is the possibility of including the ionosphere beyond the F peak in the studies.

The future development of satellites with 24-hr periods offers the possibility of carrying out the same class of studies with special advantages. This is a case where the Doppler, Faraday, and refraction effects all furnish different information, which can be observed from stations over a wide range of latitude for extended periods.

Satellites are not now the sole means of observing the outer atmosphere. There are also rocket-borne transmitters. A great advantage of firing a rocket through the ionosphere is that a detailed profile of ionization density results—one is not limited to the total electron content. Furthermore, rockets can be fired in special locations and at special times. On the other hand, a rocket experiment, being brief and localized, does not furnish data for long periods to observers all over the world as does a satellite. The role of rockets would seem to be to settle special questions left unanswered by satellites.

A further powerful technique, that of "incoherent backscatter," now promises to furnish detailed continuous ionization profiles at special localities. In this scheme proposed by Gordon [48][7], a powerful vertical beam of radio waves gives rise to scattered radiation, in much the same way as a searchlight beam. The amount of scattering from each level is proportional to the electron density, and the height of the level is determined by the usual technique of pulse radar. The profile will be available instantaneously and continuously at those locations where in due course the powerful radars can be installed. The incoherent backscatter produced by monochromatic irradiation shows a broadened spectrum from which many parameters can be extracted. The intensity gives the electron density, the Doppler displacement gives vertical drift velocities, and the spectral broadening gives temperature, collision broadening being, in general, small. Fine structure associated with the magnetic field is also to be expected. In short, incoherent backscatter opens up rich techniques resembling those of spectroscopy that have not hitherto been available in ionospheric physics.

The conventional technique of ionospheric sounding by pulse radar differs from the incoherent backscatter method in that reflection takes place from one level. The variation of this level with frequency and the critical frequencies of layer penetration displayed on the conventional ionogram have provided the basis for past ionospheric investigation. The possibility now arises of having a "topside sounder." A satellite-borne pulse transmitter would sweep from about 1 Mc to about 20 Mc while a receiver also in the satellite swept in unison. An antenna pointing downward would receive the echoes and display them on an ionogram showing echo delay time versus frequency. Such an ionogram would reveal the local plasma frequency, the F-layer critical frequency, and the electron density profile between the F peak and the satellite. Direct backscatter would be returned from all azimuths. Propagation paths involving multiple reflections or scattering from the ground and ionosphere, including possible round-the-world paths, would appear with longer delays, and a characteristic pattern would be formed by ground transmitters and cosmic noise. Since many unexpected phenomena may appear on the topside ionograms from all parts of the world, it is too early to assess their full significance for studies of the ionization in space. If the first trials are successful we may look forward to the additional refinement of the rotating oblique beam which has proved so fruitful in revealing field-aligned inhomogeneities in the ionosphere.

A closely related experiment can be carried out with the swept-frequency receiver alone. In this case attention focuses on natural sources of noise, especially galactic and solar emission. There are, however, natural sources

[7] See Laaspere [49] for additional references.

in the earth's own atmosphere also to be expected. Noise emission caused by particle bombardment may occur at frequencies at or near the local plasma frequency as happens in the solar envelope, but the noise would be prevented by the *F* layer from reaching the earth's surface. At very low frequencies, where the whistler mode permits a little leakage of energy through the dense part of the ionosphere, sporadic emission occurring at heights of several earth radii does in fact reach us. So far, however, we have been shielded from observing emission at the frequencies appropriate to less extreme heights.

A great variety of new experiments on the ionization in space by radio means can be seen opening up with the advent of satellites. There is not only the earth's ionosphere to be considered, there are also the ionospheres of Venus and Mars and of the other planets. Much could be deduced from a two-frequency transmitter in orbit around Venus. First, from frequency measurement alone, on the high frequency, all the orbital elements of the probe are deducible. Doppler shift offset of the lower frequency would give integrated electron densities along different lines of sight and open the experimental investigation of planetary ionospheres. The ionosphere of Venus in particular would be a sensitive monitor of corpuscular solar emission. Faraday effect would contribute data on magnetic fields, perhaps those around Venus, perhaps those in interplanetary space.

Satellites in circular or eccentric orbits about the sun may also be expected in due course to tell us about the solar envelope and interplanetary plasma.

The achievements so far are impressive and we look forward expectantly to the future contributions by radio means to our knowledge of the ionization in space.

List of Symbols

f_O received frequency
f_S frequency of radio source
f_C critical frequency of the *F*-layer
λ wavelength in the medium
λ_0 free space wavelength, c/f
v velocity of source
v_r velocity of recession, along ray path
c velocity of light
ϕ angle between the satellite orbit and the ray path
β angular position of the satellite measured at the center of the earth (Fig. 6). It has also been used as one of the polar coordinates describing the ray path (Section 2.4 and Fig. 12)
δ angle between the two vertical planes of the orbit and the ray path
θ angular distance between the subproximal point and subsatellite point
ρ geocentric distance
h height above the earth's surface

R straight-line distance from the observer to the satellite
G subsatellite point
ds element of path length
μ index of refraction
$\Delta\mu$ difference in the values of the refractive index for the ordinary and extraordinary modes
P phase path, meters
ΔP phase path defect; the reduction of phase path due to the dispersion of the medium, defined as $(R - P)$
N electron density, electron/meter^3
e charge of the electron
m mass of the electron
ϵ_0 permittivity of free space
μ_0 permeability of free space
ω angular wave frequency, $2\pi f$
ω_L longitudinal gyromagnetic frequency of an electron
H_L longitudinal component of earth's magnetic field
Ω number of full rotations of the plane of polarization
γ angle between the direction of propagation and the earth's magnetic field
χ angle between the ray direction and the vertical
i angle between the ray direction and the normal to the planes of stratification
ϵ angle substended by SS' at the observer (Section 2.4)
$\mathcal{R}$ ratio of the Doppler shift offset to the Doppler shift
$\mathfrak{N}$ $40.3/f^2 \int_0^{h_S} N\,dh$ = phase path difference at vertical incidence
Σ limit-of-escape surface

References

1. Brito, J. M. (1959). A correction necessary for the application of the Doppler effect to the measurements of distances to satellites. *Proc. IRE* **47**, No. 11, 2023.
2. Hibberd, F. H., and Thomas, J. A. (1959). The determination of the electron distribution in the upper ionosphere from satellite Doppler observations. *J. Atm. and Terr. Phys.* **17**, 71–81.
3. Weekes, K. (1958). On the interpretation of the Doppler effect from senders in an artificial satellite. *J. Atm. and Terr. Phys.* **12**, 335–338.
4. Hibberd, F. H. (1958). The effect of the ionosphere on the Doppler shift of radio signals from an artificial satellite. *J. Atm. and Terr. Phys.* **12**, 338–339.
5. Ratcliffe, J. A. (1959). The Magneto-ionic Theory and its Applications to the Ionosphere. Cambridge Univ. Press, London and New York.
6. Bowhill, S. A. (1958). The Faraday rotation rate of a satellite radio signal. *J. Atm. and Terr. Phys.* **13**, 175.
7. Aitchison, G. J., and Weekes, K. (1959). Some deductions of ionospheric information from the observations of emissions from satellite 1957 Alpha 2-I. *J. Atm. and Terr. Phys.* **14**, 236–243.
8. Garriott, O. K. (1960). The determination of ionospheric electron content and distribution from satellite observations. Part I. Theory of the analysis. *J. Geophys. Res.* **65**, 1139–1150.
9. Bracewell, R. N., and Garriott, O. K. (1958). Rotation of artificial earth satellites. *Nature* **182**, 760–762.

10. Bracewell, R. N. (1959). Satellite rotation. *Advances in Astronautical Sci.* **4**, 317–328.
11. Garriott, O. K., and Little, C. G. (1960). The use of geostationary satellites for the study of ionospheric electron content and ionospheric radio-wave propagation. *J. Geophys. Res.* **65**, 2025–2027.
12. Daniels, F. B. (1956). *In* "Scientific Uses of Earth Satellites" (J. A. Van Allen, ed.), pp. 276–287. Univ. of Michigan Press, Ann Arbor, Michigan.
13. Smith, F. G. (1952). The determination of the position of a radio star. *Monthly No. Roy. Ast. Soc., Geophys. Suppl.* **112**, 497–513.
14. Smith, F. G. (1952). Ionospheric refraction of 81.5 Mc/s radio waves from radio stars. *J. Atm. and Terr. Phys.* **2**, 350–355.
15. Komesaroff, M. M., and Shain, C. A. (1959). Refraction of extra-terrestrial radio waves in the ionosphere. *Nature* **183**, 1584–1585.
16. Bailey, D. K. (1948). On a new method for exploring the upper ionosphere. *Terr. Magn.* **53**, 41–50.
17. Woyk, E. (E. Chvojková). (1959). The refraction of radio waves by a spherical ionized layer. *J. Atm. and Terr. Phys.* **16**, 124–135.
18. Thomas, J. O. (1959). The distribution of electrons in the ionosphere. *Proc. IRE* **47**, 162–175.
19. Seddon, J. C. (1953). Propagation measurements in the ionosphere with the aid of rockets. *J. Geophys. Res.* **58**, 323–335.
20. Jackson, J. E., and Seddon, J. C. (1958). Ionosphere electron-density measurements with the Navy Aerobee-Hi rocket. *J. Geophys. Res.* **63**, 197–208.
21. Pfister, W., and Ulwick, J. C. (1958). The analysis of rocket experiments in terms of electron-density distributions. *J. Geophys. Res.* **63**, 315–333.
22. Gringauz, K. I. (1958). Rocket measurements of the electron concentration in the ionosphere using an ultra-shortwave dispersion interferometer. *Doklady Akad. Nauk SSSR* **120**, 1234–1237.
23. Nisbet, J. S. (1959). Electron densities in the upper ionosphere from rocket measurements. Sci. Rept. No. 126, Ionosphere Research Laboratory, Pennsylvania State University, University Park, Pennsylvania.
24. Bourdeau, R. E., Jackson, J. E., Kane, J. A., and Serbu, G. P. (1960). Ionospheric measurements using environmental sampling techniques. *In* "Space Research" (H. K. Kallmann Bijl, ed.), pp. 328–339. North-Holland Publishing Company, Amsterdam.
25. Hoffman, W. C. (1959). A report on the symposium on the use of space vehicles at the fall 1958 URSI meeting. *J. Planetary and Space Sci.* **1**, 238–248.
26. Tyas, J. P. I., Franklin, C. A., and Molozzi, A. R. (1959). Measurements of cosmic noise at low frequencies above the ionosphere. *Nature* **184**, 785–786.
27. Storey, L. R. O. (1959). A method for measuring local electron density from an artificial satellite. *J. Res. Natl. Bur. Stand.* [D] **630**, 325–340.
28. Garriott, O. K. (1960). The determination of ionospheric electron content and distribution from satellite observations. Part 2. Results of the analysis. *J. Geophys. Res.* **65**, 1151–1157.
29. Evans, J. V. (1957). The electron content of the ionosphere. *J. Atm. and Terr. Phys.* **11**, 259–271.
30. Blackband, W. T. (1960). The determination of ionospheric electron content by observation of Faraday fading. *In* "Space Research" (H. K. Kallmann Bijl, ed.), pp. 387–396. North-Holland Publishing Company, Amsterdam.
31. Briggs, B. H. (1958). A study of the ionospheric irregularities which cause spread-F echoes and scintillations of radio stars. *J. Atm. and Terr. Phys.* **12**, 34–45.

32. Martyn, D. F. (1959). The normal F region of the ionosphere. *Proc. IRE* **47**, 147–155.
33. Slee, O. B. (1958). Radio scintillations of satellite 1958 Alpha. *Nature* **181**, 1610–1612.
34. Kent, G. S. (1959). High frequency fading observed on the 40 Mc/s wave radiated from artificial satellite 1957 Alpha. *J. Atm. and Terr. Phys.* **16**, 10–20.
35. Yeh, K. C., and Swenson, G. W., Jr. (1959). The scintillation of radio signals from satellites. *J. Geophys. Res.* **64**, 2281–2286.
36. Munro, G. H. (1950). Traveling disturbances in the ionosphere. *Proc. Roy. Soc. (London)* **A202**, 208.
37. Valverde, J. F. (1958). Motions of large scale traveling disturbances determined from high-frequency backscatter and vertical incidence records. Ph.D. dissertation, Stanford University, Stanford, California.
38. Little, C. G., and Lawrence, R. S. (1960). The use of polarization fading of satellite signals to study the electron content and irregularities in the ionosphere. *In* "Space Research" (H. K. Kallmann Bijl, ed.), pp. 340–365. North-Holland Publishing Company, Amsterdam.
39. Wells, H. W. (1958). Unusual propagation at 40 Mc from the USSR satellite. *Proc. IRE* **46**, 610.
40. Garriott, O. K., and Villard, O. G., Jr. (1958). Antipodal reception of Sputnik III. *Proc. IRE* **46**, 1950.
41. Woyk, E. (E. Chvojková). (1959). The antipodal reception of Sputnik III. *Proc. IRE* **47**, 1144.
42. Paetzold, H. K. (1959). Observations of the Russian satellites and the structure of the outer terrestrial atmosphere. *J. Planetary and Space Sci.* **1**, 115–124.
43. Dewan, E. M. (1959). Unusual propagation of satellite signals. *Proc. IRE* **47**, 2020.
44. de Mendonça, F., Villard, O. G., Jr., and Garriott, O. K. (1960). Some characteristics of the signal received from 1958 δ_2. *Proc. IRE.* **48**, 2028–30.
45. Smith, N. (1939). The relation of radio sky-wave transmission to ionospheric measurements. *Proc. IRE* **27**, 332–347.
46. Ross, W. J. (1960). The determination of ionospheric electron content from satellite Doppler measurements. Part I—Method of analysis. Part II—Experimental results. *J. Geophys. Res.* **65**, 2601–2615.
47. Maeda, K. (1955). Theoretical study on the geomagnetic distortion in the *F2* layer. *Rept. Ionosphere Res., Japan* **9**, 71–85.
48. Gordon, W. E. (1958). Incoherent scattering of radio waves by free electrons with applications to space exploration by radar. *Proc. IRE* **46**, 1824–1829.
49. Laaspere, T. (1960). On the effect of a magnetic field on the spectrum of incoherent scattering. *J. Geophys. Res.* **65**, 3955–3959.

This Page Intentionally Left Blank

USE OF CONSTANT LEVEL BALLOONS IN METEOROLOGY

J. K. Angell

U. S. Weather Bureau, Washington, D. C.

1. Introduction

At the present time meteorologists base their analyses and forecasts upon data of the Eulerian type; that is, data which give the properties of different air parcels at the same (synoptic) time. An alternate possibility is to base the analyses and forecasts upon data of the Lagrangian type; that is, data which give the properties of the same parcels of air at different times. It is surprising and distressing to find that Lagrangian methods, which have occupied an important place in many hydrodynamic researches, have fallen to such a low estate in meteorology. Particularly since the advent of the frontal theory, practicing meteorologists have based their forecasts almost wholly upon change in flow pattern rather than upon the change in position of air parcels constituting the flow pattern. In recent years, however, the combination of constant level balloon (CLB) flights and nuclear weapons tests has renewed interest in the Lagrangian concept. The problem now is to incorporate Lagrangian methods into the routine of operational meteorology. This is difficult of achievement because Lagrangian concepts and methods are so unfamiliar to the practicing meteorologist of today.

This article has the following purposes:

1. To indicate how CLB flights yield data of a quasi-Lagrangian type.
2. To indicate the extent to which quasi-Lagrangian data are now being obtained from such flights.
3. To indicate the advantages that data of quasi-Lagrangian type offer in operational and research meteorology.

It will be shown that operational CLB flights are desirable because they provide meteorological data above the vast oceans, where otherwise little information would be available, and because they are capable of providing more representative winds than do conventional vertical-sounding techniques. It will be shown that CLB flights for research purposes are desirable because the acceleration of the wind (and consequently ageostrophic parameters) can be estimated from the trajectory data while a whole gamut of Lagrangian parameters can be derived which are of importance in turbulence and diffusion studies.

2. History of the Constant Level Balloon Concept

2.1. Introduction

Brief historical résumés dealing with the development and early meteorological uses of the constant level balloon (CLB) have been given by Spilhaus [1] and Haig and Lally [2]. In this section this information is collated and brought up to date.

2.2. Development Prior to World War II

The use of manned balloons for meteorological purposes extends far back in history. However, the International Air Races held during the early part of the twentieth century emphasized the importance of meteorology to the balloonist and conversely the useful information on air trajectories which the balloonist could furnish the meteorologist. After taking note of the meteorological information obtained from some of these races, Meisinger undertook a series of manned CLB flights under the auspices of the U.S. Weather Bureau and Army Air Service. During these flights Meisinger approximated air trajectories at levels of 5000–10,000 ft and compared these trajectories with the isobars drawn on surface weather maps [3]. In March 1924, on the last flight of a series of ten, Meisinger met death when his balloon was destroyed by a thunderstorm. His death, plus the rapid development of power-driven aircraft, terminated for the time being the use of manned balloons for meteorological purposes in the United States although some flights continued to be made in Europe. Recently, manned balloon flights for purposes of atmospheric exploration have again become fashionable in the United States, but not specifically for the delineation of air trajectories.

In the same way that the International Air Races stimulated the use of manned balloons for meteorological purposes, the "toy" balloon races held in England [4] suggested the utilization of unmanned CLB flights for meteorological investigations, as witness Richardson's study of diffusion by means of such "toy" balloons [5]. The first recorded CLB flights in the United States were those made by Piccard and Akerman at the University of Minnesota in 1936 [6]. They used cellophane balloons of about 70 meter^3 volume developed by the University of Minnesota and the Franklin Institute. In June 1936, one of these balloons traveled 600 miles in 10 hr at approximately constant level. The difficulty with this procedure was that the cellophane had to be kept warm, moist, and flexible in order to avoid balloon failure, and thus nighttime flights and flights during winter with this equipment were impossible. During winter, Piccard and Akerman used a train of Dewey-Almy rubber ballons which floated at constant level for only a few hours.

2.3. Development During World War II

During World War II, the CLB concept was forcefully brought to the fore by the Japanese Fugo balloon-bomb flights [7–9]. The original Japanese effort along this line in 1933 involved the construction of paper balloons which could carry bombs about 60 mi subsequent to launching from a submarine. (Perhaps this is what flew over Los Angeles shortly after the Pearl Harbor attack.) However, as the war in the Pacific developed, it became evident that the only way in which the Japanese could directly attack the United States mainland was by long-range balloon. They therefore constructed two types of nonextensible balloons, one made out of paper, coated to avoid diffusion, and the other made out of oiled silk. These balloons were about 30 ft in diameter, had automatic valving devices, and carried up to 900 lb of sand ballast distributed in 36 bags hung on a "chandelier" suspended beneath the balloon. Ballast was dropped according to the indications of an aneroid barometer. Between November 1944 and April 1945, some 9300 of these balloons were released from Japan for flight in the westerlies at levels from 20,000 to 40,000 ft. About 7% of these balloons, with their cargo of incendiary and anti-personnel bombs, were later discovered in the western United States. They had remained aloft, it is estimated, from 4 to 10 days. Because of poor armament these balloons were not effective as weapons, and since they were not tracked and positioned the flights were of little meteorological value. Nevertheless, these flights indicated the capability of the CLB and aroused interest in the United States concerning a possible meteorological role for the CLB.

In England during the war years noteworthy CLB work was done by Durst and Gilbert in 1943 and 1944 [10]. They made use of quasi-constant-level rubber balloons which were tracked and positioned by the excellent British wartime radar network. While these balloons did not stay aloft very long and did not hold level very well (variations in pressure height of 25 mb in one-half hour are reported), the results obtained were most interesting, particularly as regards the ageostrophic wind estimates derived from the balloon accelerations.

2.4. Development Subsequent to World War II

2.4.1. Project "Moby Dick." In the United States, very little CLB work was done during World War II. After the end of the war, however, New

York University entered into a contract with Watson Laboratories of the Air Material Command for the development of a CLB system [11–14]. Later, some of the New York University personnel transferred to the General Mills Aeronautical Laboratory at Minneapolis in order to carry on further work of this type [15, 16]. At about this same time the Air Force, through Project "Moby Dick," started to conduct their own CLB flights at levels from 40,000 to 80,000 ft under the cognizance of the Geophysical Research Directorate, Air Force Cambridge Research Center [17–19]. However, Project "Moby Dick" suffered greatly from the fact that few of the flights were accurately tracked and positioned, with the result that meteorological interpretations of the results have been few. At the University of California at Los Angeles the Moby Dick data were analyzed by plotting CLB latitude and longitude as a function of time followed by the "fairing-in" of a smoothed line of best fit. The velocity components were then computed at 2-hr intervals from these smoothed lines. This highly subjective method yielded data suitable for statistical analysis and resulted in two as yet unpublished manuscripts by Hovind dealing with the seasonal variation of high-level winds over the United States [20] and the diurnal and semidiurnal oscillations in the wind field between 50,000 and 75,000 ft [21]. To the writer's knowledge, most other applications of Moby Dick data involve comparisons of Moby Dick trajectories with theoretical trajectories deduced in various ways, an application not requiring great positioning accuracy. Such comparisons are discussed in Section 6.

2.4.2. Project Transosonde. In 1949, the U.S. Naval authorities became interested in the development of meteorological sounding systems which could gather data in inaccessible regions of the world. One concept of a transosonde (transocean sounding) system was placed before the American Meteorological Society in 1950 by Captain H. T. Orville, in his presidential address [22]. The responsibility for the development of the transosonde system has rested with the Mechanics, Atmosphere and Astrophysics, and Radio II Divisions of the Naval Research Laboratory (NRL) in Washington, D.C. [23–25].

Most of the transosonde flights have been made at 30,000–35,000 ft, the level of the so-called "jet stream," and the lowest levels at which the balloons could be flown under Civil Aeronautics Administration (CAA) regulations. The transosonde flights have generally been tracked by the Federal Communications Commission (FCC) radio-direction-finding (RDF) network, probably the best RDF network in the world. Positioning by this network is of sufficient quality, particularly in the vicinity of North America, to permit a good estimation of the wind velocity and a fair estimation of the wind acceleration (and ageostrophic velocity) from the successive transosonde positions. Since, to date, the transosonde system has yielded the most numerous and accurate meteorological data of a quasi-Lagrangian nature, and since the writer has been concerned with the data obtained from this system, the bulk of this article deals with the transosonde system and the meteorological data which have been derived therefrom.

3. Constant Level Balloon Instrumentation

3.1. Introduction

While the basic instrumental components of all ballasted CLB systems are the same, there are minor differences from one system to another. Space considerations make it impractical to discuss the different instrumentation used by the U.S. Navy, U.S. Air Force, New York University, and General Mills, and therefore emphasis is placed here on the U.S. Navy's transosonde instrumentation. Most of the following remarks refer to the instrumentation used in the transosonde system between 1952 and 1959. However, in Section 3.6 we briefly consider the "new" transosonde vehicle consisting of a superpressured Mylar balloon system which weighs only 12 lb. Most of the information presented in this section is culled from the various NRL reports cited in the list of references.

3.2. The Transosonde Balloon

The teardrop-shaped, constant-volume transosonde balloon was constructed by heat sealing 14 gore sections of a 2- or 2.5-mil lamination of two films of polyethylene. Polyethylene is a suitable balloon material because of its low brittle temperature (about $-80°C$), its chemical stability (apparently unaffected by ultraviolet light and ozone), its low permeability (about 7 liters of helium per mil per square meter per 24 hr), its high tensile strength (about 1900 lb/in.^{2}), its high tear resistance, and its small radiation absorption. The balloon was designed so as to minimize the circumferential stresses, thus reducing the probability of flight failure. A balloon with this design is called a "natural shape" balloon. Most of the transosonde balloons had a maximum diameter of 39 ft and a volume of about 27,000 ft^3. A balloon with these dimensions can lift about 650 lb to a height of 30,000 ft. Meridional glass filament tapes placed along the seams and center of each gore provided load-bearing members for carrying this weight.

3.3. Transosonde Equipment for Maintaining Constant Level Flight

The natural ceiling of a constant volume balloon is determined by the balance between the buoyancy of the cell (volume of balloon times the difference in density of air and helium) and the total weight supported. If the balloon ascends above this ceiling, gas is expelled through an open appendix at the base of the balloon, thus checking further ascent. After a constant volume or constant level balloon has been aloft a short time its floating equilibrium is usually disturbed due to vertical air currents, changes in temperature of the ambient air, changes in temperature of the gas within the balloon due to radiation fluxes, diffusion and solution migration through the skin of the balloon and mixing of the inflating gas with air through the open appendix. The radiation effects are particularly large at sunrise and sunset and would cause large changes in the floating level of the balloon if not compensated for by the release of gas or the dropping of ballast. This sunrise-sunset effect largely determines the practical duration of ballasted CLB flights although on some flights the turbulence effect also may be important. On the initial transosonde flights a liquid cleaning fluid was used as ballast with the ballast release being activated by an electric valve according to the indications of an aneroid barometer. On later flights small steel shot was used as ballast with the ballast release being activated by a magnetic valve. The steel shot proved superior as ballast since it had a constant flow rate, its volume requirements were less, and the magnetic valve action was more positive and used less current than the electric valve action. The ballast system was arranged so that there would be no ballast drop at ambient air pressures of 300 mb or less but so that at pressures greater than 300 mb there would be a linear increase in rate of ballast drop to a maximum value of 2.5 lb/min at ambient pressures of 325 mb or greater. Most of the transosonde balloons carried about 300 lb of ballast.

It was noted in the early transosonde flights from Japan that the floating level of the balloons tended to increase with time so that the average floating level was about 2300 ft above prescribed floating level two days after release and about 3100 ft above prescribed floating level five days after release. Obviously, some original expectations were not being realized; namely, that the increase in natural ceiling due to load decrease from helium loss and ballast drop would be balanced by air intake through the appendix as the balloon oscillated in the vertical. Consequently, it was found necessary to install an air pump on the transosonde to force air into the balloon. This air pump consisted of a 4-in.-diam fan driven by a 2½-watt motor. The fan was mounted in a Fiberglas resin duct mounted in a side duct of polyethylene which led into the balloon three-fifths of the way up its side. This fan produced a flow of air of 10–20 ft^3/min, and by adding air to the system in an amount approximately equivalent to the mass of the expended ballast, in most cases successfully maintained transosonde flight within a 2000-ft layer bracketing the prescribed floating level.

3.4. Other Airborne Transosonde Equipment

The transmitter output on the transosonde flights varied from 12 to 50 watts depending upon whether the flights were limited to the United States or were to be tracked across the Pacific and Atlantic Oceans. The power was supplied by a series-parallel arrangement of 58 BA-38 dry batteries for the 300- and 700-volt plate supply and a 66-amp-hr lead-acid storage battery for the 6-volt filament voltage. In order to obtain sufficient energy from the batteries at the low ambient temperatures met with at flight level (−40 to −50°C), thermal insulation was provided by four layers of aluminum foil, each separated by a 1-in. air space. In addition, the batteries and transmitters were placed in a kapok-padded bag and covered with a transparent polyethylene cover to provide a greenhouse effect. Transmitter frequencies utilized varied from 6420 to 19,282 kc in an attempt to find the most favorable frequencies for the RDF networks. On most of the trans-Pacific flights the transmitters were activated once every 2 hr for 5 min each at frequencies of 6693, 11,209, and 18,013 kc. The lower frequency was found to be most effective at intermediate ranges while the higher frequencies were more effective at extreme ranges. The change in frequency was brought about by a direct current timing motor activated by an 8-day aircraft clock. This motor switched filament voltages and antenna connections. The antenna contained wave traps spaced and tuned so that there was an appropriate resonant wavelength for each of the three frequencies.

On most of the transosonde flights the only parameters measured were the ambient air pressure and the transmitter pack temperature. These measurements were transmitted in Morse code by means of a modified AN/AMT-3 dropsonde. The coded signals were interrupted twice during each 5-min transmission period by 30-sec dashes for purposes of RDF bearings. Flight termination was accomplished by dual timers closing a circuit and firing an explosive squib which cut loose the load for parachute descent. In the event the balloon descended to a hazardous level prematurely or did not reach flying level within a certain time following release, the flight was similarly terminated. In order to reduce the hazard to aircraft, a flashing neon light operated during the nighttime hours.

On some short-duration transosonde flights a 35-mm camera with a 50-mm lens was attached to the balloon train to permit accurate balloon positioning by means of photographs of the terrain. On the early flights an attempt was made to retrieve the camera through radio-command, but this proved unsuccessful and on later flights a pre-set timer was used to release the camera from the balloon train.

3.5. Possible Future Instrumentation

Next to the wind, the most useful information which could be derived from CLB flights is the ambient air temperature. However, it is not as easy to obtain ambient air temperatures from CLB flights as it is from vertical soundings because the CLB moves with the air and thus there is absolutely no ventilation of the temperature instrument. Early attempts to obtain ambient air temperatures from CLB flights were not successful and it is probable that in order to obtain a reliable ambient air temperature, rather than a radiation temperature, it will be necessary to ventilate the temperature-measuring instrument with a fan as well as provide some sort of shield for the sensor. Feasible instrumentation might consist of a thermistor-type sensor forming one leg of a self-balancing bridge circuit. Since the CLB does not float exactly along a constant level or constant pressure surface, the temperature obtained will have meaning only insofar as the deviation of the CLB from average flight level is known. Thus, the possible precision in ambient air pressure measurement may indicate the accuracy of temperature measurement desirable. One use for ambient air temperatures obtained from CLB flights is in the estimation of vertical air motion by the adiabatic or isentropic method [26]. This estimate depends upon the rate of change of temperature following the CLB and the average lapse rate along the trajectory [equation (9.1)]. It can be shown that, assuming an average lapse rate, an error in temperature change of 0.5°C over a 6-hr time interval will produce an error in the computed average vertical air motion of about 1 cm/sec. Since CLB oscillations in the vertical in a region of considerable vertical temperature gradient could yield temperature changes of this magnitude, for accurate estimates of the vertical air motion it will be necessary to have an estimate of the temperature lapse rate at the CLB floating level or else utilize for temperature comparisons only temperatures obtained when the CLB is at the same (pressure) altitude.

For some problems it would be desirable to know the humidity along the CLB trajectory. Since the moisture content of the air is small at elevations where CLB flights are made, the same difficulty in humidity measurement from a CLB would exist as from a radiosonde at these elevations. The most feasible technique for measuring humidity from a CLB would appear to involve the use of a dew-point hygrometer, but it would require considerable simplification for CLB use. Knowledge of the humidity at these levels would be of direct use for contrail forecasting and of indirect use for various radiation and air transport considerations.

A most desirable piece of equipment for inclusion in the CLB system, but probably quite unfeasible economically and weight-wise, is a radio altimeter. From knowledge of the pressure height and true height of the balloon, D values could be obtained which would enable the CLB system to be tied in with the pressure-height information obtained from radiosondes. Over the ocean this would be most useful. The only alternative to this weighty and costly piece of equipment would seem to be dropsondes, by means of which one might be able to calculate the height of a given pressure surface from the surface pressure and the temperature lapse rate indicated by the dropsonde. The original transosonde concept involved the use of dropsondes but the idea was abandoned owing to the complexity of the resulting system and the large probable errors in dropsonde-derived data.

The approximate lapse rate of temperature at CLB floating level could be obtained by separate temperature measurements at top and bottom of the balloon train, by placement of thermocouples in opposition at top and bottom of the train, or by means of a dropsonde. Regardless of which technique is used, there would be telemetering problems because of the accuracy of temperature difference needed in order to obtain a useful lapse rate. The lapse rate obtained would be useful both for the estimation of the temperature at the mean CLB floating level and the estimation of vertical air motions by the adiabatic technique. It can be shown that an error in lapse rate of 0.14°C/100 meters would yield an error of 1 cm/sec in the 6-hr-average vertical air motion derived by the adiabatic technique.

The vertical wind shear could be estimated from the movement of air past the bottom of the balloon train, as measured by a thermosensory element. The wind speed shear could thus be obtained fairly easily, but the shear in direction would be much more difficult to evaluate because it would require a complicated wind-direction vane and magnetic orientation device. However, as a study [27] has shown that 80% of the velocity shear at jet stream levels is due to wind speed shear, neglect of the direction shear might not be too serious. An added degree of uncertainty in the wind shear measurements could result from the swinging of the balloon train in regions of air turbulence.

It is possible to measure directly the vertical air motion by means of an "integrating fan" attached to the CLB. This fan would give the total vertical flow of air past the balloon over a certain time interval, and since the vertical motion of the CLB itself could be estimated from the ambient pressure change, the vertical air motion could be derived. Actually, there is reason to believe that this technique would be more accurate than the adiabatic method, besides involving no approximations.

Special instrumentation of all sorts could be attached to CLB flights to yield transmissivity, conductivity, radiation balance, cloud cover, etc. It is doubtful that these parameters would ever be incorporated into an operational CLB system, however, because of the weight and cost of the additional instrumentation.

3.6. The Superpressure Balloon System

One of the major drawbacks of the ballast-dropping transosonde system utilized through 1959 is its tremendous weight which precludes flights within air space. Thus, as commercial jet aircraft fly higher and higher, the floating level of these transosonde balloons would have to be pushed higher and higher, and while flights at 40,000–50,000 ft are of interest for research purposes, it is doubtful that information from such flights would be utilized by the practicing meteorologist. Obviously, there is a vicious circle here with CLB-derived data being of most use to commercial aircraft and yet CLB flights are forbidden at levels where commercial aircraft fly. It is apparent that, if the full capabilities of the CLB technique are ever to be realized, that CLB flights will have to be made within air space. This means, in turn, that a radically new, extremely light CLB system will have to be developed whereby constant level flight is maintained without benefit of a ballast system.

The superpressured constant volume balloon system has the capability of maintaining constant level flight without release of ballast. The reason for this can be seen from the equation which equates buoyancy force and weight of the balloon system.

$$V_b(\rho_a - \rho_h) = W \tag{3.1}$$

where V_b is balloon volume, W is weight of the system, and ρ_a and ρ_h are the densities of air and helium, respectively. It is apparent from this equation that if a balloon maintains constant volume, then it will fly along a surface of constant air density. The balloon may lose its full volume due to seepage of gas through the skin of the balloon or through decrease in temperature of the gas within the balloon either due to radiation or to the lowering of the ambient air temperature. However, if initially the gas within the balloon has a considerably greater pressure than the air outside the balloon (superpressure), then, in spite of the above, the balloon will maintain its full volume and float along a constant density surface for a period of time. Only recently was a film produced which possessed properties suitable for such a pressurized balloon system. This film is Mylar which has the necessary qualities of high tensile strength, high elastic modulus, very low permeability, low brittle temperature, and transparency to solar radiation. Fortunately, there also has been developed recently a suitable adhesive and tape closure technique for balloon sealing which results in a seam strength approaching that of the Mylar itself.

During the past year, the Naval Research Laboratory has been conducting test flights over the Atlantic Ocean with such a superpressured transosonde system [28]. The instrumentation for these flights has been divided into two packages weighing 6 lb each because there is some indication that weight distributed in this way can be flown within air space. Both spherical and cylindrical Mylar balloons of 2-mil thickness and 1200 ft^3 volume have been used. The flights have been weighed-off to fly at 250 mb with a daytime superpressure of about 30 mb. The instrumentation for the superpressured balloon system is simpler and cheaper than that used previously since no ballast control devices are needed, while the balloon itself is cheaper because of its smaller size. The instrumented packages contain only a multi-frequency transmitter, a programer, a modulator for the various parameters to be telemetered, and a power supply. The power supply represents the major problem in achieving an equipment weight which will fall within the 12-lb load limit. At present, silver alkaline batteries are being used which furnish 40–50 watt-hr of energy per pound of battery. Thus, a 6-lb battery will enable a 3-min 25-watt transmission at 2-hr intervals to be made for a period of about 10 days. The major drawback of the silver cell battery is its cost (about $200) and the fact that it is necessary to maintain the battery temperature above freezing. There is hope that in the future cheap solar batteries can be obtained which could be used to charge the alkaline batteries during the sunlight hours.

The flight tests of the superpressured transosonde system have been quite successful. However, since the flights have been pre-set to terminate west of Europe the true capabilities of the system really cannot be determined. For example, it is not known how long these superpressured balloons will remain aloft, though it seems certain that circum-hemispheric flights can be achieved. Whether such circum-hemispheric flights take place in the future is dependent mostly upon international and financial considerations. It might be mentioned at this point that small superpressured balloons (tetroons) tracked at low levels by radar are now providing most interesting wind data for use in small-scale turbulence and diffusion studies. Thus, the superpressured CLB concept is slowly spreading to all scales of motion.

4. The Evaluation of Air-Flow Parameters from Constant Level Balloons

4.1. Introduction

At the present time, the basic data obtained from CLB flights consist of CLB positions as functions of time. From the distance and direction between successive positions, wind velocities may be determined, while the wind acceleration may be derived from the change in velocity with time. Treated in this section are the methods for deriving other flow parameters from CLB data, as well as some techniques for obtaining CLB positions and the resulting errors in flow parameters due to errors in such positions.

4.2. The Constant Level Balloon as a Quasi-Lagrangian System

A major advantage of horizontal sounding systems over vertical sounding systems resides in the fact that a CLB yields an approximation to the trajectory of an air parcel. Over and above the usefulness of knowing the trajectories of air parcels or airborne contaminants, this trajectory approximation makes it possible to estimate the ageostrophic components of flow and provides data in the quasi-Lagrangian frame of reference so important in turbulence and diffusion studies. A CLB yields only an approximation to an air trajectory because the CLB is confined to fly along a surface of constant height or constant pressure and thus cannot partake of the vertical air motion. Occasionally this approximation can be very important, as shown by Wexler [29] who found it impossible to trace smoke from a Canadian forest fire back to its origin unless the tracing was carried out on an isentropic surface. In spite of this drawback, the CLB comes closer to indicating the trajectory of an air parcel than does any other "tracer" in large-scale use, and its utility for this purpose should not be minimized. It is a curious thing that in the history of meteorology it was the drawing of air trajectories that first suggested the existence of surfaces of discontinuity or fronts [30], but that after the frontal system was fully established meteorologists gave very little attention to air trajectories. It now seems likely that trajectory concepts will push more and more into the forefront of meteorology as time progresses, and for this reason alone the CLB system should not be allowed to stagnate.

4.3. Determination of the Ageostrophic Wind from Constant Level Balloons

With minor approximations the vector equation of motion may be written

$$\frac{d\mathbf{V}}{dt} + w\frac{\partial \mathbf{V}}{\partial z} = f(\mathbf{V} - \mathbf{V}_g) \times \mathbf{K} + \mathbf{F} \tag{4.1}$$

where $\mathbf{V}$ is the horizontal part of the velocity, w its vertical component, $\partial\mathbf{V}/\partial z$ is vertical wind shear, f the Coriolis parameter, $\mathbf{V}_g$ the geostrophic wind, $\mathbf{K}$ the unit vertical vector, and $\mathbf{F}$ the frictional force per unit mass. The first term on the left in equation (4.1) gives the acceleration measured by a CLB. This does not represent the total acceleration experienced by an air parcel because the CLB does not partake of the vertical air motion. The second term on the left in equation (4.1) (the vertical advection of velocity) thus represents that part of the acceleration which cannot be obtained directly from CLB data. Its value may be approximated along a CLB trajectory through the estimation of vertical motion by the adiabatic technique (Section 9.2) and the estimation of vertical wind shear from rawinsonde data. On the basis of a limited number of computations from transosonde flights during 1953, Neiburger and Angell [27] estimated the magnitude of the vertical advection of velocity as about 25 % that of the CLB acceleration, whereas Giles and Peterson [31] estimated that this term averaged only 10 % of the CLB acceleration. Despite the possible importance of the vertical advection of velocity term, particularly under certain synoptic conditions, in most cases it must be assumed that the CLB acceleration represents a reasonable approximation to the total air parcel acceleration, since the bulk of the CLB positions are over the oceans where it is impractical to estimate the vertical advection of velocity by the technique described above.

In order to pass from the value of the acceleration to the value of the vector geostrophic deviation it is necessary, as seen from equation (4.1), to neglect friction. While attempts have been made to estimate the magnitude of the frictional force at jet stream levels from a combination of transosonde and radiosonde data [32, 33], the evaluations are far too crude for any but statistical treatment. With the neglect of friction and the vertical advection of velocity it is found, after taking the dot product of equation (4.1) with the unit vector along and normal to the wind, that, respectively.

$$\frac{dV}{dt} = fV_g \sin i \tag{4.2}$$

$$V\frac{d\theta}{dt} = -f(V - V_g \cos i) \tag{4.3}$$

where $d\theta/dt$ is the angular velocity and i is the angle of indraft (the angle between wind and geostrophic wind). Upon eliminating the geostrophic wind speed between equations (4.2) and (4.3), the angle of indraft can be expressed in terms of the tangential acceleration (dV/dt), normal acceleration ($Vd\theta/dt$), and Coriolis acceleration (fV), all of which are known from CLB data. From knowledge of the angle of indraft, the geostrophic wind can be evaluated from either equation (4.2) or (4.3). The cross-contour component of the ageostrophic wind ($V \sin i$) may then be determined as well as the along-contour component of the ageostrophic wind ($V \cos i - V_g$).

4.4. The Transosonde Positioning System

4.4.1. Techniques for Positioning. Positioning of CLB flights has been accomplished by means of photographs of the underlying terrain obtained from cameras attached to the balloon train, by means of aircraft "homing" on a transmitter on the balloon, and by means of bearings obtained by high frequency (HF), very high frequency (VHF), and ultra-high frequency (UHF) RDF stations. The average error in positioning by these three techniques is, respectively, about 1 mile, about 5 miles, and (as discussed below) about 20 miles if the positioning is carried out in the vicinity of North America by the high-frequency FCC RDF network. Camera positioning cannot be carried out over the oceans, at night, or in cloudy weather. Consequently, only short trajectory segments can be obtained by means of this technique, and therefore this is the method used to examine the relatively high-frequency wind oscillations in the atmosphere, as illustrated by the important work of Mantis and Huch at the University of Minnesota [34, 35]. Constant level balloon positioning by means of aircraft is inhibited by the aircraft range, weather conditions, and the considerable cost involved in keeping aircraft aloft. However, many of the early CLB flights made by New York University and General Mills were positioned in this way. Almost all of the Moby Dick and transosonde flights were positioned by RDF techniques.

4.4.2. The Federal Communications Commission Radio Direction Finding Network. The transosondes were tracked and positioned by both FCC and U.S. Navy high-frequency radio-direction-finding networks. The FCC network consists of 18 Adcock stations of which 15 are located within the contiguous United States, 2 are in Alaska, and 1 is in Hawaii. Bearings obtained from these stations were analyzed in Washington, D.C., and a most probable transosonde position was determined by means of a gnomonic plotting board. Flights over the Pacific were also tracked by the two U.S. Navy RDF networks in that area, with the most probable transosonde position being determined at Pearl Harbor, Hawaii. In addition to the most probable position, each "fix" was given a rating of accuracy (circle of merit) dependent upon the area of intercept of the bearings. Fixes were rated by the FCC as "good" if the true transosonde position was believed to be within a circle about the most probable position of radius 20 miles, "fair" if within a circle of radius 40 miles, and "poor" if within a circle of radius 60 miles.

Several studies have considered the reliability and accuracy of fixes obtained by the FCC RDF network. On the basis of transosonde flights over the United States and North Atlantic, Anderson [36] found that the probability of an FCC station getting a usable bearing is 88% at a distance of 2000 mi and 65% at a distance of 3000 miles. Anderson [37] also found from analysis of a camera-carrying Moby Dick flight tracked by the FCC, an average FCC positioning error of 19 nautical miles. However, Giles and Peterson [38] found from Moby Dick flights an average FCC positioning error of 29 nautical miles, with the average error being 10 nautical miles for fixes rated good, 29 nautical miles for fixes rated fair, 38 nautical miles for fixes rated poor, and 81 nautical miles for fixes rated "estimated," a rating to which no circle of merit is attached by the FCC. Their results indicate that 74% of the time the true balloon position is within the circle of merit indicated by the FCC. These results may be treated with caution when applied to transosonde flights since the transmitter power of the Moby Dick flights was considerably less than that of the transosonde flights, and furthermore the transosonde system utilized three frequencies with a long dash to aid direction finding operations whereas the Moby Dick system utilized only one frequency with no long dash. As a possible indication of the greater accuracy with which transosondes could be positioned, it is noted that based on a small number of comparisons of FCC positions with transosonde positions obtained by camera an average FCC position error of only about 8 nautical miles was found [27].

Ohnsorg [39] evolved a technique for estimating RDF tracking errors from comparisons of velocity variances obtained through the use of different averaging intervals. His method is based upon the finding, from CLB camera positions, that high-frequency wind oscillations are of small magnitude in comparison with the high-frequency CLB oscillations brought about by errors in RDF positioning. Thus, with little error, all high-frequency oscillations may be ascribed to errors in RDF positioning. Ohnsorg's results indicate that over the United States the average FCC position error is about 9 nautical miles whereas over the oceans the average error may range up to 35 nautical miles. However, these errors are on the conservative side because of a possible systematic bias in a whole series of position reports. Anderson and Mastenbrook [40] estimated the accuracy with which the transosondes were positioned over the North Atlantic from a comparison of individual transosonde positions with smoothed positions obtained by a type of least squares method, a technique not too different from that of Ohnsorg. In both methods all small-scale fluctuations in the trajectories are attributed to positioning error. Anderson and Mastenbrook found thereby an average FCC positioning error of 19 nautical miles over the western North Atlantic. Thus, the consensus would seem to be that in the vicinity of North America the average error in transosonde positioning by the FCC is about 20 nautical miles whereas over the western North Pacific and the eastern North Atlantic the average errors may be twice as great. However, a disillusioning note is introduced through a comparison of transosonde positions obtained simultaneously by the FCC and U.S. Navy RDF systems over the Pacific Ocean [41]. This comparison showed the average difference in transosonde position indicated by the two networks to be 228 nautical miles west of 180°, 132 nautical miles from 180–150°W, and 102 nautical miles from 150–120°W. Since the true transosonde positions were not known in this case, it is impossible to tell which of the tracking networks was responsible for most of the error in transosonde positioning.

4.4.3. Errors in Flow Parameters Due to Errors in Transosonde Positioning. If the average error in CLB positioning is $\overline{\Delta r}$, it can be shown [27, 37] that the average error in the velocity derived from the straight-line distance r between CLB positions is $\overline{\Delta r}/t$, where t is the time interval between the obtaining of the positions. In determining r, the replacement of a spherical triangle, taking into account the curvature of the earth, by a plane triangle is justified since even for velocities computed over a 6-hr time interval, the error introduced by this assumption (for latitudes less than 70°) is less than 1 knot for winds of less than 100 knots. It also may be shown that the average error in CLB acceleration due to errors in CLB positioning is $\sqrt{2}\,\overline{\Delta r}/tT$, where T is the time interval between the velocity determinations, and that the average error in ageostrophic wind obtained from the CLB positions is $\sqrt{2}\,\overline{\Delta r}/ftT$, where f is the Coriolis parameter. On the one hand it is desirable to choose large time intervals over which to compute velocity, acceleration, and ageostrophic wind in order to minimize the errors in these parameters resulting from errors in CLB positioning. On the other hand, the larger the time intervals chosen, the more the high-frequency variability of these parameters is reduced due to the filtering properties of averages [42]. For example, the variability in an oscillation of period 12 hr is reduced 36% when using a 6-hr averaging interval whereas the variability in an oscillation of period 24 hr is reduced only 10% by the use of such an averaging interval. The problem is to select optimum values of t and T which neither smooth out the high-frequency variability of interest nor permit large errors in the estimation of velocity, acceleration, and ageostrophic wind.

In Section 4.4.2 it was shown that, in the vicinity of North America, the average error in CLB position determined by the FCC is about 20 nautical miles whereas over the western North Pacific and eastern North Atlantic the average error is probably twice as great. It would be desirable to let the average error in CLB-derived velocity and ageostrophic velocity not exceed 10% of the average values of these parameters. As the average transosonde-derived values for these quantities are about 70 and 20 knots, respectively, it is seen that the choice $t = 4$ hr, $T = 8$ hr in the vicinity of North America (average velocity error 5 knots, average ageostrophic wind error 2.3 knots) and the choice $t = 6$ hr, $T = 12$ hr for flights traversing the Pacific and Atlantic Oceans (average velocity error 6.7 knots, average ageostrophic wind error 2.2 knots) approximately satisfy the criteria given the above error estimates. For comparison, the average instrumental velocity error of the rawinsonde GMD-1A is about 5 knots between 27,000 and 45,000 ft, as determined from phototheodolite data [43]. Through the use of the above averaging intervals, the velocity variability associated with a 12-hr periodicity is reduced by 17% and 36%, respectively, while the ageostrophic wind variability associated with this same periodicity is reduced by 56% and 100%, respectively. It is apparent that over the oceans the transosonde flights are not yet accurately enough positioned to permit the delineation of semidiurnal, or perhaps even inertial (18-hr period), oscillations in the ageostrophic wind, even if such oscillations exist.

5. Horizontal and Vertical Sounding Systems—A Critique

5.1. Introduction

At the present time much emphasis in meteorology is being placed upon improved knowledge of the 3-dimensional field of motion within the atmosphere, because without such knowledge it is difficult to understand the weather and even more difficult to predict it. For many years meteorologists have based their analyses and forecasts on constant level or constant pressure maps. The horizontal air flow has been estimated from these maps from a combination of contour or isobar configuration (geostrophic, gradient, balanced wind) and winds obtained from vertical soundings (rawinsondes). The fact that both sets of data are plotted on the same map suggests that neither set by itself successfully delineates the horizontal air motion which the meteorologist wishes to know. As is well known, attempts to pass from the pressure or height field to the wind field are fraught with danger owing to the theoretical approximations involved and the errors in radiosonde instruments. On the other hand, a striking example of the inadequacy of vertical soundings (or at least pilot balloons) for estimating the air flow is given by Spilhaus *et al.* [11], citing a case when a CLB launched from Alamagordo, New Mexico moved ENE at a time when Albuquerque, Roswell, and El Paso pilot balloons indicated no winds from the SW quadrant while the White Sands pilot balloon indicated only a very shallow layer moving in an ENE direction. In this section we shall consider some reasons why the types of data now entered on weather maps are not overly suitable for delineating the air flow and indicate the improvements which would result if CLB data were entered in addition or in their stead.

5.2. "Representativeness" of Horizontal and Vertical Soundings

In the present state of the meteorological art, winds above a meteorological station are obtained at six-hourly intervals by means of vertical soundings. The wind at each level is determined from the 2-min horizontal displacement of a balloon rising vertically at a rate of about 1000 ft/min. Over and above the error in wind measurement brought about by the lag in the response of a radiowind (rawin) sonde balloon to the wind in an atmospheric layer as the balloon rises through it [44], it is likely that during this 2-min interval the balloon is influenced by "eddies" which render the wind obtained for this level for this station for this time unrepresentative. For a vertical sounding it is especially difficult to take into account the possible effects of these eddies since the balloon is not affected by eddies smaller than itself and therefore as the balloon goes higher in the atmosphere and expands, different size eddies are effective in displacing the balloon [45]. The magnitudes of these eddies have been estimated by serial vertical soundings in quick succession [46–49] by the study of smoke puffs [50] and by the study of accurately positioned constant level balloon flights [35]. It has been found that the local standard deviation of the wind is not inconsiderable over small time intervals, with values ranging from 1–5 knots within time intervals of 5–15 min. The CLB flights indicate wind direction changes exceeding 20° within a few miles with transverse velocity fluctuations as large as 20 knots in 1 hr. Thus, it seems likely that a certain unrepresentativeness will always creep into winds determined by means of vertical soundings.

For the determination of the atmospheric stability, a vertical sounding is a natural technique. However, logically, it is a curious thing to find winds averaged in the vertical plotted on quasi-horizontal maps. How much more reasonable to plot on a quasi-horizontal map a wind averaged in the horizontal with respect to both space and time. This is just the type of average wind that is derived from CLB trajectories. If the positioning of the CLB is sufficiently precise, winds can be averaged over any desired time interval (30 mins, 2 hr, 6 hr, etc.), whichever is most desirable for the purpose at hand. Certainly winds averaged over such time intervals would be more representative than winds averaged over 2 min regardless of the spectral details of atmospheric oscillations. Moreover, the very fact that CLB-derived winds can be averaged over different time intervals permits experimentation with the averaging period needed to delineate those fluctuations in the air flow which have a bearing on the forecast. This can never be done with vertical soundings. Thus it may be found that oscillations of 30-min period have no bearing on future weather conditions whereas oscillations of 6-hr period do have such a bearing. On the other hand, it may be found that oscillations of all periods are important in future weather prediction. Only the CLB yields wind data continuous in space and time so that one may judge the continuity of oscillations and their importance in future weather developments. In fact, with a sufficient number of CLB flights in the air at one time, one can foresee an experiment whereby CLB-derived winds, averaged over different time intervals, are fed into electronic computers for the purpose of delineating that averaging period which best serves as an indicator of tomorrow's flow pattern and weather.

5.3. Validity of Climatological Data Derived from Horizontal and Vertical Soundings

Since the positioning of a CLB is not dependent upon single-station line-of-sight transmission systems, successive CLB positions yield winds without regard for the magnitude of the wind speed or balloon height in the atmosphere. This is a goal not yet achieved by even the most modern rawinsondes under conditions of great tropospheric wind speed. It is of interest to note that in an attempt to avoid low elevation angles (under conditions of great wind speed) with present-day rawinsonde equipment, the ascent rate of the conventional balloon is being increased. This means either that the wind at a level is determined by a horizontal displacement of less than 2-min duration or that the spatial averaging interval in the vertical is made greater. Obviously, in the vertical sounding field *"you can't have your cake and eat it too."* On the other hand, the only bias in CLB-derived data involves the launching limitations. A CLB launch in winds of more than 25 knots is difficult, while precipitation and icing conditions must be avoided because they inhibit the ascent of the balloon. If the release point is near an airport, the hazard to aircraft will prevent releases if more than one-half the sky is covered by clouds below 10,000 ft. In a few cases a temperature inversion has inhibited the balloon ascent so that it had to be terminated prematurely. Once aloft, however, the more rapid the movement of the CLB the percentually more accurate the value of wind speed determined therefrom as the errors in positioning have relatively little influence if the CLB has traveled a great distance during the averaging period. Thus the horizontal sounding furnishes the best wind information under just those conditions when the vertical sounding furnishes the poorest wind information, if it furnishes any information at all. A recent study [51] gives some statistics on different mean wind speeds obtained by means of vertical and horizontal soundings which are highly suggestive on this point.

Another important attribute of the CLB is its ability to furnish a complete wind climatology in space and time. Thus the two years of routine transosonde flights from Japan yielded wind statistics near the tropopause level over the Pacific Ocean which could hardly be obtained in any other way. Even the establishment of rawinsonde equipment on merchant vessels will not solve the problem of upper-air coverage above the Pacific Ocean because the merchant ships will traverse only certain designated routes, and consequently, upper-air data gaps will still exist. The CLB probably offers the most feasible means, if not the only reasonable means, of obtaining upper-air data over the tropical regions, the great water masses of the Southern Hemisphere, and the continent of Antarctica [52]. In addition, the fact that CLB-derived velocity data are obtained at relatively closely spaced geographical intervals suggests many advantages of these data in the computation of parameters such as the meridional transport of zonal momentum, while the fact that CLB-derived velocity data are obtained at frequent intervals of time (not at 6- or 12-hr intervals) suggests many advantages of these data in the investigation of phenomena such as atmospheric tidal oscillations [53].

6. The Constant Level Balloon as a Verifier of Trajectory Hindcasts and Forecasts

6.1. Introduction

The development of the CLB has made possible the verification of various techniques for forecasting or hindcasting trajectories in the atmosphere. This verification is possible only for trajectories constructed on surfaces of constant pressure or constant height since these are the surfaces most nearly adhered to by present-day CLB systems. The establishment of the most accurate method of trajectory construction is important for a variety of meteorological purposes including the forecasting of where CLB flights will go if released from a certain locale during a given time interval [54]. Along this line, it was gratifying to find that "representative" trajectories exist in the atmosphere as shown by the small deviations in the trajectories of simultaneously released CLB flights at the same elevation [15].

Techniques for trajectory construction conveniently may be divided into kinematic and dynamic methods. In the kinematic method the trajectory is estimated from knowledge of the wind field. The wind field, in turn, may have been estimated from the pressure field through the geostrophic, gradient, or other pressure gradient-wind approximations, but the important thing is that it is the wind field which is directly utilized. In the dynamic method of trajectory determination, on the other hand, the trajectory is estimated from comparison of the pressure gradient force and Coriolis force, the difference between which yields the air parcel acceleration [equation (4.1)]. In the dynamic method nearly everything depends upon accurate knowledge of the pressure field, and consequently, reversing the procedure utilized in the kinematic method, the winds are used as an aid to better analysis of the pressure field.

6.2. Verification of Kinematic Trajectory Methods

There are at least eight well-known kinematic trajectory methods; namely, Petterssen's method, George's method, central tendency method, consecutive streamline method, linear interpolation method, relative trajectory method, General Mills method, and U.S. Navy method. General discussions of these methods have been presented by Gustafson [55] and Johannessen [56] while specific techniques are discussed in references [57–59]. The interesting thing is that while CLB trajectories have indicated that each of the above methods has certain advantages under certain synoptic conditions, in the average they are all about equally successful (or unsuccessful). Thus, Machta [60], utilizing for verification 200-mb CLB flights from New Mexico and South Dakota during 1949–50, found little difference in the accuracy of hindcasts obtained by means of the central tendency method, consecutive streamline method, and linear interpolation method with, in all three cases, the average error in terminal position amounting to about 32% of the flight range of some 700 miles. It should be pointed out, however, that since 1950 wind coverage and wind accuracy over the United States have very much improved so that better results might be expected now. Machta also tested various kinematic methods in mathematically defined flow patterns [61] and found the difference in accuracy among techniques to be small.

An Air Weather Service project in 1952 determined the accuracy with which various kinematic techniques yielded trajectory forecasts at 30,000 ft using Moby Dick CLB flights for verification. Again there was no significant difference among techniques with the average error in terminal position being 32.5% of the distance traveled. It might seem surprising that the hindcast error of Machta is no less than the forecast error of the Air Weather Service, but this result is partially confirmed in [15] wherein it is reported that by the use of various kinematic techniques the average forecast error in terminal position was 23% of the distance traveled whereas the average hindcast error was 20% of the distance traveled. Apparently there are times in meteorology when the blame for an incorrect result does not rest solely on the forecaster.

6.3. Verification of Dynamic Trajectory Methods

The two dynamic methods most utilized for trajectory forecasting or hindcasting are the Freeman-Franceschini method [62–64] and the Air Weather Service method [56]. There also have been a few trajectory constructions using Sasaki's Hamiltonian method [65]. The first two methods are basically the same and involve iteration of the simplified equation of motion [equation (4.1)]. Thus, an initial estimate of the geostrophic deviation leads to a value for the acceleration which in turn leads to an estimate of the new position for the air parcel where a new value for the geostrophic deviation is obtained, etc. Initially this method works well, probably giving better results than the kinematic methods. Sooner or later, however, these dynamic trajectories tend to "blow up," due in no small part to the ever increasing amplitude of the inertial oscillations (see, e.g., p. 45 of [56]). Various mechanisms for damping the inertial oscillations have been proposed but the physical reason why such damping is required has never been obvious. Recently Mantis [35], in a most interesting work, has presented some fine-scale trajectory information obtained from camera positioning of CLB flights. These data indicate the existence of small-scale pressure patterns which cannot be detected on the synoptic scale and Mantis suggests that it is these small-scale pressure oscillations which damp the inertial geostrophic deviations and prohibit large deviations between wind and geostrophic wind in the atmosphere. If such small-scale pressure fields actually exist, then, regardless of the complexity of the model used, it will be impossible to improve dynamic trajectory forecasting or hindcasting, until one obtains knowledge of this small-scale field.

Customarily, in using the dynamic method of trajectory construction, the term involving the vertical advection of velocity is neglected. Since it has been found that the magnitude of this term is 10–25 % of the magnitude of the CLB acceleration, its neglect might be harmful in some trajectory determinations. This was tested by hand-iterating the equation of motion (basically the Freeman-Franceschini method) along transosonde flight 993 with and without use of this term [66]. The magnitude of the vertical motion in the vertical advection of velocity term was estimated by the adiabatic method using radiosonde temperatures interpolated to the CLB position while the vertical wind shear was estimated from surrounding rawin stations. It was found that the inclusion of the vertical advection of velocity term brought about a considerable improvement in the hindcast trajectory in this particular case. Since it is believed that the velocity fluctuations along this flight are mostly inertial in character (Section 8.3), the possibility arises that the term involving the vertical advection of velocity enters in some systematic way into inertial fluctuation and that, the neglect of this term is at least partially responsible for the unrealistic amplitudes of inertial oscillations determined by the Freeman-Franceschini and Air Weather Service dynamic methods.

6.4. Verification of Electronic Computer Methods

The increased use of electronic computers for the preparation of prognostic upper-air maps makes trajectory computations fairly simple. The type of trajectory resulting from such a computation is dependent upon the assumption (geostrophic wind, balanced wind, etc.) utilized in the preparation of the prognostic chart. Examples of such trajectories have been given by Djuric and Wiin-Nielsen [67] and Hubert *et al.* [68]. The latter based the trajectory forecast upon 72-hr equivalent barotropic forecasts made by the Joint Numerical Weather Prediction Unit (JNWP) at Suitland, Maryland. These trajectory forecasts are particularly interesting because they were initiated at the same time that transosondes were launched from Japan or California. On the basis of 11 comparisons, the average trajectory forecast error in terminal transosonde position amounted to about 25 % of the trajectory length for flights of 72-hr duration, thus being of the same order of magnitude as the errors derived from the kinematic and dynamic methods. This error might have been even less except for the fact that the numerical forecast was made at the 500-mb level and thus extrapolations to

300 mb had to be carried out for the purpose of comparison with transosonde trajectories. A discussion of an unsuccessful transosonde trajectory forecast by electronic computer is given in Section 8.5.

7. Climatological Analysis of Operational Transosonde Flights

7.1. *Introduction*

Under the transosonde development program, flights were made from Tillamook, Oregon in the fall of 1952 [69]; from Minneapolis, Minnesota in the spring of 1953 [70]; from Vernalis, California in the spring of 1955 [71]; from the Naval Air Facility at Oppama, Japan in the winter of 1956 [72]; and from Fallon, Nevada in the summer of 1956 [73]. The continuous improvement with time in the quality of these transosonde flights warranted the planning of an operational transosonde system for 1957. In preparation for this event, two manuscripts were issued for the guidance of meteorological personnel who would make use of the resulting trajectory-type data [74, 75]. Operational transosonde flights from the Naval Air Facility at Iwakuni, Japan were carried on during the fall, winter, and spring of 1957–58 and 1958–59. Since these flights furnish by far the most homogeneous and extensive data yet obtained from horizontal sounding systems, this section treats the meteorological information derived from these two years of flights.

7.2. *Trajectories and Trajectory Dispersion*

Most of the flights from Japan during the first year of operation were at 300 mb (30,000 ft) while most of the flights during the second year of operation were at 250 mb (34,000 ft) owing to a change in the pressure surface analyzed by National Weather Analysis Center (NAWAC). Figure 1 shows the trajectories of all the transosonde flights for the 2-yr period which were tracked and positioned by the RDF networks for at least one day. Detailed summaries of these trajectories are given in [76, 77] while a climatological summary of the data derived therefrom is given in [78]. Originally the flights were set to terminate 7 days after release from Japan, but in November 1957 the transosondes began to approach Europe only four days after release with average speeds along the trajectories exceeding 100 knots. Because of the complaints of certain nations with regard to overflying Eurasia with such balloons, the flight duration during the winter months of 1957–58 was pre-set at five days. During the winter of 1958–59 the transosonde flights were limited to four days' duration in order to avoid interference with the activities of commercial jet aircraft over the Atlantic Ocean. Of the 230 transosondes released during these two years, 179 were tracked and positioned by RDF stations for at least 1 day, 152 for at least 2 days, 116 for at least 3 days, and 83 for at least 4 days.

Figure 2 shows the (smoothed) percentage of transosondes released which

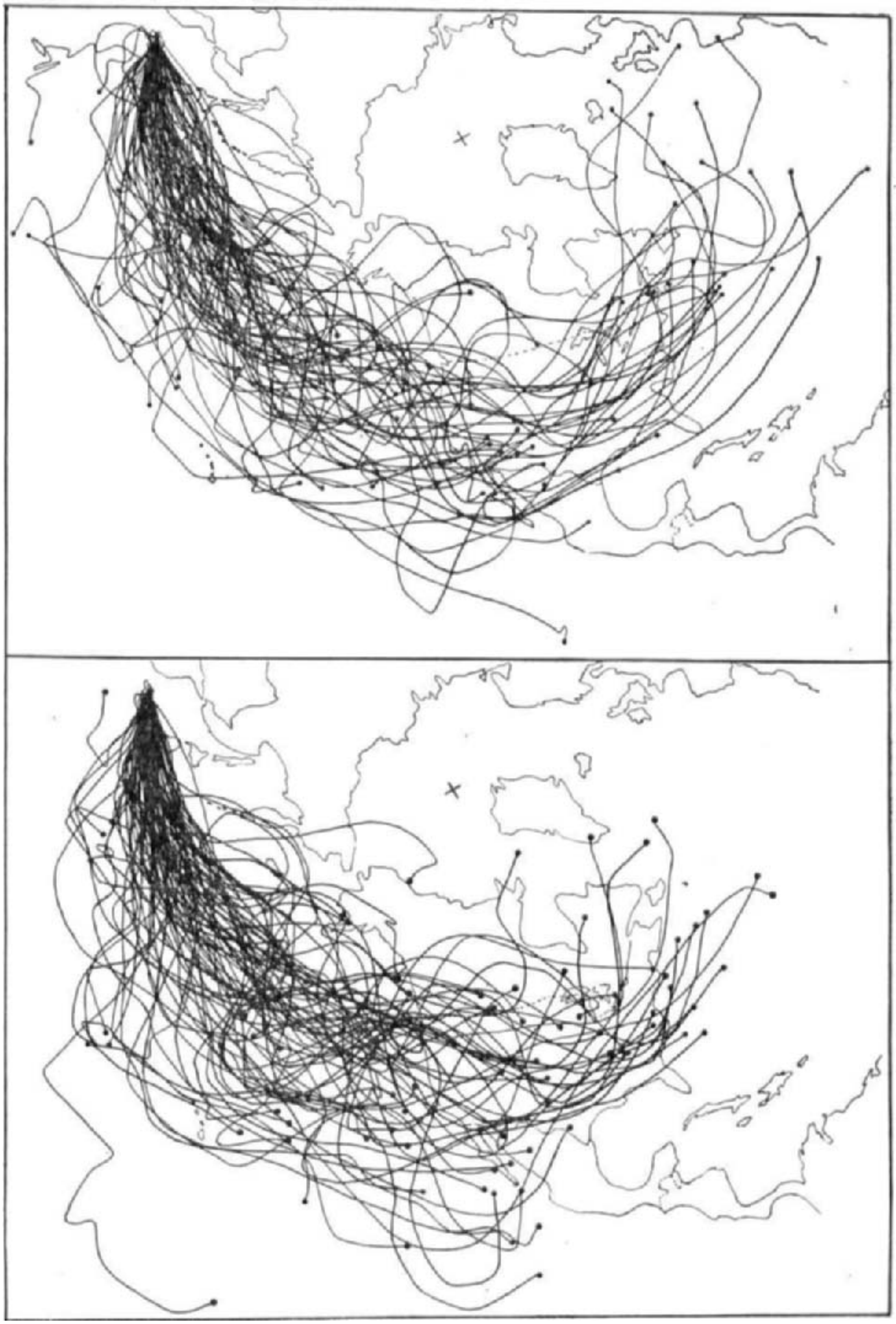


FIG. 1. Transosonde trajectories, for flights of one or more days' duration, between September 1957 and June 1958 (top) and between September 1958 and April 1959 (bottom).

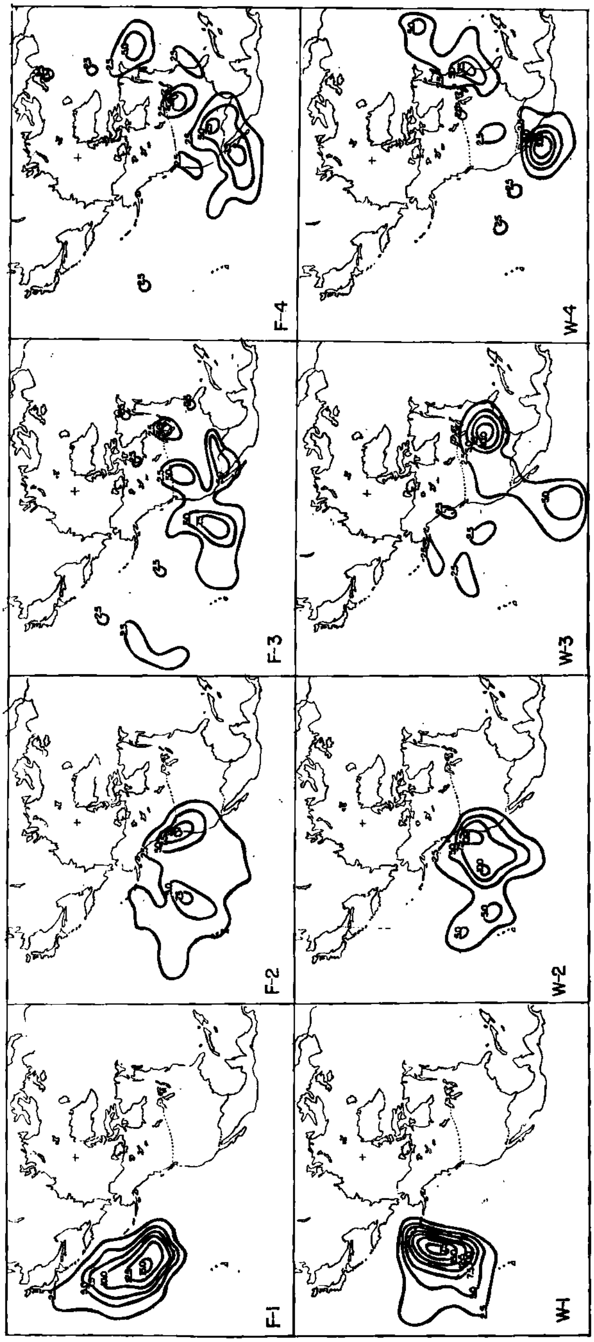


FIG 2.

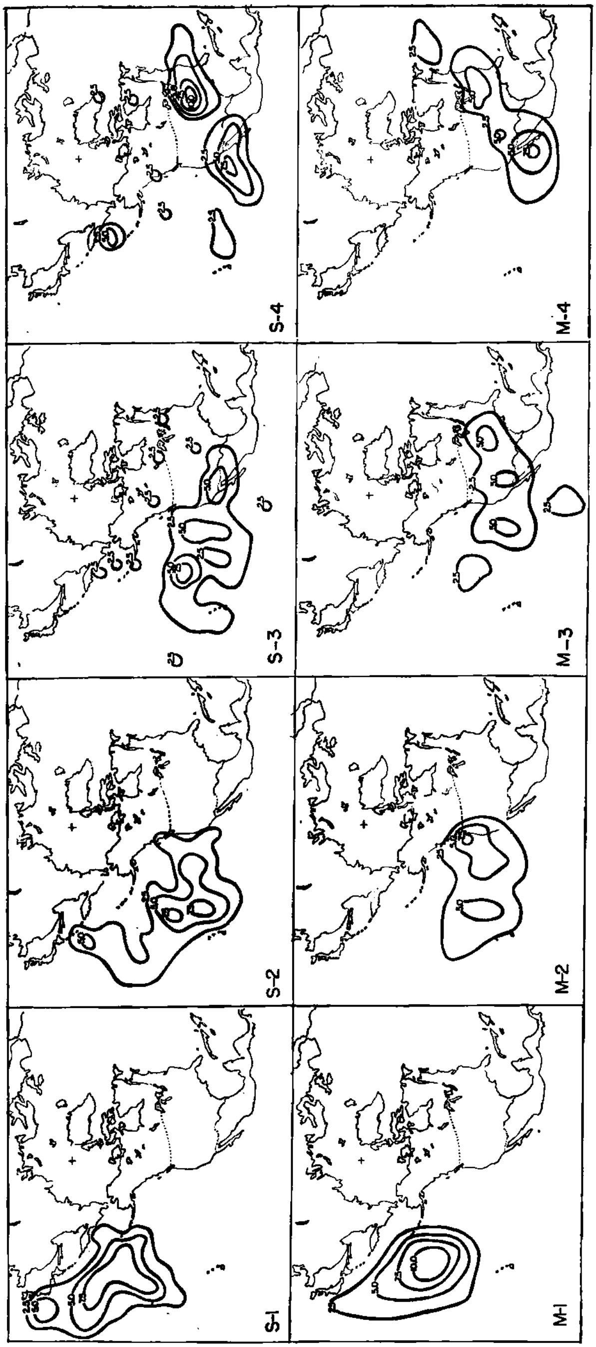


FIG. 2. Percentage of transosondes located within 10-degree latitude-longitude "boxes" a given number of days after release from Japan for fall(F), winter(W), spring(S), and the mean(M) of all 1957–59 flights of one or more days' duration.

were located within 10° latitude-longitude "boxes" from one to four days after release from Japan. In the mean for all flights, the maximum percentages are located near 180° longitude one day after release and just to the west of the State of Washington two days after release. Three days after release, in the mean for all flights, the band of maximum percentage extends from 140°W to the Mississippi River, while four days after release the maximum percentages are centered over Kentucky and an area somewhat to the west of Baja California. One would judge from Fig. 2 that in all three seasons there are two dominant circulation patterns over the northeastern Pacific and North America, one of which carries the transosondes eastward across the United States while the other carries the balloons into a transosonde "graveyard" to the west of Baja California. This tendency for transosondes to congregate near Baja California is of practical importance since many storms influencing southern California and Arizona are difficult to position owing to the lack of conventional upper-air data in this region. Also of interest in Fig. 2 is the concentration in transosonde position one day after release during winter which indicates the steadiness of the wintertime winds over and to the east of Japan.

Values of the latitudinal and longitudinal standard deviation of transosonde position a given number of days after release were determined by direct evaluation of the root mean square of the latitudinal and longitudinal distances between mean and individual flights positions for all 300 and 250 mb flights of four or more days' duration. Figure 3 shows these lati-

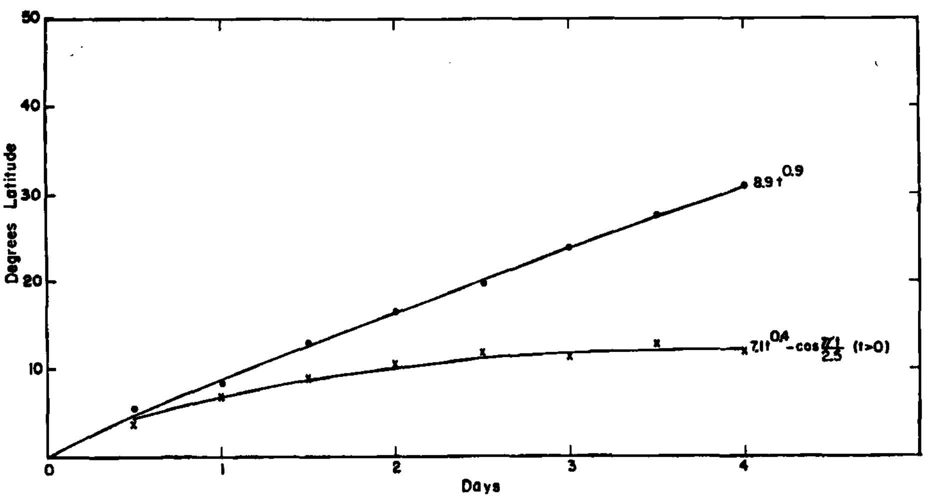


FIG. 3. Mean latitudinal (crosses) and longitudinal (dots) standard deviations of transosonde position as a function of time after release from Japan for all 1957–59 flights of four or more days' duration. The lines represent the fit of the indicated analytic expressions to the observed values (t in days).

tudinal (crosses) and longitudinal (dots) standard deviations as a function of time after transosonde release. The values plotted are the means of the standard deviations (in degrees latitude) determined separately for fall, winter, and spring. The lines in the figure represent the fit of analytic expressions to these values. It is seen that the increase of the longitudinal standard deviation (σ_l) with time is well approximated by the expression

$$\sigma_l = 8.9t^{0.9} \tag{7.1}$$

where t is in days, while the increase of the latitudinal standard deviation (σ_L) with time is well approximated by the expression

$$\sigma_L = 7.1t^{0.4} - \cos(\pi t/2.5) \qquad (t > 0) \tag{7.2}$$

where t is in days. Thus the longitudinal standard deviation increases almost linearly with time after release, whereas the latitudinal standard deviation increases more slowly with some evidence for the superposition upon the power variation of a small-scale sinusoidal fluctuation. This sinusoidal fluctuation would be associated with the tendency for latitudinal bunching of the transosondes as they approach the mean trough position near the east coast of North America. With the aid of these analytic expressions the standard deviations of position can be estimated for time intervals after transosonde release exceeding four days. Figure 4 shows, by means of ellipses, the areas within which these standard deviations indicate 50% of the transosondes would be located a given number of days after release from Japan. The ellipses are centered on a mean wintertime trajectory determined from the 1957–59 transosonde flights for the first four days following release and from mean 300-mb maps prepared by Brooks *et al.* [79] for the remainder of the trajectory. This figure makes more obvious the great dispersion in the longitudinal direction compared with the dispersion in the latitudinal direction when the flights are grouped by season. As a result, 10 days after release it is estimated that the latitudinal extremes of the ellipses would not fall outside the zone of prevailing westerlies, whereas the longitudinal extremes embrace all of Eurasia. On the basis of earlier transosonde flights, launched from Minneapolis, Minnesota and Japan, Edinger and Rapp [80] found larger latitudinal and smaller longitudinal standard deviations of position as a function of time after release than found here.

In turbulence theory the relations

$$\sigma_l^2 = 2K_l t \tag{7.3}$$

$$\sigma_L^2 = 2K_L t \tag{7.4}$$

may be derived where K_l and K_L are the *austausch* coefficients in the longitudinal and latitudinal directions. It is thus possible to estimate the value of the large-scale *austausch* from values of standard deviations of posi-

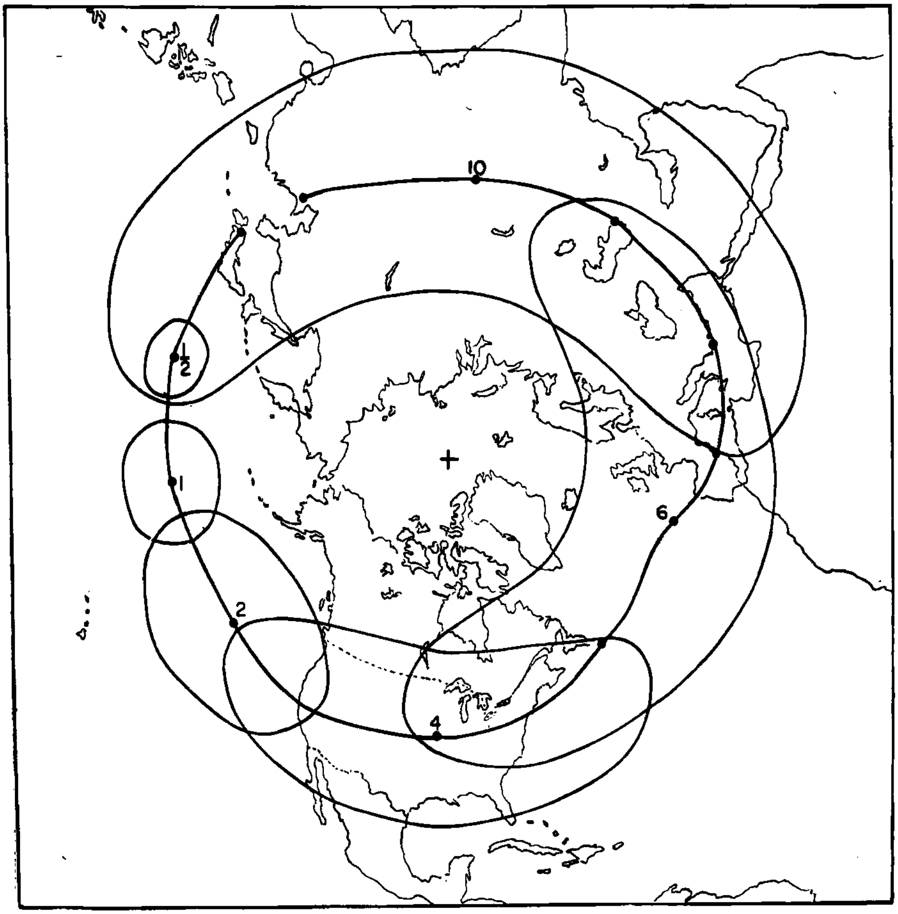


FIG. 4. Mean trajectory from Iwakuni, Japan and areas (ellipses) within which observations and computations indicate 50% of the transosondes would be found a given number of days after release.

tion as functions of time obtained from the 1957–59 transosonde flights of four or more days' duration. Table I gives these values based on the standard deviations presented in Fig. 3. The latitudinal *austausch* averages about 5.5×10^{10} cm^2/sec whereas the longitudinal *austausch* averages about 3 times this value but increases steadily with time after transosonde release. For comparison, based on moisture variations along isentropic surfaces, Miller [81] and Grimminger [82] found average *austausch* values of about 3×10^{10} cm^2/sec while Defant [83], on the basis of the heat flux associated with cyclones, estimated the large-scale *austausch* to be of somewhat smaller magnitude.

TABLE I. Latitudinal and Longitudinal *Austausch* Values Derived from Standard Deviations of Position as Functions of Time after Release for all 1957–59 Transosonde Flights from Japan of Four or More Days' Duration.

Time (hr)	Latitudinal *austausch* (cm^2/sec) $\times 10^{10}$	Longitudinal *austausch* (cm^2/sec) $\times 10^{10}$
12	3.4	6.8
24	5.7	8.2
36	6.2	13.0
48	6.6	15.5
60	6.6	17.9
72	5.2	21.8
84	5.5	25.0
96	4.3	28.2

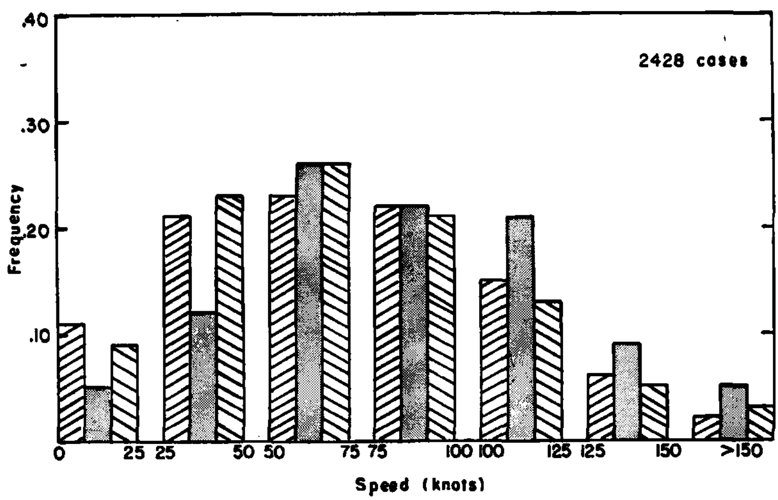


FIG. 5. Distribution of 6-hr-average wind speeds derived from 1957–59 transosonde flights from Japan. Within each class interval the left-hand, middle, and right-hand columns give, respectively, the frequencies for flights during fall, winter, and spring.

7.3. *Velocity and Ageostrophic Velocity Statistics*

Because of the difficulty in obtaining accurate transosonde positions over the vast reaches of the Pacific Ocean, the transosonde-derived wind velocity was obtained from the distance and direction between smoothed transosonde positions 6 hr apart, where the smoothed positions were obtained by the averaging of three successive latitude and longitude determinations 2 hr apart. Figure 5 shows the distribution by season of the 6-hr-average 300- and 250-mb wind speeds so obtained. For all three seasons the speed mode is 50–75 knots. However, in winter 35 % of the wind speeds

TABLE II. Variation of Transosonde-Derived Wind Speed (Knots) with Latitude and Longitude Based on 1957–59 Flights from Japan.

Latitude	Fall	Winter	Spring	Mean	Longitude	Fall	Winter	Spring	Mean
60	85	53	64	62	150°E	82	113	71	88
50	85	75	79	79	170°E	72	103	78	84
40	83	95	81	87	170°W	76	81	82	80
30	50	82	65	66	150°W	70	74	69	71
20	44	63	52	54	130°W	63	65	66	65
					110°W	60	67	58	62
					90°W	72	89	69	76
					70°W	84	86	78	83

exceed 100 knots whereas in fall and spring only about 23 % of the speeds exceed this value. In winter, 5 % of the speeds exceed 150 knots. This percentage would doubtless be higher except for the considerable reduction in peak wind speeds brought about by the use of 6-hr-average winds derived from smoothed positions.

Table II shows the transosonde-derived wind speed at 300 and 250 mb as a function of latitude and longitude for the three seasons and the mean of all flights. In the mean for all flights the wind speed is a maximum at latitude 40°, considerably north of latitude 30° where the zonal geostrophic wind apparently has its hemispheric maximum [84]. This discrepancy could be a result of comparing wind speed with zonal wind speed, the result of a geostrophic wind bias (wind in the trough always subgeostrophic), a result of the fact that the transosondes, launched from a single location, sample ridges at latitude 40° and troughs at latitude 30°, or the result of a geographical bias brought about by the comparison of winds over a portion of the hemisphere with winds around the entire hemisphere. The variation with longitude of the transosonde-derived wind speed may be slightly biased because it is determined over a limited latitude range. Nevertheless, the magnitude of the wind speed just to the east of Japan during the winter is impressive. Early evidence for unusually strong wind speeds in this region of the hemisphere was obtained from balloon releases upwind from Tateno, Japan [85]. Rather surprisingly in spring the maximum transosonde-derived wind speed is located near 165°W rather than near Japan. It is of interest that from 180° longitude to the west coast of North America the wind speed is nearly the same during all three seasons, whereas the table illustrates the well-known fact that over North America the wind speed is greater in winter than in fall or spring.

In general, previous attempts to estimate the magnitude of the ageostrophic wind involved comparisons of pilot balloons or rawins with the geostrophic wind obtained from an isobaric analysis [86, 87]. Godson evaluated the ageostrophic wind by estimating the values of the partial derivatives in the equation of motion [88]. One of the advantages of CLB data is that an approximation to the individual change of velocity with time is obtained directly, making the estimation of the ageostrophic wind a comparatively simple task.

Figure 6 gives the distribution of 12-hr-average "natural" ageostrophic components obtained from the 2 years of transosonde flights. These components were evaluated from the equation of motion utilizing accelerations

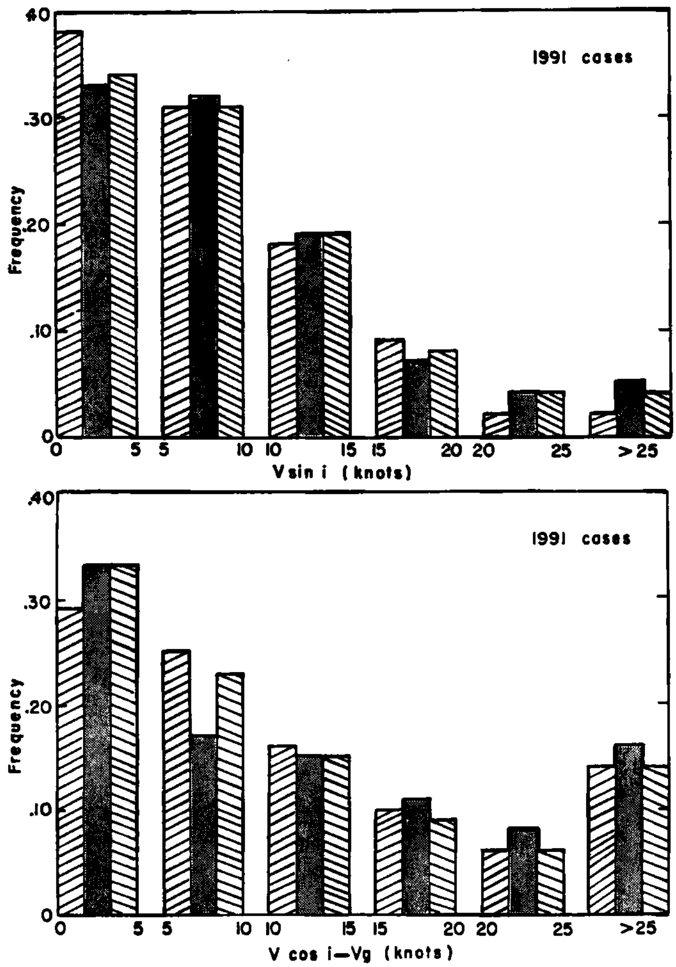


FIG. 6. Distribution of 12-hr-average cross-contour (V sin i) and along-contour (V cos i $-$ V_g) components of ageostrophic wind derived from 1957–59 transosonde flights from Japan. Columns indicate frequencies for fall, winter, and spring from left to right, as in Fig. 5.

derived from transosonde velocity changes in 12 hr. It is seen that for both ageostrophic components the mode is 0–5 knots while about 4% of the cross-contour ($V \sin i$) and 15% of the along-contour ($V \cos i - V_g$) components of the ageostrophic wind exceed 25 knots. In the mean, the cross-contour component of the ageostrophic wind is about two-thirds the magnitude of the along-contour component. From comparison of the average values of these ageostrophic components with the values of the wind speed shown in Fig. 5, it is estimated that, in the average, the ageostrophic wind is nearly one-fourth the magnitude of the wind itself. However, since the magnitude of the ratio of ageostrophic wind and wind varies somewhat depending upon the averaging interval utilized to obtain both parameters, the true magnitude of this ratio is somewhat uncertain. Information on the magnitude of the natural ageostrophic wind components as functions of wind speed and latitude is given in [89]. It also is shown therein that through the use of the geostrophic wind equation probably 50% of the time an error exceeding 29% will be introduced into derived results whereas through use of the gradient wind equation probably 50% of the time an error exceeding 11% will be introduced into derived results. These percentages vary considerably with wind speed and latitude.

In addition to computing the natural ageostrophic components ($V \sin i$, $V \cos i - V_g$), one may compute zonal and meridional ageostrophic wind components ($u_{ag} = u - u_g$, $v_{ag} = v - v_g$). Of particular interest is the product of these zonal-meridional ageostrophic parameters with each other ($u_{ag}v_{ag}$) and with geostrophic wind components ($u_g v_{ag}$, $u_{ag}v_g$) since such products are associated with the ageostrophic meridional transport of momentum. Since values of the zonal-meridional wind products obtained from transosonde data apparently agree well with similar products obtained from synoptic data, it may be hoped that the values of $u_{ag}v_{ag}$, $u_g v_{ag}$, $u_{ag}v_g$ obtained from transosonde data are also representative of values which would be obtained from synoptic data.

The zonal and meridional ageostrophic components were evaluated from the simplified equations of motion,

$$u_{ag} = -\frac{1}{f}\frac{dv}{dt} \tag{7.5}$$

and

$$v_{ag} = +\frac{1}{f}\frac{du}{dt} \tag{7.6}$$

where du/dt and dv/dt are 12-hr-average values of zonal and meridional acceleration obtained from the transosonde flights and f is the Coriolis parameter. Table III gives the mean seasonal values for the products of in-

TABLE III. Magnitude of Zonal-Meridional Wind and Geostrophic-Ageostrophic Wind Products (Knots²) Based on 1957–59 Transosonde Flights from Japan.

	$\overline{uv}$	$\overline{u_g v_{ag}}$	$\overline{u_{ag} v_g}$	$\overline{u_{ag} v_{ag}}$	Sum of geostrophic-ageostrophic products	Cases
Fall	284	2	29	−9	22	611
Winter	260	−11	3	−20	−28	743
Spring	393	41	32	−49	24	637
Mean	310	12	21	−26	7	1991

TABLE IV. Correlation Coefficients among Velocity and "Natural" Ageostrophic Velocity Components Based on 1957–59 Transosonde Flights from Japan.

Correlation between:	Fall	Winter	Spring	Mean
Zonal and meridional wind	0.258	0.135	0.288	0.211
Zonal wind and along-contour component of ageostrophic wind	0.236	0.055	0.085	0.113
Meridional wind and cross-contour flow	0.321	0.136	0.056	0.138
Cross-contour flow and along-contour component of ageostrophic wind	−0.040	0.003	−0.057	−0.025
Cases	611	743	637	1991

terest. The mean value of the zonal-meridional wind product for the two years of data is 310 knots², corresponding to a mean correlation between zonal and meridional wind of 0.21. Integrating around the hemisphere at latitude 35°N during January, Mintz [90] found an average (geostrophic) value of 285 knots². Especially noteworthy in Table III is the sum of the products involving ageostrophic terms, as shown in the next to the last column. These sums are almost 10% of the wind products themselves and suggest that in winter the ageostrophic products counteract the northward meridional flux of momentum associated with the geostrophic winds whereas in fall and spring they augment this flux. This result may be associated with the seasonal shift of the jet stream as suggested by the theoretical work of Lorenz [91].

Table IV shows the correlation coefficients among certain velocity and ageostrophic velocity parameters for the season and the mean of all flights. According to the "*Z* test" presented in Hoel [92, p. 89], with the number of cases here available, a correlation of 0.05 is significant at the 95% level in the mean and a correlation of 0.08 is significant at the 95% level for the seasonal data. As would be expected, the zonal and meridional wind components are significantly positively correlated in all three seasons with the correlation most pronounced in fall and spring. The zonal wind and along-contour component of the ageostrophic wind are positively correlated in all three seasons (fall and spring significant) as are the meridional wind and the cross-contour component of the ageostrophic wind (fall and winter significant). The correlation between the two ageostrophic wind components is negative in fall and spring and slightly positive in winter, but none of the correlations is significant. As along an idealized wave-shaped trajectory, the meridional wind component would be a maximum at the pre-trough inflection point and the along-contour component of the ageostrophic wind would be a maximum at the trajectory crest, the relative magnitudes of the above correlations yield estimates of where in the wave-shaped flow pattern the maximum zonal wind and maximum flow toward low pressure occur on the average. Based on such reasoning, Fig. 7 shows that during all three seasons the maximum zonal wind is located about halfway between pre-trough inflection point and trajectory crest while the maximum flow toward low pressure occurs near the pre-trough inflection point, with some doubt as to whether it occurs just upstream from or just downstream from this point. The association of the maximum zonal wind speed with southerly flow is not unexpected, because thereby zonal momentum is transported northward, but the fact that the zonal wind tends to be a maximum on the trajectory ridge rather than in the trajectory trough is somewhat surprising. The erroneous meteorological impressions on this latter score probably result from a subjective tendency to compare isobar or contour spacing (geostrophic wind) in ridge and trough rather than the

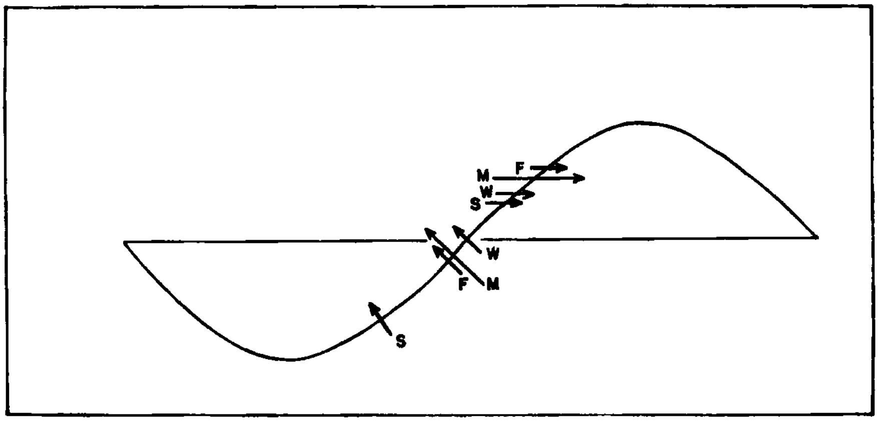


FIG. 7. Positions along schematic wave-shaped trajectory of maximum zonal wind (horizontal arrows) and maximum flow toward low pressure (arrows normal to trajectory) for fall(F), winter(W), spring(S), and the mean(M) of all 1957–59 transosonde flights from Japan.

wind speed itself. The manner in which these positions of maximum zonal wind and maximum flow toward low pressure vary for different periods of oscillations is presented in Section 7.4.

7.4. Periodicities in the Flow and Phase Lags between Parameters

Preliminary study of the Lagrangian periodicity in the flow involved the calculation of the (Lagrangian) autocorrelation coefficients of the zonal and meridional wind components based on the 1957–58 300-mb transosonde flights. All flights of one or more days' duration were used in this computation yielding 1136 products upon which to base the values of the autocorrelation coefficients for 6-hr time lags and 40 products upon which to base the values of the autocorrelation coefficients for 120-hr time lags. The denominator in the expression for the autocorrelation coefficient was assumed constant in the computations. Figure 8 shows that the autocorrelation coefficient of the zonal wind is well represented by the analytic expression

$$R_u(t) = e^{-(7t+1)/6} \tag{7.7}$$

where t is in days. The departures after four days of the transosonde-derived autocorrelations from this analytic expression probably have reality and represent the presence of a long-term sinuosity in the autocorrelation of the zonal wind due to strong winds over Japan being followed, after five or more days, by strong winds off the east coast of North America. The autocorrelation coefficient of the meridional wind is well represented by the analytic expression

$$R_v(t) = e^{-t} \cos \pi t \tag{7.8}$$

where t is in days. Thus, the autocorrelation of the meridional wind is basically a superposition of an exponential function and a sinusoidal term of 2-day period. This 2-day period is an expression of the average time it takes the transosondes at 300 mb to traverse the long waves in the westerlies. Substitution of these autocorrelation coefficients in Taylor's equations for dispersion yielded transosonde standard deviations of position after time of release almost identical to those obtained by direct calculation (Section 7.2.).

Calculations also show, on the basis of these data, that for time lags up to 18 hr

$$1 - R_u(t) \cong 0.75t^{1/2} \tag{7.9}$$

$$1 - R_v(t) \cong 1.80t \tag{7.10}$$

Thus, in the case of the meridional (transverse) wind, one minus the La-

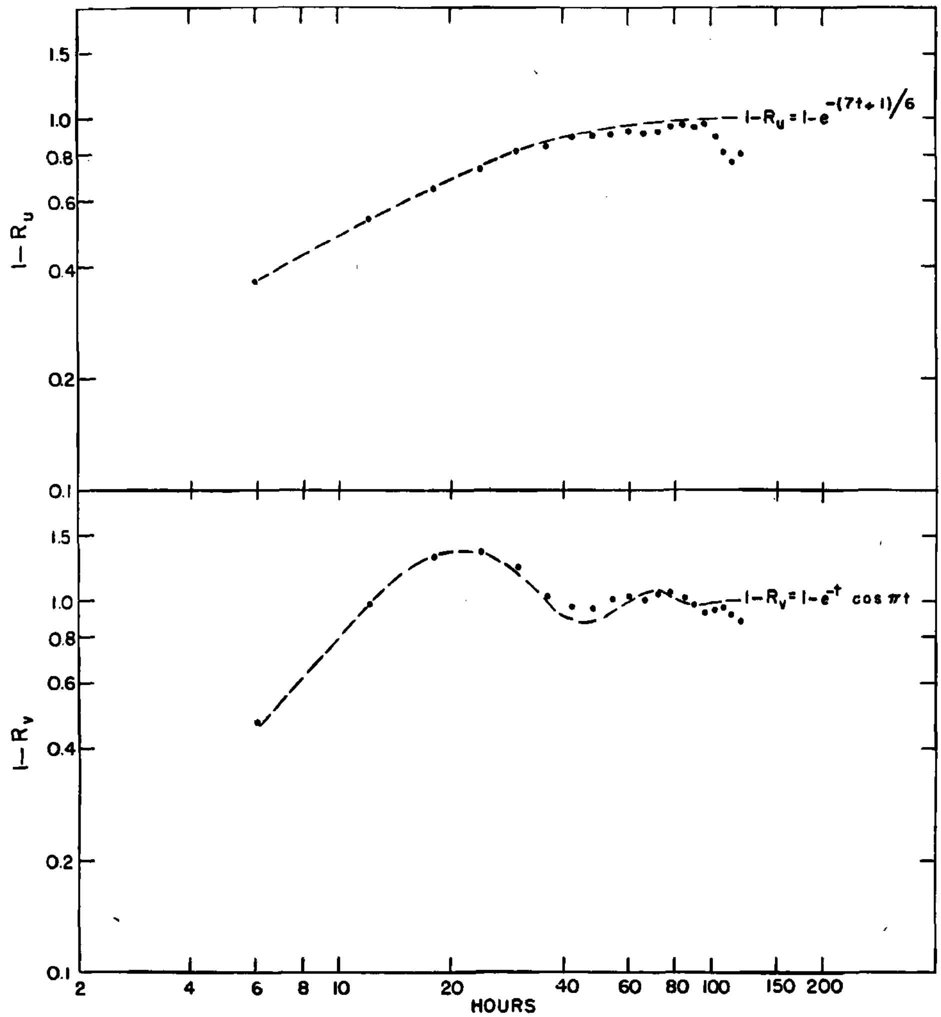


FIG. 8. Mean autocorrelation coefficients of zonal wind (top) and meridional wind (bottom) derived from 1957–58 transosonde flights from Japan. The dots represent the observed autocorrelation coefficients at 6-hr intervals while the dashed lines represent the fit of analytic expressions to these values (t in days). Note that the data are plotted on log-log diagrams and that the ordinate is expressed in terms of one minus the autocorrelation coefficient.

grangian autocorrelation coefficient is proportional to time for the large-scale synoptic circulations as well as for the small scale "turbulent" fluctuations investigated by Inoue [93] and others. Taking the ratio of equations (7.9) and (7.10), it is found that, for time lags up to 18 hr

$$\frac{1 - R_u(t)}{1 - R_v(t)} \cong 0.60 \tag{7.11}$$

This is a value rather close to the two-thirds value to be expected from Eulerian-type data [94].

As a second step in the study of the periodicity of the flow, for each 1957–58 transosonde flight of four or more days' duration, the contribution of oscillations of various frequency to the variance of the series was determined for the wind speed (V), the zonal and meridional velocity components (u, v), and the "natural" ageostrophic components ($V \sin i$, $V \cos i - V_g$) by applying the cosine transform to the serial products of the respective parameters [95] [96]. When this spectral technique is applied to Lagrangian-type data, the peaks in the spectra indicate the preferred periods or frequencies of oscillation [97].

Figure 9 gives the spectra for velocity and ageostrophic velocity parameters based on the 1957–58 transosonde flights of four or more days' duration. These values were obtained by averaging the variances per unit frequency interval obtained from individual flights. The spectra were extrapolated to periods of oscillation of 240 hr through the use of flights of five or more days' duration. In the case of the wind speed V and the zonal velocity component u the greatest variance is associated with oscillations of a period of 5 days or more, as also suggested by the autocorrelation coefficients of the zonal wind. In winter there is some indication of a secondary peak in the wind speed variance at a period of about 60 hr. It should be emphasized that even if oscillations of a 60-hr period in V and u were prominent in nature, they would be difficult to resolve by this technique owing to the large variances associated with the oscillations of period greater than 5 days. In the case of the meridional wind v, in the mean for all flights the variance is a maximum at a period of about 54 hr, as approximately indicated by the autocorrelation coefficient of the meridional wind. The seasonal diagrams of the variance of the meridional wind suggest that these long wave fluctuations are of slightly larger period in winter than in spring or fall, intimating that an increase in geographical wavelength during winter more than compensates for the increased wind speed during that time of year.

In the case of the cross-contour flow ($V \sin i$), the variance per unit frequency interval has, in the mean, only a very indistinct peak at a period of about 37 hr. In winter, however, there is a more pronounced peak at a period of 38 hr and a secondary peak at a period near 18 hr. Despite the absence of a pronounced spectral peak, it seems likely that $V \sin i$ has a considerably shorter period of oscillation than does the meridional wind with, on the average, about three regions of flow toward low pressure

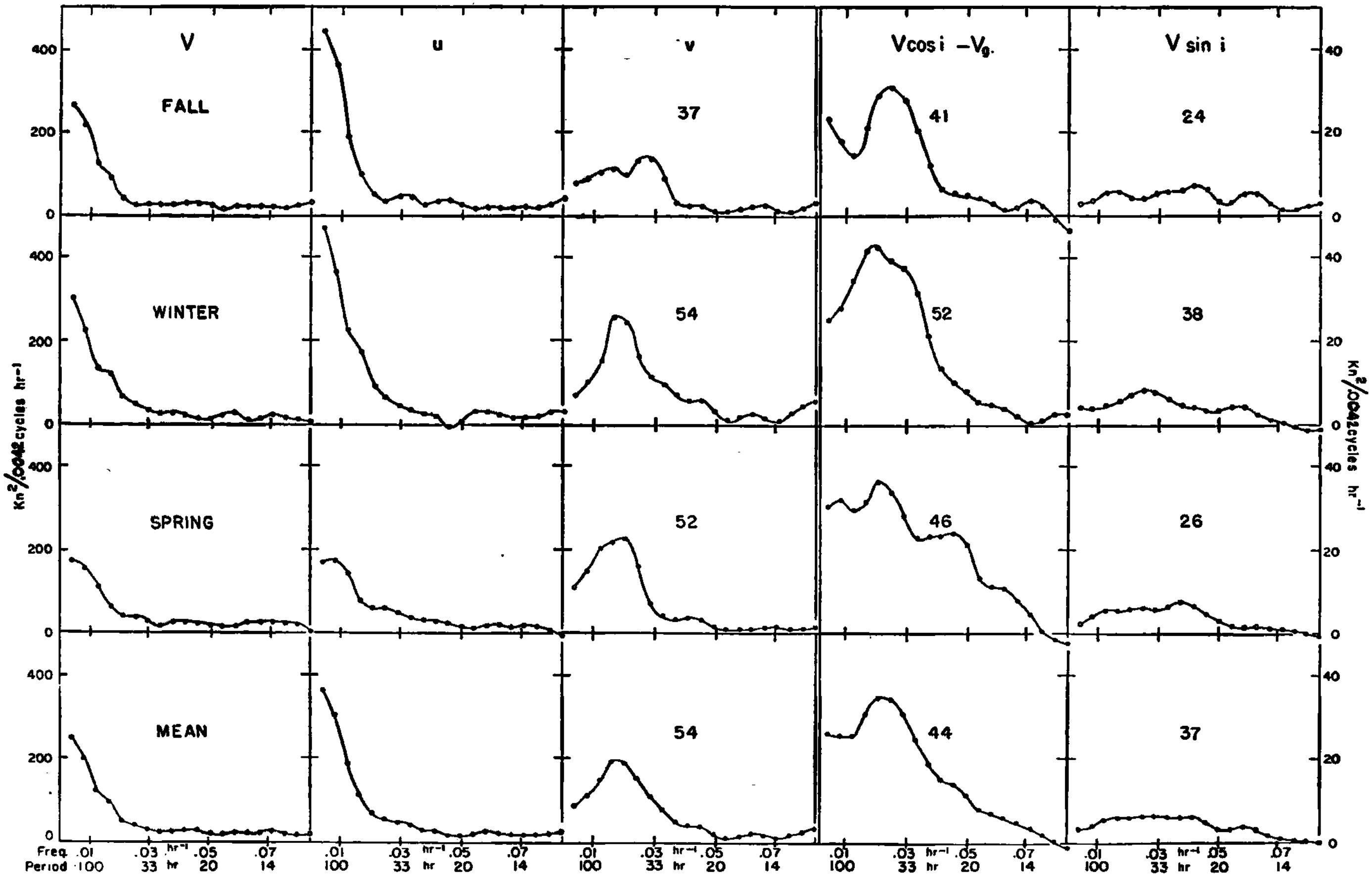

V
u
v
Vcos i −Vg.
V sin i
FALL
WINTER
SPRING
MEAN
37
41
24
54
52
38
52
46
26
54
44
37
Kn²/.004 cycles hr⁻¹
400
200
0
40
20
Freq. .01 .03 hr⁻¹ .05 .07
Period 100 33 hr 20 14

FIG. 9

included in two wavelengths of the long waves in the westerlies. In the case of the along-contour component of the ageostrophic wind ($V \cos i - V_g$), the variance per unit frequency interval is a maximum at a frequency little different from that found for the meridional wind, but shaded somewhat toward shorter periods of oscillation as if influenced by the periodicity in $V \sin i$. For both $V \sin i$ and $V \cos i - V_g$ the periods of maximum variance are somewhat longer in winter than in spring or fall, in agreement with the findings for the meridional wind.

Cospectra and quadrature spectra of the zonal-meridional and natural ageostrophic wind components were determined by applying the cosine and sine transform to, respectively, the sum and difference of forward and backward lagged serial products in the manner suggested in [98]. From the spectra, cospectra, and quadrature spectra, it is possible to obtain correlation coefficients between the zonal-meridional and natural ageostrophic wind components as a function of frequency of oscillation. Table V shows that the correlation between zonal and meridional wind and between the ageostrophic components varies in a systematic manner with respect to the period of oscillation. Thus, the zonal and meridional winds are negatively correlated for oscillations of period from 30 to 16 hr and positively correlated for other periods whereas the natural ageostrophic wind components are positively correlated for oscillations of a period of 40–20 hr and negatively correlated for other periods. The tendency for the zonal wind to be a maximum in the northerlies (difluent troughs) for fluctuations of a period of 30–16 hr while for longer period fluctuations the maximum zonal wind is in the southerlies (confluent troughs) indicates, in general, that the difluent structures have shorter geographical wavelengths than the confluent structures or that the average wind speed is greater when difluent structures exist than when confluent structures exist. Since difluent structure is often associated with an increase in wave amplitude, we may say that an increase in wave amplitude tends to take place when the wavelength is shorter than normal or the wind speeds are above normal or both conditions obtain.

By finding the arctangent of the ratio of quadrature spectra and cospectra, the phase lags between the oscillations in the zonal and meridional

---

FIG. 9. Variance per unit frequency interval as a function of frequency for the wind speed (V), the zonal (u) and meridional (v) velocity components, and the natural ageostrophic components ($V \sin i$, $V \cos i - V_g$) for fall, winter, spring, and the mean of all 1957–58 transosonde flights from Japan of four or more days' duration. The units indicated on the left-hand ordinate refer to the speed and velocity component diagrams while those on the right-hand ordinate refer to the ageostrophic component diagrams. The numbers in some diagrams give the period of oscillation (hours) at which the variance per unit frequency interval is a maximum (the preferred periodicity).

TABLE V. Correlation Coefficients between Zonal-Meridional and Natural Ageostrophic Wind Components as a Function of their Period of Fluctuation based on 1957-58 Transosonde Flights from Japan of Four or more Days' Duration.

Period of fluctuation (hr)	Correlation between u and v	Correlation between $V \sin i$ and $V \cos i - V_g$
240	0.12	−0.19
120	0.13	−0.26
80	0.24	−0.29
60	0.36	−0.16
48	0.37	−0.04
40	0.16	0.03
34	0.03	0.11
30	−0.11	0.05
27	−0.20	−0.02
24	−0.16	0.10
22	−0.17	0.17
20	−0.31	0.07
18	−0.38	−0.12
17	−0.18	−0.15
16	0.00	−0.07
15	0.18	−0.10
14	0.35	−0.16
13	0.28	−0.12
12	0.45	−0.06

wind and the oscillations in the natural ageostrophic components can be computed for oscillations of various frequency. Further, if it is assumed that the maximum value of the meridional wind along the trajectory occurs at the pre-trough inflection point and the maximum value of $V \cos i - V_g$ occurs at the crest of the trajectory, the percentage frequency with which the positions of maximum zonal wind and maximum flow toward low pressure occur within 1/12 wavelength sectors along wave-shaped trajectories can be determined. Figure 10 shows, for example, that for oscillations of a 96-hr period, on 9.6% of the flights the maximum zonal wind occurred in the 1/12 wavelength sector downstream from the pre-trough inflection point (top diagram) while 8.1% of the flights had the maximum flow toward low pressure in the same sector (bottom diagram). From an inspection of these diagrams it is seen that for long-period oscillations there is a pronounced tendency for the maximum zonal wind to be located between the pre-trough inflection point (ordinate of zero) and the trajectory crest (ordinate of 90 deg) and the maximum flow toward low pressure to be located between the trough-line (ordinate of −90 deg) and the pre-trough inflection point. For shorter period oscillations, the differences in per-

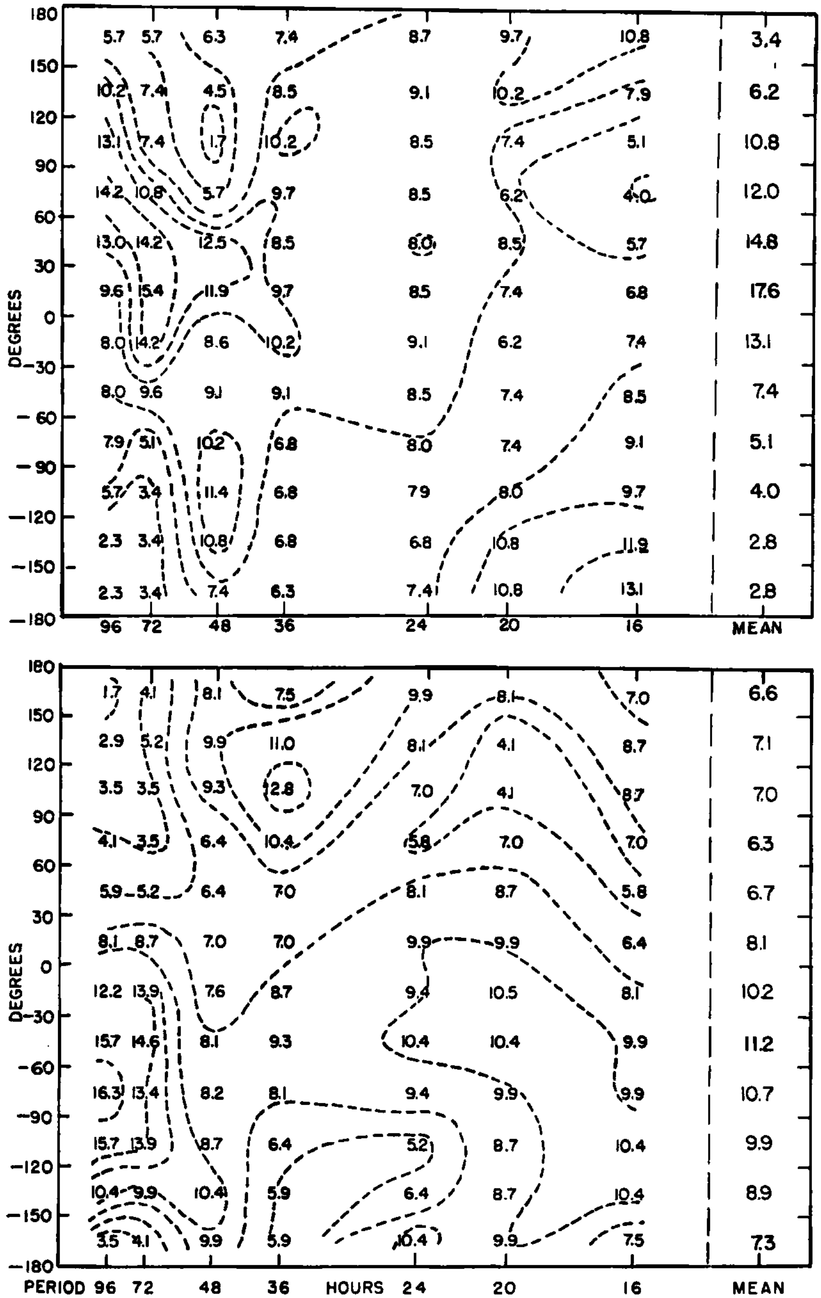


Fig. 10. Percentage of 1957–58 transosonde flights from Japan of four or more days' duration for which the maximum zonal wind (top) and maximum flow toward low pressure (bottom) occurred within ½ wavelength sectors along wave-shaped trajectories with oscillations of different period and oscillations with all periods from 12–240 hr (to the right of dashed vertical line). Along the ordinate 0° refers to the pre-trough inflection point, 90° to the trajectory crest, −90° to the trajectory trough line.

centages are not as pronounced. It is interesting to note that for oscillations of all periods from 240 to 12 hr (right-hand column in fig. 10), the maximum flow toward low pressure occurs with greatest frequency about halfway between trough-line and pre-trough inflection point, but that, in addition, there is a secondary peak in the frequency of occurrence of flow toward low pressure somewhat downstream from the trajectory crest. This secondary peak is largely the result of maximum flow toward low pressure occurring with great frequency slightly downstream from the trajectory crest for oscillations of a 36-hr period. This region of maximum frequency may be associated with the Bjerknes concept concerning the inability of air parcels flowing around ridges to follow contour channels with small radii of curvature [99]. As a further test of this concept, the total variance of $V \cos i - V_g$ (proportional to trajectory curvature) for each flight was plotted as a function of the position along the wave-shaped trajectory of the maximum flow toward low pressure for each flight. It was found that, when the anticyclonic (or cyclonic) curvature of the trajectory is greater than normal, the region of maximum flow toward low pressure is more likely than normal to be associated with the trajectory crest, a result certainly not in contradiction with the Bjerknes concept.

7.5. Some Evidence for a Mean Meridional Flow

The existence of a mean meridional cell in the tropics (Hadley cell) has been confirmed by wind measurements and estimates of the angular momentum balance [100]. However, the sense, or even existence, of a meridional cell in temperate latitudes (Ferrel cell) is still in dispute. It is shown below how the changes in height of a constant pressure surface at successive transosonde positions permits the estimation of the mean meridional flow over a limited portion of the hemisphere.

The change in height of a constant pressure surface along a transosonde trajectory can be estimated according to the equation

$$\Delta Z = \int_{t_1}^{t_2} \frac{\partial Z}{\partial t}\, dt - \frac{V_2^{\,2} - V_1^{\,2}}{2g} - \frac{1}{g}\int_{t_1}^{t_2} Vw \frac{\partial V}{\partial z}\, dt \tag{7.12}$$

where ΔZ is the height change following the transosonde, $\partial Z/\partial t$ is the local height change along the trajectory, V_2 and V_1 are transosonde speeds at two different times t_2 and t_1, g is the acceleration of gravity, w is vertical air motion, and $\partial V/\partial z$ is vertical wind shear. The term involving the local height change may be large on individual flights, but as it is not likely to have an average much different from zero for flights released almost randomly from Japan, it will be set equal to zero in this discussion. Moreover, while the term involving the vertical velocity may be important along isolated trajectory segments, it was found to be of negligible magni-

TABLE VI. Computed and Observed Trans-Pacific Height Changes (in feet) of Constant Pressure Surfaces along Transosonde Trajectories and Derived Mean Meridional Ageostrophic Flow.

	Fall	Winter	Spring	Mean
Height at 120°W minus height at Iwakuni derived from change in transosonde speed	170	450	180	290
Height at 120°W minus height at Iwakuni obtained from 300-mb and 250-mb NAWAC maps	−390	220	−100	−50
Height discrepancy in radiosondes (Japanese minus United States)	130	130	130	130
Observed height difference minus computed height difference	−430	−100	−150	−210
Derived mean meridional flow (knots)	3.1	0.9	1.2	1.6
Number of evaluations	27	40	27	94

tude in the mean for many flights due to the very small value for the trans-Pacific vertical velocity (0.1 cm/sec) determined from the transosonde flights by the adiabatic method.

The top row in Table VI gives the change in height of the constant pressure surface between Iwakuni and the place where the transosonde transits 120°W longitude which would be expected due to the change in speed along the individual transosonde trajectories [middle term in equation (7.12)]. Since the transosondes slow down while crossing the Pacific, one would expect the height of the constant pressure surface at the transosonde position at 120°W to be higher than the height over Iwakuni at the time of transosonde release, as shown in the top row of Table VI. The second row in Table VI gives the average observed height difference between Iwakuni and the place where the transosonde transits 120°W as obtained from National Weather Analysis Center (NAWAC) maps. It is most surprising that in the mean the height at Iwakuni is greater than that at 120°W in complete contradiction with what would be anticipated from the change in transosonde speed. In the mean for all flights the discrepancy between what is observed and what would be computed from the middle term in equation (7.12) is 340 ft. Part of this discrepancy is due to the incompatibility of United States and Japanese radiosondes, as found at Payerne and reported by Harmantas [101]. According to the Payerne tests, the Japanese 300- and 250-mb heights should average about 130 ft more than those reported by United States radiosondes. However, since the existence of a mean floating level for the transosondes slightly above the constant pressure surface in question was found to have no influence on the above calculations, unless the incompatibility of Japanese and United States radiosondes is greater than determined at Payerne, the above data suggest the presence of a mean northward ageostrophic flow at 300 and 250 mb above the Pacific Ocean. The bottom row of Table VI gives the magnitude of this mean meridional ageostrophic flow based on the height-change discrepancies and the average meridional distance between 300 mb contours over the Pacific Ocean as evaluated from the transosonde flights. Over the Pacific Ocean the mean meridional flow during all seasons is estimated to be from the south with a speed of 1.6 knots, with the flow being somewhat stronger in fall than in winter or spring. Such a northward drift is counter to the concept of the Ferrel cell but would be in agreement in sense, if not in magnitude, with the concept of a one-cell meridional circulation between equator and pole. It seems very likely, however, that if a northward ageostrophic flow of this magnitude actually does exist over the Pacific Ocean, that over much of the remainder of the hemisphere there would exist a southward ageostrophic flow which would nearly cancel it. The mean hemispheric meridional cell may therefore be the small difference between two cells of considerable strength but with different senses; the one cell extending at least across the Pacific Ocean and the other cell extending around the remainder of the hemisphere.

7.6. Transosonde Density as a Function of Time since Nonsimultaneous Release

For the purpose of studying the transosonde density a given number of days after release for transosondes released from the same point at different times, the number of transosonde pairs at various distances apart was determined by season for the 1957–59 transosonde flights from Japan. The approach resembles Richardson's distance-neighbor technique [102] except that, instead of determining the temporal variation of the distance separating pairs of air parcels in the air at the same time, here we determine the variation with time since release of the distance separating pairs of transosondes in the air at different times. Certainly results obtained utilizing data of the latter type need not be representative of the results which would be obtained using data of the former type.

Figure 11 shows the percentage of possible transosonde pairs which were located (within 5-deg latitude bands) a given number of degrees latitude apart a certain number of days after release from Japan, as obtained by averaging the results derived for the three seasons of fall, winter, and spring. This diagram shows, for example, that one day after transosonde release 23% of the possible transosonde pairs are 10–15° latitude apart, while 4 days after transosonde release 8% of the possible pairs are separated by this distance. With an average error in frequency of 0.03 in

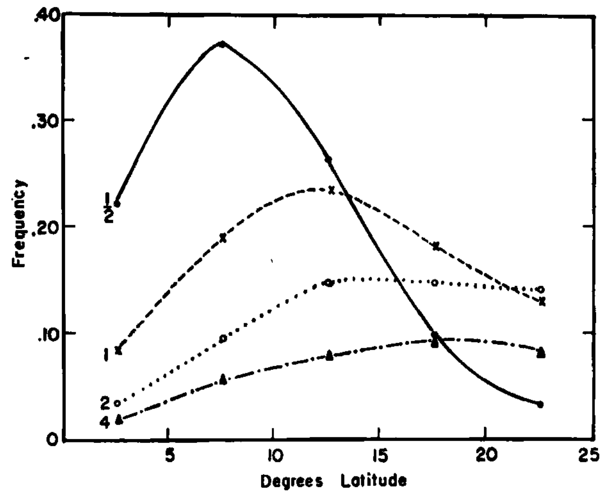


FIG. 11. Distribution of distances between all possible transosonde pairs as a function of time in days after their (nonsimultaneous) release. The values plotted are the means of values obtained separately for fall, winter, and spring as derived from the 1957–59 transosonde flights from Japan.

the 20 plotted positions given in Fig. 11, the percentage of possible transosonde pairs P, separated by a given number of degrees latitude R, a given number of days after release from Japan t, is given analytically by

$$P = 40\, tR^{-2.2} \exp\,(-25t/R^{1.2})\, dR \tag{7.13}$$

where dR is the latitude interval under consideration (5° in Fig. 11).

Differentiating equation (7.13) with respect to R and setting the resultant equal to zero it is found that the most frequent distance (mode) between balloon pairs (R_m) is given by

$$R_m = 8.8\, t^{0.833} \tag{7.14}$$

where t is in days and R_m is in degrees latitude. Thus, seven days after transosonde release it is estimated that the most frequent separation of pairs of transosondes would be 45° of latitude with, from equation (7.13), 11 % of the possible transosonde pairs being separated by 40–50° latitude. It is interesting that, with the replacement of the time coefficient in equation (7.14) by the value 5.8, this equation almost exactly satisfies the average distance between 700-mb geostrophic-trajectory pairs initiated from a point at 2-day intervals (similar to the transosondes) during the period April–June 1948, as reported in a paper by Durst *et al.* [103]. Thus there is independent evidence in support of the power of the time given in equation (7.14). Theoretical considerations presented by Durst *et al.* indicate that, if it is assumed that the autocorrelation coefficient is exponential in form, the average distance between trajectory pairs should increase nearly linearly with time initially and increase according to the square root of time two or more days after initiation. If anything, the experimental results suggest a quicker trajectory separation with time than does the theory based upon the assumption of an exponential form for the autocorrelation coefficient.

Integrating equation (7.13) with respect to R it is found that the percentage of possible transosonde pairs separated by less than the distance $R(P_R)$ is given by

$$P_R = \tfrac{4}{3} \exp\left(-25t/R^{1.2}\right) \tag{7.15}$$

Thus, seven days after transosonde release it would be estimated that 27% of the transosonde pairs would be separated by less than 50° latitude or, conversely, that 73% of the transosonde pairs would be separated by more than this amount.

The above equations are not exact since the average distance between the elements of a transosonde pair increases with the time interval between the release of the transosondes making up the pair. However, if the transosondes are released more than six days apart there is little increase in the average separation as a function of time between release. As the results presented in Fig. 11 are based upon seasonal data, the average interval between the release of the components of a transosonde pair is 45 days. If the transosondes had been released in quick succession so that the interval between transosonde release was, on the average, only two days say, then a larger percentage of the transosonde pairs in Fig. 11 would be separated by smaller distances.

The above analysis has application to problems of areas habitually void of observations. For example, suppose that a balloon were at a certain geographical location two days after release and it were desired to know the probability of getting another balloon, launched from the same site, within 10° latitude of this location two days after release. According to Fig. 11 there would be only a 13% chance of getting a balloon within the desired area unless the additional balloons were released soon after positioning of the first balloon, in which case the probability might jump to 15% or so. Thus, if seven or eight additional balloons were released, on the average it would be expected that one of these balloons would be within the required area two days after release. For further extension in time and space reference may be made to equation (7.15). For example, this equation would suggest that in order to get a second balloon within 50° latitude of an earlier balloon seven days after release, on the average four additional balloons would have to be released.

7.7. Use of Transosonde Data in Daily Map Analysis and Forecasting

One of the purposes of the transosonde system is to furnish upper-level wind data over the vast Pacific and Atlantic Oceans where such data are nonexistent or can be obtained only at great expense through the use of weather ships. The desirability of obtaining such data has been indicated by the Division of Meteorology of the Pacific Science Congress at the New Zealand meeting in 1949 [104] and by a recent resolution of the World Meteorological Organization [105].

The usefulness of transosonde data for meteorological analysis and forecasting is forcefully illustrated by flight 27, launched from Japan in February 1956. Figure 12 shows a portion of the 300-mb trajectory of this flight over the eastern North Pacific. The transosonde trajectory clearly delineates a trough off the west coast of the United States. The NAWAC analysis, drawn without knowledge of the transosonde trajectory, gives very little indication of such a trough. In fact, on the map for February 4, 1500Z, the geostrophic wind derived from the contour analysis indicates a

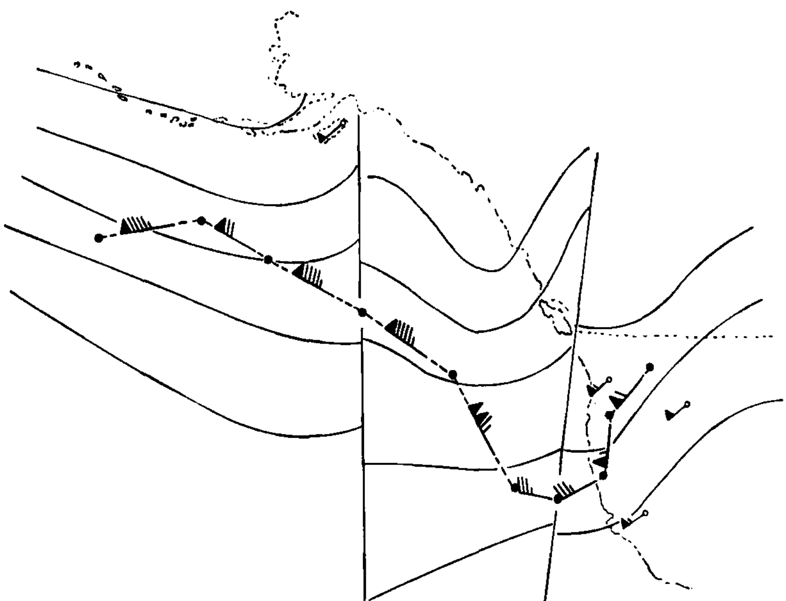


FIG. 12. Portion of 300-mb trajectory of transosonde flight 27 in February 1956. Transosonde positions (dots) and transosonde-derived winds (conventional wind barbs) are indicated at 4-hr intervals along the trajectory. Superimposed on the trajectory are segments of 12-hr NAWAC 300-mb maps with contours drawn at 400-ft intervals.

50-knot wind from the west at a point where the transosonde-derived wind indicates a 115-knot wind from the north-northwest. This represents a vector error in the wind of 100 knots only a few hundred miles off the west coast of the United States. On February 5, 1956, very heavy rains were reported along the northwest coast of the United States, rains which were poorly predicted owing to the underestimate of the trough intensity on the 1500Z map of February 4. It is clear that if the transosonde trajectory had been available to the map analyst, a considerable improvement in analysis and forecasting would have resulted. A statistical comparison of transosonde-derived winds and geostrophic winds obtained from maps analyzed without benefit of transosonde data is given by the author [106].

The operational transosonde flights from Japan were positioned at 2-hr intervals by means of U.S. Navy and Federal Communications Commission (FCC) radio direction finding stations. The positions obtained, and the winds derived therefrom, were then placed on the meteorological teletypewriter circuits in Pearl Harbor, Hawaii, and Norfolk, Virginia for transmission to analysis centers where the data could be used by map analysts and forecasters. Figure 13 shows the 6-hr average transosonde-derived winds plotted on NAWAC 250-mb maps from November 1958 to March 1959. In the five-month period a total of 1058 transosonde-derived winds were plotted, yielding wind data over vast reaches of the Pacific Ocean where such data had previously been unavailable. Since the tracking, communications, and plotting procedures were only operating at 50% efficiency during this time, twice as many winds could have been derived from the flights.

In order to indicate the usefulness of transosonde data for the analysis of individual synoptic maps, let us take a case where three transosondes were aloft over the Pacific Ocean at the same time. Figure 14 represents a copy of the NAWAC 250-mb map for 0000Z February 28, 1959. Transosonde flights 216, 217, and 218 were released from Iwakuni, Japan within 24 hr of each other, and in Fig. 14 they are shown spread out across the Pacific Ocean in a condition of nearly zonal flow. The transosonde positions are represented by dels, with the flight number plotted to the upper left of the del, the message number to the lower left, the pressure at which the transosonde was flying to the upper right, and the time of the position to the lower right. The transosonde-derived winds are plotted at the appropriate latitudes and longitudes with a message number attached. In general, smoothed transosonde positions 6 hr later than map time and transosonde-derived winds 3 hr later than map time were available to plot on the 250-mb NAWAC map, as indicated in Fig. 14. Thus, the transosonde data were reasonably well centered (temporally) on the NAWAC 250-mb map.

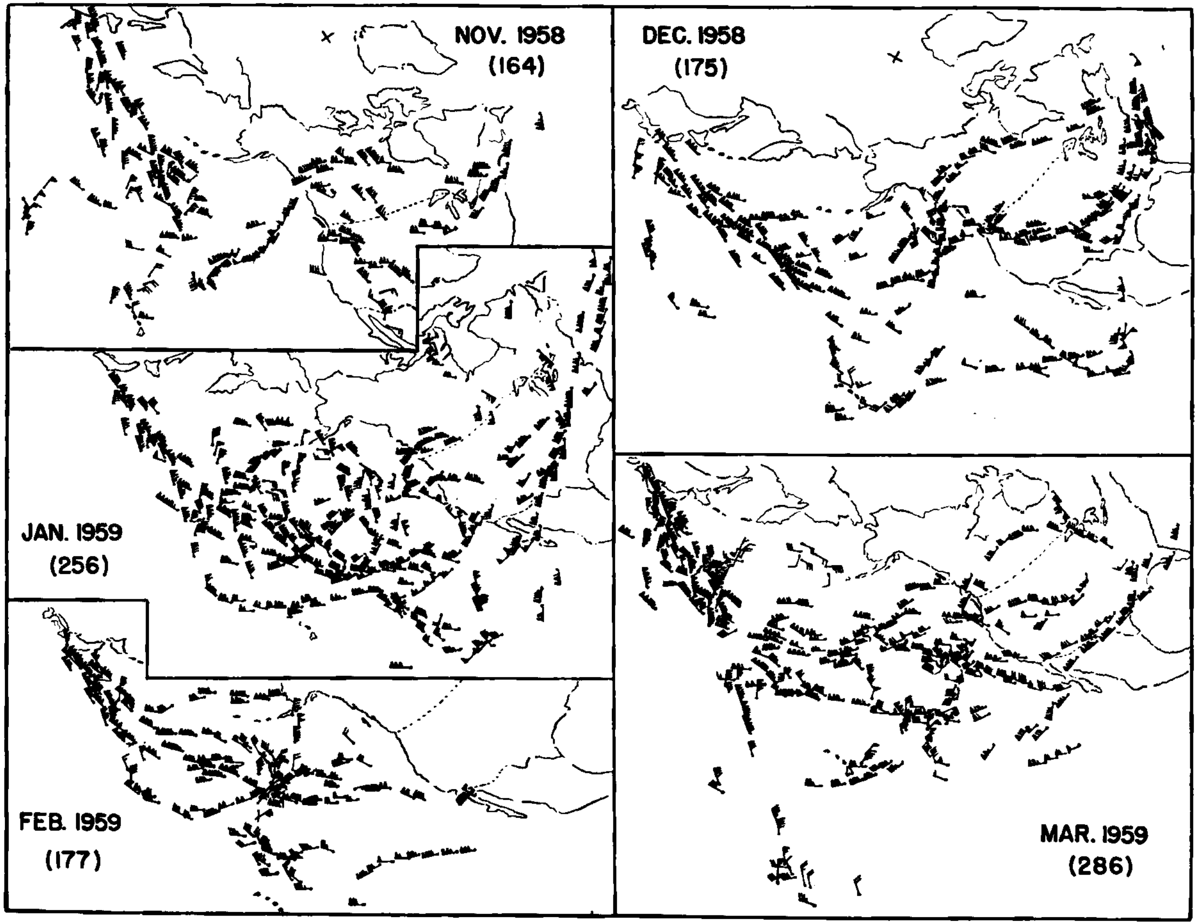


FIG. 13. Transosonde-derived winds plotted on NAWAC 250-mb maps from November 1958 to March 1959. The numbers in parentheses give the total plottings for each month.

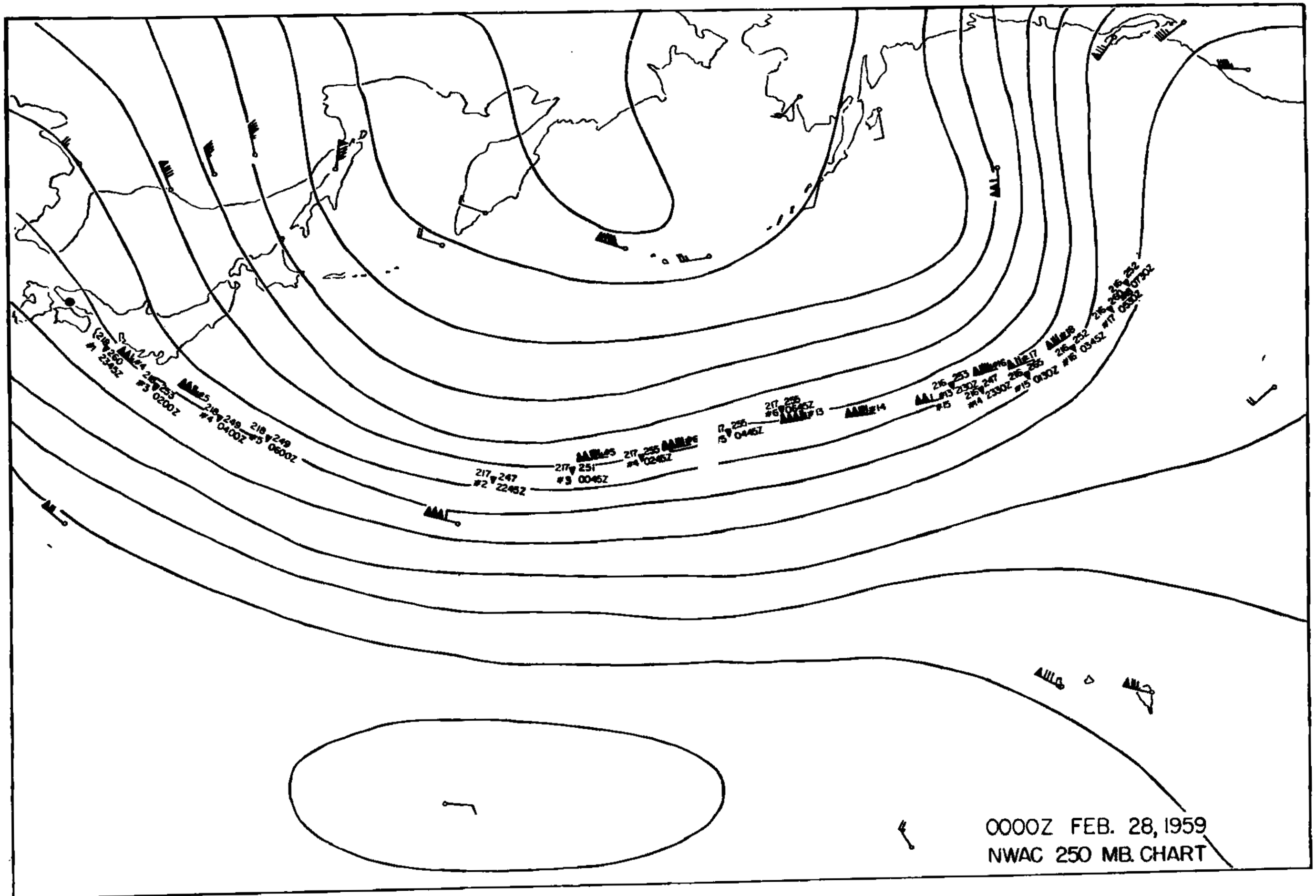


FIG. 14

In general, the analyst has appeared to use the transosonde data to good advantage in making this 250-mb analysis. Note that on flight 216, the transosonde-derived wind speed decreases from 175 knots to 70 knots in 8 hr. Such a deceleration would be associated [equation (4.2)] with an angle of indraft of -18 degrees (flow toward high pressure) so that, whether by design or accident, the analyst has properly suggested that the transosonde is moving toward high pressure. What is not explained from the contour pattern, as drawn, is why the balloon should be experiencing such a strong deceleration to begin with since the pressure gradient is not shown as weakening downstream. A more complete discussion of the usefulness of transosonde data for analysis and forecasting, including a cost estimate for the data so obtained, is given in [107]. An excellent discussion of the usefulness of one particular transosonde flight for the delineation of the trans-Pacific flow has been presented by Mastenbrook [108]. In this report it is shown how flight 202 provided a more or less zonal sweep of the Pacific, delineating in a precise manner two major troughs, indicating more exactly the point of division of northern and southern circulation branches, and in general "firming-up" the Pacific analysis throughout. Arakawa [109] has also emphasized the usefulness of transosonde data to the practicing meteorologist through a discussion of a case where the 300-mb flow indicated by a transosonde presaged the unusual movement of a surface cyclone.

For the planning of any future transosonde operation it is desirable to know what percentage of the time a transosonde released from a certain point would pass over a given geographical region and provide meteorological data there. Solot and Darling [110] have shown how such estimations can be obtained from statistical manipulations of Eulerian wind data. Based on the 1957–59 transosonde flights at 300 and 250 mb, Fig. 15 shows the percentage of transosondes released from Iwakuni which transited 5-deg latitude-longitude "boxes" within a certain number of days following release. This diagram shows, for example, that one-tenth of the transosondes released from Iwakuni would, within 3 days following release, furnish meteorological information within the 5-deg latitude-longitude box embracing most of Wyoming. It may be noted that the highest percentages are found in just that latitude belt of the Pacific Ocean where conventional upper-air data are almost totally lacking. It thus is obvious that transosonde releases from Iwakuni, Japan would provide the means for obtaining

FIG. 14. Transosonde flights 216, 217, and 218 plotted on the NAWAC 250-mb map for 0000Z, February 28, 1959. The dels represent transosonde positions with the flight number plotted to the upper left of the del, the message number to the lower left, the pressure altitude of the transosonde to the upper right, and the time of the position to the lower right of the del. Transosonde-derived winds are plotted with message numbers attached. The NAWAC contours are drawn at 400-ft intervals.

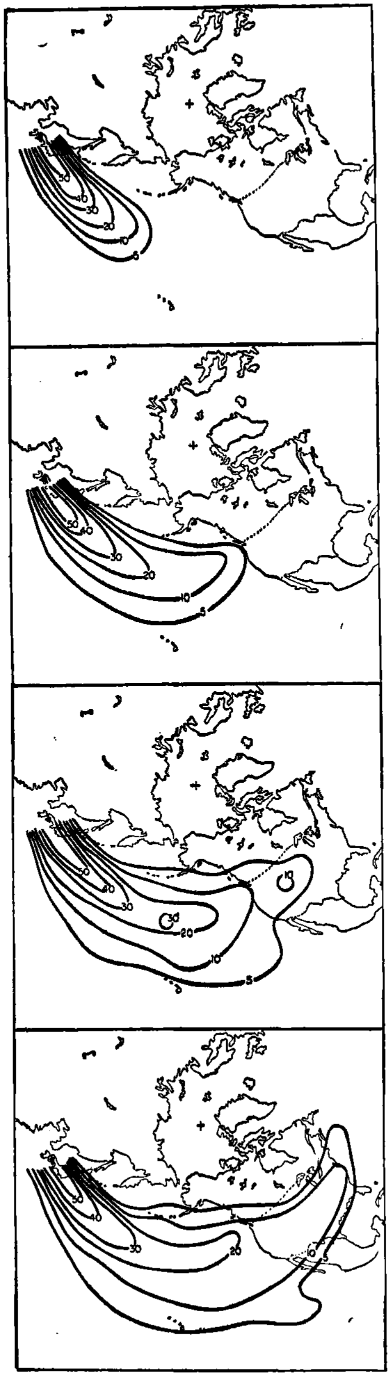


Fig. 15. Percentage of 1957–59 transosonde flights from Japan passing through 5-degree latitude-longitude boxes within (from top to bottom) one, two, three, and four days following release.

much-needed upper-wind data and other meteorological data over the Pacific Ocean.

8. Singularities and Periodicities Along Transosonde Trajectories

8.1. Introduction

A study of individual transosonde flights reveals the presence of some unusual trajectory patterns in the atmosphere, the existence of which could only be hinted at by means of customary vertical-sounding data. It is the purpose of this section to illustrate and discuss some of these trajectories with special emphasis on trajectory singular points or cusps, inertial oscillations, and abnormal or anomalous flow.

Inertial oscillations are the natural oscillations of an air parcel not in geostrophic (or gradient) equilibrium [111]. For example, if it is assumed that the geostrophic wind following an air parcel is constant ($d\mathbf{V}_g/dt = 0$) then from equation (4.1) we may write

$$\frac{d\mathbf{V}_a}{dt} = f\,\mathbf{V}_a \times \mathbf{K} \tag{8.1}$$

where $\mathbf{V}_a = \mathbf{V} - \mathbf{V}_g$ is the ageostrophic wind. It can be seen from this equation that, if an ageostrophic wind exists at some place and time, then under these special conditions the terminal point of the ageostrophic wind vector (and the terminal point of the wind velocity associated with it) will trace out a circle with a period of $2\pi/f$, or about 18 hr at latitude 40°. At lower latitudes the period of oscillation will be greater while at higher latitudes it will be smaller. This is the period of a so-called "pure inertial oscillation." However, if the air parcel moves through a region of cyclonic geostrophic vorticity the inertial period is reduced while if it moves through a region of anticyclonic geostrophic vorticity the inertial period is increased [64, p. 46]. If the anticyclonic geostrophic vorticity exceeds the Coriolis parameter, the inertial period becomes infinite, i.e., the air parcel will no longer oscillate about the geostrophic trajectory but rather will move continuously toward high or low pressure [112]. The period of inertial oscillation of a CLB also will be influenced by the vertical motion of air parcels in a region of vertical geostrophic wind shear (vertical advection of velocity term) in the sense that the inertial period exhibited by the CLB is increased if descending air motion is associated with CLB movement toward high pressure in a region where the geostrophic wind increases with height.

"Abnormal" flow is most simply defined as flow which possesses anticyclonic angular velocity in space; that is, the anticyclonic angular velocity of the air parcel in the atmosphere is larger than the cyclonic angular velocity of the earth about the local vertical [113]. Substitution in the normal equation of motion will show that abnormal flow is characterized by $V > 2V_g \cos i$. Thus, abnormal flow is fast flow. Gustafson [114] has identified regions of abnormal flow in the atmosphere by use of the above inequality. The transosonde trajectories furnish a more direct way of sensing regions of abnormal flow since all that is required is knowledge of the angular velocity along the trajectory.

8.2. Trajectory Singular Points or "Cusps"

Figure 16 shows the 300-mb trajectory of transosonde flight 36, launched from Oppama, Japan on February 14, 1956. Of particular interest on this flight are the large wind decelerations along the trajectory to the west of Baja California (145 knots in 12 hr) and in mid-Atlantic (180 knots in 8 hr). In both cases these large decelerations culminate in cusp-shaped trajectories and in both cases there is considerable evidence, both from the 300-mb NAWAC contour analysis and from rawin data, that during the period of deceleration the balloon is traversing a region of anticyclonic wind shear of sufficient magnitude to satisfy the criterion for inertial or dynamic instability (anticyclonic wind shear greater than the Coriolis parameter). Interestingly enough, Gustafson [115], through the use of analytic flow models and an analog computer, obtained similar cusped trajectories in regions of large anticyclonic wind shear downstream from the trajectory crest. Before flight 36 passed through the pronounced ridge over the Gulf of Alaska, there was a rather obvious 12-hr periodicity in the transosonde speed, but subsequently the only obvious periodicity in transosonde speed is one of 36 hr associated with the time interval between cusp points. Since it is believed that inertial instability occurs along this flight, the apparent lack of inertial oscillations is not surprising. A spectral analysis of the velocity variability along this flight is given in [116].

Figure 17 shows the 300-mb trajectory of transosonde flight 993, launched from Minneapolis, Minnesota on May 3, 1953. Of interest on this flight is the degree to which the changes in speed of the transosonde agree with flow toward high or low pressure as estimated from the independently analyzed contour field. In the northeast portion of the trajectory there is a trajectory cusp subsequent to a deceleration of 100 knots in 4 hr. In this case there is some evidence that before tracing out the cusp the transosonde passes from the cyclonic to the anticyclonic side of the jet stream. While doing so, the balloon transits four contours drawn at 100-ft intervals. Cusps in the trajectory also appear earlier in the flight to the north and west of Colorado. The similarity between the cusps in this actual trajectory and the trajectory cusps obtained theoretically by Forsythe [117] for the case of circular isobars is striking indeed, and tends to substantiate the

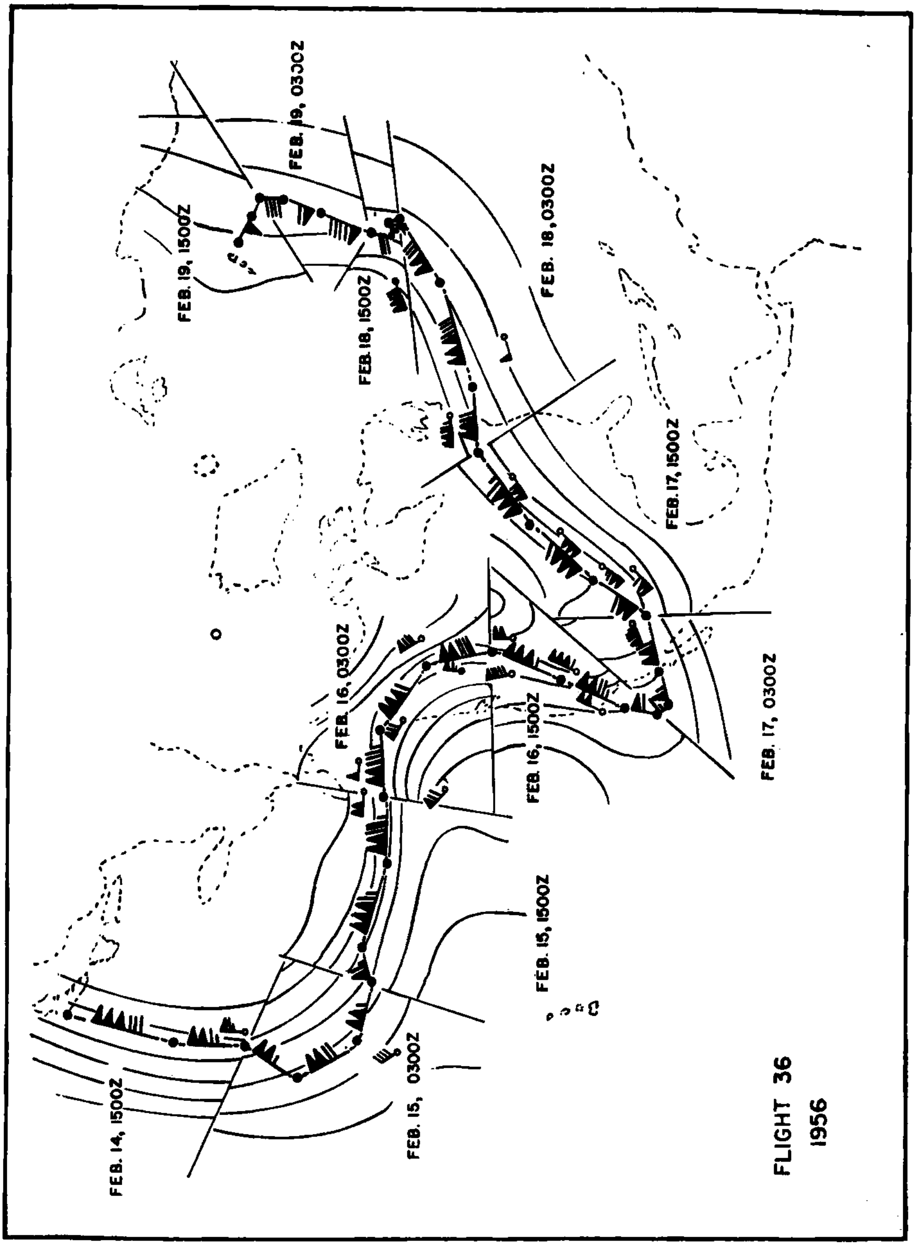


FIG. 16. Trajectory of 300-mb transosonde flight 36 in February 1956. Transosonde positions and transosonde-derived winds are indicated at 4-hr intervals along the trajectory. Superimposed on the trajectory are segments of 12-hr NAWAC 300-mb maps with contours drawn at 400-ft intervals.

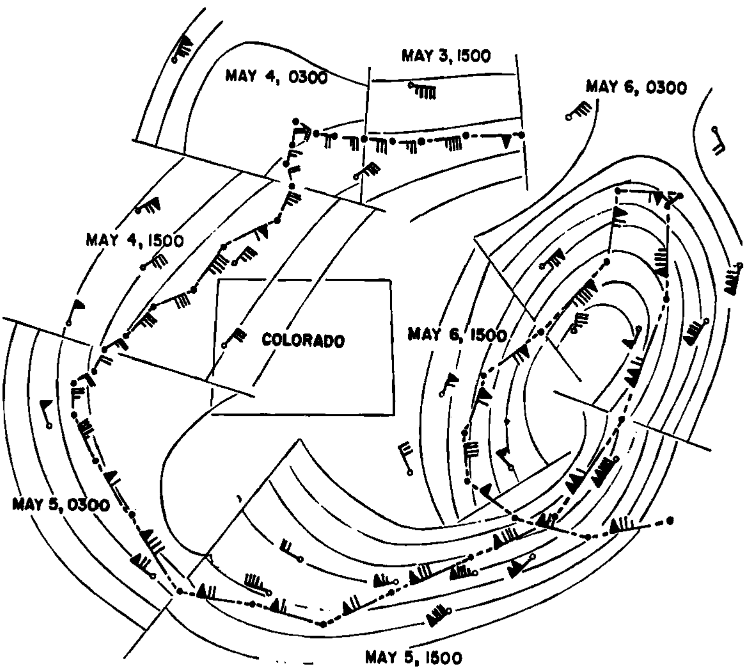


FIG. 17. Trajectory of 300-mb transosonde flight 993 in May 1953. Transosonde positions and transosonde-derived winds are indicated at 2-hr intervals along the trajectory. Superimposed on the trajectory are segments of 12-hr 300-mb maps (analyzed by a person who had not seen the trajectory) with contours drawn at 100-ft intervals.

hypothesis that the cusps along the trajectory of flight 993 are associated with inertial oscillations, as discussed in Section 8.3.

8.3. *Inertial Oscillations*

Figure 18 shows the wind speed as a function of time along flight 993. It is apparent that the wind speed tends to increase and decrease in a rather systematic fashion. The numbers in the figure, which give the hours between speed maxima and minima indicate that, on the average, the wind speed has a period of 15 hr, but that the period decreases as the flight progresses. Consideration of the influence of the geostrophic vorticity and the vertical advection of velocity upon the inertial period makes it appear likely that these are inertial oscillations [118]. The period of oscillation becomes shorter as the flight progresses because the geostrophic vorticity becomes greater along the trajectory.

While oscillations of apparently inertial character have been found on some individual transosonde flights, statistical verification of such

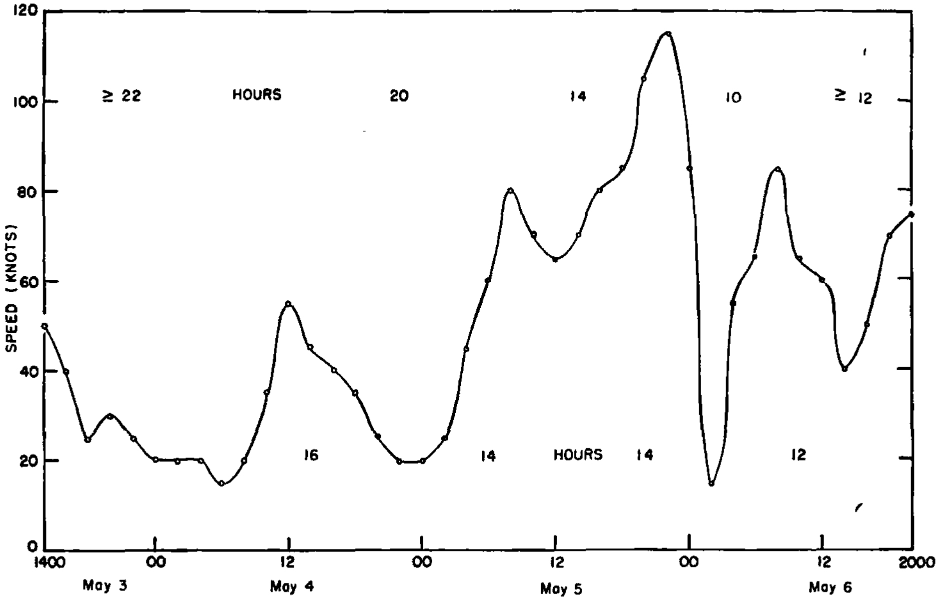


FIG. 18. Two-hour average wind speed as a function of Greenwich time along transosonde flight 993. The numerals indicate the number of hours between wind speed maxima and wind speed minima.

oscillations has not been successful. Perhaps this is not too surprising when it is realized that the inertial period will vary along each trajectory and from flight to flight, and that the transosonde positioning is none too accurate for the detection of such small-scale oscillation. A good example of the first difficulty is found on flight 993 where a spectral analysis of the speed fails to point up any pronounced periodicity in the inertial range because the period of inertial oscillation varies along the flight.

Part of the problem of detecting inertial oscillations by spectral means lies in the fact that for transosonde flights of the given duration, it is difficult to resolve oscillations of long-wave period (about 48 hr) and oscillations of inertial period (about 18 hr). If inertial oscillations exist, they should be most evident in the spectrum of $V \sin i$ since this parameter has relatively little power at low frequencies and consequently the problem of spectral resolution is not as great. However, Fig. 9 gives no indication of a peak in the variance of $V \sin i$ at the appropriate inertial period (20 hr) except during winter, although the fact that $V \sin i$ was averaged over a 12-hr period would not facilitate the detection of oscillations of inertial period. In order to eliminate this problem of averaging interval and at least partly eliminate the problem of positioning accuracy, all the 1957–59 transosonde flights from Japan which were positioned by the

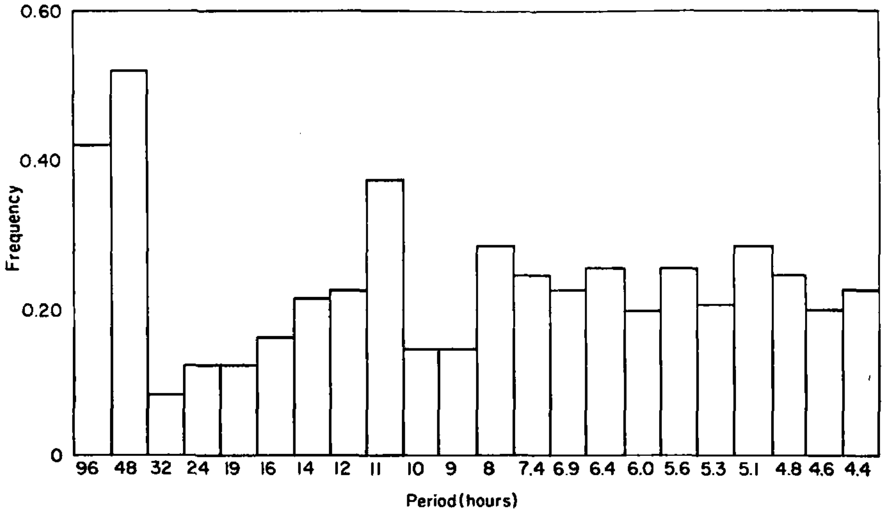


FIG. 19. Percentage of FCC-positioned transosonde flights of two or more days' duration whose velocity spectra show a peak at the given period of oscillation.

FCC for more than two days were selected, and a spectral analysis was made of 2-hr-average speed and zonal and meridional velocity components. Figure 19 shows the percentage of flights for which a spectral peak in any of these 3 parameters occurred at a given period of oscillation. Other than the peak associated with the long waves in the westerlies, the only peak with any significance is one of about a 11-hr period. There is absolutely no evidence for a predominance of oscillations of pure inertial period, although this might be due to the resolution problems mentioned previously. An interesting article by Newton [119], however, suggests that one would not expect to find "inertial" oscillations of an 18- or 20-hr period from trajectory data, but rather that oscillations of a period considerably greater than 18 or 20 hr might represent inertial oscillations owing to a systematic variation of wind speed and pressure gradient along wave-shaped trajectories. In any event, before the problem of inertial oscillations is settled once and for all, it probably will be necessary to make longer duration CLB flights with more accurate positioning, a goal most difficult to achieve.

8.4. Abnormal or Anomalous Flow

Figure 20 shows the trajectory of flight 48 in November 1957 with 300-mb contours extracted from NAWAC maps superimposed. Note the anticyclonic loop performed by the transosonde between California and Hawaii. The transosonde makes a full circle in 36 hr, corresponding to an angular velocity of -0.48×10^{-4}/sec. At this latitude of 30°N, the angular

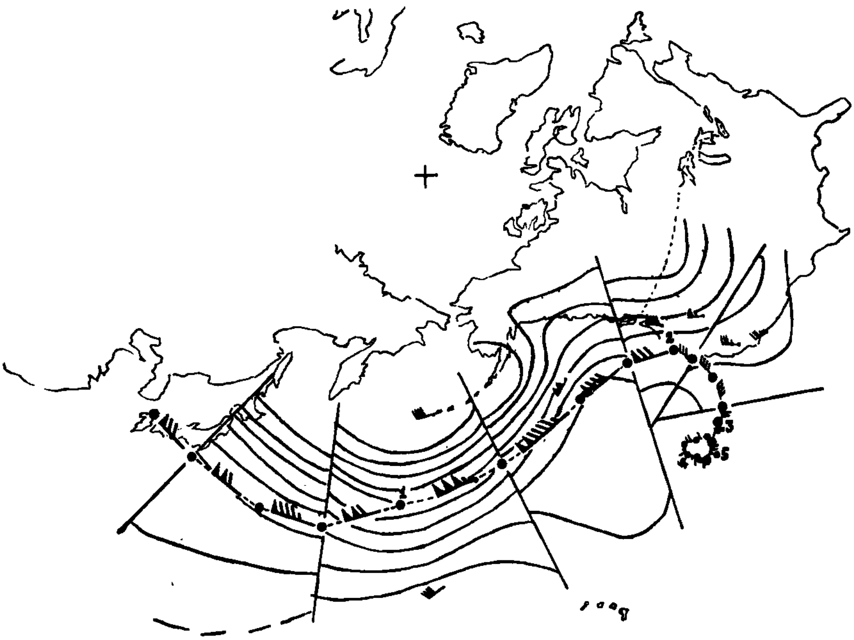


FIG. 20. Trajectory of 300-mb transosonde flight 48 in November 1957. Transosonde positions and transosonde-derived winds are indicated at 6-hr intervals along the trajectory. Superimposed on the trajectory are segments of 12-hr 300-mb NAWAC maps with contours drawn at 400-ft intervals.

velocity of the earth about the local vertical is 0.36×10^{-4}/sec. Thus, the transosonde, which we assume moves with an air parcel, has a negative angular velocity in space, and we can state that the flow is probably abnormal. From previous transosonde flights, it had been noted that abnormal flow tended to occur downstream from regions of inertial instability. This is precisely the case on flight 48, with the balloon 30 to 36 hr after release imbedded in the strongest anticyclonic geostrophic shear measured along any of the 1957–58 flights—a shear satisfying the above condition. The consequent loss of geostrophic control permitted the transosonde to move continuously toward high pressure rather than perform inertial oscillations. During the initial 18 hr of deceleration, the mean flow toward high pressure amounted to 17 knots. Certainly, mass depletion brought about by friction in the lower levels of an anticyclone would be strongly counteracted by such a strong flow toward high pressure at upper levels.

Transosonde flight 228 in March 1958 yields another striking example of the trajectory resulting from the passage of a transosonde through a region of strong anticyclonic wind shear [120]. After the transosonde passes through this region of anticyclonic shear, it traces out part of an anticyclonic loop before termination of the flight. While tracing out this loop, the transosonde-derived wind direction changes from NNW to SW in 12 hr. If this rate of rotation continued, the transosonde would complete the loop in 21.9 hr. The inertial period at this latitude is 20.9 hr. Thus, there is some CLB evidence for inertial flow ($d\theta/dt = -f$) as well as abnormal flow ($d\theta/dt < -f/2$) in the atmosphere.

Of perhaps more interest than the occurrence of abnormal flow in a somewhat unusual situation is the question of whether abnormal flow occurs on the large-scale ridges of the long waves in the westerlies. A study of the angular velocity possessed by transosondes as they passed through such ridges suggests that abnormal flow is rarely present on large-scale ridges—that, in fact, the dividing line between normal and abnormal flow represents a limit for the angular velocity on large-scale ridges in the same way that the dry adiabatic lapse rate represents a limit on the vertical lapse of temperature [121]. As an example, on flight 36 (Fig. 16) the anticyclonic angular velocity of the transosonde as it passes through the ridge in the Gulf of Alaska is just sufficient to counterbalance the cyclonic angular velocity of the earth about the local vertical at this latitude, so that for a short time the flow is just on the borderline between normal and abnormal flow.

8.5. Anticyclonic Wind Shear and the Trajectory Forecast Problem

On the basis of the above discussion, it might be anticipated that the forecasting of trajectories in regions of strong anticyclonic wind shear is dangerous. As proof of this danger, Fig. 21 shows the actual (dashed line) and numerically forecasted (dotted line) trajectories for transosonde flight NL-1, launched from Vernalis, California on November 13, 1957. This is one of the forecast trajectories presented in [68]. In Fig. 21 the NAWAC 300-mb map nearest to launch time is superimposed on the trajectories to show the region of anticyclonic wind shear into which the transosonde was launched. It is seen that the forecast trajectory closely follows the contour pattern on the 300-mb map, whereas in actuality the transosonde moved quickly toward high pressure and the trajectory became anticyclonically curved. The error in the trajectory forecast in this case is tremendous and probably is due to the fact that the transosonde was launched into a region of strong anticyclonic shear.

8.6. Critique

Inspection of two years of operational transosonde flights shows that in all cases where the trajectories are cusp shaped or the flow is abnormal, the transosonde has recently passed through a region of strong anticyclonic

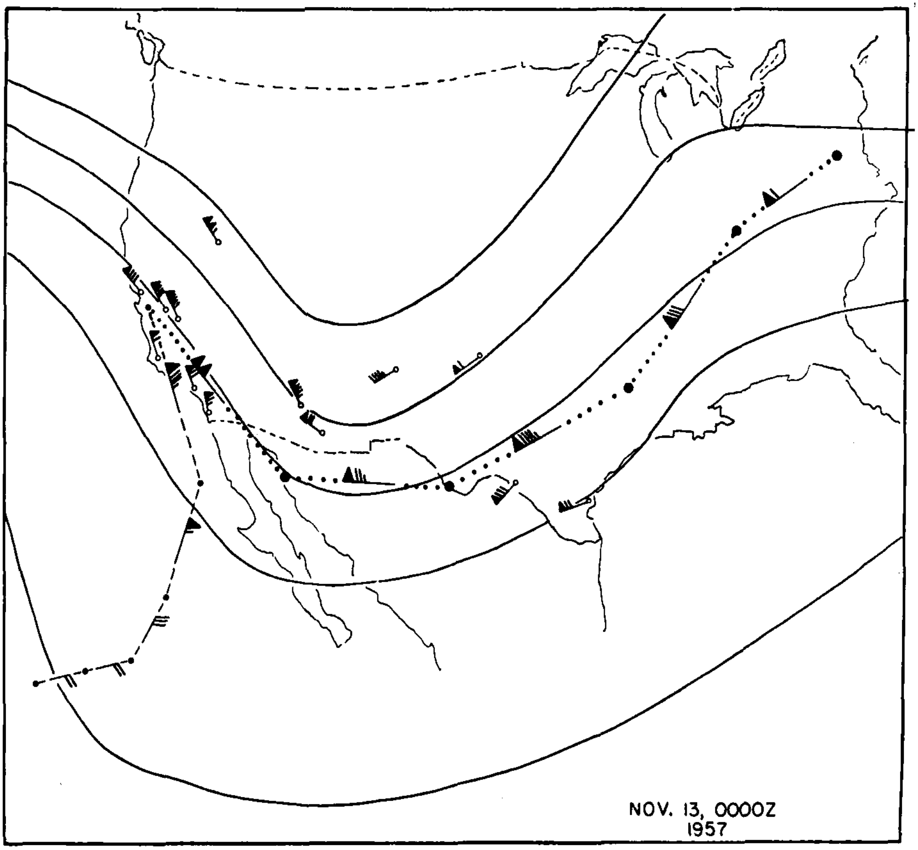


FIG. 21. Numerically forecasted trajectory (dotted line) and actual trajectory (dashed line) for 300-mb transosonde flight NL-1 in November, 1957. Indicated are 6-hr observed and forecasted transosonde positions and transosonde-derived winds. Superimposed on the trajectories is the NAWAC 300-mb map most nearly corresponding to the time of transosonde release.

shear, a shear which usually approaches that associated with dynamic or inertial instability. However, while the transosonde usually passes through such a region of strong anticyclonic shear immediately preceding the tracing out of the cusp or anticyclonic circle, it is by no means obvious that it is embedded in a region of anticyclonic shear during its entire path. For example, as transosonde flight 36 enters the United States from Canada (Fig. 16) it appears to be located near the jet stream core, while on transosonde flight 993 (Fig. 17) the balloon appears to pass from the cyclonic to the anticyclonic shear side of the jet stream while moving northward on May 6. Nevertheless, it cannot be denied that on the basis of experience so far, only those flights whose paths are at least in the proximity of strong anticyclonic shear exhibit such anomalous trajectories, while those flights firmly embedded in cyclonic shear describe quite regular paths.

Interesting theoretical treatments of air parcel trajectories in the vicinity of jet streams with parabolic speed profiles and the resulting periodicities in wind are given in [122–124], and it is hoped that further theoretical and observational work with CLB trajectories will soon clarify the problem of trajectory singularities.

9. Derivation of Kinematic Properties Other Than the Wind from Constant Level Balloon Data

9.1. Introduction

Other than the wind, no kinematic parameters can be determined precisely from CLB data. The vertical air motion may be estimated fairly accurately through the adiabatic assumption, but the field properties of the flow (divergence, vorticity, deformation) can never be determined accurately from a single CLB trajectory. This section considers some exact and inexact methods for estimating these kinematic parameters by means of one or more CLB trajectories.

9.2. The Estimation of Vertical Motion

The vertical air motion w may be estimated by the so-called "adiabatic method" according to the equation

$$w = -(d\Theta/dt)/(\gamma_p - \gamma) \tag{9.1}$$

where $(d\Theta/dt)$ is the change in temperature following an air parcel or CLB on a constant pressure surface, γ is the lapse rate, and γ_p is the process lapse rate, usually assumed dry adiabatic at 300 mb. So far, no reliable measurements of air temperature have been obtained from thermometers placed on CLB flights so that evaluation of the vertical air motion has been dependent upon interpolation of temperatures and lapse rates determined from radiosonde data to the successive positions of the CLB. Transosonde flight 993, which remained within the confines of the United States radiosonde network for its entire duration, is a good flight to illustrate the obtaining of vertical motions by this technique.

In Fig. 22, the heavy line shows the trajectory of flight 993 superimposed on the temperature fields of 12-hr, 300-mb maps. On the basis of the isotherms drawn on these maps, temperatures at the transosonde positions at map time have been estimated. Based on a linear interpolation, temperatures at the position of the transosonde midway between map times also have been estimated. From the estimation of the 300-mb temperature

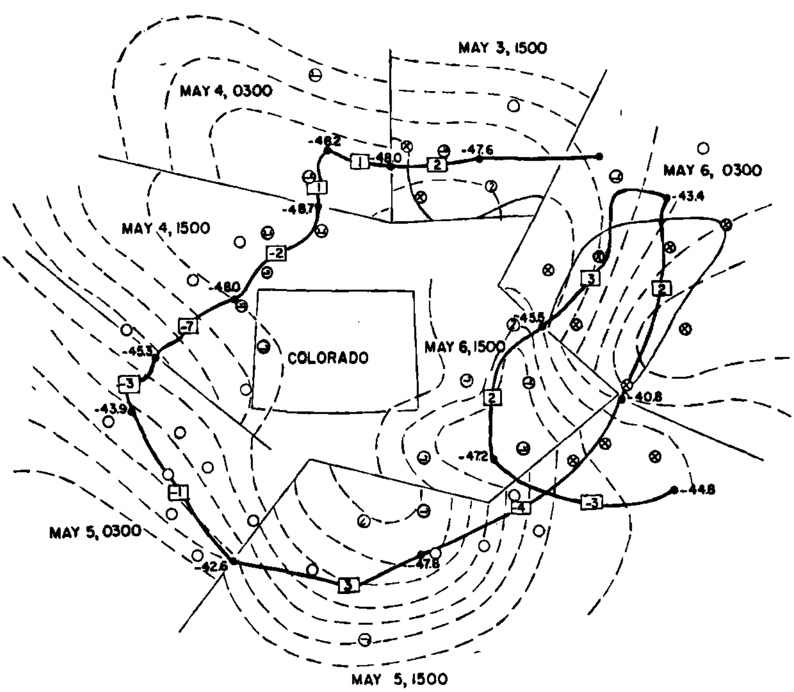


FIG. 22. Trajectory of 300-mb transosonde flight 993 in May 1953 with the 300-mb temperature interpolated from radiosonde data indicated at 6-hr intervals along the trajectory and the vertical air motion derived by the adiabatic technique indicated in boxes (cm/sec). The dashed lines represent isotherms at 1°C intervals, the stippled areas regions of precipitation, and within the small station circle is shown the type of high cloudiness present (if any), or the inability to estimate same due to low overcast (circle with a cross).

change along the trajectory and the mean 300-mb lapse rate along the trajectory (not shown), 6-hr average vertical motions (cm/sec) have been computed, as shown by the boxed numbers along the trajectory. To the west of Colorado this 6-hr average vertical motion is as large as −7 cm/sec based on a 6-hr increase in temperature at the transosonde position of 2.7°C. It may be shown that there is considerable correlation between the height of the air parcel initially associated with flight 993 (as estimated from the vertical motions along this flight) and the high cloudiness reported in the station circles [27].

The great advantage of the CLB for the estimation of the vertical motion by the adiabatic technique resides in the fact that the CLB indicates the exact trajectory on a constant pressure surface, and thus it is not required to estimate the individual derivative of temperature from the sum of local temperature change and horizontal temperature advection, as usually done in synoptic work.

9.3. *The Estimation of Horizontal Divergence*

9.3.1. Derivation from Comparison between Constant Absolute Vorticity and Transosonde Trajectories. The use of constant absolute vorticity (CAV) trajectories as an aid to weather forecasting almost disappeared with the advent of numerical forecasting. However, CAV trajectories are still of use for comparison with CLB trajectories because thereby an estimate of the horizontal divergence along the CLB trajectory may be obtained. This estimate depends on the validity of neglecting the solenoidal, "twisting," frictional, and vertical advection of vorticity terms in the vorticity equation.

The comparison of transosonde and CAV trajectories is made difficult by the fact that few of the transosonde inflection points are located at places of no horizontal wind shear, one of the prerequisites for the initiation of CAV trajectories. However, since it would be expected that in a statistical treatment the effect of nonzero initial wind shears would cancel out, for the purposes of the following analysis a CAV trajectory was originated at every inflection point along a smoothed transosonde trajectory. The tables used for the determination of the CAV trajectories are presented in reference [125] and apply to a spherical earth.

The transosonde flights used for comparison with CAV trajectories were made at 300 mb. Since 300 mb should be well above the mean surface of nondivergence, one would anticipate horizontal divergence in the southerlies and horizontal convergence in the northerlies. With such a distribution of divergence the transosonde amplitudes should be less than the CAV amplitudes. On the basis of 30 comparisons it was found that, on the average, the transosonde amplitude was 1.2 deg of latitude less than the CAV amplitude, the transosonde quarter wavelength was 2.5 deg of longitude less than the CAV quarter wavelength, and the transosonde quarter period was 2.5 hr less than the CAV quarter period. The existence of a smaller transosonde amplitude than CAV amplitude tends to confirm the existence, at 300 mb, of horizontal divergence in the southerlies between inflection point and trajectory crest and/or horizontal convergence in the northerlies between inflection point and trajectory trough line. The fact that the transosonde quarter wavelengths and quarter periods are also smaller than those of the CAV trajectories suggests that, on the average, the horizontal divergence at 300 mb extends somewhat downstream from the trajectory crest and/or the horizontal convergence extends somewhat downstream from the trajectory trough line.

In order to exploit the comparisons further, the differences between transososonde and CAV quarter periods, amplitudes, and quarter wavelengths were plotted as functions of transosonde-derived wind speed at the inflection point and deviation of transosonde-derived wind direction from west. It was found that the transosonde quarter period is less than the CAV quarter period except when the inflection point wind is nearly meridional, that the transosonde amplitude is less than the CAV amplitude except when the inflection point wind is nearly zonal, and that the transosonde quarter wavelength is less than the CAV quarter wavelength except when the inflection point wind deviates more than 45 deg from west. These results suggest that for 300-mb waves of small amplitude (wind at the inflection point nearly zonal) horizontal convergence tends to occur immediately downstream from the pre-trough inflection point followed by horizontal divergence near the trajectory crest, whereas for 300-mb waves of large amplitude horizontal divergence tends to occur immediately downstream from the pre-trough inflection point followed by horizontal convergence near the trajectory crest. It must be emphasized, however, that since the above results are based on relatively few comparisons of transosonde and CAV trajectories, the results may not be too representative.

9.3.2. Derivation through the Vorticity Equation. The relative geostrophic vorticity ζ_g may be estimated at the position of a constant level balloon by determining the height of the constant pressure surface at the CLB position and at 4 surrounding grid points according to the equation

$$\zeta_g = \frac{g}{f} \frac{(Z_x + Z_{-x} + Z_y + Z_{-y} - 4Z_0)}{d^2} \tag{9.2}$$

where g is the acceleration of gravity, f is the Coriolis parameter, and Z_x, Z_{-x}, Z_y, Z_{-y} are values of contour height along Cartesian coordinates at distances d from the CLB position where the contour height is Z_0. In order to obtain the heights at the grid points with sufficient accuracy for vorticity calculations, it is necessary to use a proportional (Gerber) ruler. With the neglect of the solenoidal, twisting, and frictional terms in the vorticity equation, this equation becomes

$$\frac{d\zeta_a}{dt} = -\zeta_a \nabla \cdot \mathbf{V} \tag{9.3}$$

where ζ_a is the absolute (geostrophic) vorticity

$$\zeta_a = \zeta_g + f + \frac{u \tan \phi}{a} \tag{9.4}$$

In equation (9.4) u is the zonal wind speed, ϕ is the latitude, and a is the earth's radius. The third term on the right in equation (9.4) arises because of the cyclonic vorticity possessed by an air parcel due to its zonal movement around the spherical earth. It is seen that, on the basis of these equations, one may estimate the horizontal divergence along a trajectory segment from the change in absolute (geostrophic) vorticity between the end points of the segment. Table VII shows the values of ζ_g, f, $u \tan \phi/a$, ζ_a, and $\nabla \cdot \mathbf{V}$ obtained from the segments of map analyses along transosonde flight 48 in November, 1957 (Fig. 20). At two points along this trajectory the (geostrophic) absolute vorticity is almost zero; consequently, the values of horizontal divergence derived from the changes in vorticity are large. Along this particular flight the absolute vorticity and horizontal divergence appear to fluctuate with a period near 24 hr.

9.3.3. Derivation from Changes in Vertical Stability. If a balloon could be instrumented so as to fly along an isentropic surface, then the horizontal divergence could easily be determined (assuming isentropic flow) from the temporal change in lapse rate (change in weight between two isentropic surfaces) at the position of the balloon. For a balloon floating along a constant level or constant pressure surface the change in lapse rate can still be used to estimate the horizontal divergence, but a complication is introduced due to the likelihood of vertical advection of lapse rate by the vertical air motion. While it has been found that the vertical advection of velocity is a rather small part of the total acceleration (Section 4.3), a study [126] has shown that, particularly near the tropopause, the vertical advection of lapse rate represents a large part of the total change in lapse rate following a CLB, and consequently it is difficult to estimate the horizontal divergence in this way. In the stratosphere large changes in lapse rate are associated with small values of the horizontal divergence and it is here, far removed from the tropopause, that this technique should be of value. Since transosonde flights have not yet been made well within the stratosphere over the United States, where a good radiosonde network

TABLE VII. Evaluation of Vorticity and Horizontal Divergence along the Trajectory of Transosonde Flight 48 in November 1957.

Date		$\zeta_g(\text{sec}^{-1}) \times 10^{-4}$	$f(\text{sec}^{-1}) \times 10^{-4}$	$\frac{u \tan \phi}{a}$ $(\text{sec}^{-1}) \times 10^{-4}$	$\zeta_a(\text{sec}^{-1}) \times 10^{-4}$	$\nabla \cdot \mathbf{V}(\text{sec}^{-1}) \times 10^{-4}$
Nov. 25	0000Z	−0.15	0.82	0.04	0.70	
	1200	−0.72	0.79	0.04	0.11	0.34
Nov. 26	0000	−0.34	0.90	0.08	0.64	−0.33
	1200	−1.09	1.05	0.09	0.05	0.40
Nov. 27	0000	−0.18	0.98	0.03	0.83	−0.41
	1200	−0.31	0.84	0.00	0.53	0.10
Nov. 28	0000	−0.58	0.76	0.00	0.18	0.23
	1200	0.19	0.71	0.00	0.90	−0.28

exists, no work has been done on this problem since that presented in [126].

9.3.4. Derivation from Transosonde Triads. It is well known that the horizontal divergence $\nabla \cdot \mathbf{V}$ may be expressed as the percentual rate of change of area A of an infinitesimal fluid element according to the formula

$$\nabla \cdot \mathbf{V} = \frac{1}{A}\frac{dA}{dt} \tag{9.5}$$

Consequently, the temporal change in triangular area delineated by a triad of CLB flights permits an approximation of the horizontal divergence. For example, the upper left diagram of Fig. 23 shows the triangles delineated by three transosondes (flights 216, 217, and 218) at times 6 hr apart, with the solid lines delineating the initial triangle. Planimetric measurement indicates that the area so delineated increased from 7.3 units to 8.2 units in 6 hr. Substitution of these values in equation (9.5) yields a horizontal divergence of 0.54×10^{-5}/sec. It should be noted that since the area A appears in both numerator and denominator of equation (9.5), the variation in map scale is not critical in the planimetric evaluations.

9.4. The Estimation of Vorticity and Deformation

In simplified form, the vertical component of relative vorticity ζ is given by

$$\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \tag{9.6}$$

where u and v are zonal and meridional wind components and x and y are zonal and meridional directions. Letting $u' = v$ and $-v' = u$ it is found that

$$\zeta = \frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y} = \frac{1}{A'}\frac{dA'}{dt} \tag{9.7}$$

Thus the relative vorticity also may be estimated from the temporal change in area of two triangles delineated by CLB flights if the second triangle is obtained by rotating the CLB displacements per unit time 90° to the right. The upper right diagram of Fig. 23 shows that the area of the second triangle obtained by the 90° rotation is 2.3 units in comparison with an area of 7.3 units for the initial triangle. From these values it is found that the relative vorticity indicated by this triad of transosondes is -4.82×10^{-5}/sec.

The stretching deformation D_{st} is given by

$$D_{st} = \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \tag{9.8}$$

Letting $u'' = u$ and $v'' = -v$ it is found that

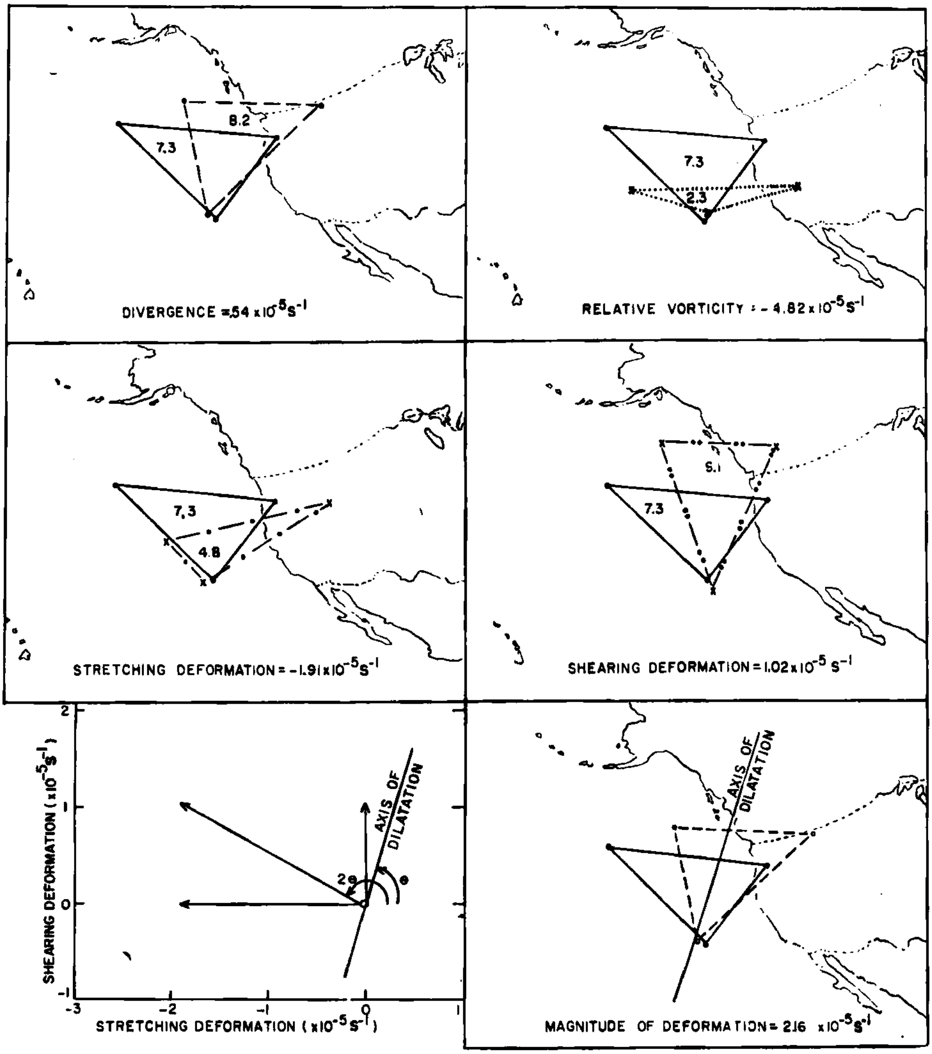


FIG. 23. The triangle delineated by the solid line indicates a transosonde triad, consisting of flights 216, 217, and 218, on 0600Z, March 1, 1959. The upper left diagram shows the change in area of the triangle due to 6-hr movement of the transosondes. From this percentual change in area the horizontal divergence is estimated. By varying the direction of the displacements per unit time, different triangles are formed from which the vorticity (upper right diagram), stretching deformation (middle left diagram), and shearing deformation (middle right diagram) can be obtained from the percentual change in area. The lower left diagram shows the obtaining of the axis of dilatation while the lower right diagram shows the superposition of this axis upon the transosonde triad.

$$D_{st} = \frac{\partial u''}{\partial x} + \frac{\partial v''}{\partial y} = \frac{1}{A''}\frac{dA''}{dt} \tag{9.9}$$

Thus the stretching deformation also may be estimated from the temporal change in area of two triangles delineated by CLB flights if the second triangle is obtained by the above substitutions of displacements per unit time. The left middle diagram in Fig. 23 shows that the area of the second triangle so obtained from the transosondes is 4.8 units in comparison with the initial area of 7.3 units. From these values it is found that the magnitude of the stretching deformation indicated by this triad of transosondes is -1.91×10^{-5}/sec.

The shearing deformation D_{sh} is given by

$$D_{sh} = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \tag{9.10}$$

Letting $u''' = v$ and $v''' = u$ it is found that

$$D_{sh} = \frac{\partial u'''}{\partial x} + \frac{\partial v'''}{\partial y} = \frac{1}{A'''}\frac{dA'''}{dt} \tag{9.11}$$

Thus the shearing deformation also may be estimated from the temporal change in area of two triangles delineated by CLB flights if the second triangle is obtained by the above substitutions of displacements per unit time. The right middle diagram of Fig. 23 shows that the area of the second triangle so obtained from the transosondes is 9.1 units in comparison with the initial area of 7.3 units. From these values it is found that the magnitude of the shearing deformation indicated by this triad of transosondes is 1.02×10^{-5}/sec.

The magnitude of the deformation equals the square root of the sum of the squares of the stretching and shearing deformation. As shown by Saucier [127], the axis of dilatation is located halfway between the positive axis of stretching deformation and the vector resultant of stretching and shearing deformation (lower left diagram of Fig. 23). The lower right diagram of Fig. 23 shows this axis of dilatation superimposed on the successive transosonde triads.

The values of horizontal divergence and vorticity to be expected on the synoptic scale are well known. Less well known are the values to be expected for the stretching and shearing deformation and total deformation. Based on 28 calculations from triads of transosondes it was found that the average value for the stretching deformation (2.17×10^{-5}/sec) slightly exceeded the average value for the shearing deformation (1.68×10^{-5}/sec). The average value for the total deformation determined from the transosonde triads was 3.04×10^{-5}/sec. These few data thus suggest that the horizontal deformation is intermediate in magnitude to the vertical component of vorticity and the horizontal divergence.

10. Meteorological Applications of Constant Level Balloon Oscillations in the Vertical

10.1. Introduction

One of the parameters telemetered from the transosondes at 2-hr intervals is the pressure at which the transosonde is flying. It is desirable to make a study of these flight pressures even though experience would dictate that the repeated dropping of ballast and valving of gas by a ballasted CLB system would introduce difficulties in the utilization of such pressures for meteorological investigations. In this section some interesting meteorological implications resulting from analysis of such flight pressures are considered.

10.2. The Estimation of Atmospheric Turbulence

For the purpose of analyzing the variations in pressure height of the transosondes, it is desirable to treat these pressures in the same way that the position fixes were treated. Accordingly, a pressure p was obtained at 6-hr intervals following transosonde release by averaging 3 successive pressures each 2 hr apart. The pressures so obtained were again averaged by finding the average of three successive pressures each 6 hr apart, yielding a mean value of the pressure $\bar{p}$. The difference between the pressure and the mean pressure $p - \bar{p}$ was taken as a measure of the large-scale vertical "turbulence" which the transosonde was undergoing.

The top four rows in Table VIII give the correlation between the 6-hr

Table VIII. Correlation Coefficients and their Probability of Chance Occurrence between Meteorological Parameters and Vertical Displacements of Transosondes (Δp) and between Meteorological Parameters and the Deviation of Transosonde Pressure Height from Mean Value of Pressure Height $|p - \bar{p}|$.

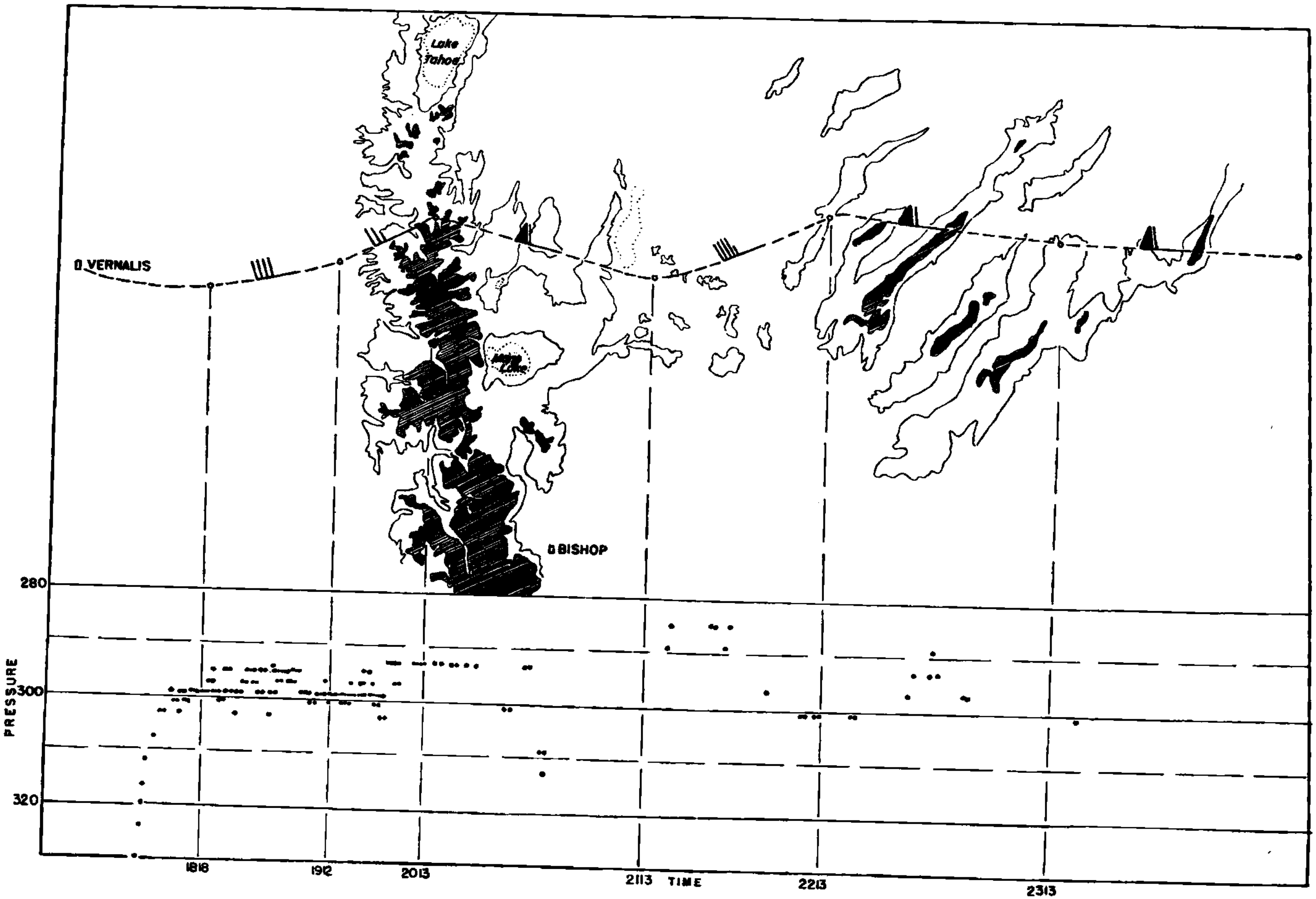
Parameter	Correlation coefficient	Probability of chance occurrence (%)
$(\Delta p)v$	−0.004	95
$(\Delta p)V\dot{\theta}$	0.039	20
$(\Delta p)\dot{V}$	−0.034	25
$(\Delta p)\partial V_g/\partial n$	−0.004	95
$\|p - \bar{p}\| v$	−0.007	85
$\|p - \bar{p}\| V\dot{\theta}$	−0.014	65
$\|p - \bar{p}\| \dot{V}$	−0.094	<1
$\|p - \bar{p}\| \partial V_g/\partial n$	0.114	<1

change in pressure height of the transosonde and, respectively, the meridional wind, the normal acceleration, the tangential acceleration, and the geostrophic wind shear along the trajectories. The probability that the correlation determined would occur by chance was determined by the "Z test" as given by Hoel [92]. The fact that the correlation between the meridional wind and the change in the pressure at which the transosondes are flying is of no significance probably proves that the transosondes do not follow the vertical air motion, except, perhaps, under extreme conditions. The indicated tendency for the transosondes to ascend on ridges and when the balloon accelerates is also of little significance.

The last four rows in Table VIII give the correlation coefficients (and their significance) between the absolute deviation from the mean pressure height $|p - \bar{p}|$ and the meteorological parameters mentioned previously. Interestingly enough, two of these correlations are significant at the 99% level; namely, there is a pronounced tendency for the deviations or the vertical turbulence to be large in regions of wind deceleration and in regions where the geostrophic wind shear is anticyclonic. The former tendency has been noted independently by Mastenbrook [128] during a study of transosonde flights launched from Vernalis, California. The latter tendency is not contradicted by the predominance of "clear air" turbulence on the cyclonic-shear side of the jet stream because, with the averaging procedure employed, the transosonde would be responding to a much larger scale of turbulence than that represented by "clear air" turbulence. In fact, the existence of large-scale vertical turbulence on the anticyclonic-shear side of the jet stream probably would be expected since it is in such regions that geostrophic control tends to be lost. For a discussion of vertical CLB oscillations on a much smaller (gravity wave) scale, the reader is referred to an article by Emmons and his co-workers [129].

10.3. The Delineation of Mountain Waves

Transosonde flight N-2 was released from Vernalis, California on May 12, 1955, a time of severe Sierra Wave activity at Bishop, California [130]. This flight was positioned at hourly intervals, permitting a more detailed trajectory analysis than usually possible. The upper part of Fig. 24 shows a portion of the horizontal trajectory of this flight superimposed on the topography of the Sierra Nevada. The trajectory is sinusoidal with one of the trajectory crests situated directly above the crest of the Sierra Nevada. Note that the trajectory wind speed is greater downstream than upstream from the two trajectory crests, which distribution would be associated with southward transport of momentum. The bottom part of Fig. 24 shows the pressure at which the transosonde was flying as a function of time. The time has been fitted to the geographical longitude, as shown by the vertical


Lake Tahoe
VERNALIS
Mono Lake
BISHOP
280
300
320
PRESSURE
1818
1912
2013
2113
TIME
2213
2313

dashed lines in the figure. Noteworthy is the above-normal elevation of the transosonde at 2013Z as it transited the crest of the Sierra Nevada, followed by a most unusual descent of the balloon to a pressure surface of 312 mb, a surface far below the ballast release surface (300 mb). Since this height fluctuation took place in the middle of the day when strong radiation changes would not be expected, it is suggested that the transosonde was forced to this low elevation by descending air currents of such strength that the balloon could not maintain flight level by release of ballast. The subsequent quick ascent of the transosonde is not necessarily associated with ascending air motion because, after the release of so much ballast, an ascent of the transosonde would be expected. If the vertical oscillations of the transosonde are in some degree related to the vertical air motions, it may be noted that the presence of descending air motion at a time when the zonal wind components along the trajectory are above average would be a means for transporting momentum downward. A detailed account of a mountain influence on another transosonde trajectory is given in the report by Mastenbrook [128].

Statistical indications of the influence of mountains on transosonde flights may be obtained from the termination points of transosondes. It has been found that, density-wise, the largest percentage of transosonde flights released from Japan terminate between the Sierra Nevada and the eastern slopes of the Rocky Mountains. It is presumed that the high density of transosonde terminations in this region is associated with depletion of the transosonde ballast supply caused by mountain-induced vertical air motions.

11. Conclusion

In this article an attempt has been made to illustrate the usefulness of CLB data from both operational and research points of view. On the one hand, CLB flights provide a means for obtaining meteorological data over the vast oceans and other unpopulated regions of the world. On the other hand, these balloons provide the means for realizing important research progress along certain frontiers of meteorological knowledge. In order to reap the full benefit of the CLB system, however, it is essential that this system be recognized as a truly international, meteorological probe deserving of international acceptance. The obtaining of such recognition should be but the first step on the road toward development of a system for weather

---

Fig. 24. Portion of 300-mb trajectory of transosonde flight N-2 on May 12, 1955, a time of severe Sierra Wave activity at Bishop, California. The transosonde positions and transosonde-derived winds are indicated at 1-hr intervals along the trajectory. In the bottom half of the figure the pressure (in mb) at which the transosonde was flying is indicated as a function of Greenwich time and longitude.

analysis and forecasting based upon Lagrangian rather than Eulerian concepts. Such a change in emphasis may be the step needed to achieve a higher plateau of knowledge in the fields of weather diagnosis, prognosis, and control. The CLB system offers the means by which this change in emphasis can be made a reality.

List of Symbols

a radius of earth
A area of infinitesimal fluid element
d grid distance in geostrophic vorticity determination
D_{sh} shearing deformation
D_{st} stretching deformation
f Coriolis parameter
g acceleration of gravity
i angle between wind and geostrophic wind
K_L latitudinal austausch coefficient
K_l longitudinal austausch coefficient
p pressure-altitude of a CLB
$\bar{p}$ mean pressure-altitude of a CLB
Δp change in pressure-altitude of a CLB
P percentage of CLB pairs separated by a given distance
P_R percentage of CLB pairs separated by less than distance R
r distance between successive CLB positions
$\overline{\Delta r}$ mean error in CLB positions
R distance between CLB pairs
R_m modal distance between CLB pairs
$R_u(t)$ autocorrelation coefficient of zonal wind
$R_v(t)$ autocorrelation coefficient of meridional wind
t time; in particular, time interval for which CLB velocity determined
T time interval for which CLB acceleration determined
u zonal wind speed
u_g zonal geostrophic wind speed
u_{ag} zonal ageostrophic wind speed
v meridional wind speed
v_g meridional geostrophic wind speed
v_{ag} meridional ageostrophic wind speed
V wind speed
V_g geostrophic wind speed
V_b volume of balloon
w vertical velocity
W weight of balloon system
x, y, z Cartesian coordinates, x toward the east, y toward the north, z toward the zenith
Z height of a constant pressure surface
γ lapse rate
γ_p process lapse rate
ζ vertical component of vorticity
ζ_g vertical component of geostrophic vorticity
ζ_a vertical component of absolute vorticity

θ wind direction
Θ air temperature
ρ_a density of air
ρ_h density of helium
σ_L latitudinal standard deviation of CLB position
σ_l longitudinal standard deviation of CLB position
ϕ latitude
$\mathbf{F}$ frictional force per unit mass
$\mathbf{K}$ unit vertical vector
$\mathbf{V}$ wind velocity
$\mathbf{V}_g$ geostrophic velocity
$\mathbf{V}_{ag}$ ageostrophic velocity
∇ vector differential operator
$\nabla \cdot \mathbf{V}$ horizontal divergence

References

1. Spilhaus, A. F. (1950). Progress in meteorological instrumentation, 1920–1950. *Bull. Am. Meteorol. Soc.* **31**, 358–364.
2. Haig, T. O., and Lally, V. E. (1958). Meteorological sounding systems. *Bull. Am. Meteorol. Soc.* **39**, 401–409.
3. Meisinger, C. L. (1919). The constant elevation free-balloon flights from Fort Omaha. *Monthly Weather Rev.* **47**, 535–538.
4. Harries, H. (1923). Toy balloon "races" from Brighton. *Meteorol. Mag.* **58**, 30–32.
5. Richardson, L. F., and Proctor, D. (1925). Diffusion over distances ranging from 3 to 86 kilometers. *Mem. Roy. Meteorol. Soc.* **1**, 1–16.
6. Akerman, J. D., and Piccard, J. F. (1937). Upper air study by means of balloons and radio meteorograph. *J. Aero. Sci.* **4**, 332–337.
7. Arakawa, H. (1956). Basic principles of the balloon bomb. *Papers in Meteorol. and Geophys.* **6**, 239–243.
8. McKay, R. W. (1945). Japanese paper balloons. *Engineering J.* **28**, 563–567.
9. Jacobs, W. C. (1947). Wartime developments in applied climatology. *Meteorol. Monographs* **1**, No. 1., 52 pp.
10. Durst, C. S., and Gilbert, G. H. (1950). Constant-height balloons—Calculations of geostrophic departures. *Quart. J. Roy. Meteorol. Soc.* **76**, 75–88.
11. Spilhaus, A. F., Schneider, C. S., and Moore, C. B. (1948). Controlled-altitude free balloons. *J. Meteorol.* **5**, 130–137.
12. Smith, J. R., Moore, C. B., and Murray, W. D. (1949). Summary of flights of constant level balloons. Technical Report 93.02, Research Division, College of Engineering, New York University, New York.
13. Murray, W. D., Smith, J. R., and Emmons, G. (1950). High altitude balloon trajectory study, 34 pp. Technical Report 121.05, Research Division, College of Engineering, New York University, New York.
14. Murray, W. D., Schneider, C. S., and Smith, J. R. (1950). Development and utilization of constant-level balloons. *Transact. Am. Geophys. Un.* **31**, 843–848.
15. Moore, C. B., Smith, J. R., and Gaalswyk, A. (1954). On the use of constant level balloons to measure horizontal motions in the atmosphere. *J. Meteorol.* **11**, 167–172.
16. Giles, K. C., and Peterson, R. E. (1957). The usefulness of constant pressure level balloons. Mechanical Division Report No. 1725, General Mills Inc., Minneapolis, Minnesota.

17. Haig, T. O., and Craig, R. A. (1953). Results of controlled-altitude balloon flights at 50,000 to 70,000 feet during September 1952. Air Force Surveys in Geophys., No. 30, Air Force Cambridge Res. Center, Massachusetts.
18. Doty, E. A. (1955). Constant altitude balloon flights at 39,000 feet from October through December 1954, 133 pp. Appendix I, Phase VI, Functional Development Test Report, System 124A, Air Force Cambridge Res. Center, Massachusetts.
19. Haas, M. L., ed. (1956). Compilations of meteorologically useful data from Project Moby Dick, (ten sections). Air Force Cambridge Res. Center, Massachusetts.
20. Hovind, E. (1957). High level winds over the western United States. Unpublished manuscript, Department of Meteorology, Univ. of California at Los Angeles.
21. Hovind, E. (1957). Diurnal and semidiurnal oscillations in the wind field at 50,000 feet to 75,000 feet. Unpublished manuscript, Department of Meteorology, Univ. of California at Los Angeles.
22. Orville, H. T. (1950). Exploring the upper atmosphere with transosondes (transoceanic-sondes). Presidential address by the retiring President of the American Meteorological Society, Saint Louis, Missouri, January 4, 1950.
23. Mastenbrook, H. J., Anderson, A. D., and Cubbage, H. D. (1951). Evaluation of the transosonde system, 13 pp. Naval Res. Laboratory Letter Report to the Bureau of Aeronautics Ser. R-3900-66/51, Washington, D. C.
24. Mastenbrook, H. J., and Anderson, A. D. (1954). The transosonde. *Weatherwise* 7, 79–81.
25. Anderson, A. D., Mastenbrook, H. J., and Cubbage, H. D. (1955). The transosonde-A new meteorological data-gathering system, 18 pp. Naval Res. Laboratory Report 4649, Washington, D. C.
26. Panofsky, H. (1946). Methods of computing vertical motion in the atmosphere. *J. Meteorol.* **3**, 45–49.
27. Neiburger, M., and Angell, J. K. (1956). Meteorological applications of constant-pressure balloon trajectories. *J. Meteorol.* **12**, 166–194.
28. Mastenbrook, H. J. (1958). A transosonde system for operation within air space. Naval Res. Laboratory Quarterly Report, September 1959, Washington, D. C.
29. Wexler, H. (1950). The great smoke pall, September 24–30, 1950. *Weatherwise* **3**, 129–134.
30. Shaw, W. N., and Lempfert, R. G. K. (1906). The life history of surface air currents. M. O. 174 Meteorological Office, London.
31. Giles, K. C., and Peterson, R. E. (1957). A study of wind velocities measured from constant-pressure balloons. *J. Meteorol.* **14**, 569–572.
32. Angell, J. K. (1959). Estimating the frictional force at 300 mb by means of transosonde and radiosonde data. *J. Meteorol.* **16**, 216–217.
33. Rao, K. (1956). Evaluation of the order of magnitude of frictional dissipation by transosonde observations. Unpublished manuscript, Department of Meteorology, Univ. of California at Los Angeles.
34. Mantis, H. T. (1954). Meteorological section in Progress report on high altitude plastic balloons December 22, 1952–December 3, 1953, 149 pp. Department of Physics, Univ. of Minnesota, Minneapolis, Minnesota.
35. Mantis, H. T., and Huch, W. F. (1959). "Meteorological studies using constant altitude balloons," 63 pp. Technical Report No. AP-14, School of Physics, Univ. of Minnesota, Minneapolis, Minnesota.
36. Anderson, A. D. (1955). An estimate of the tracking effectiveness of high frequency radio direction finding for transosonde balloons, 10 pp. Naval Res. Laboratory Memorandum Report 502, Washington, D. C.

37. Anderson, A. D. (1955). A study of the accuracy of winds derived from transosonde trajectories, 10 pp. Naval Res. Laboratory Memorandum Report 498. Washington, D. C.
38. Giles, K. C., and Peterson, R. E. (1958). Accuracy of RDF position fixes in tracking constant level balloons," 11 pp. Air Force Surveys in Geophys. No. 107, Air Force Cambridge Res. Center, Massachusetts.
39. Ohnsorg, F. (1957). Error-estimates for constant-level balloon tracking. *J Meteorol.* **14,** 81–83.
40. Anderson, A. D., and Mastenbrook, H. J. (1956). A new upper air data system—the transosonde. *Bull. Am. Meteorol. Soc.* **37,** 342–350.
41. Angell, J. K. (1958). Transosonde tracking and communication networks—An evaluation, 7 pp. Unpublished typescript, U. S. Weather Bureau, Washington, D. C.
42. Jones, R. A. (1957). Studies of eddy structure in the first few thousand feet of the atmosphere, 18 pp. Porton Technical Paper No. 588, Chemical Defense Experimental Station, Porton, Wilts, England.
43. Smith, L. B. (1954). A study comparing winds aloft measuring equipment at Salton Sea Test Base, 9 pp. Report SC-3512(TR), Sandia Corporation, Albuquerque, New Mexico.
44. Perkins, D. T. (1952). The response of balloons to the wind. *Bull. Am. Meteorol. Soc.* **33,** 135–139.
45. Brier, G. W. (1950). The statistical theory of turbulence and the problem of diffusion in the atmosphere. *J. Meteorol.* **7,** 283–290.
46. Great Britain Meteorological Office (1940). Variation of wind with distance and variation of wind with time, 81 pp. H. M. Stationery Shop, London.
47. Brooks, E. M. (1947). The accuracy and representativeness of the 10,000 foot pibal report. *Bull. Am. Meteorol. Soc.* **28,** 405–408.
48. Gabriel, J. E., and Bellucci, R. (1951). Time variation of winds aloft. *J. Meteorol.* **8,** 422–423.
49. Rapp, R. R. (1952). The effect of variability and instrument error on measurements in the free atmosphere, 41 pp. Meteorol. Papers, Vol. 2, No. 1, New York Univ., New York.
50. Durst, C. (1948). The fine structure of wind in the free air. *Quart. J. Roy. Meteorol. Soc.* **74,** 349–356.
51. Angell, J. K. (1959). Comparisons of rawins and transosonde-derived winds, 2 pp. Unpublished typescript, U. S. Weather Bureau, Washington, D. C.
52. Crary, A. P. (1959). Antarctica. *Transact. Am. Geophys. Un.* **40,** 331–339.
53. Angell, J. K. (1958). Some evidence for tidal oscillations obtained from 300 mb transosonde data. *J. Meteorol.* **15,** 566–567.
54. Anderson, A. D. (1952). Meteorological trajectory study for test flights to be made from San Francisco, 11 pp. Naval Res. Laboratory Memorandum Report No. 23, Washington, D. C.
55. Gustafson, A. F. (1950). The computation of air trajectories, 30 pp. Air Weather Service Technical Report 105–62, Washington, D. C.
56. Johannessen, K. R. (1956). Constant-pressure trajectories, 86 pp. Air Weather Service Manual 105-47, Washington, D. C.
57. Mantis, H. T., and Ohnsorg, F. R. (1952). Development of methods for forecasting balloon trajectories. Final Report, Trajectory Forecasting Project, Institute of Technology, Univ. of Minnesota, Minneapolis, Minnesota.
58. Gaalswyk, A. (1952). Balloon trajectory forecasting. Report 1072, Part 1, General Mills Aero. Res. Laboratories, Minneapolis, Minnesota.

59. Bureau of Aeronautics, Project AROWA (1956). Task 19, Interim trajectory forecasting manual, 33 pp. U. S. Naval Air Station, Norfolk.
60. Machta, L. (1951). Meteorological trajectories. Unpublished manuscript, Sci. Services Division, U. S. Weather Bureau, Washington, D. C.
61. Machta, L. (1954). A mathematical verification of trajectory methods. Unpublished manuscript, Sci. Services Division, U. S. Weather Bureau, Washington, D. C.
62. Franceschini, G. A., and Freeman, J. C. (1955). Computation of horizontal trajectories in the atmosphere, 188 pp. Sci. Reports No. 5, and No. 1, on Projects 57 and 108, Department of Oceanog., the A & M College of Texas, College Station, Texas.
63. Franceschini, G. A. (1957). Results of recent trajectory studies. *J. Meteorol. Soc. Japan, 75th Anniversary Vol.*, pp. 268–270.
64. Franceschini, G. A. (1958). The theory of analysis and prediction of trajectories, 94 pp. Final report on Project 108, Department of Oceanog. and Meteorol., the A & M College of Texas, College Station, Texas.
65. Sasaki, Y. (1958). Atmospheric field equations based on a variational principle with application to the motion of closed systems, 23 pp. Sci. Report No. 3, on Project 155, Department of Oceanog. and Meteorol., the A & M College of Texas, College Station, Texas.
66. Angell, J. K. (1956). The hindcasting of trajectories. Unpublished manuscript, Department of Meteorol., Univ. of California at Los Angeles.
67. Djuric, D., and Wiin-Nielsen, A. (1956). On computation of meteorological trajectories with the aid of electronic computers, 20 pp. Preliminary Report, International Meteorological Institute in Stockholm.
68. Hubert, W. E., Wolff, P. M., and Cave, C. L. (1958). A comparison of JNWP trajectory forecasts with transosonde flights. *Monthly Weather Rev.* **86**, 53–59.
69. Mastenbrook, H. J., and Anderson, A. D. (1953). Evaluation of the transosonde system—Report for period January 1952 to March 1953, 22 pp. Naval Res. Laboratory Memorandum Report No. 171, Washington, D. C.
70. Mastenbrook, H. J., and Anderson, A. D. (1953). Evaluation of transosonde system—Data report for period April 1953 to August 1953, 99 pp. Naval Res. Laboratory Memorandum Report 240, Washington, D. C.
71. Mastenbrook, H. J. (1956). Transosonde flights for 1955—Data report, 61 pp. Naval Res. Laboratory Memorandum Report 612, Washington, D. C.
72. Mastenbrook, H. J. (1956). Transosonde flights for January and February 1956—Data report, 25 pp. Naval Res. Laboratory Memorandum Report 632, Washington, D. C.
73. Mastenbrook, H. J. (1956). Transosonde flights for June and July, 1956—Data report, 14 pp. Naval Res. Laboratory Memorandum Report 628, Washington, D. C.
74. Angell, J. K. (1957). The application of routine transosonde data to analysis and forecasting, 33 pp. U. S. Weather Bureau Training Paper No. 13, Washington, D. C.
75. Gaalswyk, A. (1958). Suggested uses of meteorological data collected by constant-pressure level balloon systems, 32 pp. Sci. Report No. 1 on Project 80070, General Mills, Inc., Minneapolis, Minnesota.
76. Angell, J. K. (1959). A summary of navy-sponsored 300 mb constant level balloon (transosonde) flights from Japan in 1957–58, 53 pp. Unpublished Manuscript, U. S. Weather Bureau, Washington, D. C.
77. Angell, J. K. (1959). A summary of navy-sponsored 250 mb constant level bal-

loon (transosonde) flights from Japan in 1958–59, 53 pp. Unpublished manuscript, U. S. Weather Bureau, Washington, D. C.

78. Angell, J. K. (1959). A climatological analysis of two years of routine transosonde flights from Japan. *Monthly Weather Rev.* **87**, 427–439.
79. Brooks, C. E. P., Durst, C. S., and Carruthers, N. (1950). Upper winds over the world. *Geophys. Memoirs No.* **85**, 149 pp. Meteorol. Office, Air Ministry, London.
80. Edinger, J. G., and Rapp, R. R. (1957). Dispersion in the upper atmosphere. *J. Meteorol.* **15**, 180–183.
81. Miller, J. E. (1948). Studies of large scale vertical motions of the atmosphere. Meteorol. Papers, Vol. 1, No. 1, New York Univ., New York.
82. Grimminger, G. (1941). The intensity of lateral mixing in the atmosphere as determined from isentropic charts. *Bull. Am. Meteorol. Soc.* **22**, 227–228.
83. Defant, A. (1921). Die Bestimmung der Turbulenzgrössen der atmosphärischen Zirkulation aussertropischer Breiten. *Sitzber. Akad. Wiss. Wien, Abt. IIa* **130**, 383–403.
84. Mintz, Y. (1954). The observed zonal circulation of the atmosphere. *Bull. Am. Meteorol. Soc.* **35**, 208–214.
85. Oishi, W. (1930). Vento super Tateno. *J. Aerol. Obs. at Tateno, Japan* **1**, No. 5.
86. Neiburger, M., Sherman, L., Kellogg, W. W., and Gustafson, A. F. (1948). On the computation of wind from pressure data. *J. Meteorol.* **5**, 87–92.
87. Machta, L. (1948). A study of the observed deviations from the geostrophic wind. Doctoral dissertation, Department of Meteorology, Massachusetts Institute of Technology, Cambridge, Massachusetts.
88. Godson, W. L. (1950). A study of the deviations of wind speeds and directions from geostrophic values. *Quart. J. Roy. Meteorol. Soc.* **76**, 3–15.
89. Angell, J. K. (1959). Some statistics on the magnitude of the ageostrophic wind obtained from constant level balloon data. *Monthly Weather Rev.* **87**, 163–170.
90. Mintz, Y. (1951). The geostrophic poleward flux of angular momentum in the month of January 1949. *Tellus* **3**, 195–200.
91. Lorenz, E. N. (1955). Geostrophic and ageostrophic poleward flux of angular momentum. Article X in Investigations of the General Circulation of the Atmosphere, Final Report, General Circulation Project, Department of Meteorology, Univ. of California at Los Angeles.
92. Hoel, P. G. (1947). "Introduction to Mathematical Statistics," 258 pp. Wiley, New York.
93. Inoue, E. (1952). On the Lagrangian correlation coefficient for turbulent diffusion and its application to atmospheric diffusion phenomena. *Geophys. Res. Pap. No.* **19**, USAF, 397–411.
94. Hutchings, J. W. (1955). Turbulence theory applied to large-scale atmospheric phenomena. *J. Meteorol.* **12**, 263–271.
95. Tukey, J. W. (1950). The sampling theory of power spectrum estimates. In Symposium on the application of autocorrelation analysis to physical problems, pp. 47–67, Office of Naval Res., Washington, D. C.
96. Panofsky, H. A. (1955). Meteorological applications of power-spectrum analysis. *Bull. Am. Meteorol. Soc.* **36**, 163–166.
97. Gifford, F. (1958). The interpretation of meteorological spectra and correlations. *J. Meteorol.* **16**, 344–346.
98. Van der Hoven, I., and Panofsky, H. A. (1954). Final report on statistical properties of the vertical flux and kinetic energy at 100 meters, 55 pp. Department of Meteorology, the Pennsylvania State Univ., University Park, Pennsylvania.

99. Bjerknes, J. (1951). Extratropical cyclones. *In* "Compendium of Meteorology" (T. Malone ed.), pp. 577-598. Am. Meteorol. Soc., Boston, Massachusetts.
100. Palmén, E. (1955). On the mean meridional circulation in low latitudes of the northern hemisphere in winter and the associated meridional and vertical flux of momentum. Article VII in Investigations of the General Circulation of the Atmosphere, Final Report, General Circulation Project, Department of Meteorol., Univ. of California at Los Angeles.
101. Harmantas, C. (1957). Notes on the world comparison of radiosondes at Payerne, Switzerland, May–June 1956. Unpublished manuscript, U. S. Weather Bureau, Washington, D. C.
102. Richardson, L. F. (1926). Atmospheric diffusion on a distance-neighbor graph. *Proc. Roy. Soc. (London)* **110,** 709–737.
103. Durst, C. S., Crossley, A. F., and Davis, N. E. (1959). Horizontal diffusion in the atmosphere as determined by geostrophic trajectories. *J. Fluid Mechanics* **6,** 401–422.
104. Proceedings of the Seventh Pacific Science Congress of the Pacific Science Association (1949), Division of Meteorology, New Zealand.
105. World Meteorological Organization, Commission on Aerology (1957). Draft Recommendation A-11 (CAe-II/Doc. 79-E, Annex II).
106. Angell, J. K. (1958). An analysis of routine 300 mb transosonde flights from Japan. *Monthly Weather Rev.* **86,** 335–343.
107. Angell, J. K. (1959). The use of transosonde data as an aid to analysis and forecasting during the winter of 1958–1959. *J. Geophys. Res.* **64,** 1845–1853.
108. Mastenbrook, H. J. (1959). Utilization of transosonde data, 10 pp. Naval Res. Laboratory Report No. 5310, Washington, D. C.
109. Arakawa, H. (1957). Case study of erratic movement of well-developed cyclones and upper air circulation over the NW-Pacific. *J. Meteorol. Soc. Japan, 75th Anniversary Vol.*, pp. 187–191.
110. Solot, S. B., and Darling, E. M. (1958). Theory of large-scale atmospheric diffusion and its application to air trajectories, 784 pp. (3 volumes). *Geophys. Res. Paper No. 58,* Air Force Cambridge Res. Center, Massachusetts.
111. Hesselberg, Th. (1915). Über oszillatorische Bewegungen der Luft. *Ann. Hydrog.* **33,** 311–317.
112. Van Mieghem, J. M. (1953). Hydrodynamic instability. In "Compendium of Meteorology" (T. Malone, ed.), pp. 434–453. Am. Meteorol. Soc., Boston, Massachusetts.
113. Holmboe, J., Forsythe, G. E., and Gustin, W. (1945). "Dynamic Meteorology," 378 pp. Wiley, New York.
114. Gustafson, A. F. (1953). On anomalous winds in the free atmosphere. *Bull. Am. Meteorol. Soc.* **34,** 196–201.
115. Gustafson, A. F., (1949). Final report on the upper-level winds project, 7 pp. Department of Meteorology, Univ. of California at Los Angeles.
116. Angell, J. K. (1958). Lagrangian wind fluctuations at 300 mb derived from transosonde data. *J. Meteorol.* **15,** 522–530.
117. Forsythe, G. E. (1949). Exact particle trajectories for non-viscous flow in a plane with constant Coriolis parameter. *J. Meteorol.* **6,** 337–346.
118. Angell, J. K. (1956). The flow pattern at 300 mb as derived from controlled-altitude balloon data, 177 pp. Doctoral dissertation, Department of Meteorology, Univ. of California at Los Angeles.
119. Newton, C. W. (1959). Axial velocity streaks in the jet stream: ageostrophic "inertial" oscillations. *J. Meteorol.* **16,** 638–645.

120. Angell, J. K. (1959). The application of constant level balloon data to weather analysis, forecasting, and research, 69 pp. Unpublished Manuscript, U. S. Weather Bureau, Washington, D. C.
121. Angell, J. K. (1960). An analysis of operational 300 mb transosonde flights from Japan in 1957–58. *J. Meteorol.* **17**, 20–35.
122. Kao, S. K., and Neiburger, M. (1959). Oscillations and trajectories of air parcels in some pressure systems. *J. Geophys. Res.* **64**, 1283–1291.
123. Kao, S. K., and Wurtele, M. G. (1959). The motion of a parcel in a constant geostrophic wind field of parabolic profile. *J. Geophys. Res.* **64**, 765–777.
124. Kao, S. K. (1960). Nonlinear oscillations of a particle in a long wave in a rotating fluid. *J. Geophys. Res.* **65**, 279–286.
125. Office of the Chief of Naval Operations (1952). Practical methods of weather analysis and prognosis. (NAVAER 50-110R-51), 193 pp.
126. Angell, J. K. (1956). Determination of the horizontal divergence along the path of a controlled-altitude balloon. Unpublished manuscript, Department of Meteorology, Univ. of California at Los Angeles.
127. Saucier, W. J. (1955). "Principles of Meteorological Analysis," 438 pp. Univ. Chicago Press, Chicago, Illinois.
128. Mastenbrook, H. J. (1959). Environmental factors affecting altitude performance and control of constant-volume balloons, 35 pp. Naval Res. Laboratory Report No. 5322, Washington, D. C.
129. Emmons, G., Haurwitz, B., and Spilhaus, A. F. (1950). Oscillations in the stratosphere and high troposphere. *Bull. Am. Meteorol. Soc.* **31**, 135–138.
130. Holmboe, J., and Klieforth, H. (1957). Investigations of mountain lee waves and the air flow over the Sierra Nevada. Final Report, Sierra Wave Project, Department of Meteorology, Univ. of California at Los Angeles.

This Page Intentionally Left Blank

PALEOMAGNETISM*

Richard R. Doell and Allan Cox

U.S. Geological Survey, Department of the Interior, Menlo Park, California

1. Introduction

Because of the phenomenon of magnetic hysteresis, the magnetization of ferromagnetic minerals, including those in rocks, depends on the history as well as the present value of the ambient magnetic field. Moreover, the remanent magnetization of some rocks can be shown to have been acquired during a short interval of time early in the rock's history. Thus the geomagnetic field, of all of the force fields of interest to geophysicists, has a unique importance. For while the history of stress and gravitational force patterns in the mantle and crust can at best be inferred only in a very general way, the entire history of the geomagnetic field can theoretically be mapped in great detail by analyzing the remanent magnetizations of suitable rocks.

Results from this comparatively recent field of geophysics have already greatly extended our knowledge of the geomagnetic field. Especially impor-

* Publication authorized by the Director, U.S. Geological Survey.

tant to geophysicists are the findings that the geomagnetic field has had an axial, nontransient dipole component during the past half a million years, that the intensity of the main dipole has been decreasing for several thousand years, and that the field has undergone reversals in polarity. In the present review, emphasis is given to these topics, as well as to a discussion of the ways in which rocks become magnetized. Other topics covered include interpretations of paleomagnetic results in terms of the theories of polar wandering, continental drift, and an expanding earth. For additional discussions of these latter topics, reference is made to the following recent reviews [1–5].

2. Results from Direct Observation of the Geomagnetic Field

2.1. Spherical Harmonic Analysis

Before reviewing the results of paleomagnetic investigations, it will be helpful to recall some of the salient features of the geomagnetic field established from direct observation and analysis. Since Gilbert's classic study in 1600, it has been recognized that during the period of direct observation and analysis, extending back four centuries at several observatories, the earth's field has been approximately dipolar. Comparing the observed magnetic field at a point on the earth's surface with the field of a theoretical geocentric dipole that best fits the observed field over the entire earth, we find that for every epoch for which records are available, there is no sizable region where the two vectors differ in direction by more than 30° or in intensity by more than 50 %, and generally the differences are much smaller. (Small regions where highly magnetic deposits cause large local anomalies are exceptions.)

In a spherical harmonic analysis of the geomagnetic field, the dipole and nondipole components appear as different terms in the series expression for the geomagnetic potential due to sources within the earth, which in contemporary analyses of the geomagnetic field is usually written in the form:

$$V = a \sum_{n=1}^{\infty} \sum_{m=0}^{n} (a/r)^{n+1} \ (g_n^m \cos m\bar{\phi} + h_n^m \sin m\bar{\phi}) P_n^m(\bar{\theta}) \tag{2.1}$$

where a is the earth radius, r is the distance from the earth's center (the origin of coordinates), $\bar{\theta}$ is the colatitude, and $\bar{\phi}$ the east longitude. Here, $P_n^m(\bar{\theta})$ is the spherical harmonic function of Schmidt of degree n and order m. An additional potential series, similar except for the substitution of $(r/a)^n$ for $(a/r)^{n+1}$, corresponds to sources outside the earth. Two important properties of the geomagnetic field have been established by spherical harmonic analysis. First, the field is derivable from a potential and hence contributions to the total field from currents flowing across the earth-air interface are negligibly small. Secondly, 2 to 5% of the field at most is of external origin and is due to movements of charged particles in the space around the earth; these currents cause daily angular deflections of the total field vector of the order of 10 min of arc. During magnetic storms, deflections of the total field vector with periods of the order of one hour usually do not exceed $\frac{1}{2}°$ of arc in amplitude; brief deflections as large as 2° have occurred in only a few large storms. Although these relatively small transient phenomena are of great contemporary interest, their contribution to the grosser morphology of the geomagnetic field and its longer period variations is negligible.

The three terms of degree $n = 1$ in equation (2.1) specify the orientation of a magnetic dipole which has the property of being the geocentric dipole that best fits, in the sense of least squares, all of the field data used in the spherical harmonic analysis. Since the late 19th century, when sufficient world-wide geomagnetic data for accurate spherical harmonic analysis first became available, the orientation of this dipole or, equivalently, the geographical coordinates of the corresponding geomagnetic pole, has probably not changed significantly, as may be seen from Fig. 1. Earlier analyses of fewer data, including the original memoir of Gauss, suggest a possible westward movement of the geomagnetic pole of 3.4° of longitude (0.68° of arc) between 1829 and 1880. Bullard and others [6] have concluded, however, that in view of the lack of movement shown in later analyses and the small amount of data available for the earlier studies, any change in orientation of the dipole during the time covered by analyses is uncertain.

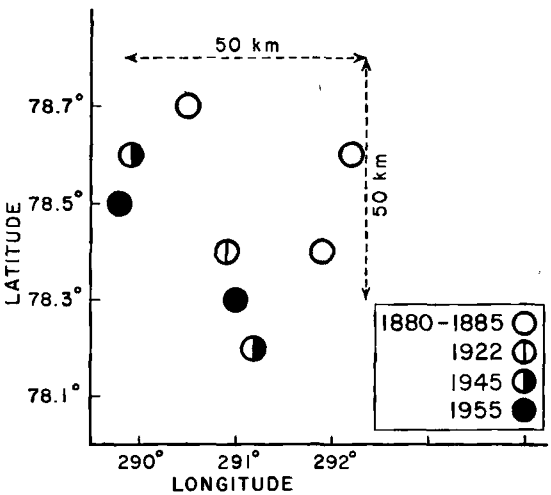


FIG. 1. Geomagnetic pole positions based on spherical harmonic analyses for epochs 1880–1955. (Data from [7–10]).

TABLE I. Intensity M of Inclined Dipole and its Equatorial Field H_0, 1835–1955.

Epoch	M	H_0
1835	8.553	3309
1839	8.449	3269
1845	8.483	3282
1880	8.359	3234
1885 (Fritsche)	8.343	3228
1885 (Schmidt)	8.351	3231
1885 (Neumayer)	8.333	3224
1922	8.165	3159
1945 (Afanasieva)	8.005	3097
1945 (Vestine)	8.062	3119
1955 (Finch)	8.064	3120
1955	8.041	3111
	10^{25} gauss cm^3	10^{-4} cgs units

The moment M of the theoretical geocentric dipole is equal to

$$M = H_0 a^3 \tag{2.2}$$

where a is again the earth's radius and H_0, which is the theoretical dipole field at the geomagnetic equator, is given by

$$H_0 = [(g_1^0)^2 + (g_1^1)^2 + (h_1^1)^2]^{1/2} \tag{2.3}$$

where g_1^0, g_1^1, and h_1^1 are the spherical harmonic coefficients of degree one. Values of H_0 for epochs 1835–1945, calculated by equation (2.3) from the data of Vestine and earlier workers [8], are listed in Table I. The first value of H_0 listed for epoch 1955 is that of Finch and Leaton [9], and the second was found by adding the amount of secular change for the period 1945–55 to the values of g_1^0, g_1^1, and h_1^1 for 1945 [8, 10]. Different values of H_0 for approximately the same epoch are due in part to the use of somewhat different sets of data and in part to methods of analysis which weigh differently the data from various regions. Although the different analyses cannot, for these reasons, be profitably compared in great detail, there can be little doubt (Fig. 2) that during the past century H_0 has decreased at a rate of the order of 15 gammas (1 gamma = 10^{-5} cgs units) per year. A sharp decrease in the rate of decrease in recent decades appears possible but far from certain.

The nondipole component of the geomagnetic field, represented by spherical harmonic coefficients of degree $n > 1$, consists physically of irregularly distributed regions of high and low field intensity which range in diameter from 25° to 100°. Observed rates of change of the nondipole field suggest an average life for an individual cell of the order of 100 years. The pattern of the nondipole field, although constantly changing, shows a striking drift in a westward direction. Bullard [6] finds that the rate of this westward drift is 0.18° of longitude per year in all latitudes. At the equator the nondipole features are thus moving at the remarkably rapid rate of 2000 km per century. The great difference in the rates of change of the dipole and nondipole field components suggests that they may have their respective origins in distinguishable phenomena.

The secular variation or time derivative of the total field is due predominantly to changes in the nondipole field. Maps of secular variation show broad regions of increase and decrease with dimensions similar to those of the nondipole field. Part of the observed secular variation is due to movement of the nondipole field with respect to the mantle and part to the growth and decay of features of the nondipole field [11]. The rate of westward drift of isoporic contours showing equal rates of change has been estimated by Bullard [6] to be 0.3° of longitude per year.

2.2. Origin of the Earth's Magnetic Field

Until very recently, theories about the origin of the earth's magnetic field have been largely theories of desperation. For example, the theory that the earth's field is due to ferromagnetic minerals in the earth, which was widely accepted at one time, encounters several serious difficulties. The Curie point temperatures of minerals in rocks and of iron and nickel are all below 800°C, a temperature reached at a depth probably not much greater

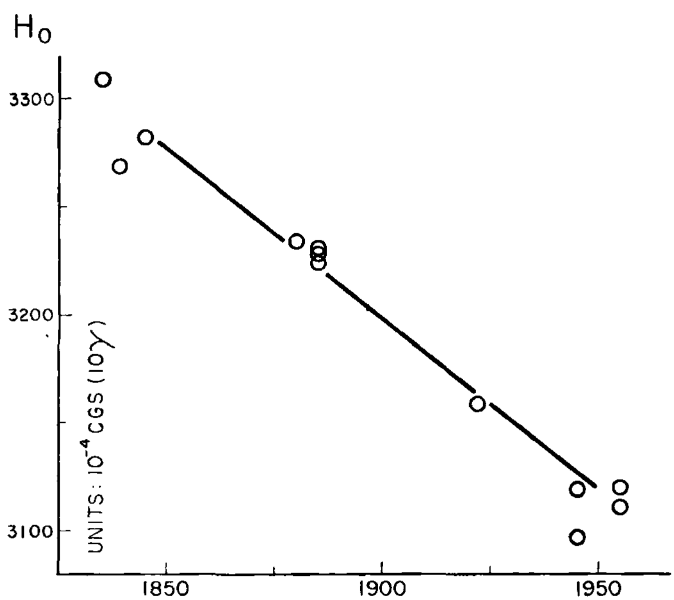


FIG. 2. Equatorial field of inclined geocentric dipole, 1835–1955.

than 25 or 30 km. To account for the magnetic moment of the earth, a crust 30 km thick would have to have a remanent magnetization of 5 emu/cc. However, the maximum saturation remanent magnetization that can be given to most igneous rocks, even those with abundant magnetite, usually ranges between 0.5 and 2.5 emu/cc, and natural remanent magnetizations of igneous rocks are usually between 0.001 and 0.05 emu/cc. Attempts to escape from this dilemma by discovering an increase of Curie point temperature with depth have failed. Not only does this theory fail to explain the main dipole field, but it also is unable to explain the rapid nondipole variations; transport of magnetic material within the crust and mantle or alternatively, movement through the crust of isotherms corresponding to the Curie point temperature of ferromagnetic minerals, at velocities of 20 km per year is very improbable. Jacobs [12] gives an excellent survey of this and other geomagnetic field theories.

A major advance in understanding the origin of the geomagnetic field was made with the self-exciting dynamo theory of Larmor, Elsasser, and Bullard. This theory, which attributes the geomagnetic field to induction effects accompanying fluid motions in a conducting but nonmagnetic core, has been extensively reviewed by Elsasser [13, 14], Bullard [15], and Jacobs [12], and only a few salient features of the theory will be recalled here. The theory requires, in addition to the existence of an electrically conducting fluid, a source of energy to keep the fluid in convective motion. Heat from the decay of radioactive elements in the earth's core was originally thought to be a likely source of energy for convection; however, if the abundance of radioactive elements in the core is similar to that in iron meteorites, as has been frequently argued, then the more recent chemical analyses of meteorites indicate this source is insufficient. Urey [16] has suggested crystallization or chemical change as an energy source, and Verhoogen [17] points out that heat provided by solidification of the inner core would be sufficient to maintain convection during the past several billion years.

The dynamo theory also requires some process that will impart order to the otherwise random convective movements. In the case of the earth, sun, and other rotating bodies with magnetic fields, this ordering is thought to be the result of the Coriolis force. Because of the essential role of rotation in this model, the resulting magnetic field should, in general, have axial symmetry about the rotation axis; the average declination should be zero and the average inclination should be a function only of latitude [18].

The rapid changes in the nondipole field and its westward drift find a simple explanation in the dynamo theory. The large, irregular highs and lows of the nondipole field reflect the occurrence of fluid eddies near the surface of the core; the core's high electrical conductivity would screen deep-seated inductive processes occurring at rates as rapid as those observed in the nondipole field. The westward drift is due to a magnetic coupling between the conductive core and the much less conductive mantle which effectively causes the outer part of the core to rotate more slowly than the mantle and hence move relatively westward. The vastly differing rates of movement and change of the dipole and nondipole fields may be due to the generation of the dipole component by fluid motions deeper in the core, where differential motions between core and mantle are presumably smaller.

2.3. Unsolved Problems

For every generalization about the geomagnetic field established during the short span of direct human observation several questions remain unanswered. The more important properties of the earth's field might be summarized as follows:

1. For several centuries the geomagnetic field has been predominantly dipolar.
2. The orientation of the geocentric dipole that best fits the observed field has, during the period covered by adequate measurements, remained fixed with respect to the earth's crust, lying in the plane of the 110th meridian and inclined 11½° from the rotation axis.
3. The intensity of the inclined geocentric dipole has decreased at an average annual rate of 1 part in 2000 during the past 120 years.
4. The nondipole field changes more rapidly than the dipole field, both as measured by rates of change of the respective spherical harmonic coefficients and as measured by relative rates of change of the intensity and direction of the dipole versus nondipole field vectors at the earth's surface.
5. The nondipole component exhibits a general westward displacement with respect to the mantle at an average rate of 0.18° per year.

Many important unanswered questions about the earth's field are related to the problem of how far the trends of the past century or two may be safely extrapolated:

1. For how long a time has the intensity of the inclined geocentric dipole been decreasing at its present rate?
2. Has the orientation of the inclined geocentric dipole always remained fixed with respect to the crust or does it move like the nondipole component but at a much slower rate? When averaged over a longer period of time, does the dipole axis coincide in direction with the rotation axis, as suggested by the dynamo theory?
3. Has the geomagnetic field always been predominantly dipolar? The present predominance of the dipole term at the surface of the earth is in part an accident depending on the distance from the surface to the mantle-core boundary; Elsasser [14] shows that as we approach the center of the

earth, the ratio of the rms nondipole field to the rms dipole field increases from a value of 0.05 at the earth's surface to a value of about 1 at the core-mantle boundary. If the average intensity of the nondipole field were to remain unchanged for several thousand years while the dipole component decreased at its present rate, the resulting field would become much more irregular in shape and would change in direction more rapidly. Although the relative motion of the core and mantle may cause all components without symmetry about the rotation axis to cancel when averaged over a sufficiently long time [18], a decrease in the dipole field might well result in the dominance of nondipole components with axial symmetry, corresponding to a series of spherical harmonic coefficients of order $m = 0$ but with degree $n > 1$.

3. Magnetization of Rocks

3.1. Origin of Natural Remanent Magnetization

Remanent magnetizations acquired by ferromagnetic minerals in the laboratory bear a systematic relationship to the ambient magnetic fields during magnetization, and this presumably is true also of the natural remanent magnetizations found in rocks. The possibility of learning something about the prehistoric geomagnetic field from rocks has long been recognized by geophysicists, but until comparatively recently two obstacles have stood in the way. The comparatively minor problem of instrumentation has now been solved quite satisfactorily, and perhaps a dozen instruments of both the astatic and the spinner or rock generator type, capable of measuring the remanent magnetizations of all rocks likely to be of interest in paleomagnetic studies, have been constructed in various laboratories. Much more intractable has been the problem of understanding the processes by which rocks become magnetized. Do rocks acquire their remanent magnetizations continuously during their entire histories and thus present an integrated picture of the past geomagnetic field, or do the more important magnetizing processes occur during brief intervals of time? When, if ever, is it safe to assume that the direction of the natural remanent magnetization found in a rock is parallel to the geomagnetic field at the time of formation of the rock? What is the relationship between intensity of remanent magnetization and intensity of the ancient geomagnetic field? Complete answers to all these questions have not yet been found, but much progress has been made in understanding many of the important processes by which rocks become magnetized, thanks to the extensive field and laboratory investigations by Pockels [19], Koenigsberger [20], Chevallier [21], Thellier [22–24], Nagata [25–28], Graham [29–31], Balsley and Buddington [32, 33], Petrova [34, 35], Uyeda [36], Akimoto [37], Haigh [38], Kobayashi [39], Kawai [40, 41], and others. Many techniques are now available for testing whether the remanent magnetization found in a rock is indicative of the direction or intensity of the geomagnetic field at some known time in the past.

At least five distinct processes by which rocks may acquire remanent magnetization in nature are now known. Although very commonly the remanent magnetizations are exactly parallel to the ambient magnetic field, this is not always true. Moreover, since the rocks of greatest paleomagnetic interest may be very old, sufficient time has been available for processes with rates so slow as to be scarcely detectable in the laboratory and for processes which occur very infrequently to have taken place. Since the superposition of several components of magnetization acquired at different times is also quite possible, techniques for analyzing natural remanent magnetizations into components corresponding to separate, well-defined magnetizing processes are of great importance in paleomagnetism. Equally important is the development of criteria for determining which of the magnetic components gives reliable information about the geomagnetic field in the past.

Before reviewing these paleomagnetic problems, a word should be said about the importance of remanent magnetization in another geophysical application, the interpretation of magnetic anomaly maps. Until recently it has generally been assumed in geophysical exploration that the magnetization causing anomalies is *induced magnetization*. By definition, induced magnetization exists only in the presence of an applied field. Except for anisotropic materials it is parallel to the applied field and in weak fields proportional to it, the proportionality constant being the susceptibility; occasionally the susceptibility is further assumed to be proportional to the amount of ferromagnetic ore mineral present. However, as more and more measurements of susceptibility and remanent magnetization are made on a wide variety of lithologic types, it becomes increasingly clear that this assumption is not always justified. Remanent magnetizations are frequently larger than induced and moreover may not be parallel to the present geomagnetic field. Especially important is the fact that the amount of remanent magnetization per unit volume of a given magnetic mineral may vary by more than a factor of 10, depending on such factors as rates of cooling, impurities, grain size, and time elapsed since magnetization; moreover, the variation between different mineral species may be even greater. Thus the assumption that the magnetization causing an anomaly is parallel to the present field or that the intensity of magnetization is proportional to the per cent of ferromagnetic mineral present may lead to serious interpretive errors.

3.2. Types of Natural Remanent Magnetization

The natural remanent magnetization of rocks would be quite simple if each rock contained only one ferromagnetic mineral occurring in grains of identical size and shape with all crystallographic axes parallel. In actuality, rocks often contain several ferromagnetic minerals occurring in grains of widely varying size and shape; crystallographic axes may or may not be randomly oriented; chemical composition may vary from grain to grain, as may the extent of lattice imperfections. Many magnetic properties, especially the important property of coercive force, are affected by these factors, and as Graham [30] pointed out, to describe the coercive properties of most rocks one must specify, not a coercive *force*, but a coercive *force spectrum.*

The following rather generalized model based on a rock containing single domain grains will be used in describing the different magnetizing processes; the theory of multidomain systems is much more complex, but most of the experimental results are the same. The smallest unit of magnetization will be understood to be the elementary Weiss domain. The intensity of magnetization within the domain, termed the "spontaneous magnetization" J_s, is uniform throughout all domains of the same composition; J_s decreases with increasing temperature, and vanishes at the Curie point temperature. Within each domain the direction of magnetization is, in the absence of an applied field, parallel to a so-called "easy direction of magnetization" characteristic of the domain and determined by its shape or the orientation of its crystallographic axes. Separating easy directions of magnetization are barriers which may be characterized by either of two quantities: the coercive force H_c, which is the magnetic field necessary to change the direction of magnetization from one easy direction to another; or the energy $v\ J_sH_c/2$ needed to surmount an energy barrier specified by coercive force H_c, spontaneous magnetization J_s, and volume v.

3.2.1. Isothermal Remanent Magnetization. Magnetization acquired at constant temperature in a steady magnetic field is termed "isothermal remanent magnetization" (IRM). For macroscopically isotropic materials, including aggregates of randomly oriented anisotropic crystals, IRM is always found experimentally to be parallel to the applied field. Isothermal remanent magnetization is the remanence of the familiar hysteresis curve used to describe magnetic properties of materials and, although it is perhaps the simplest type of remanent magnetization, its structure can be as complex as the history of the ambient magnetic field.

After a rock has been placed briefly in a dc magnetic field H_1, all domains with coercive forces $H_c < H_1$ will be aligned along the easy direction of magnetization in each domain nearest the direction of H_1; if the easy directions are randomly oriented, the vector sum of many such domains will be essentially parallel to H_1. If subsequently the rock is placed in a field $H_2 < H_1$, where H_2 is not parallel to H_1, then all domains with coercive forces $H_c < H_2$ will constitute a component of IRM directed along H_2, while those for which $H_2 < H_c < H_1$ will still be directed along H_1. Thus, to completely describe the IRM in the sample we must know not only the total direction and intensity, but also how the magnetization is distributed throughout the coercive force spectrum. An IRM concentrated in domains with low coercive force can easily be reoriented by weak disturbing fields, while IRM of the same intensity concentrated in high coercive force domains will be very stable.

The intensity of IRM acquired by most rocks in the laboratory in fields as weak as the earth's is very small, in many cases too small to measure. However many rocks have very strong natural remanent magnetizations which, on the basis of extensive laboratory analysis, appear almost certainly to be IRM acquired in fields several thousand times stronger than the earth's magnetic field today. Recent detailed investigations by Momose *et al.* [42], Schmucker [43], Khan [44], and Cox [45] confirm the views of earlier workers that these unusually strong IRM's were acquired in the large fields associated with lightning discharges, with peak currents ranging between 10,000 and 200,000 amp.

3.2.2. Thermoremanent Magnetization. Of very great importance in paleomagnetic studies is the remanent magnetization that minerals acquire on cooling from their Curie point temperatures in weak fields of the order of one oersted. Three salient characteristics of thermoremanent magnetization (TRM) have emerged from thousands of experiments. (*1*) In equidimensional isotropic materials, TRM is exactly parallel to the ambient magnetic field during cooling. (The only exceptions are the rare minerals which acquire TRM exactly opposite to the applied field, discussed in Section 5.3.) (*2*) TRM is extremely stable, even when acquired in fields as weak as the earth's. Rimbert [46] reports the occurrence of natural TRM that is not destroyed in demagnetizing fields up to 900 oe, and the present authors have extended these experiments up to 2100 oe and find a substantial amount of TRM still present. (*3*) The intensity of the TRM acquired by rocks in fields as weak as the earth's is commonly only several per cent of the limiting saturation IRM acquired isothermally in strong fields.

An elegant theoretical treatment explaining these observations has been developed by Néel [47] for single domain grains, and it has been extended to the much more complex case of multidomain grains by Verhoogen [48]. In the single domain theory, the spontaneous magnetization J_s appears at the Curie point temperature and increases on further cooling. At temperatures just below the Curie point temperature the thermal energy, KT, is comparable with the magnetostatic energy vhJ_s of a domain of volume v and spontaneous magnetization J_s in a weak field h. Thus the magnetization moves readily from one easy direction to another. However, h establishes a slight bias in favor of domains parallel to it, and the average or equilibrium magnetization at any temperature is proportional to (tanh hvJ_s/KT). A *relaxation time* τ may be defined as the time interval during which the thermal energy will cause the magnetization of a domain to cross a magnetic energy barrier with probability $1/e$. In Néel's model,

$$\frac{1}{\tau} = A\left(\frac{v}{T}\right)^{1/2} \exp\left(\frac{-vH_cJ_s}{2KT}\right) \tag{3.1}$$

where A is a constant characteristic of the mineral, v is the volume of the domain, H_c its coercive force, J_s its spontaneous magnetization, and T is the temperature. As T decreases, H_c and J_s increase and, because of the exponential relationship, there is an extremely narrow temperature interval within which τ increases very steeply and effectively freezes the equilibrium magnetization. This temperature is termed the "blocking temperature" of the domain, and since v and other properties vary from domain to domain, the blocking temperatures also vary. Thus for most rocks, the range over which TRM is acquired extends at least some tens of degrees below the Curie point temperature. On further cooling below the blocking temperature, τ continues to increase, so that at room temperature it is long even on the geologic time scale. The theory thus explains how weak fields of the order of 1 oe can give a preferential alignment to domains with coercive forces of several thousand oersted. It is important to note that all domains frozen in each temperature interval are not parallel to the weak field h, but only a number proportional to the quantity tanh hvJ_s/KT, which, for small values of h, is simply proportional to h.

In his theory for the TRM of multidomain grains Verhoogen [48] indicates that less stable components of TRM may be of multiple origin. The most stable part of the TRM, which is of greatest interest in paleomagnetism, probably resides in strained regions surrounding dislocations, and these regions behave essentially as single domain particles for which the theory outlined above is applicable.

A final remarkable property of TRM of considerable interest in paleomagnetic research, especially in field intensity studies, is the law of additivity of partial thermoremanent magnetization described by Nagata [49] and Thellier [22]. If, on cooling a rock from its Curie point temperature T_c, a weak field h is applied only over the temperature interval T_2 to T_1 with zero field applied at all other temperatures, the resulting magnetization is known as the partial thermoremanent magnetization (PTRM). Extensive experiments by Nagata [49] and Thellier [22] show that the PTRM acquired in any temperature interval depends only on the field applied over that temperature interval and is independent of the field at all other temperatures. Therefore, the total TRM acquired on cooling from the Curie point temperature T_c to room temperature T_0 in a steady field h is equal to the sum of all PTRM's acquired in temperature intervals T_c to T_n, T_n to T_{n-1}, $\cdots$, T_1 to T_0.

In Thellier's words, the rocks preserve a magnetic memory of each temperature interval: "*Vis-à-vis des élévations de température, l'A.T.R. partielle normale acquisé en champ faible, est remarquable par le fait de la mémoire magnétique: une élévation de température de T_2 à T_1 détruisant spécifiquement le moment acquis pendant le refroidissement initial de T_1 à T_2.*" (With respect to increases in temperature, 'the magnetic memory' of normal partial TRM is a remarkable fact: an increase in temperature from T_1 to T_2 destroys only the magnetization that was originally acquired in cooling from T_2 to T_1.)

This remarkable additive property of TRM is also explained by the single domain theory. The PTRM in each temperature interval T_2 to T_1 resides in a distinct group of domains having blocking temperatures between T_2 and T_1. At temperatures greater than T_2, the relaxation times of domains in this group will be very short and the equilibrium magnetization will be able to follow variations in the applied field; therefore, the history of the applied field at temperatures above T_2 does not affect the magnetization acquired by this group of domains. At temperatures below the lower blocking temperature T_1, the relaxation times in this group will be long with respect to the duration of the laboratory experiments and changes in the field will not be reflected by changes in magnetization. The additivity relationship of PTRM would probably be somewhat blurred if either the individual temperature intervals were extremely small or if the rate of cooling were extremely slow.

3.2.3. Depositional Magnetization. Sediments consisting of magnetic and nonmagnetic particles acquire a remanent magnetization when deposited in the laboratory in a weak field [25, 50–52]. In contrast with IRM and TRM, however, this depositional magnetization may not be exactly parallel to the applied field. King [52] finds that if the surface of deposition is level and the applied field is neither perfectly vertical nor horizontal, then the declination of the depositional magnetization agrees with that of the applied field but the inclination may be 20° to 30° smaller. If the surface of deposition is not level, both declination and inclination may be changed. Additional deviations are caused by bottom water currents.

Answers to a number of questions about depositional magnetization await future investigations. Perhaps the most important is the effect of changes in the applied field after deposition. Clegg *et al.* [51] found that sediments deposited in zero applied field and then placed in a weak field acquired a remanent magnetization as long as the water content of the sediments was greater than 50% by weight; King [52] found that changes in the applied field after deposition of sediments containing 45% water did not change the magnetization. Thus the mechanism by which depositional magnetization is acquired, and possibly the importance of inclination and declination errors such as those found by King, may depend on the rate at which sediments are consolidated. Griffiths *et al.* [53] have recently published the full results of their extensive research on this subject, and the reader is referred to this paper for more details.

3.2.4. Crystallization or Chemical Magnetization. Sediments may contain ferromagnetic minerals formed after deposition either by crystallization from circulating solutions or by alteration of iron-rich mineral grains present in the original sediment. Martinez and Howell [54] and Doell [55] found several sedimentary formations with natural remanent magnetizations due to post-depositional crystallization or chemical magnetization, and subsequently Haigh [38] and Kobayashi [39] conclusively demonstrated the process in the laboratory. Crystallization or chemical magnetization (CRM) has magnetic properties similar to those of TRM, and of special importance in paleomagnetism is its stability when produced in weak magnetic fields. A substantial part of CRM produced in the laboratory in weak fields is not destroyed in fields up to 500 oe.

A theoretical treatment of CRM is already available in Néel's single domain theory of TRM, as pointed out by Haigh [38]. During chemical magnetization T remains approximately constant and the volume v of individual grains increases. The relaxation time of each grain thus increases almost exponentially with increasing volume [equation (3.1)], and the equilibrium magnetization is effectively frozen at a critical blocking diameter which is analogous with the critical blocking temperature in the theory for TRM.

3.2.5. Viscous Magnetization. Only a very small IRM is acquired by most rocks in magnetic fields as weak as the earth's. Over long intervals of time, however, rocks acquire a magnetization, termed "viscous magnetization" by Thellier [56], which is proportional to the logarithm of the time and is parallel to the weak applied field. Viscous magnetization has its origin in the thermal energy which, because of the Boltzmann distribution, consists in part of energy levels high enough to realign the directions of magnetization of even those domains with rather high-energy barriers. As was the case for TRM, the weak field of the earth acts to bias the direction of these jumps.

The stability and intensity of viscous magnetization depends on the strength of the applied field as well as the time spent in the field. Rimbert [57] found that the alternating magnetic field $\approx\!\!\!\!H_d$ necessary to destroy viscous magnetization acquired by extrusive volcanic rocks followed the relationship:

$$(3.2) \qquad \overset{\approx}{H}_d = -100 + 75 \log h + 10\,(2 + \log h) \log t$$

where h is the strength of the applied dc field in oersteds and t is the time in seconds during which the viscous magnetization was acquired. Here, $\tilde{\tilde{H}}_d$ is the root mean square (rms) demagnetizing field. If these results are applicable to longer intervals of time, the alternating fields necessary to destroy viscous magnetization acquired in a field of ½ oe in ½ million, 1 million, 100 million, and 1000 million years are, respectively, 101, 106, 141, and 157 oe (all rms). Brynjólfsson [58] and Cox [45] have found that viscous magnetizations acquired by volcanic rocks in the earth's field over roughly the past ½ million years are destroyed in alternating fields with peak intensities of 50 to 100 oe (or 35 to 70 oe rms). Irving [59] finds that viscous magnetizations in volcanic rocks about 250 million years old are destroyed in fields of 210 oe peak or 148 oe rms. The evidence available thus indicates that viscous magnetization acquired in weak fields is comparatively unstable, and that even in old rocks it is removed by magnetic fields that destroy less than half of the TRM or CRM acquired in the same weak fields.

3.2.6. Magnetization due to Magnetostriction. Since many rocks have complex stress histories, an important problem in paleomagnetism is the determination of whether stress may substantially affect the natural remanent magnetization measured in rocks. Graham [31, 60], the first paleomagnetist to be concerned with the problem of magnetostriction, applied uniaxial stress up to 186 kg/cm^2 to a variety of rocks and measured the remanent magnetizations while the rocks were stressed, but in zero applied field. The directions and intensities of magnetization changed while the rocks were under stress, and on release of the stress most but not all of the magnetizations returned to their original values. Powell [61] has applied uniaxial stresses up to 636 kg/cm^2 to volcanic rocks in magnetic fields of 0.6 to 2.4 oe directed along the stress axes; magnetizations were measured before and after stressing. The response of different samples from the same formation was variable, but all appear to have acquired a component of magnetization parallel to the applied field.

A second group of experiments is in apparent conflict with the above. Stott and Stacey [62] reasoned that the history of many intrusive rocks is one of cooling while under stress at depth followed by relief of stress as the rock becomes exposed at the surface. They cooled samples from above their Curie point temperatures while under constant axial stress and measured the remanent magnetization at room temperature after removal of the stress. For the entire range of stresses used, 0 to 350 kg/cm^2, the magnetizations acquired were always parallel to the applied field.

Kern [63], in his theoretical treatment of magnetostrictive remanent magnetization, indicates how these apparently contradictory experiments may be reconciled. During magnetostriction, the domains with lowest magnetic energy barriers should be most easily realigned by stress-induced effects. Magnetostriction will therefore probably not change the TRM component of magnetization, which is concentrated in domains with high coercive force, rather it will add a new component concentrated in domains with low coercive force.

Many occurrences of natural remanent magnetization are consistent with this theory. For example, Cox [64] measured the direction of magnetization of samples from different columnar joints radiating from the center of an ellipsoidal pillow in a submarine basalt flow. It may be reasonably inferred that the stresses had different orientations in different parts of the pillow and that they approached the maximum tensile stress attainable in basalt, since the pillow was fractured. The original directions of magnetization in different parts of the pillow were highly variable; but after partial demagnetization, which randomized the less stable components, all the directions of magnetization were closely parallel. The original magnetization was clearly a superposition of an unstable component with varying direction and a stable component with the same direction throughout the pillow.

3.3. *Methods for Establishing Paleomagnetic Reliability*

From the knowledge we have of magnetizing processes, it is clear that the natural remanent magnetization measured in rocks is not always parallel to the geomagnetic field present at the time rocks were formed. In some cases, the original remanent magnetization may not have been parallel to the geomagnetic field, and in others one or more magnetization processes may have subsequently added a large secondary component of magnetization not parallel to the original field. Special laboratory techniques and criteria for assessing paleomagnetic reliability are therefore an essential part of every careful paleomagnetic investigation.

3.3.1. Parallelism of Magnetization within a Formation. One criterion for the absence of a secondary component of magnetization is parallelism of directions of magnetization throughout a large rock unit, provided the average direction is not parallel to the present field direction. The maximum distance apart that samples can be collected for a paleomagnetic study from one geologic formation rarely exceeds 100 miles, and changes in direction of the present geomagnetic field usually do not exceed several degrees over this distance. A large angular variation of directions of magnetization in such a unit is therefore indicative of the presence of varying amounts of secondary magnetization. Conversely, parallelism of directions of magnetization throughout a large rock unit reduces the probability of late, secondary magnetization. This criterion for the absence of a secondary component is strengthened if factors affecting either the intensity of the original magnetization or the direction and intensity of a possible secondary

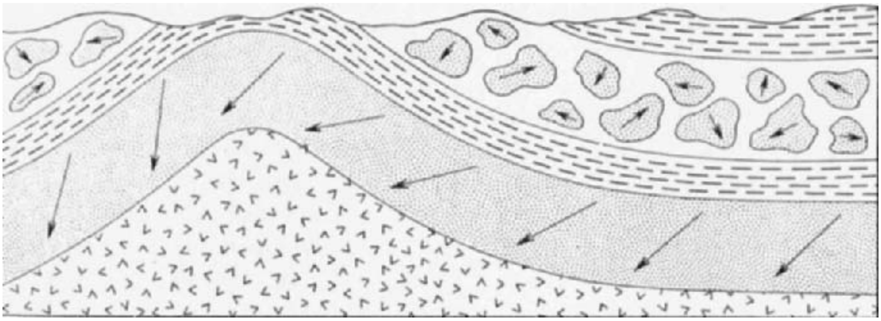


FIG. 3. Graham's fold and conglomerate tests, indicating that magnetizations were originally parallel in the stratum and are magnetically stable in the pebbles.

magnetization are not uniform throughout the formation. For example, a typical baked contact between igneous and sedimentary rocks is more magnetic than nearby unbaked sediments by several orders of magnitude, and usually contains ferromagnetic minerals different from those in the igneous rock. It is thus very likely that both received a TRM at the time of cooling. Parallelism of the respective directions of magnetization indicates a lack of a secondary component of magnetization as it is extremely unlikely that both could have identical secondary components. Thus, extensive sampling throughout a large body of rock exhibiting variation in the size of ferromagnetic grains, their composition, their shape, or their cooling or stress histories leads to a much more reliable paleomagnetic datum than does limited sampling from homogeneous material.

The natural remanent magnetization of many rock units is, however, approximately parallel to the present direction of the geomagnetic field, raising the question of whether their magnetization is a reliable indication of an ancient geomagnetic field direction or merely a "soft" component acquired comparatively recently. Two methods for detecting large soft magnetizations have been used in paleomagnetic studies. If the directions of magnetization of fragments of rock in a conglomerate are parallel within each fragment but are randomly oriented between fragments (Fig. 3), then a large, soft component parallel to the present field has not been added in the fragments and, presumably, not in the parent body; this criterion is sometimes referred to as "Graham's conglomerate test." Extremely soft components can also be detected by remeasuring samples after storage in the present geomagnetic field (Thellier's test).

In rocks which have been deformed, Graham's classical fold test [29] may be used to determine whether the directions of magnetization were parallel before deformation. The test is simple but has great significance. If a body of rock originally acquired a remanent magnetization parallel throughout, and if subsequently different parts of the body were deformed by quasi-rigid-body rotations about one or more axes not all parallel to the original magnetization, then the resulting *in situ* magnetizations will not be parallel (Fig. 3). The amount of rotation or folding and the orientation of the fold axes may sometimes be found using standard geologic techniques, and each part of the body may be rotated back to its original orientation. If the magnetic directions all come into agreement, then the directions of magnetization must originally have been parallel throughout the body; moreover, no secondary components of magnetization can have been added subsequent to the deformation. An example of an application of this test is shown in Fig. 4.

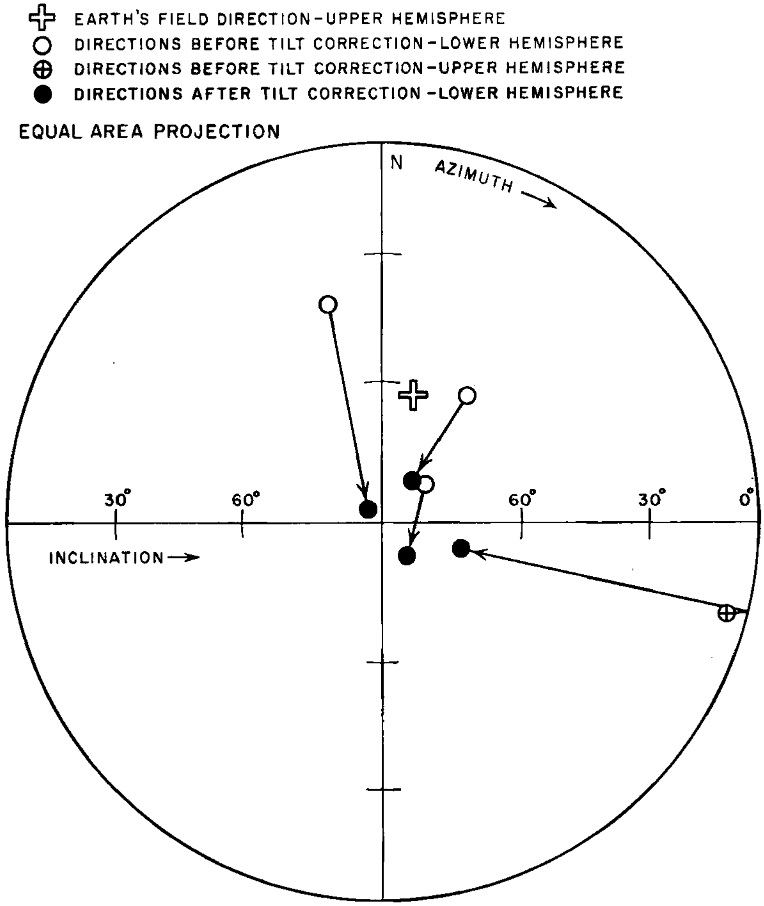


FIG. 4. Directions of magnetization in folded strata before and after correction for post-depositional rotations. (Data from Irving [65].)

The consistency-of-reversals test is also a useful criterion of whether the original magnetization was parallel throughout a formation. This test makes use of the fact that the directions of magnetization of a suite of rock units frequently fall into two distinct groups 180° apart. Two theoretical explanations for these "reversals" will be discussed in a later section; it is sufficient for the present purpose to point out that whichever explanation of the origin is applicable to a particular suite of rocks, the original magnetizations of the "reversed" and "normal" groups very probably were 180° apart. Subsequent addition of a secondary component is indicated if normal and reversed groups are not now 180° apart, and conversely, finding normal and reversed groups in the same formation with closely opposing magnetizations indicates the absence of a secondary magnetization. This test is very powerful because it does not depend on the intensity or direction of possible secondary magnetizations and will also detect a secondary magnetization that is uniform throughout the formation.

The criteria discussed in this section are primarily designed to establish whether the natural remanent magnetizations measured in rocks were originally parallel throughout a geologic unit, and in particular, whether a secondary component is present.

3.3.2. Parallelism of Remanent Magnetization and Ancient Geomagnetic Field. Even after parallelism within a formation has been established two very important questions remain: (*1*) Is the direction of the natural remanent magnetization parallel to the ancient field that produced it? (*2*) At what point in the history of the rock was the magnetization acquired?

One approach to these problems involves the study of rocks which acquired their remanent magnetizations simultaneously but which have different mineralogies or were magnetized under different conditions. A very good example of this criterion is in the congruency of directions of magnetization in a baked sedimentary rock and the baking igneous rock that was discussed above. It would be very unlikely, in such a case, that the two magnetizations could be acquired in a direction other than that of the geomagnetic field at the time they cooled.

These questions may more often be approached by analyzing the natural magnetization into components, each corresponding unambiguously to a clearly understood process which occurred at a known time during the history of the rock. As previously mentioned, abundant experimental evidence indicates that some magnetizing processes common in nature result in extremely stable magnetizations parallel to the original applied field, while others are much less stable and tend to follow the direction of a changing field. These tests are therefore concerned with the coercive properties of the measured magnetizations.

Early paleomagnetic studies used the coercive force of the usual hysteresis curve, and more recently the *remanent coercive force* (Uyeda's H_{cr}) [36]

has come into use. The H_{cr} is the value of the dc field required in order that the remanent magnetization be equal to zero when the field is removed. As previously discussed, remanent magnetization may reside in domains with different coercive forces, and therefore, H_{cr} indicates only that there are sufficient domains with coercive force less than H_{cr} to balance those carrying the original remanent magnetization with coercive forces greater than H_{cr}. This test does not tell us how the stabler and paleomagnetically important part of the remanent magnetization is distributed.

In a variation of this test used by Creer *et al.* [66], the rock sample is first given its saturation remanent magnetization in a strong field; the *coercivity of remanence* is the opposing field necessary to produce zero total remanent magnetization. The coercivity of remanence is, in effect, the median point of the coercive force spectrum. It is thus a property of the rock and is not used as a means of analyzing a remanent magnetization of unknown origin.

In contrast with the above one-parameter coercive force tests for stability, partial demagnetization experiments give very detailed information about how the total remanent magnetization is distributed among different parts of the coercive force spectrum. The two partial demagnetization processes used in paleomagnetic analysis, heating and demagnetizing in alternating fields, were largely developed by Thellier's group [23, 24] and by Nagata's group [25–28]. (Examples of the application of these techniques in paleomagnetic studies are given in the following references [42–45, 67–70].

In alternating field demagnetization experiments, the sample is placed in an alternating field with peak intensity $\overset{\approx}{H}_1$, and the field is then slowly reduced to zero. With suitable precautions the process destroys, by randomization, all magnetic components with coercive forces less than $\overset{\approx}{H}_1$. Thus, by repeating the experiment with higher peak values of the ac field one may determine the intensity and direction of a component of magnetization with any given range of coercivity.

In the other commonly used technique of partial demagnetization, a rock sample is heated in a series of steps to its Curie point temperature. The sample is then cooled in field-free space and its remanent magnetization measured after each increase in temperature. The amount of magnetization remaining at a given temperature resides in domains with relaxation times which, at that temperature, are long with respect to the duration of the experiment. As may be seen from equation (3.1), this depends primarily on the magnetization energy, $vJ_sH_c/2$, rather than on the coercive force alone as in ac field demagnetization. However, partial thermal demagnetization resembles partial demagnetization in alternating fields by selectively destroying the less stable components of magnetization.

From the experimental evidence now available, the demagnetization curves for TRM and CRM appear to be quite similar. Both indicate a large proportion of stable magnetization, and it would be difficult to distinguish CRM from TRM on the basis of demagnetization experiments. Distinguishing either of them from less stable IRM or viscous magnetization should generally be possible, however, and IRM can probably also be distinguished from viscous magnetization.

The origin of the natural remanent magnetization of a rock can sometimes be determined by comparing the demagnetization curves of the natural remanent magnetization with curves for the same rock after it has first been demagnetized and then remagnetized by known magnetizing processes. Figure 5 shows the demagnetization curve of the natural remanent magnetization of a lava sample (sample 1) as well as the demagnetization curve after this same sample had been demagnetized and then placed in a dc magnetic field of 300 oe. The congruence of the curves above 50 oe suggests that most of the natural remanent magnetization is IRM. From a detailed study of 42 specimens collected very near to this sample Cox [45] has shown that this sample was very probably magnetized by a lightning discharge. The demagnetization curve for sample 2, which was collected in an undisturbed part of the same flow, shows comparatively greater sta-

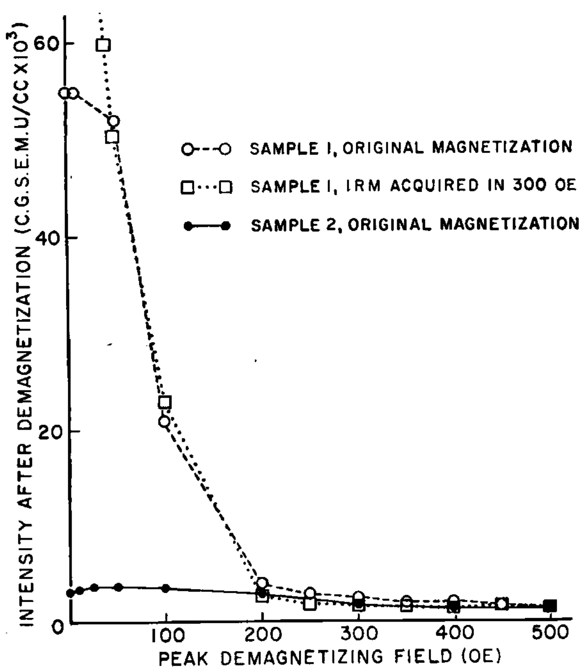


FIG. 5. Alternating field demagnetization curves for two basalt samples; sample 1 was near lightning discharge. (Data from Cox [45].)

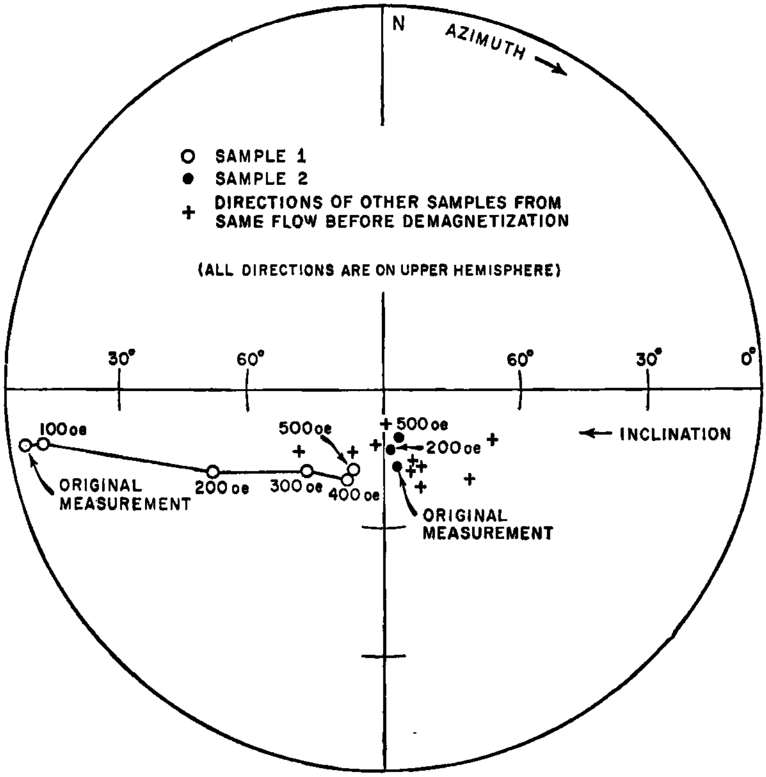


FIG. 6. Directions of magnetization of the two basalt samples of Fig. 5 during ac demagnetization.

bility in low fields, a typical characteristic of TRM. This sample has a direction about 180° away from that of the earth's field today and the initial increase in intensity is due to the removal of a small viscous component parallel to the present field. A comparison of Figs. 5 and 6 clearly points to the presence of two components of magnetization in sample 1, an unstable IRM superimposed on a stable TRM that is parallel to the TRM in undisturbed parts of the same lava flow. Similar demagnetization curves have been reported by other workers [42–44, 68, 70].

A similar approach for ascertaining whether natural magnetization measured in Japanese volcanics was simple, stable TRM has been given by Nagata *et al.* [28]. In these tests, the ac demagnetization curves, the thermal demagnetization curves, and the intensities of the natural magnetizations were compared with these same properties after the rocks had been heated to their Curie point temperatures and cooled in the earth's field.

After determining the process by which a rock became magnetized, the time of magnetization should also be established. In relatively young igneous rocks that show no evidence of reheating there can be little doubt that the TRM was acquired at the time of cooling, as shown by the studies of Chevallier [21] and Minakami [71] on historic lava flows. In old rocks, however, the assumption that a stable component of magnetization is TRM parallel to the geomagnetic field at the time of cooling may not always be justified. Many minerals undergo chemical or phase changes at extremely slow rates. Exsolution phenomena are common in ferromagnetic minerals, for example, and these may have occurred long after cooling [72]. The magnetization of an exsolving mineral may be controlled by that of the older host minerals, but it is also possible that the magnetization may be controlled by the ambient field during exsolution.

The problem of determining when rocks acquired natural CRM is much more difficult than for TRM. Almost all the CRM of interest in paleomagnetic research at the present time is that in sediments, and chemical changes may occur throughout the history of such rocks. Considerable geologic acumen must be exercised, therefore, to determine exactly when a given chemical reaction occurred. Sometimes a fold test or other geologic evidence sets an upper limit to the time which elapsed between deposition and chemical magnetization; but in most cases, the age of the magnetization cannot be unambiguously established. Several examples, however, of congruent paleomagnetic results between igneous and sedimentary rocks of the same geologic age indicate that chemical magnetization may take place not too long after deposition.

As previously mentioned, TRM and CRM produced in isotropic materials in laboratory experiments are always parallel to the applied field; the CRM of α-Fe_2O_3 and γ-Fe_2O_3 grown under uniaxial stress [41] appears to be an exception. However, to the authors' knowledge, the experiments have not been performed on noticeably anisotropic rocks and remanent magnetization in such materials would certainly not be expected to be parallel to the applied field. Magnetic anisotropy occurs when either (*1*) minerals occur as nonequidimensional grains or as planar or linear aggregates of grains, or (*2*) minerals with large magnetocrystalline energies have preferred orientations of crystallographic axes.

Instruments suitable for testing for magnetic susceptibility anisotropy are presently in use in several laboratories [73, 74, 75], and the reliability of paleomagnetic results based either on natural TRM or CRM in rocks is greatly enhanced if the absence of anisotropy has been demonstrated. Of considerable interest in this regard is the observation by Granar [76] that another phenomenon which may produce a remanent magnetization not parallel to the applied field (namely, the inclination error in depositional magnetization) is also reflected in magnetic susceptibility anisotropy.

3.4. Instruments

The natural intensities of magnetization of the igneous rocks which have proved useful for paleomagnetic studies are generally in the range 1×10^{-5} to 1×10^{-2} cgs emu/cc and those of useful sediments between 5×10^{-7} and 5×10^{-5}. In order to carry out partial demagnetization experiments, magnetometers are required capable of making measurements on samples with intensities one to two orders of magnitude below these ranges.

For strongly magnetized samples, any of the dozens of standard methods for measuring magnetic moments may be used. However, for the weakly magnetized rocks which are often of greatest interest in paleomagnetic research, the two types of instrument that have proved to be most successful are the astatic and the rock spinner magnetometers.

3.4.1. Astatic Magnetometers. The detecting unit of these instruments consists of two magnets of equal moment rigidly mounted one above the other. The magnets are oriented horizontally and antiparallel in the rigid mount, which is suspended from a torsional fiber. The rock specimen to be measured is placed in various positions and orientations below the lower magnet, and the angular deflection of the magnet system is measured using standard methods. In order to achieve the required sensitivity, the torsional force of the fiber must be the predominant restoring force; this is accomplished by carefully balancing the two magnets and by further reducing the effects of imperfect balancing by placing the entire system in coils canceling the earth's field. The theory of these instruments and many helpful practical details are well described by Blackett [77] and by Collinson and others [78].

3.4.2. Rock Spinner Magnetometers. Rock spinner or rock generator magnetometers [27, 79] are basically alternating current generators. As the rock specimen is rotated at high speed near a multiturn pickup coil, its magnetic moment generates an alternating current in the coil. The rotating system simultaneously generates a signal with the same frequency and of known phase which is used as a reference. The phase and intensity of the signal from the specimen are analyzed electronically to give the direction and intensity of the component of magnetization normal to the rotation axis, the total moment of the sample being found by rotation about several axes. Ultimate sensitivity is limited by the speeds of rotation achievable in the rotor, electrostatic noise, and thermal noise in the pickup coil.

The ultimate sensitivity of the two instrument types is comparable, and each has advantages for certain types of experiments. The astatic instruments essentially measure the magnetic field gradient due to the rock specimen and hence are quite sensitive to constant field gradients. If high sensitivity is needed, astatic magnetometers must be housed in nonmagnetic buildings removed from ferromagnetic materials or dc magnets. Spinner magnetometers, on the other hand, are sensitive to electromagnetic field gradients such as those generated by the laboratory power circuits, and extensive precautions are needed to minimize these effects.

4. Paleomagnetic Studies of Field Intensity

4.1. General Methods

Fewer than 10 accurate determinations of the intensity of the ancient geomagnetic field have been made, and the oldest only goes back to about 600 B.C. Both the number of intensity measurements and the time spanned are in striking contrast with the paleomagnetic studies of ancient field directions, of which there are now more than 250 going back to Precambrian time. The few valid field intensity measurements are principally the result of two decades of research of É. Thellier and O. Thellier; their recent review [69] of this classic work should be read by anyone interested in this aspect of paleomagnetism.

Reasons for the small number of intensity determinations are readily apparent. Few magnetized materials satisfy all of the tests necessary to insure that an intensity determination is valid; moreover, once suitable material has been found, the intensity determination is much more laborious than a determination of direction. In addition, a continuous record of the ancient geomagnetic field would undoubtedly show sizable short period fluctuations comparable to those of the recent geomagnetic field, so that many "spot" intensity determinations are necessary to be certain of detecting possible slow, steady changes. The considerable amount of effort required is, however, amply justified by the great importance of data on intensity in furthering our understanding of the geomagnetic field. Of particular interest is the problem of whether, in the period immediately before the beginning of observatory measurements, the intensity of the main dipole field was decreasing, as it has been during the past century.

Because of the importance of this subject, and the fact that the Thelliers' work has only appeared in French and Russian, we will review the subject in some detail.

In determining ancient field intensities, use is made of the relationship between the ambient magnetic field and the intensity of the magnetization due to one of the magnetizing processes occurring in nature. At the outset it might appear that if we know the intensity of remanent magnetization as a function of the applied field for different magnetizing processes and different mineral species, we can simply determine the amount of ferromagnetic mineral in a sample and calculate the intensity of the original field. Many complications make such a naive approach unworkable. Perhaps the most serious is the great variability of the magnetic properties of an individual mineral. As discussed more completely in Section 3, the amount of remanent magnetization per unit volume may vary by an order of magnitude for one mineral, depending on such factors as grain size, grain shape, cooling and stress history, and small variations in minor constituents present in solid solution. Moreover, some types of natural remanent magnetization decay over long periods of time [27, 80]. Another practical difficulty is that of obtaining a quantitative separation of fine-grained minerals which, as the petrologist well knows, may be all but impossible.

The other methods for determining field intensity have been classified by the Thelliers [69] into two groups, indirect and direct. The indirect methods are based on assumed or empirical equations relating intensity of remanent magnetization to the ancient field intensity by means of easily measurable magnetic properties of the rock such as coercive force, susceptibility, or saturation isothermal remanent magnetization. These indirect methods, the Thelliers conclude, all have the following very serious shortcoming: the ferromagnetic minerals or the parts of them giving rise to the original remanent magnetization may not be the same as those responsible for the other magnetic properties measured.

In the direct method, the sample is remagnetized in different known fields by the same process that caused the original natural remanent magnetization, and the resultant remanent magnetizations compared with the original. In order to obtain a valid determination of the ancient field direction, the following conditions must be satisfied: (1) The natural remanent magnetization must be due to a single, known, reproducible process. (2) This natural remanent magnetization must not have changed since it was acquired. (3) Chemical or phase changes affecting magnetic properties must not have occurred either since the original magnetization or during the remagnetizing process. Up to now, only the thermoremanent magnetizations in some baked clays and a few volcanic rocks have yielded results satisfying these conditions.

4.2. Thellier's Experimental Method

By far the most successful material for paleomagnetic studies of intensity has been baked clay from archeologically dated sites, including bricks from fireplaces or kilns, tiles, and fired pottery. The remanent magnetization is entirely TRM aside from a possible viscous component acquired since the last firing of the clay. Several of the properties of TRM discussed in Section 3 make TRM especially useful for intensity studies. It is extremely stable. Moreover, because of the additivity property of PTRM, separate determinations of field intensity can be made by investigating the TRM in different temperature intervals of the same specimen, and the results compared for consistency. Finally, the intensity of the TRM acquired in each small temperature interval should theoretically be proportional to the quantity tanh hvJ_s/KT (where h is the applied field, v is the volume of the grains, J_s the spontaneous magnetization, K Boltzmann's constant, and T the absolute temperature); in weak fields, therefore, TRM is theoretically nearly proportional to the field, and this fact has been experimentally verified for fields up to 1 oe.

To test for the presence of a viscous component of magnetization and for magnetic stability, Thellier uses the following procedure. After the sample is collected and transported to the laboratory, it is immediately placed in its original orientation with respect to the geomagnetic field and left for 2 weeks. Possible soft components of magnetization that may have been altered during collection and transportation will be re-established in the sample during this time. The remanent magnetization is then measured and the sample rotated 180° about a horizontal axis normal to the local magnetic meridian. After storage in the geomagnetic field for 2 weeks or more the remanent magnetization is remeasured, the vector difference between the two measurements being the amount of viscous magnetization acquired since the 180° rotation. Using the general relationship that viscous magnetization is proportional to the logarithm of the time, an estimate of the total viscous magnetization acquired since the last firing is calculated. If this probable viscous component turns out to be more than a few per cent of the total natural remanent magnetization, it is assumed that stability is not sufficiently high to insure that some of the original TRM may not have decayed and the experiment is discontinued.

If the amount of viscous magnetization is sufficiently small, the next step is to heat the sample to 100°C and cool it in the earth's field in a known direction. After measurement of the remanent magnetization, the sample is rotated 180° about a horizontal axis normal to the local magnetic meridian, heated to 100°C, cooled in the same field, and remeasured. The vector sum of the two measurements is twice the PTRM acquired in the ancient geomagnetic field between the Curie point temperature and 100°C and the difference is twice the PTRM acquired in the present field between 100°C and room temperature. The heating to 100°C also removes the small amount of comparatively unstable viscous magnetization which may be present.

The next step in the simplest form of Thellier's procedure is to heat the sample to the Curie point temperature, cool it in the earth's field, measure the remanent magnetization, rotate 180° as before, heat to 100°C, cool in the earth's field, and remeasure. The vector sum of these last two measurements is twice the PTRM acquired in the present field on cooling from the Curie point temperature to 100°C; from this quantity together with the corresponding PTRM acquired in the ancient field and the present intensity of the earth's field, the ancient field intensity can be calculated.

Remaining to be investigated is the possibility that the ferromagnetic minerals may have undergone chemical or other changes during the last heating. To test this, Thellier reheats the sample to its Curie point temperature and repeats the last set of measurements, holding the sample at the Curie point temperature for a longer time than on the original cycle. Only if the PTRM is the same in the two experiments is the intensity determination accepted as valid. The reasoning is that if any change occurs at a sufficiently high rate to affect the results of the measurement, then different amounts of change will have occurred during the two cycles and will be detected by different PTRM's. Other magnetic properties such as coercive force and susceptibility might be measured before and after heating as further tests for significant changes.

A final question remaining is whether chemical or other changes in the magnetic minerals have occurred since the original firing. This possibility cannot be completely eliminated, but for the materials investigated by Thellier it appears to be unlikely. The magnetic mineral present in the baked clays is hematite; this mineral almost certainly reached its present completely oxidized state during firing and since then has been in an oxidizing environment. Chemical changes since firing are thus rather unlikely. Moreover slow changes occurring in the oxidizing environment since firing would likely be accelerated during the reheating cycle and thus detected. The oldest clay that has yielded consistent results was last fired about 600 B.C., however, and this problem may prove a formidable obstacle in extending intensity measurements to older rocks.

As an additional check on this method, Thellier divides the temperature interval between the Curie point temperature and 100°C into several temperature subintervals and determines the ancient field intensity from the PTRM of each. Only if the results from all subintervals agree is the ancient field intensity accepted as valid.

4.3. Discussion of Results

Seven paleomagnetic determinations of past field intensities have been made by Thellier on clays from France and Carthage baked between 600 B.C. and 1750 A.D. These data meet Thellier's rigorous criteria for reliability as described above, and the intensities are plotted as open circles in Fig. 7 with the age of the last baking indicated along the upper scale.

The older baked clays in general indicate higher field intensities, but before this can be interpreted as indicative of a long-term decrease in the dipole moment of the geomagnetic field, two factors must be considered. Many of the sampling sites were at different latitudes, and therefore comparisons between different sites are meaningful only after correction has been made for the expected change of intensity with latitude. Moreover,

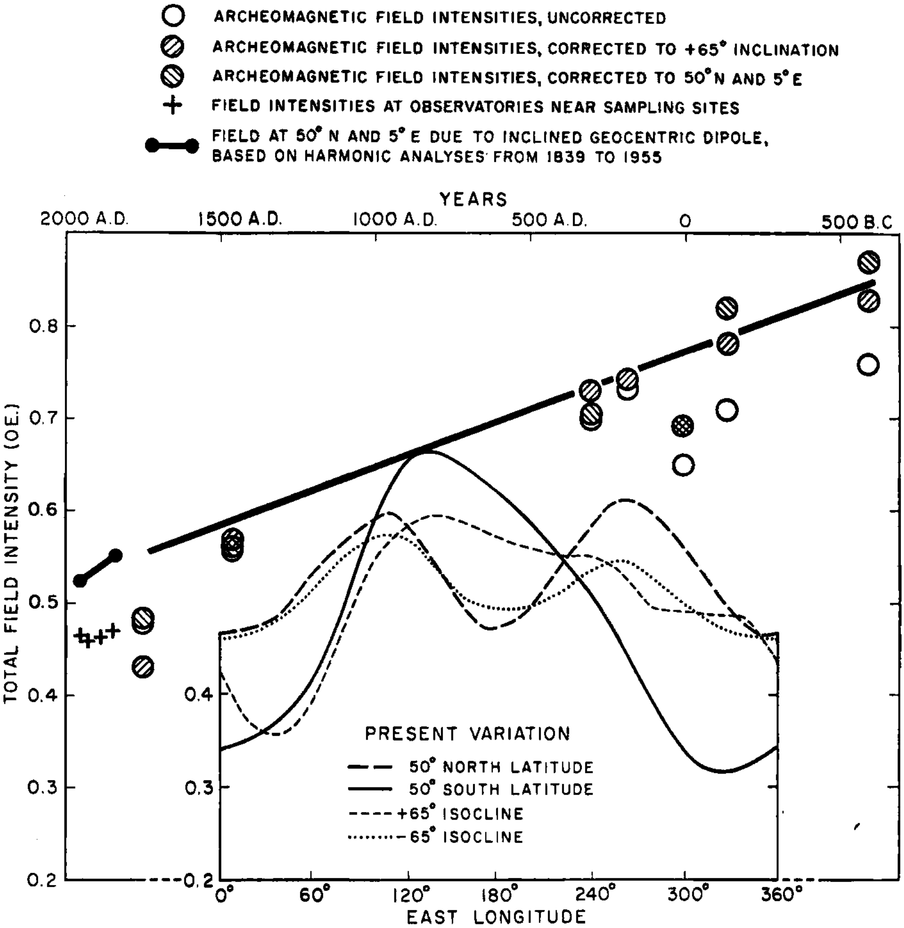


FIG. 7. Intensity of geomagnetic field, F_0, in France during past 2500 years (data from Thellier [69]) and along 50° latitude circles and 65° isoclines in 1945.

irregularities due to comparatively rapid changes in the nondipole field comparable with directly observed changes, or due to possible changes in the dipole field orientation, will probably be superimposed on any possible slow steady change of the dipole field intensity.

Reduction of intensities from sites at different localities and of different ages would be entirely straightforward if the geomagnetic field were due entirely to a dipole at the earth's center and if the orientation of the dipole for all times in the past were known. The actual geomagnetic field, however, has a substantial nondipole component, and the orientation of the dipole during the time covered by these intensity measurements is uncertain.

Hence, in order to reduce magnetic intensity values for comparison, assumptions about the past geomagnetic field are necessary.

Thellier [69] uses the paleomagnetically determined inclination of the ancient field in reducing his intensities; the reduced intensity is the value it would have if the field were entirely dipolar and if the inclination were 65° instead of the observed value. The following equation is used:

$$F_{65} = F(1 + 3\cos^2 I)^{1/2}(1 + 3\cos^2 65°)^{-1/2} \tag{4.1}$$

where F is the measured field intensity, I is the associated inclination, and F_{65} is the equivalent reduced intensity. This method of reduction assumes that the measured inclination is that of the main dipole field; nondipole variations in inclination contribute an additional "noise" signal. Intensities reduced by equation (4.1) are shown in Fig. 7.

An alternative method of reduction is to reduce all intensity values to the same geographic position, assuming the field to be dipolar and of known orientation. The reduced field is then given by

$$F_0 = F(1 + 3\cos^2 p_0)^{1/2}(1 + 3\cos^2 p)^{-1/2} \tag{4.2}$$

where F is the measured intensity, p is the distance from the sampling locality to the magnetic pole, p_0 is the distance to the pole from the standard reference locality, and F_0 is the reduced field intensity. In Fig. 7, Thellier's intensity values are shown reduced to an average reference position at 50°N 5°E calculated for a geomagnetic pole fixed at its present position. Variations in this reduced field value would be exactly proportional to variations of the main dipole intensity only if the main dipole did not move during the period covered; as with the other method of reduction, nondipole intensity variations will appear as a superimposed noise.

As discussed in Section 2, changes in the orientation of the main dipole have been scarcely if at all detectable during the period covered by direct observation. For intensity determinations young enough so that the main dipole had its present orientation, reduction by equation (4.2) is the more appropriate, and for older rocks for which the dipole orientation is unknown, reduction by equation (4.1) gives the best measure of changes in dipole intensity. Comparison of the two sets of reduced values shows the same general decrease in field intensity with time and indicates little sensitivity to the method of reduction used. The rate of decrease is greater after reduction by either method than it was for the uncorrected intensities, simply reflecting the fact that the older sampling sites are in lower latitudes.

Let us now consider the possibility that the paleomagnetic field intensities are due to comparatively rapid secular changes in the nondipole field and not to a steady change in the main dipole moment. Thellier [69] suggests as a measure of the amplitude of the nondipole field the present intensity variations along the $+65°$ isocline (Fig. 7). At present the nondipole field in the southern hemisphere is greater than in the northern, but this may not always have been true. The present intensity variation along the $-65°$ isocline is therefore also plotted in Fig. 7. An alternative measure of the expectable variation is the present total intensity along the circles of latitude 50°N and 50°S. This is the variation in intensity which an observer at about the average locality of the samples would see if the present geomagnetic field were rotated 360° about the earth's rotation axis; it is in part due to the nondipole field and in part to the inclination of the main dipole. Note that the ordinate for these curves is the same as that for the paleomagnetic determinations. Thus, although the amplitude of these nondipolar changes may be almost as large as the total paleomagnetic decrease, the maximum expected intensity for nondipolar changes falls far short of the ancient intensities.

The rate of westward drift of the secular variation foci and of the nondipole field poses an additional difficulty in explaining the high ancient intensities as due to secular changes in the nondipole field. During the period of direct observation, individual features of the nondipole field extend across at most 100° of longitude, and at a rate of westward drift of 0.18° per year, these features would move across a fixed point in the mantle in less than 600 years. An alternative estimate of the maximum period of a nondipole fluctuation may be made from the rate of westward drift of secular variation foci. During the past century, regions within which the rate of change of total intensity is negative are larger than positive regions, reflecting the steady decrease of the main dipole field. The largest regions of increase in the vicinity of the circles of latitude 50°N and 50°S span about 190° of longitude, and the largest regions of decrease about 300°. Bullard *et al.* [6] estimate the rate of westward movement of secular variation patterns at 0.3° of longitude per year. Thus the maximum time that an observer at one locality could expect the intensity to either increase or decrease in response to changes in the nondipole field alone may be estimated at about 800 years. Since the time interval spanned by Thellier's measurements is more than 2500 years, the period as well as the absolute intensity of expected secular changes in the nondipole field do not appear to be adequate to explain the paleomagnetic results.

A final point to be considered is whether these data may not simply indicate that the main geomagnetic dipole in earlier times had the same intensity but an orientation such that the geomagnetic pole was nearer the sampling sites. Brynjólfsson [58] has described four lava flows from Iceland ranging in age from 1000 B.C. to 200 A.D. which have directions of magnetization consistent with a magnetic pole in the vicinity of Norway or Novaja Zemlja; he further cites a passage from Aristotle indicating that the ancient Greeks saw the northern lights much more frequently than they are seen at present in Greece. However, the magnetic inclinations found by Thellier do not support this interpretation, which would require a steady increase in inclination going from the youngest to the oldest samples; actually the samples from France show an increase in inclination going back 200 years, before which they decrease to less than the present value for the oldest samples. Moreover, if a change in the orientation of the geocentric dipole had occurred, Thellier's method of reduction would compensate for it so that changes in F_{65} would still be proportional to changes in the intensity of the main dipole. In summary, the high early geomagnetic field intensities found by the Thelliers cannot be accounted for either by changes in the nondipole field as large as those observed during the past century or by a change in orientation of the main dipole. They most probably indicate a decrease in the main dipole moment of the geomagnetic field during the period studied.

In order to compare these results with those from spherical harmonic analyses, the theoretical field intensity at the reduced sampling site with coordinates 50°N and 5°E due to the main geocentric dipole has been calculated from the following equation:

$$F_0 = H_0(1 + 3\cos^2 p_0)^{1/2} \tag{4.3}$$

where F_0 is the intensity at the reduced sampling site, p_0 is the distance from the site to the geomagnetic pole, and H_0 is the theoretical dipole field at the equator as evaluated from the first degree Gaussian coefficients [equation (2.3)]. Corresponding to a rate of change (dH_0/dt) of 15 gammas per year at the equator, as indicated in Fig. 2, is a corresponding rate (dF_0/dt) of 25 gammas per year at 50°N, 5°E, as is shown in Fig. 7.

As an estimate of the average rate of change of the dipole field at this latitude, a line is drawn in Fig. 7 between the reduced ancient intensities and the present theoretical dipole field intensity in the sampling area. This estimate is lower than that which would be obtained using the present total field intensity in the sampling area; the total field is smaller than the theoretical dipole field because of the present negative nondipole component in France. The indicated rate of change of the dipole field at 50°N and 5°E, dF_0/dt, is 12 gammas per year, which is equivalent to a rate of change of the equatorial dipole field, dH_0/dt, of 7 gammas per year and of the main dipole moment of 1.8×10^{22} gauss cm^3 per year, approximately half the rate of decrease found from spherical harmonic analyses during the past century.

The effect of a decrease in field intensity on the rate of production of carbon 14 has been considered by Elsasser *et al.* [81]. They conclude that an exponential decrease of the main dipole field during the past 2000 years sufficient to reduce the field by 65 %, preceded by a long interval when the field was constant, would result in apparent C^{14} "ages" that would be 240 years too old for 2000 year old objects. Extrapolation of this exponential decrease back 4000 years would result in C^{14} ages 1000 years too old, which is greater than the error in C^{14} ages of objects dated historically. An exponential decrease this large thus appears unlikely. An extension of this method to a calculation of C^{14} age discrepancies for a linearly decreasing field and for a field varying periodically would be of great interest in setting limits to the amplitude and the waveform of past intensity changes. A more complete discussion of these results is given in Thellier's review [69].

5. Paleomagnetic Studies of Field Direction

5.1. Special Methods of Analysis

It is apparent at the outset that the standard method of spherical harmonic analysis used in contemporary investigations of the geomagnetic field is not applicable to the analysis of paleomagnetic data. Such an analysis uses, as its basic data, measurements of field direction and intensity from a world-wide net of observatories all reduced to the same time. To obtain similar data from paleomagnetic studies, it would be necessary that a group of suitable rocks with world-wide distribution should originally have become magnetized over a time interval short compared with field changes, that these rocks and their magnetizations should have been preserved through the ages, and that we should now have geologic techniques for establishing that the rocks in question are of the same age. Desirable as such paleomagnetic data would be, it will probably rarely if ever prove possible to assemble them.

In a typical paleomagnetic investigation, measurements are made of the average directions of magnetization of perhaps several scores of rock units in a geologic formation covering only a very small fraction of the earth's surface. Although where possible a sampling area is sufficiently large to render unlikely a rigid-body rotation or displacement of the sampling area with respect to the land mass of which it is a part, it is rarely possible to trace and sample a single geologic formation over a sufficiently long distance to detect changes of inclination with latitude. Moreoever, although the sequence of the units studied in a given formation is often known, the time interval between the units can rarely be determined exactly. Thus, in contrast with the data used for a spherical harmonic analysis, the data obtained from a paleomagnetic study are, at best, a series of measurements of declination and inclination, but not intensity, at a single "observatory." The order in which the "measurements" were made is often known, but the exact times at which they were made are not.

5.1.1. Virtual Geomagnetic Poles. The first special technique used in paleomagnetic analysis is the representation of the direction of the earth's field as measured at some observatory or sampling area by the theoretical geocentric dipole which, in the absence of nondipole components, would produce the field observed at that locality. This method of presenting magnetic data has the advantage of permitting a direct comparison of data from different magnetic latitudes. The orientation of the fictitious geocentric dipole, specified in terms of the geographical coordinates (θ', ϕ') of the corresponding south geomagnetic pole, is related to the declination and inclination of the earth's field at the paleomagnetic sampling area by the following equations:

$$(5.1) \quad \theta' = \sin^{-1} (\sin \theta \cos p + \cos \theta \sin p \cos D), \qquad (-90° \leq \theta' \leq 90°)$$

$$(5.2) \quad \begin{aligned} \phi' &= \phi + \beta, & &(\cos p \geq \sin \theta \sin \theta') \\ \phi' &= \phi + 180° - \beta, & &(\cos p < \sin \theta \sin \theta') \end{aligned}$$

where

$$(5.3) \quad \beta = \sin^{-1} (\sin p \sin D/\cos \theta'), \qquad (-90° \leq \beta \leq 90°)$$

and p, the angular distance along the great circle from the sampling area to the geomagnetic pole, is given by the "dipole" equation describing the inclination of a dipole magnetic field on the surface of a sphere as a function of distance from the magnetic pole:

$$(5.4) \quad p = \cot^{-1} (\tfrac{1}{2} \tan I), \qquad (0° \leq p \leq 180°)$$

Here D is the declination of the north-seeking magnetic field vector, in degrees east or clockwise from north, and I is the inclination, positive values indicating dips below and negative signs dips above the horizontal. The θ and ϕ are the latitude and longitude of the locality where the field direction was measured, and θ' and ϕ' are the coordinates of the corresponding magnetic pole; positive signs indicate latitudes 0° to 90° north and longitudes 0° to 180° east and negative signs latitudes 0° to 90° south and longitudes 0° to 180° west.

Equations (5.1) to (5.4) establish a one-to-one mapping relationship between all possible field directions at a locality and corresponding geomagnetic pole coordinates. Since "poles" can be calculated for any set of field directions over the surface of a sphere whether or not the field is due to a geocentric dipole, these calculated poles are called "virtual geomagnetic poles." Only when the magnetic field over the surface of a sphere is entirely dipolar will the virtual geomagnetic poles calculated from field directions at different localities coincide, and the amount of dispersion or scatter of virtual geomagnetic poles calculated from simultaneous world-wide field

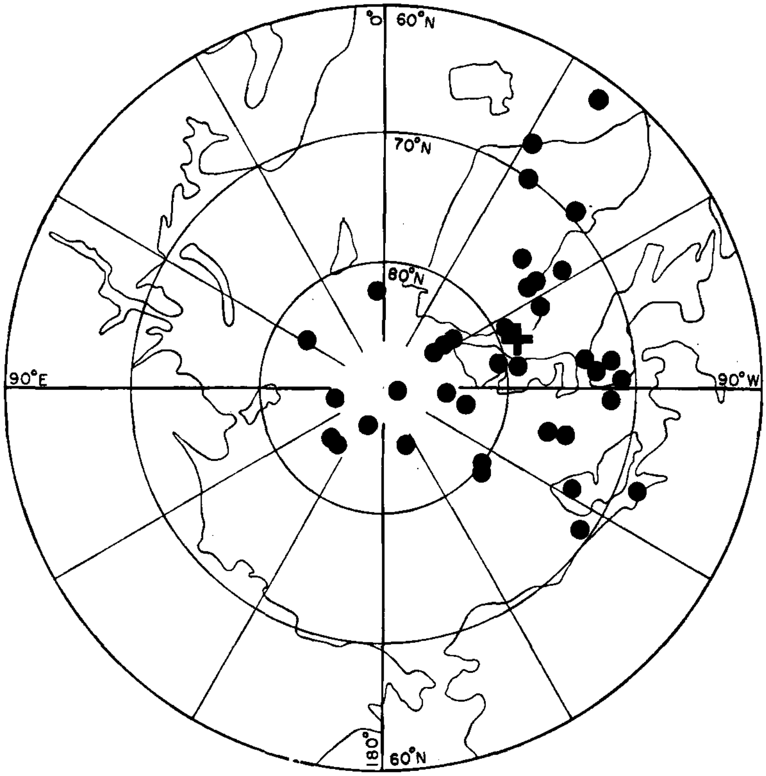


FIG. 8. Virtual geomagnetic poles calculated by equations (5.1) to (5.4) from present field directions at magnetic observatories.

measurements is thus a measure of the relative magnitude of the nondipole field. As an example, the present-day dispersion of virtual geomagnetic poles calculated from observatories with world-wide distribution is shown in Fig. 8. Unless the relative magnitude of the nondipole field has been different in the past, this figure indicates an order of magnitude for the expected "noise signal" in ancient virtual geomagnetic pole positions calculated from paleomagnetic data.

5.1.2. Statistical Analysis by Fisher's Method. A second special technique used in analyzing paleomagnetic data is a statistical method developed by Fisher [82]. Although this method and its extensions by Watson, Irving, and other workers [83, 84] has to date been used only in paleomagnetic studies, it may also be applicable to other types of geophysical data consisting of sets of directions [5].

Paleomagnetic data consisting either of a set of directions of magnetization or of the corresponding set of virtual geomagnetic poles may be represented as a set of unit vectors with common origin or, equivalently, as a set of points on a unit sphere. The set of data is assumed to be a sample randomly drawn from a population having the following two properties: (1) the vectors or points are distributed with azimuthal symmetry about their mean; (2) the density of vectors decreases with increasing angular displacement ψ from the mean according to the probability density function:

$$\mathcal{P} = (\kappa/4\pi \sinh \kappa) \exp(\kappa \cos \psi) \tag{5.5}$$

This density function is analogous with a symmetrical Gaussian distribution on a plane. Here, κ, termed the "precision parameter," is a positive number characteristic of a population of vectors; large values indicate tight grouping about the mean, while 0 indicates uniform distribution of points over the unit sphere. For a sample consisting of N vectors drawn from such a population, Fisher shows that k, the best estimate of κ, is given for $\kappa > 3$ by:

$$k = N - 1/N - R \tag{5.6}$$

where R is the length of the vector sum of the N unit vectors. As the size N of the sample drawn from a given population increases without limit, the value of k converges to κ.

The measure of central tendency adopted in Fisher's method is the direction of the vector sum of the N unit vectors constituting the sample. At a probability level of $(1 - P)$, the true mean direction of the population from which the sample was drawn lies within a circular cone around the vector sum with radius $\alpha_{(1-P)}$ given by

$$\cos \alpha_{(1-P)} = 1 - \frac{N - R}{R}\left[\left(\frac{1}{P}\right)^{1/N-1} - 1\right] \tag{5.7}$$

In most paleomagnetic analyses, P is taken as 0.05. The following approximate relationships, valid for small values of α, are frequently useful:

$$\alpha_{95} \cong 140°/(kN)^{1/2} \tag{5.8}$$

$$\alpha_{50} \cong 67.5°/(kN)^{1/2} \tag{5.9}$$

As the size N of the sample increases without limit, α converges to zero. For a more complete discussion of Fisher's statistical method and its application to paleomagnetic data, reference is made to papers by Watson [83], Watson and Irving [84], Runcorn [85], and Cox and Doell [5].

5.2. *Paleomagnetic Data*

In presenting a compilation of the rather large volume of paleomagnetic data now available, the first decision facing the reviewer is whether to be selective or inclusive. Should data be included only from geologically well-dated rocks for which there exists compelling evidence that the rocks were originally magnetized parallel to the earth's field, that the rocks have not been subsequently remagnetized, and that the rocks represent a time interval long enough (10^4 to 10^5 years) to insure that transient effects have been averaged out? Or should all published data be presented in the interest of not inadvertently discarding unproven but significant data?

The first approach would be preferable if sufficient data meeting the standards outlined were available, but an initial attempt to prepare a compilation of select data for the present review proved impractical for several reasons. Fewer than half of the paleomagnetic studies made to date contain adequate discussions of the paleomagnetic reliability of the results, partly, at least, because some of the techniques for investigating magnetic stability and otherwise assessing paleomagnetic reliability are quite recent. For parts of the earth's history where reliable data are sparse, all existing data are of interest, since past experience indicates that even reconnaissance paleomagnetic studies without field or laboratory evidence of reliability may be useful indicators of the general pattern which later emerges from more careful investigations; such preliminary results are especially useful in suggesting areas in which to concentrate future research. An additional advantage of listing all the available data is that reconnaissance studies of two contemporaneous formations with different lithologies (for example, lava flows and sandstones) have in some cases yielded concordant results; two such otherwise unsubstantiated results may constitute a reliable datum, as discussed in Section 3. For these reasons, all the data currently available have been included in this review.

Occasionally a very thorough paleomagnetic study has been made of a rock formation that is poorly dated. Rather than discard such data or arbitrarily assign it to one of the possible periods within its range of uncertainty, we have, in the illustrations, included each such poorly dated paleomagnetic result in all geologic periods within the range of uncertainty of its age.

A second decision facing the reviewer is how best to present the paleomagnetic data from widely separated sampling areas in a manner that will enable comparisons to be made easily but will not bias the interpretation of the data in favor of some particular hypothesis. One approach is to reduce all paleomagnetic measurements of the geomagnetic field to a common base station [1]. To make this reduction, some configuration must be assumed for the geomagnetic field, and in practice this is invariably the dipole configuration.

An alternative way of representing paleomagnetic data, permitting easy comparison of data from widely separated sampling areas, is to represent each average field direction by a virtual geomagnetic pole (Section 5.1.1). As before, an ancient field configuration must be assumed, and in practice this is usually the dipole configuration. The virtual geomagnetic pole representation appears to the authors to be more convenient for interpreting paleomagnetic results than reduction of data to a common base station, and therefore, poles will generally be used in this review. The choice of this method does not imply an *a priori* assumption that the field was dipolar; the dipolar hypothesis is, in fact, one of the hypotheses that can be tested most conveniently by plotting virtual geomagnetic poles.

5.2.1. Paleomagnetic Reference Table. All the geologic formations studied paleomagnetically up to about April, 1960, that are known to the authors, are listed in Table II. The information listed includes the following. (1) Lithology of the formation. (2) Geographic location of the sampling site. (3) Name and age of the formation; the geologic period with its range of uncertainty is listed as given in the original reference and subsequently modified; the geologic period designations have been translated into a range in millions of years using Kulp's time scale [86]. (4) Original references are listed, together with subsequent references giving modifications or extensions of the original data. (5) A number is assigned to each entry to identify the poles plotted on the diagrams to follow; the numbering sequence goes approximately from younger to older, the lower limit to the range of possible ages for a given formation determining its position in the list; however, many of the ranges of uncertainty in ages overlap, and no attempt has been made to place formations in exact stratigraphic sequence.

For additional details about individual paleomagnetic studies such as extent of sampling, tests for paleomagnetic reliability, and numerical results in terms of magnetic declination, inclination, virtual geomagnetic poles, and statistical parameters, reference is made to catalogs recently compiled by Irving [87] and by Cox and Doell [5] and to the original sources.

5.2.2. Atlas of Maps of Virtual Geomagnetic Poles. The paleomagnetic data for different geologic periods are shown in this atlas as sets of virtual geomagnetic poles on maps of half of the earth (Figs. 9–23). The following conventions are used:

Magnetic polarity: each possible orientation of a geocentric dipole can be represented by either of two virtual geomagnetic poles of opposite polarity, 180° apart on the globe. The convention used here is that south magnetic poles are shown as solid symbols and north magnetic poles as open symbols; the present pole near Greenland would appear solid by this convention. In many paleomagnetic investigations directions of magnetization within one geologic formation are found to fall into two groups 180° apart; virtual geomagnetic poles for these data are shown as half-solid symbols. Where polarity is not known symbols are shaded (Fig. 9).

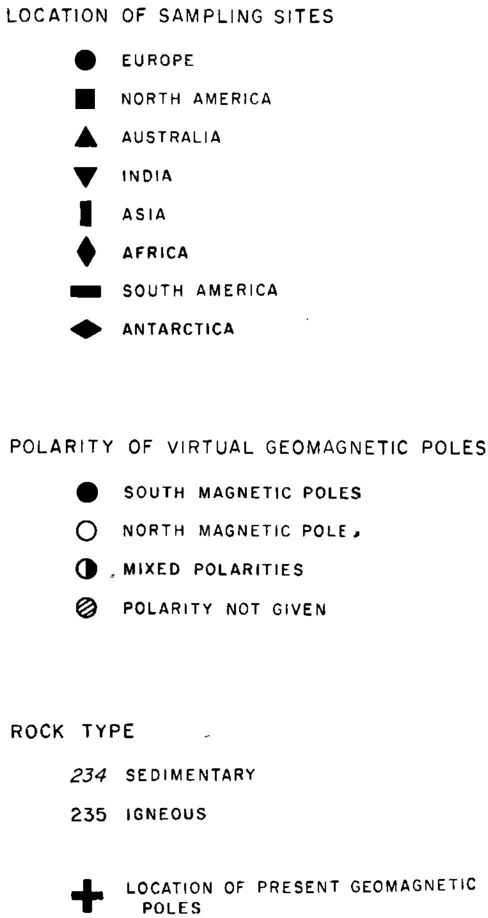


FIG. 9. Legend for maps in this atlas.

Locality of sampling site: the continent where samples were collected is indicated by the shape of the symbol (Fig. 9).

Rock type: igneous rocks are indicated by roman type and sediments by italics (Fig. 9).

Numbers: refer to entries in Table II.

Present geomagnetic pole: shown by a cross (Fig. 9).

All the maps are made on Schmidt equal-area projections and the poles of the individual projections were chosen so as to place most of the virtual geomagnetic poles near the center of the projection, where distortions are minimal.

Table II. Paleomagnetic Reference List.

No.	Rock type	Locality	Age, period, formation	References
			Late Pleistocene to Recent (0 to ½ m.y.[a])	
1	Lavas	Mt. Etna, Sicily	394 B.C. to A.D. 1911, historic lava flows	[5, 21, 88]
2	Fired clays	England	1st through 4th, 12th, 13th, 15th centuries	[5, 89]
3	Glacial varves	Central Sweden	0 to A.D. 1000, Prästmon varves	[5, 88, 90]
4	Fired clays	North Africa	146 B.C. and A.D. 300, Carthage kiln samples	[5, 91]
5	Glacial	Central Sweden	1100 B.C. to A.D. 750, Ångerman River varves	[5, 92]
6	Lavas	Iceland	3400 B.C. to A.D. 1950, post glacial lavas	[5, 58]
7	Fired clays	Japan	5600 to 4400 B.C. and A.D. 300 to 1800	[5, 93]
8	Glacial varves	New England, USA	13,000 to 7000 B.C.	[5, 50, 88]
9	Sands	Gulf Coast, USA	Recent sands	[94]
10	Volcanics	Japan	Pleistocene to recent, lava flows	[5, 95]
11	Lavas	Iceland	Post glacial, lava flows	[5, 88]
12	Lavas	France	Late Pleistocene, Chain des Puys lavas	[5, 96]
13	Glacial varves	Sweden	Glacial and post-glacial, varves	[5, 76]
14	Sandstone	California, USA	Post early Pleistocene, Neroly formation	[5, 97]
			Pliocene to Early Pleistocene (½ to 12 m.y.[a])	
15	Volcanics	Japan	1 to ? m.y., mostly Pleistocene and younger	[5, 98]
16	Volcanics	Japan	1 to ½ m.y., Quaternary, North Izu and Hakone volcanics	[5, 99]
17	Lavas	Argentina	1 to ½ m.y., Quaternary, Neuquen lavas	[5, 68]
18	Volcanics	South Australia	12 to less than ½ m.y., Pliocene to Recent, newer volcanics of Victoria	[100]
19	Lavas	Iceland	1 to ¾ m.y., early Quaternary	[5, 88]
20	Lavas	France	1 to ¾ m.y., early Quaternary, Plateau basalts	[5, 96]
21	Basalt	Japan	1 to ¾ m.y., early Pleistocene, Kawajiri basalt	[2]

22	Lavas	France	6 to ½ m.y., Pliocene and Pleistocene (oldest are Villafranchien)	[5, 88, 101]
23	Lavas	Iceland	12 to ½ m.y., Pliocene and Pleistocene	[5, 102]
24	Lavas	Iceland	12 to ½ m.y., Pliocene and Pleistocene	[5, 102]
25	Sediments	Kazakh SSR, USSR	12 to ½ m.y., Pliocene and Pleistocene, sediments from Chelekan Peninsula	[5, 103]
26	Sediments	Kazakh SSR, USSR	12 to ½ m.y., Pliocene and Pleistocene, sediments from Chelekan Peninsula	[5, 103]
27	Basalt	New Zealand	6 to ¾ m.y., late Pliocene or early Pleistocene, Timaru tholeiitic basalt	[104]
28	Lavas	Japan	6 to 1 m.y., late Pliocene, Enrei formation	[42]
29	Ignimbrites	New Zealand	12 to 1 m.y., Pliocene	[5, 105]
30	Volcanics	Japan	12 to 1 m.y., late and early Pliocene	[5, 106]
31	Lavas	France	16 to 1 m.y., Miocene and Pliocene (oldest are Pontien)	[5, 88, 101]
32	Sediments	Idaho, USA	26 to 1 m.y., Miocene and Pliocene, Payette formation	[5, 107]
33	Sediments	Washington, USA	17 to 6 m.y., late Miocene and early Pliocene, Ellensburg formation	[5, 107]
34	Sediments	Turkmen SSR, USSR	26 to 1 m.y., late Tertiary	[108]
35	Basalts	West Canada	26 to 1 m.y., late Tertiary	[5, 109]
36	Volcanics	Japan	26 to 1 m.y., late Tertiary	[5, 95]
37	Sediments	Utah, USA	65 to 1 m.y., Tertiary, Duchesne River formation	[110]
38	Volcanics	Anarctica	65 to 1 m.y., Cenozoic, Cape Hallet volcanics	[111]
			Oligocene to Miocene (12 to 35 m.y.[a]*)*	
39	Volcanics	New Zealand	19 to 12 m.y., late middle to upper Miocene, Dunedin volcanic complex	[104]
40	Lavas	Iceland	26 to 12 m.y., Miocene	[5, 88]
41	Basalts	Germany	26 to 12 m.y., Miocene, Vogelsberg basalts	[5, 112]
42	Lavas and intrusions	Germany	26 to 12 m.y., Miocene, Vogelsberg basalts	[113]
43	Lavas and intrusions	Germany	26 to 12 m.y., Miocene, Vogelsberg basalts	[113]

TABLE II—*Continued*

No.	Rock type	Locality	Age, period, formation	References
44	Basalts	Washington, USA	26 to 12 m.y., Miocene, Columbia River basalts	[5, 114]
45	Sediments	South Dakota, USA	26 to 12 m.y., Miocene, Arikaree formation	[5, 107]
46	Volcanics	Japan	26 to 12 m.y., late, middle, and early Miocene	[5, 106]
47	Sediments	Gulf Coast, USA	26 to 17 m.y., middle and early Miocene	[94]
48	Lavas	Scotland	19 to 26 m.y., early Miocene(?) Skye lava flows	[44]
49	Sediments	Gulf Coast, USA	19 to 26 m.y., early Miocene	[94]
50	Basalts	Ireland	65 to 12 m.y., Eocene or Oligocene—may be late Miocene, Antrim basalts	[5, 88, 115]
51	Basalts	Ireland	65 to 12 m.y., Eocene or Oligocene—may be late Miocene, Antrim basalts	[5, 116]
52	Lavas	Scotland	65 to 12 m.y., early Tertiary—probably Eocene—may be Oligocene or Miocene, Mull lavas	[5, 117]
53	Intrusives	Scotland	65 to 12 m.y., early (?) Tertiary, Mull intrusions	[5, 118]
54	Dikes	Scotland	65 to 12 m.y., early (?) Tertiary, Arran dykes	[2]
55	Dikes	England	65 to 12 m.y., early (?) Tertiary, Lundy dykes	[5, 119]
56	Gabbro	Scotland	65 to 12 m.y., early Miocene(?) Rhum gabbro	[44]
57	Gabbro	Scotland	65 to 12 m.y., early Miocene(?) Ardnamurchan gabbro	[44]
58	Tuffs	India	65 to 12 m.y., post Eocene, Mt. Pavagadh acid tuffs	[5, 120]
59	Intrusions	Germany	35 to 12 m.y., Miocene to Oligocene, Bonn intrusions	[113]
60	Dikes	England	35 to 12 m.y., Oligocene or Miocene, North West dykes	[5, 121]
61	Basalt	France	30 to 26 m.y., Oligocene (Aquitanian) Limagne basalt	[5, 122]
62	Intrusives	France	35 to 26 m.y., Oligocene	[5, 88, 123]
63	Sediments	California, USA	35 to 26 m.y., Oligocene, Sespe formation	[94]
64	Lavas and intrusives	New Zealand	50 to 30 m.y., late Eocene to early Oligocene, pillow lavas and dolerite intrusives	[104]
65	Sediments	Turkmen, USSR	65 to 26 m.y., early Tertiary (Paleogene). Calculated from "uncorrected" *D* and *I* values.	[108]
66	Sediments	Turkmen, USSR	65 to 26 m.y., early Tertiary (Paleogene). Calculated from "uncorrected" *D* and *I* values.	[108]

67	Volcanics	Australia	65 to 26 m.y., early Tertiary, probably Eocene, Older Volcanics of Victoria	[100]
68	Basalt	Tasmania	65 to 26 m.y., early Tertiary	[5, 124]
			Eocene (35 to 65 m.y.[a])	
69	Shale	Wyoming, USA	65 to 35 m.y., Eocene, Laney shale member of the Green River formation	[5, 107]
70	Sediments	Colorado, USA	65 to 35 m.y., Eocene, Green River formation	[5, 107]
71	Sediments	Wyoming, USA	65 to 35 m.y., Eocene, Wasatch formation	[5, 107]
72	Volcanics	Oregon, USA	65 to 45 m.y., early middle to early Eocene, Siletz River volcanics	[5, 64]
73	Sediments	Gulf Coast, USA	65 to 50 m.y., early Eocene, Wilcox formation	[94]
74	Lavas	India	135 to 35 m.y., Cretaceous to Eocene, upper Deccan traps	[5, 120, 125, 126]
75	Lavas	India	135 to 35 m.y., Cretaceous to Eocene, lower Deccan traps	[5, 120, 125, 126]
76	Lavas	India	135 to 35 m.y., Cretaceous to Eocene, Deccan traps	[5, 127]
			Cretaceous (70 to 135 m.y.[a])	
77	Sandstone	Arizona, USA	135 to 70 m.y., early and late Cretaceous, Dakota sandstone	[5, 128]
78	Lavas and dikes	Madagascar	100 to 85 m.y., Cretaceous (Turonian stage)	[5, 129]
79	Shale	Japan	90 to 135 m.y., early to middle Cretaceous, Inkstone red shales	[106]
80	Dolerites	Tasmania	70 to 180 m.y., Jurassic or Cretaceous	[130]
81	Dolerites	Tasmania	70 to 180 m.y., Jurassic or Cretaceous	[5, 124]
82	Lavas	India	180 to ? m.y., Jurassic or later, upper Rajmahal traps	[5, 131]
83	Basalt and sandstone	Uraguay and Brazil	220 to 70 m.y., Mesozoic, Parana basin basalts and baked sandstones	[5, 68]
84	Dolerite sills	Antarctica	220 to 70 m.y., Mesozoic(?)	[132]
85	Dolerite sills	Antarctica	220 to 70 m.y., Mesozoic	[111]
86	Dolerite sills	Antarctica	220 to 70 m.y., Mesozoic (Jurassic ?)	[133]
87	Sediments	England	100 to 135 m.y., early Cretaceous, Wealden sediments	[5, 116]
88	Sediments	England	100 to 135 m.y., early Cretaceous, "Iron Grit"	[113]

TABLE II—*Continued*

No.	Rock type	Locality	Age, period, formation	References
89	Sediments	England	100 to 135 m.y., early Cretaceous, "Iron Grit"	[113]
90	Sediments	Wyoming, USA	100 to 135 m.y., early Cretaceous	[94]
			Jurassic (135 to 180 m.y.[a]*)*	
91	Sediments	Arizona, USA	180 to 135 m.y., Jurassic, Carmel formation	[110]
92	Radiolarites	Austria	160 to 150 m.y., middle Jurassic	[5, 134]
93	Limestone	Austria	180 to 160 m.y., early Jurassic	[5, 134]
94	Sediments	Arizona, USA	180 to 160 m.y., early Jurassic (Glen Canyon group), Kayenta formation	[5, 110, 135]
95	Sandstone	England	180 to 160 m.y., early Jurassic (upper Lias), Midford sands	[5, 136]
96	Sediments	England	180 to 160 m.y., early Jurassic (upper Lias), Cotswald sediments	[5, 136]
97	Sediments	England	180 to 160 m.y., early Jurassic (middle Lias–lower Corallian)	[5, 135, 137]
98	Lavas	France	180 to 160 m.y., early Jurassic (lower Lias), Pyrenees lavas	[5, 136]
99	Sediments	Scotland	180 to 160 m.y., early Jurassic	[5, 137, 138]
100	Dolerites	Africa	200 to 160 m.y., late Triassic or early Jurassic, Karoo dolerites	[5, 139]
101	Dolerites	Africa	200 to 160 m.y., late Triassic or early Jurassic, Karoo dolerites	[5, 139]
102	Basalt	Africa	200 to 160 m.y., late Triassic or early Jurassic, Karoo basalts	[5, 135]
			Triassic (180 to 220 m.y.[a]*)*	
103	Sediments	Africa	200 to 180 m.y., late Triassic, Forest sandstone	[140]
104	Sandstone	Africa	200 to 180 m.y., late Triassic, Bechuanaland Cave sandstone	[138, 141]
105	Sediments	England	200 to 180 m.y., late Triassic, Keuper marles	[5, 51]
106	Sediments	England	200 to 180 m.y., late Triassic, Keuper marles	[67]

107	Sediments	Eastern USA	200 to 180 m.y., late Triassic (top of Newark group), Brunswickian formation	[5, 142]
108	Sediments	Eastern USA	200 to 180 m.y., late Triassic (bottom of Newark group), New Oxford formation	[5, 143]
109	Sandstone	Arizona, USA	200 to 180 m.y., late Triassic, Dinosaur Canyon sandstone member of the Moenave formation	[5, 144]
110	Sediments	Southwestern USA	200 to 180 m.y., late Triassic	[94]
111	Sediments	Southwestern USA	200 to 180 m.y., late Triassic, Chinle formation	[5, 110, 143, 144]
112	Sandstone	Southwestern USA	200 to 180 m.y., late Triassic(?), Springdale sandstone member of the Moenave formation	[5, 128]
113	Sandstone	Spain	220 to 180 m.y., Triassic, Villaviciosa sandstones	[5, 126]
114	Sandstone	Spain	220 to 180 m.y., Triassic, Alcolea and Aguilar sandstones	[5, 126]
115	Lavas	Eastern USA	220 to 180 m.y., Triassic, Connecticut Valley lavas	[5, 142]
116	Lavas and sediments	Eastern USA	220 to 180 m.y., Triassic, Connecticut lavas and sediments	[5, 142[
117	Lavas	Nova Scotia, Canada	220 to 180 m.y., Triassic	[5, 145]
118	Tuffs	Tasmania	220 to 180 m.y., Triassic(?), volcanic tuffs	[5, 124]
119	Tuffs	Australia	220 to 200 m.y., probably early Triassic, Brisbane tuff	[146]
120	Sediments	Southwestern USA	220 to 190 m.y., early to middle(?) Triassic, Moenkopi formation	[5, 110, 128, 144]
121	Sandstone	France	220 to 200 m.y., early Triassic, Vosges sandstone	[5, 126]
122	Sediments	Western USSR	220 to 200 m.y., early Triassic, Vetlujskij sediments	[5, 147]
123	Sediments	Southwestern USA	220 to 200 m.y., early Triassic	[94]
124	Sediments	France	215 to 205 m.y., Triassic, Grés Vosgien	[113]
125	Sandstone	Germany	215 to 205 m.y., Triassic, Pfalz Buntsandstein (Sm)	[113]
126	Sandstone	Germany	215 to 205 m.y., Triassic, Lörrach Buntsandstein (S)	[113]
127	Sandstone	Germany	215 to 205 m.y., Triassic, Baden Baden-Freiburg Buntsandstein (Sm)	[113]
128	Sandstone	Germany	215 to 205 m.y., Triassic, Vogelsberg Buntsandstein (Sm)	[113]
129	Sandstone	Germany	215 to 205 m.y., Triassic, Pfalz Buntsandstein (Sm-O)	[113]

Table II—*Continued*

No.	Rock type	Locality	Age, period, formation	References
130	Sediments	Wyoming, USA	240 to 180 m.y., Triassic and Permian, Chugwater formation	[5, 110]
130a	Olivinites	North Siberia	300 to 200 m.y., late Carboniferous to early Triassic, Kugda olivinites	[148]
130b	Basalts	North Siberia	300 to 200 m.y., late Carboniferous to early Triassic, Guli diabasic basalts	[148]
130c	Olivinites	North Siberia	300 to 200 m.y., late Carboniferous to early Triassic, Odikhincha olivinites	[148]
130d	Intrusives	North Siberia	300 to 200 m.y., late Carboniferous to early Triassic, Guli ore pyroxinites and peridotites	[148]
130e	Dikes	North Siberia	300 to 200 m.y., late Carboniferous to early Triassic, Kotuy ultrabasic dikes	[148]
130f	Meymechites	North Siberia	300 to 200 m.y., late Carboniferous to early Triassic, Guli meymechite	[148]
			Permian (220 to 275 m.y.[a]*)*	
131	Sediments	Western USSR	245 to 220 m.y., late Permian, Ufimskij and Kazanskij sediments	[5, 147]
132	Sediments	Western USSR	245 to 220 m.y., late Permian, Tartarskij sediments	[5, 147]
133	Sediments	Africa	245 to 220 m.y., late Permian, Maji Ya Chumvi formation	[141]
134	Sediments	France	235 to 260 m.y., middle Permian (Saxonian), Montcenis sediments	[5, 149]
135	Porphyry	France	235 to 260 m.y., middle Permian (Saxonian ?), Niedeck porphyry	[5, 149]
136	Volcanics	England	275 to 220 m.y., Permian, Exeter volcanic series	[127, 150]
137	Sediments	Scotland	275 to 220 m.y., Permian, Mauchline sediments	[5, 151]
138	Lavas	Scotland	275 to 220 m.y., Permian, Mauchline lavas	[5, 151]
139	Volcanics	France	275 to 220 m.y., Permian, Esterel pyromeride R4	[5, 152]

140	Dolerite	France	275 to 220 m.y., Permian, Esterel dolerite	[5, 152]
141	Rhyolite	France	275 to 220 m.y., Permian, Esterel rhyolite	[5, 153]
142	Volcanics	France	275 to 220 m.y., Permian, Esterel volcanics	[5, 70]
143	Trachyandesite	Norway	275 to 220 m.y., Permian, Oslo graben trachyandesite	[5, 153]
144	Volcanics	Scotland	275 to 220 m.y., Permian, Ayshire kylites	[5, 154]
145	Volcanics	Germany	275 to 220 m.y., Permian, Nahe volcanics	[43]
146	Sediments	Southwestern USA	275 to 220 m.y., Permian, Cutler formation	[5, 143]
147	Sediments	New Mexico, USA	275 to 220 m.y., Permian (Leonard), Yeso formation	[5, 143]
148	Volcanics	Australia	275 to 220 m.y., Permian, Upper Marine volcanic series	[146]
149	Volcanics	Australia	275 to 220 m.y., Permian, Lower Marine volcanic series	[146]
150	Sediments	Africa	275 to 250 m.y., early Permian, Taru grit	[141]
151	Sediments	France	275 to 250 m.y., early Permian (Autonian), Saint-Wendel sediments	[5, 149]
152	Sandstones and lavas	Germany	275 to 250 m.y., early Permian, Saar Rotliegendes (R)	[113]
153	Sandstones	Germany	275 to 250 m.y., early Permian, Freiburg Rotliegendes (R)	[113]
154	Sandstones and lavas	Germany	275 to 250 m.y., early Permian, Baden Baden Rotliegendes (Ro-Rm)	[113]
155	Sandstones and lavas	Germany	275 to 250 m.y., early Permian, Pfalz Rotliegendes (Ro)	[113]
156	Sill	England	330 to 250 m.y., late Carboniferous or early Permian, Great Whin sill	[66]
157	Sediments	Southwestern USA	275 to 250 m.y., early Permian, Abo formation	[5, 143]
158	Sediments	Southwestern USA	305 to 250 m.y., early Permian and middle or late Pennsylvanian, Sangre de Cristo formation	[5, 143]
159	Sediments	Southwestern USA	330 to 220 m.y., Permian and Pennsylvanian, Supai formation	[5, 110, 143, 155, 156]
160	Sediments	Southwestern USA	330 to 220 m.y., Permian and Pennsylvanian, Abo, Supai, and Cutler formations	[94]
161	Granite	England	330 to 220 m.y., Permian or Carboniferous, Lundy granites	[5, 119]
162	Dikes	Antarctica	600 to 220 m.y., Paleozoic (?), basic dikes	[132]

TABLE II—*Continued*

No.	Rock type	Locality	Age, period, formation	References
			Carboniferous (275 to 355 m.y.[a]*)*	
163	Sandstones	England	330 to 275 m.y., late Carboniferous (late Coal measures), Glouster Pennant sandstone	[5, 51]
164	Baked sediments	England	330 to 275 m.y., late Carboniferous baking, Tideswelldale samples	[5, 126]
165	Sandstones	France	300 to 275 m.y., Carboniferous (Stephanian), St. Étienne sandstones	[113]
166	Sandstones	France	300 to 275 m.y., Carboniferous (Stephanian), Montceaux-les-Mines sandstones	[113]
167	Sandstones	Germany	300 to 275 m.y., Carboniferous (Stephanian), Saar sandstones	[113]
168	Sediments	Southwestern USA	330 to 275 m.y., Pennsylvanian, Naco formation	[5, 128]
169	Sediments	Southwestern USA	330 to 275 m.y., Pennsylvanian, Naco formation	[110]
170	Sandstones	Oklahoma, USA	330 to 275 m.y., Pennsylvanian	[94]
171	Sediments	Australia	330 to 275 m.y., late Carboniferous, Kuttung varvoid sediments	[65]
172	Lavas	Australia	330 to 275 m.y., late Carboniferous, Kuttung lavas	[146]
173	Clays	Africa	330 to 275 m.y., late Carboniferous, Dwyka varved clays	[141]
174	Clays	Africa	330 to 275 m.y., late Carboniferous, Dwyka varved clays	[5, 141]
175	Baked sediments and igneous rocks	England	330 to 275 m.y., probably late Carboniferous, Clee Hill samples	[5, 126]
176	Sediments	England	330 to 275 m.y., late (?) Carboniferous, Lancashire samples	[5, 157]
177	Sediments	Southwestern USA	330 to 275 m.y., Pennsylvanian (?), Madera formation	[94]
178	Sandstones	England	340 to 300 m.y., Carboniferous (late Westphalian or early Stephanian), Staffordshire Keele beds	[113]

179	Sediments	Canada	340 to 275 m.y., Pennsylvanian and late Pennsylvanian or early Mississippian, Bonaventure, Kennebecasis, and Bathhurst formations	[5, 158]
180	Sediments	England	355 to 275 m.y., Carboniferous, Lancashire Pendle monocline	[5, 157]
181	Sandstones and siltstones	England	355 to 275 m.y., early and late Carboniferous, Derbyshire sandstones and siltstones	[5, 157]
182	Intrusives	England	355 to 275 m.y., Carboniferous, Shatterford intrusion	[5, 126]
183	Basinite	Scotland	355 to 275 m.y., Carboniferous, Southdean basinite plug at Jedburgh	[113]
184	Grit	England	345 to 330 m.y., late early Carboniferous, Lancashire Millstone grit	[5, 157]
185	Lavas	England	355 to 330 m.y., early Carboniferous, Derbyshire "Toadstones"	[5, 157]
186	Lavas	Scotland	355 to 330 m.y., early Carboniferous, Kinghorn lavas	[5, 126]
187	Sediments	Texas, USA	355 to 330 m.y., Mississippian, Barnett formation	[5, 159]
188	Sediments	Texas, USA	355 to 330 m.y., Mississippian, Barnett formation	[5, 159]
189	Sediments	Eastern Canada	355 to 330 m.y., Mississippian, Codroy group	[5, 160]
			Devonian (355 to 410 m.y.[a])	
190	Sandstones	Germany	395 to 375 m.y., early middle Devonian, Eifel sandstones	[113]
191	Sandstones	England	410 to 355 m.y., Devonian, Old Red Sandstone sandstones	[5, 150]
192	Sandstones	Scotland	410 to 355 m.y., Devonian, upper Old Red Sandstone sandstones	[113]
193	Sandstones	England	410 to 355 m.y., Devonian, lower Old Red Sandstone sandstones	[5, 51]
194	Lavas	Scotland	410 to 355 m.y., Devonian, lower Old Red Sandstone lavas	[113]
195	Limestone	Eastern USA	410 to 380 m.y., early or middle Devonian (late Ulsterian), Onondaga limestone	[31]
196	Volcanics	Australia	430 to 355 m.y., Silurian or Devonian, Ainslie volcanics	[146]
			Silurian (410 to 430 m.y.[a])	
197	Sediments	England	420 to 410 m.y., late Silurian, Ludlow series	[161]
198	Prophyry	Australia	420 to 410 m.y., late Silurian, Mugga porphyry	[5, 146]
199	Peridotite	Urals, USSR	430 to 410 m.y., Silurian	[162]

TABLE II—*Continued*

No.	Rock type	Locality	Age, period, formation	References
200	Iron ore	Eastern USA	425 to 415 m.y., Silurian (Niagara series), Clinton iron ore	[5, 75]
201	Siltstones	Yuman, China	425 to 415 m.y., middle Silurian, red siltstones	[5, 163]
202	Sediments	Eastern USA	425 to 420 m.y., early middle Silurian (lower Niagara series), Rose Hill formation of Swartz, 1923	[5, 29]
			Ordovician (430 to 490 m.y.[a])	
203	Sediments	Eastern USA	460 to 430 m.y., late Ordovician, Juniata formation	[5, 110]
204	Sediments	New York, USA	470 to 450 m.y., middle Ordovician, Trenton group	[5, 31]
205	Sediments	New York, USA	470 to 450 m.y., middle Ordovician, Trenton group	[5, 164]
206	Basalt	Ukraine, USSR	490 to 430 m.y., Ordovician, Ukrainian basalts	[5, 162]
207	Basalt	Ukraine, USSR	490 to 430 m.y., Ordovician, Ukrainian basalts	[5, 162]
208	Sands and clays	Northwestern USSR	490 to 460 m.y., early Ordovician, Leningrad red sands and brown clays	[5, 147]
			Cambrian (490 to 600(?) m.y.[a])	
209	Sediments	Texas, USA	540 to 490 m.y., late Cambrian, Point Peak shale member of the Wilberns formation	[159]
210	Dolomite	Colorado, USA	450 to 490 m.y., late Cambrian, Sawatch quartzite sandy dolomite	[5, 159]
211	Sandstone	Australia	525 to 575 m.y., middle Cambrian, Elder Mountain sandstone	[5, 146]
212	Sandstone	England	600 to 490 m.y., Cambrian, lower Caerfai series, Caerbwdy sandstone	[5, 127]
213	Sediments	Utah, USA	600 to 490 m.y., Cambrian, Lodore formation	[110]
214	Sediments	Wyoming, USA	600 to 490 m.y., Cambrian, Deadwood formation	[110]
215	Basalts	Australia	550 to 600 m.y., early Cambrian, Antrim plateau basalts	[146]
			Precambrian (greater than 600 m.y.[a])	
216	Sandstone	Scotland	Late Precambrian, Aultbea, Applecross, and top Diabaig formations	[165]
217	Sandstone	Scotland	Late Precambrian, lower Diabaig formation	[165]

218	Sandstone	England	Late Precambrian, Wentnor series of the Longmynd	[4, 5]
219	Sparagmites	Norway	Late Precambrian	[166]
220	Sandstone	Wisconsin, USA	Late Precambrian, Chequamegon sandstone	[5, 151]
221	Sandstone	Michigan, USA	Late Precambrian, Jacobsville sandstone	[5, 151]
222	Sandstone	Ontario, Canada	Late Precambrian, Jacobsville sandstone	[167]
223	Sandstone	Wisconsin, USA	Late Precambrian, Orienta sandstone	[167]
224	Sandstone	Wisconsin, USA	Late Precambrian, Eileen sandstone	[167]
225	Sediments	Michigan, USA	Late Precambrian, Freda sandstone and Nonesuch shale	[168]
226	Sandstone	Ontario, Canada	Late Precambrian, Sault Freda(?) sandstone	[167]
227	Sandstone and lavas	Michigan, USA	Late Precambrian, Copper Harbor conglomerate	[5, 151]
228	Shale	Arizona, USA	Late Precambrian, Hakatai shale	[5, 128]
229	Shale	Arizona, USA	Late Precambrian, Hakatai shale	[110]
230	Shale and limestone	Arizona, USA	Late Precambrian, Hakatai shale and Bass limestone	[5, 156]
231	Limestone	Arizona, USA	Late Precambrian, Bass Limestone	[110]
232	Quartzite	Arizona, USA	Late Precambrian, Shinumo quartzite	[110]
233	Sediments	Montana, USA	Late Precambrian, McNamara formation (Belt series)	[110]
234	Sediments	Montana, USA	Late Precambrian, Miller Peak formation (Belt series)	[110]
235	Shale	Montana, USA	Late Precambrian, Spokane shale (Belt series)	[5, 110]
236	Shale	Montana, USA	Late Precambrian, Spokane shale (Belt series)	[110]
237	Sediments	Montana, USA	Late Precambrian, Grinell formation (Belt series)	[110]
238	Argillite	Montana, USA	Late Precambrian, Appekunny argillite (Belt series)	[110]
239	Quartzite	Montana, USA	Late Precambrian, Bonito canyon quartzite (Belt series)	[110]
240	Quartzite	Australia	Late Precambrian, Buldiva quartzite	[146]
241	Lavas	Australia	Late Precambrian, Nullagine lavas	[146]
242	Volcanics	Australia	Late Precambrian, Edith River volcanics	[146]
243	Lavas	Ontario, Canada	Precambrian, Mamainse Point lavas	[167]
244	Lavas	Michigan, USA	Precambrian, Keweenawan lavas (Portage Lake lava series)	[5, 151]
245	Dikes	Ontario, Canada	Precambrian, Logan diabase dykes	[167]

TABLE II—*Concluded*

No.	Rock type	Locality	Age, period, formation	References
246	Sills	Ontario, Canada	Precambrian, Logan diabase sills	[167]
247	Lavas	Ontario, Canada	Precambrian, Alano Bay lavas	[167]
248	Intrusive	Minnesota, USA	Precambrian, Duluth gabbro	[167]
249	Metamorphics	Quebec and Ontario, Canada	Precambrian, Grenville gneisses	[167]
250	Dikes	Michigan, USA	Precambrian, Michigan diabase dikes	[4, 30]
251	Intrusive	Ontario, Canada	Precambrian (1.0–1.2 b.y.[b]), Boulter gabbro	[169]
252	Intrusive	Ontario, Canada	Precambrian (1.0–1.2 b.y.), Thanet and Umfraville gabbros	[5, 169]
253	Intrusive	Ontario, Canada	Precambrian (1.0–1.2 b.y.), Tudor gabbro	[5, 169]
254	Intrusive	Ontario, Canada	Precambrian (1.2–1.8 b.y.), Sudbury basin norite	[5, 169]
255	Tillites	Greenland	Precambrian	[166]
256	Sediments	Texas, USA	Precambrian, Hazel formation (red beds)	[5, 75]
257	Sediments	Texas, USA	Precambrian, Hazel formation (red beds)	[5, 75]
258	Sandstone	Newfoundland, Canada	Precambrian, Signal Hill sandstone	[5, 160]
259	Sandstone	Newfoundland, Canada	Precambrian, Blackhead sandstone	[160]
260	Dikes	Africa	Precambrian (1.3 b.y.), Pilansberg dykes	[5, 170]
261	Intrusive	Africa	Precambrian (2.0 b.y.), Bushveld gabbro	[5, 171]
262	Gneiss	Antarctica	Precambrian, Ongul Island gneiss	[172]

[a] Here, m.y. means million years.
[b] Here, b.y. means billion years.

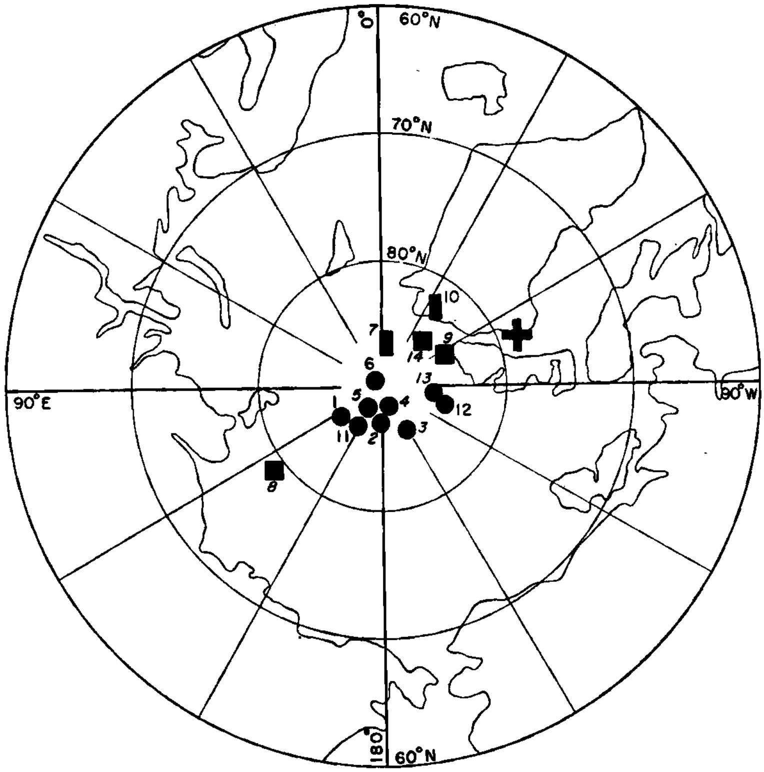


FIG. 10. Recent and late Pleistocene (0 to ½ m.y.) virtual geomagnetic poles. (Schmidt equal area polar projection.)

5.3. Reversals

About half of the rocks that are 30 million years old or younger have directions of magnetization within several tens of degrees of the present field direction, and approximately an equal number have directions of magnetization 180° away from those of the first group. The former are said to be *normally* magnetized and the latter to be *reversed.* Although in older rocks directions of magnetization also commonly fall into two groups 180° apart, neither of the two groups usually has a direction near the present field.

Two theories have been advanced to explain reversed magnetization. Matuyama [98], noting that rocks in Japan and Korea which were younger than early Pleistocene were normal while older rocks were reversed, put forward the hypothesis that the geomagnetic field had undergone a complete 180° reversal. Several decades later Graham [29] found opposing directions of magnetization in apparently contemporaneous sediments,

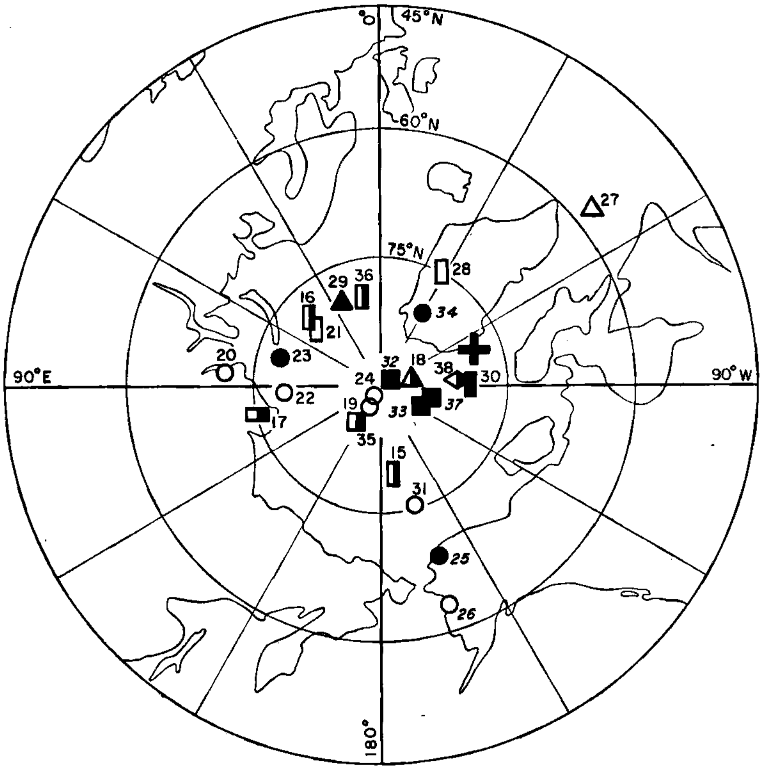


FIG. 11. Early Pleistocene and Pliocene (½ to 12 m.y.) virtual geomagnetic poles. (Schmidt equal area polar projection.)

which induced Néel [173] to make a thorough theoretical investigation of the possibility of a remanent magnetization having a direction exactly opposing that of the ambient magnetic field in which the magnetization was acquired. He found several theoretical mechanisms for such self-reversals. Nagata and Uyeda [28, 36, 174] subsequently discovered a volcanic rock which reproducibly acquires a thermoremanent magnetization opposing the field applied during cooling. Moreover, Balsley and Buddington [32, 33] have found a strong correlation between the polarity of remanent magnetizations of metamorphic rocks and their mineralogy, which also suggests a control other than that of the geomagnetic field. However, although the occurrences of self-reversal cannot be doubted, the majority of specialists in this field of research hold the view that reversals of the geomagnetic field account for most reversals in rocks.

Two approaches are available for determining whether the reversals encountered in a group of rocks are due to self-reversal or geomagnetic field reversal. If self-reversal is the cause, then reversely magnetized rocks should

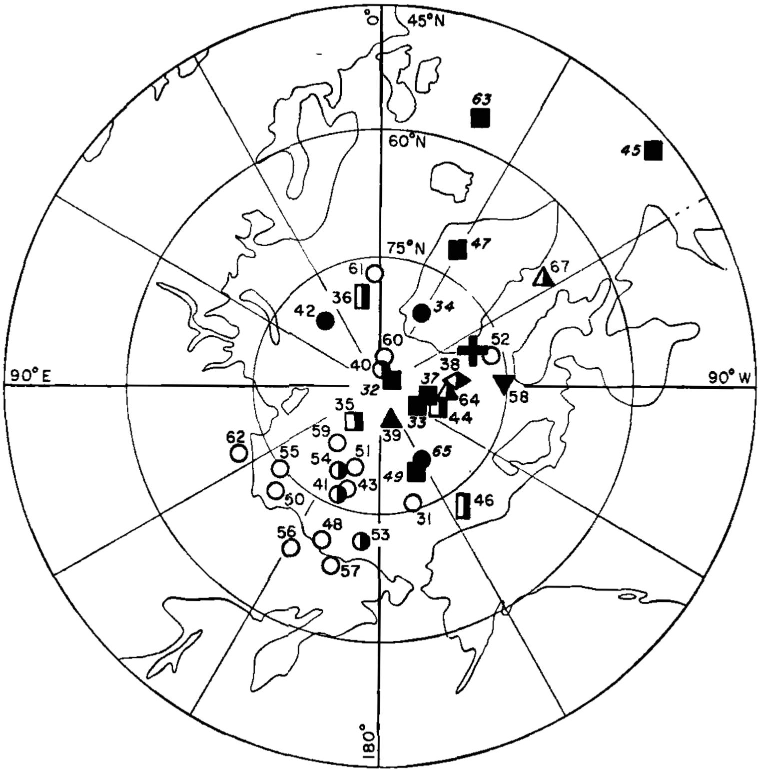


FIG. 12. Miocene and Oligocene (12 to 35 m.y.) virtual geomagnetic poles. See Fig. 13 for poles 66 and 68. (Schmidt equal area polar projection.)

have physical properties different from those of normal rocks, and laboratory tests based on the various theoretical mechanisms of self-reversal may be designed to detect these differences. A second approach makes use of the fact that the distribution in time of normally and reversely magnetized rock units depends strongly on the origin of the reversals.

5.3.1. Laboratory Tests for Self-Reversal. At the outset, the definitive test for self-reversal in igneous rocks might appear to be to heat them up, cool them in a known applied field, and measure the polarity of the resulting TRM. This experiment has been performed on perhaps several thousand igneous rocks by a score of different workers, and to date fewer than 10 self-reversals or even tendencies toward self-reversal have been reported. However, the fact that TRM acquired in the laboratory is normal does not preclude the possibility that the natural TRM of a rock is self-reversed. The reasons for this will appear on examination of some mechanisms of self-reversal.

In one mechanism of self-reversal, the direction of the spontaneous mag-

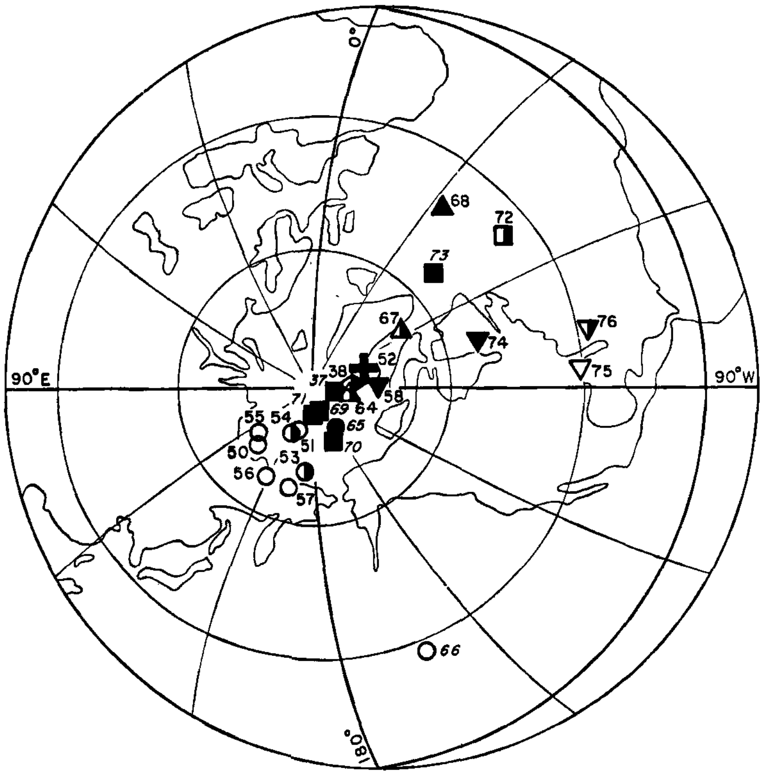


FIG. 13. Eocene (35 to 70 m.y.) virtual geomagnetic poles. (Schmidt equal area projection with pole at 75°N and 90°W.)

netization of a single ferromagnetic mineral reverses during cooling. This phenomenon occurs in ferrites in which two sublattices have different temperature coefficients, and the component with the lower Curie point temperature has the largest spontaneous magnetization after cooling [47]. Gorter and Schulkes [175] have found this type of self-reversal in synthetic minerals, but it has not been found in rocks. A reversal of this type would be reproducible in the laboratory and should be detectable in one heating cycle.

In some ferrites the distribution of cations on different sites is temperature dependent, and when materials of this type have been quenched in a magnetic field, the slow movement of cations toward the low-temperature equilibrium configuration may generate a self-reversal which is strongly time dependent [47]. This type of self-reversal, involving a single ferromagnetic mineral, would not be reproducible in the laboratory. Verhoogen [176] concludes that reversals of this type are possible for magnetite containing substantial amounts of aluminum, the time required for reversal

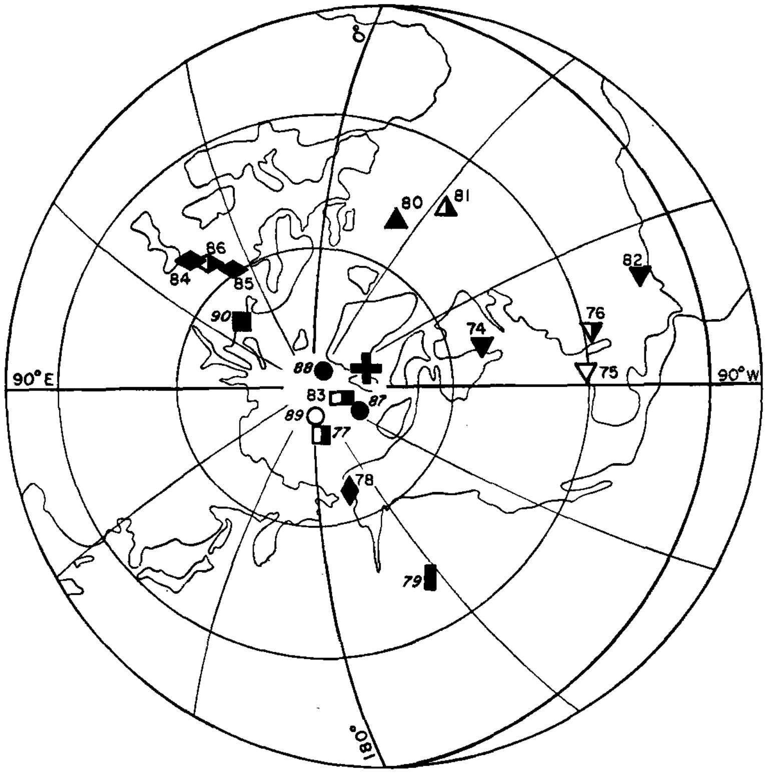


FIG. 14. Cretaceous (70 to 135 m.y.) virtual geomagnetic poles. (Schmidt equal area projection with pole at 75°N and 90°W.)

being of the order of 10^5 to 10^6 years. To test for this type of reversal in a suite of normal and reversed rocks, a comparison of chemical analyses of the ferromagnetic minerals of the two groups would be necessary.

A second group of self-reversal mechanisms requires the presence of two separate magnetic phases, A and B, with different Curie point temperatures. On cooling, constituent A, with the higher Curie point temperature, acquires a TRM parallel to the applied field. On further cooling through the Curie point temperature of constituent B, the previously magnetized A constituent directs the magnetization of B by one of several interaction mechanisms so that it becomes magnetized in a direction opposite to that of A, and hence, to the applied field. If, as the temperature is further lowered, the total magnetization of B becomes larger than that of A, or if A is selectively removed by subsequent chemical action, the material will be reversely magnetized [30, 36, 47].

Magnetostatic interaction may cause constituent B to become reversely

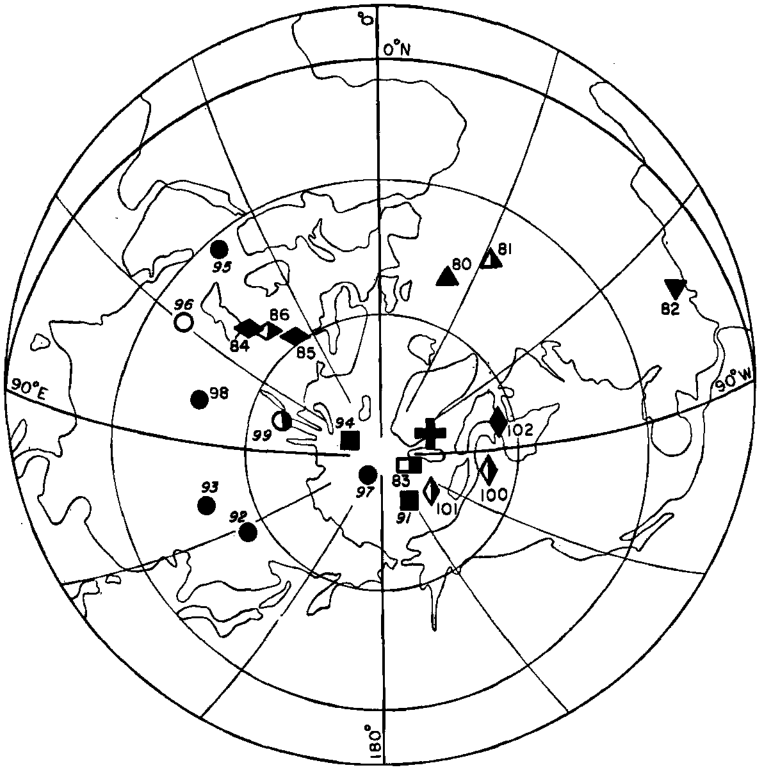


FIG. 15. Jurassic (180 to 135 m.y.) virtual geomagnetic poles. (Schmidt equal area projection with pole at 75°N and 0°E.)

magnetized if the local magnetic field in which B cools, because of the proximity of previously magnetized A grains, has a direction opposed to that of the original applied field. Certain spatial distributions of the A and B constituents are required for this mechanism, and for materials with the spontaneous magnetizations of most minerals found in rocks, it is possible only in very weak applied fields. Uyeda [36] concludes that this type of self-reversal in rocks is possible but not probable in fields as strong as the earth's field today, and no examples have been found in nature.

A second type of interaction between two constituents is by exchange interaction across a common boundary with good registry between the crystal lattices of the two phases [36]. Since the force tending to align spin moments of the B phase is of the order of the Weiss-Heisenberg molecular field, self-reversal due to exchange interaction should occur even in very strong applied magnetic fields, in contrast with magnetostatic interaction. Again self-reversal requires that the total magnetization of B after cooling

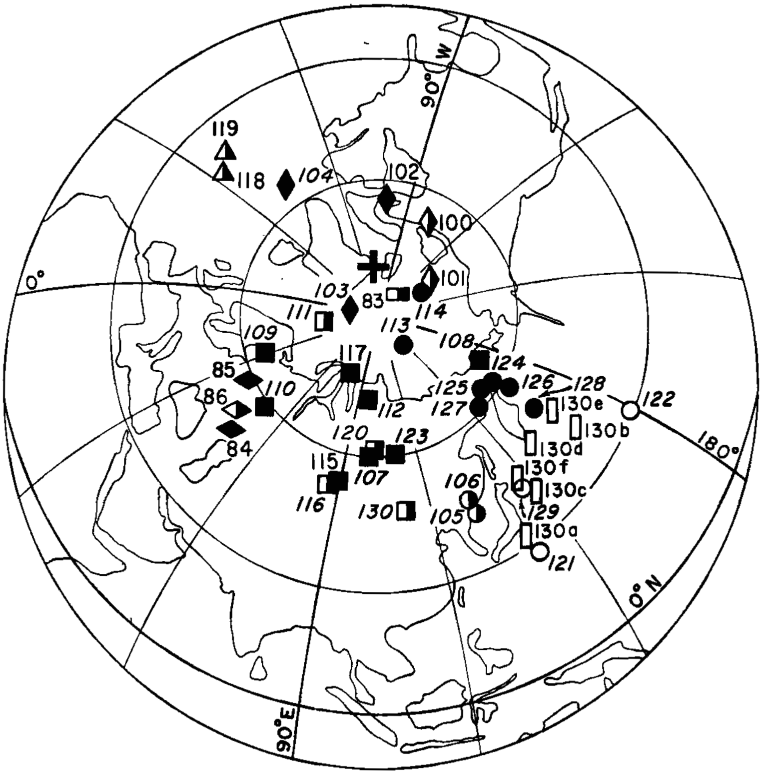


FIG. 16. Triassic (180 to 220 m.y.) virtual geomagnetic poles; Asian data is probably Permian. (Schmidt equal area projection with pole at 75°N and 105°E.)

exceed that of A; for example, if B is ferromagnetic or ferrimagnetic and A is parasitically ferromagnetic, the total magnetization of A would be only a small fraction of that of B [36].

The self-reversal of the Haruna dacite [36], which is the most carefully studied of the naturally self-reversing rocks, appears to be due to this last mechanism. The A and B constituents appear to be, respectively, a parasitically ferromagnetic disordered phase and a ferrimagnetic ordered phase, both having rhombohedral crystal structure and the composition of ilmenohematite. Extensive experiments with synthetic minerals of the ilmenite hematite series indicate that only members with compositions in a narrow range have the self-reversal property; self-reversal occurs in applied fields up to 17,000 gauss.

The two constituent mechanisms for self-reversal are not intrinsically time dependent. However, the development of a two constituent system from a one constituent system by slow exsolution is quite possible, and

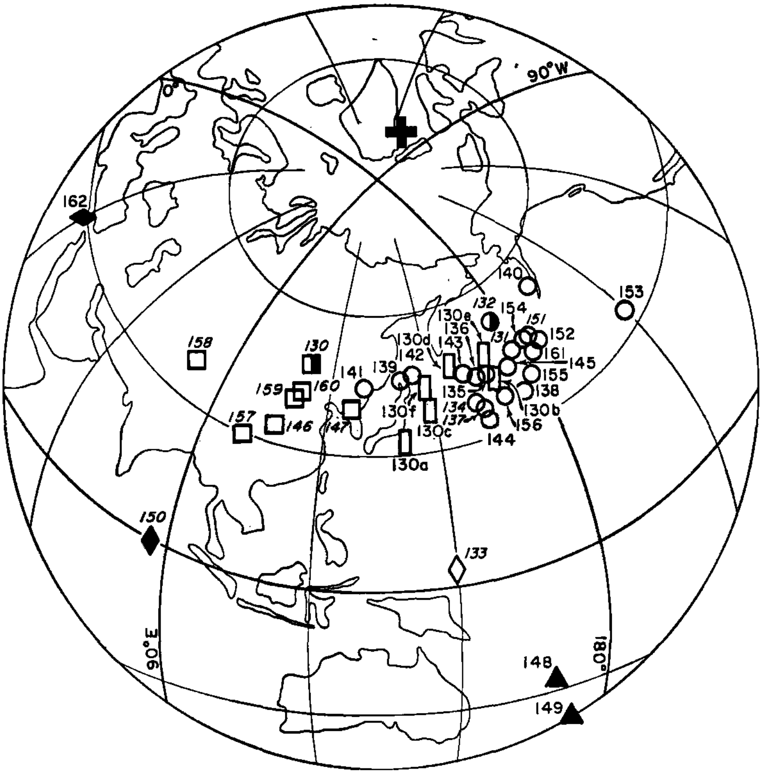


FIG. 17. Permian (220 to 275 m.y.) virtual geomagnetic poles. Data on which pole 130 is based are in part Triassic. (Schmidt equal area projection with pole at 45°N and 135°E.)[a]

exchange interaction might align the exsolving mineral in a direction antiparallel to the original applied field. The time required for the appearance of self-reversals would in this case depend on the rate of exsolution. Such self-reversals might not be reproducible in a heating and cooling cycle, but petrographic study should reveal systematic differences in the mineralogy of normal and reversed samples. For discussions of other laboratory tests useful in studying reversals, see references [26, 28, 36].

5.3.2. Geologic Relations of Normal and Reversed Rocks. Many of the theoretical self-reversal processes are impossible to reproduce and difficult to detect in the laboratory. Fortunately, time plays an essential and different role in the various theoretical self-reversal processes and in the geomagnetic field reversal hypothesis. Therefore, the distribution of reversals in a sequence of rocks of different ages reflects the origin of the reversals.

[a] The atlas continues on p. 282.

If the origin of reversals is one of the instantaneous self-reversal mechanisms such as that of the Haruna dacite, the normally and reversely magnetized rocks will be randomly distributed throughout a group of rocks of different ages. If reversals are due to one of the time-dependent self-reversal mechanisms such as the growth of a second ferromagnetic phase by exsolution, reversals will be increasingly abundant in older rocks; for a given group of rocks, there may be an overlap in the ages of normally and reversely magnetized rocks if the exsolution rates are not uniform, and the age of a reversal horizon will generally vary between groups of rocks containing different assemblages of minerals.

If reversals in rocks are due to geomagnetic field reversals, normal and reversely magnetized groups of rocks will be exactly the same age over the entire earth. The proportion of reversed magnetizations will not be greater among older rocks unless it happens that the earth's field was reversed more at that time. These relations are summarized in Table III.

The stratigraphic distribution of normally and reversely magnetized rocks studied paleomagnetically strongly supports the field reversal hypothesis. All the paleomagnetically studied sediments and lava flows of the past half-million years are normally magnetized, as may be seen in Fig. 10. Underlying this normally magnetized group of rocks is invariably found a reversely magnetized group. The age of the reversal horizon at sampling localities in Japan [98, 99], Iceland [177, 178], France [179], Russia [103], New Zealand [104], and North America [180] is shortly after the beginning of the Pleistocene. Whether the horizons are exactly stratigraphically equivalent depends on the method of intercontinental correlation employed,

TABLE III. Hypothetical Stratigraphic Distribution of Reversals Due to Self-Reversals and Field Reversals.[a]

Instantaneous self-reversal				Time-dependent self-reversal				Geomagnetic field reversal				
N	N	R	R	N	N	N	N	N	N	N	N	
N	R	N	R	N	N	N	N	N	N	N	N	
N	N	R	R	N	N	N	N	N	N	N	N	
R	N	R	N	R	N	N	N	N	N	N	N	
R	N	N	R	R	R	N	N	R	R	R	R	
N	R	N	R	R	N	N	R	R	R	R	R	Time →
R	N	R	N	R	R	N	R	R	R	R	R	
N	N	N	R	R	R	R	R	R	R	R	R	
R	R	N	R	R	R	R	R	N	N	N	N	
R	N	R	N	R	R	R	R	N	N	N	N	

[a] The expected distribution and normal (N) and reversed (R) magnetizations for four different geologic sections are shown under the three different hypotheses.

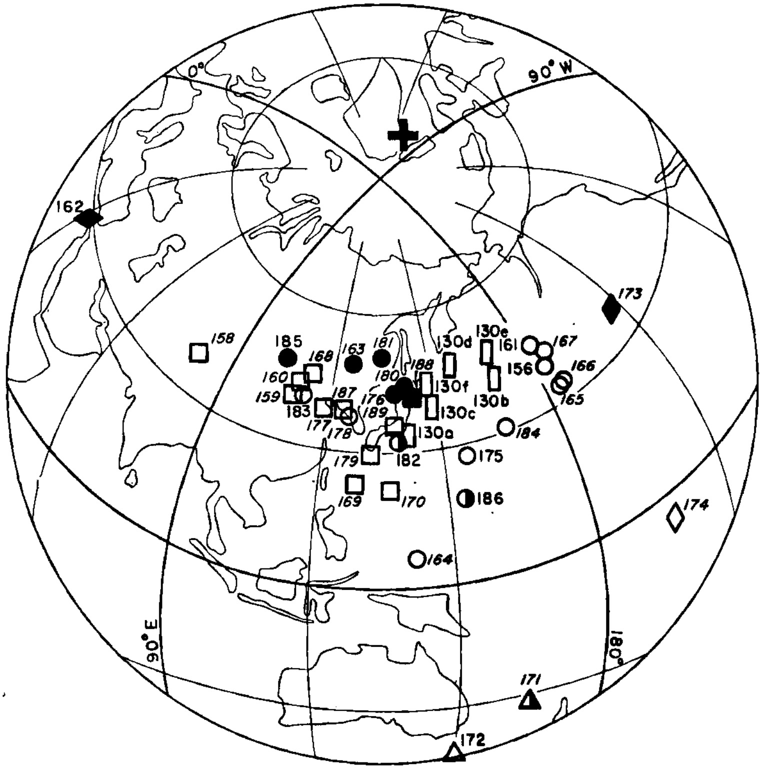


FIG. 18. Carboniferous (275 to 355 m.y.) virtual geomagnetic poles. (Schmidt equal area projection with pole at 45°N and 135°E.)

with disagreements lying well within the limits of geologic uncertainty. Moreover, in the many alternations between normal and reversed groups of strata going back 30 million years, the reversely magnetized zones are in general neither more abundant nor of longer duration.

Other geologic evidence relevant to the origin of reversals is available when rocks with different assemblages of minerals were magnetized contemporaneously. For example, lava flows and the sediments baked by them acquire magnetizations in essentially the same field, and they usually contain vastly different assemblages of ferromagnetic minerals. If their polarities are different, the self-reversal of one is established; if they agree, self-reversal is unlikely. Several such baked zones have been described, and in all cases the polarities agree, regardless of whether the magnetization was normal or reversed [181, 182].

The absence of reversals from all rocks of Permian age also bears on the origin of reversals. Reversals are abundant in rocks younger or older than those from this period, but are completely absent from all rocks which

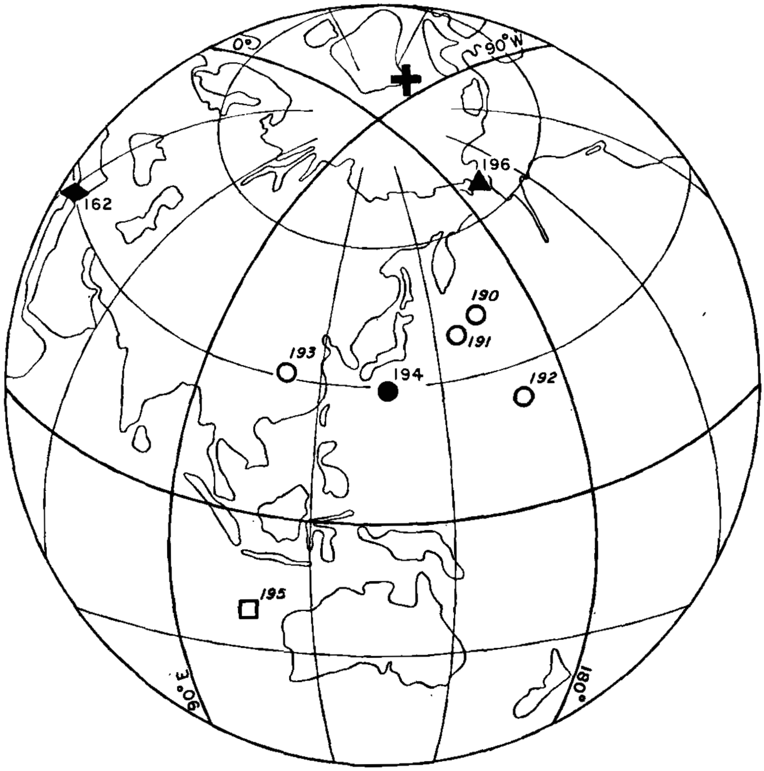


FIG. 19. Devonian (410 to 355 m.y.) virtual geomagnetic poles. (Schmidt equal area projection with pole at 30°N and 135°E.)

cooled or were deposited during the several tens of millions of years spanned by the Permian. (The one exception seen in Fig. 17 is based on a very doubtful interpretation of highly erratic data.) An explanation by means of the self-reversal hypothesis requires that self-reversing minerals were absent from igneous rocks and sediments for several tens of millions of years, but were abundant before and after this time. That the field did not alternate during this period would seem to be a more plausible explanation.

In summary, there can be no doubt that self-reversal occurs in some rocks. However, the number of reproducibly self-reversing rocks is no more than 1 or 2% of the total number investigated in the laboratory, whereas about half of all natural remanent magnetizations are reversed. The stratigraphic distribution of normal and reversed zones, the evidence from baked zones, and the absence of reversals from Permian rocks all indicate that self-reversals are rare in nature and that most reversals in igneous rocks and sediments are due to reversals of the geomagnetic field. Of interest with respect to this conclusion is the fact that the dynamo theory of the

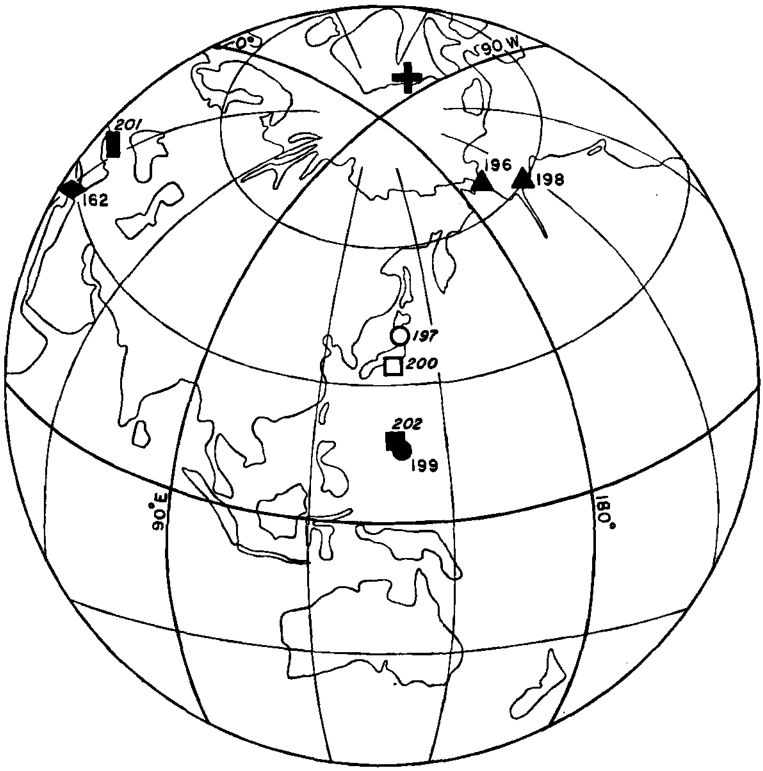


FIG. 20. Silurian (410 to 430 m.y.) virtual geomagnetic poles. (Schmidt equal area projection with pole at 30°N and 135°E.)

origin of the geomagnetic field has no difficulty in accomodating alternations in the polarity of the geomagnetic field [14].

5.3.3. Characteristics of the Reversing Geomagnetic Field. The following summary of the reversal characteristics of the geomagnetic field is based on paleomagnetic evidence from rocks 30 million years old or younger. Characteristics of reversals in earlier periods are in most cases less well known, but there can be little doubt that reversals occurred during the very earliest geologic period studied (Precambrian) and are present in all subsequent periods with one notable exception, the Permian.

Age of most recent reversal horizon: First to second glacial stage of the Pleistocene, approximately 0.2 to 1 million years ago. This estimate is based on the reversal horizon in Japan [98, 99], Iceland [177, 178], France [179], New Zealand [104], Russia [103], and the United States [180]. In some regions the stratigraphic position of this horizon is subject to considerable uncertainty, and inclusion of the full range of geologic opinion about each horizon would broaden the geologic range cited. The broad

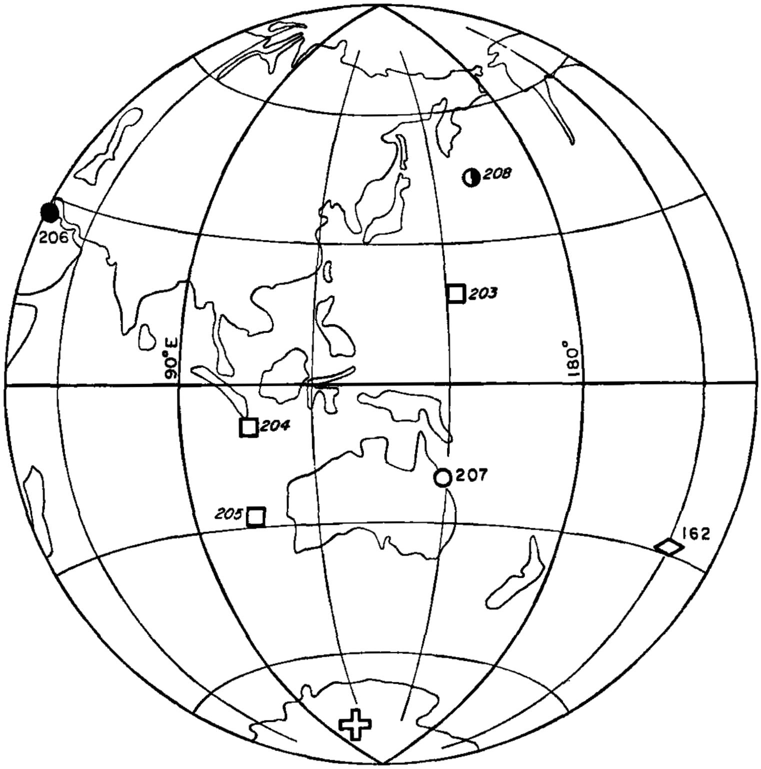


FIG. 21. Ordovician (430 to 490 m.y.) virtual geomagnetic poles. (Schmidt equal area projection with pole at 0°N and 135°E.)

range of 0.2 to 1 million years for the absolute age of the horizon reflects uncertainties in the absolute age of the Pleistocene as determined radiometrically [183].

Periodicity: Whether successive normal and reversed groups represent equal lengths of time is uncertain. In sections of volcanic rock, successive groups of opposing polarity commonly differ greatly in thickness [182], but unequal time intervals are not necessarily implied since rates of extrusion of lava may vary. In a study of sedimentary cores with no apparent breaks in sedimentation, Khramov [103] finds that the two most recent normal zones and the youngest reversed zone are all of approximately equal thickness, suggesting periods of equal length. There is no evidence that periods of either sign are systematically longer or shorter.

Number of magnetic periods during past 30 million years: At least 15 successive groups of rock with opposing magnetic polarity have been described in Iceland, and the total number for this period may exceed 30 [178].

Length of normal and reversed magnetic periods: About half a million

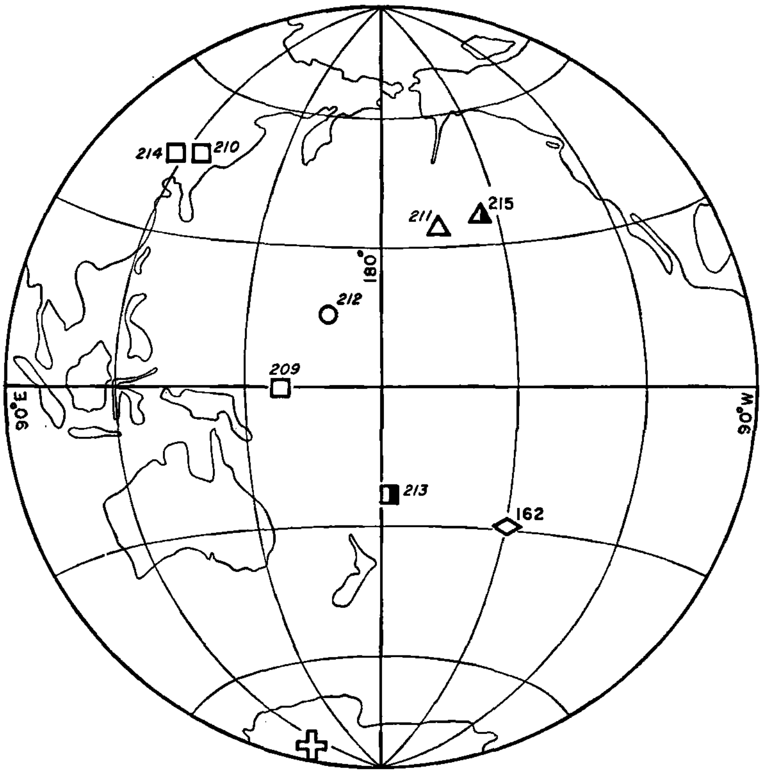


FIG. 22. Cambrian (490 to 600? m.y.) virtual geomagnetic poles. (Schmidt equal area projection with pole at 0°N and 180°E.)

years, based on various geologic estimates [184]. The number of magnetic periods during the past 30 million years sets an upper limit of 2 to 3 million years for the average duration of a magnetic period.

Relative intensity of normal and reversed fields: No rigorous paleomagnetic field intensity measurements have been made on rocks this old. However, the average remanent magnetizations of normal and reversed groups show no systematic intensity difference, suggesting that the normal and reversed field strengths were not vastly different.

Transitions between normal and reversed polarity: Between normally and reversely magnetized strata are sometimes found strata with intermediate directions of magnetization; conversely, strata with intermediate directions are usually found only between normal and reversed groups. The duration of the transition is estimated from direct geologic evidence to be several thousand years [102]; another estimate of 5000 to 10,000 years may be made from Sigurgeirsson's report that rocks having intermediate directions constitute only a few per cent of the total, together with the estimated

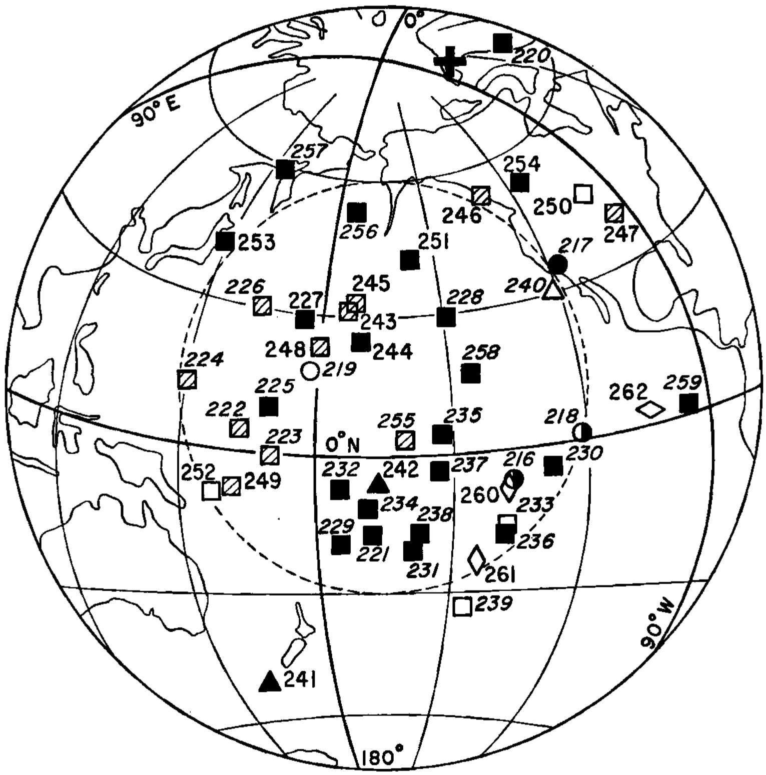


FIG. 23. Precambrian (>600 m.y.) virtual geomagnetic poles; dotted circle encloses one-third area of hemisphere. (Schmidt equal area projection with pole at 15°N and 165°W.)

duration of magnetic periods of half a million years. Field intensities during transition have not been determined accurately; however, Sigurgeirsson [102] finds that lava flows with intermediate directions have intensities about half those of the normal and reversed flows that bracket them, and Brynjólfsson [58] describes one intermediate set of flows with one-fifth the intensity of associated normal and reversed flows. The directions of magnetization of adjacent strata in an intermediate zone typically lie along a smooth curve connecting the normal and reversed directions [58, 102]. Whether this represents a spiraling around of the main geomagnetic dipole or the dominance of nondipole terms as the main dipole collapses before building up in the opposite direction cannot be determined from paleomagnetic directions at one locality. A comparison of virtual geomagnetic poles for the same transition zone, found from paleomagnetic measurements at two widely separated localities, will undoubtedly resolve this interesting problem.

5.4. Interpretations of Paleomagnetically Determined Field Directions

The results of paleomagnetic investigations have been interpreted as evidence for or against a number of geophysical and geological hypotheses, including (1) the dynamo theory of the origin of the geomagnetic field, which requires that the average field directions have axial symmetry about the rotation axis; this average axial field may or may not be dipolar; (2) the hypothesis that the earth's axis of rotation has moved widely with respect to the crust; (3) the hypothesis that large segments of the earth's crust have undergone relative displacements; and (4) the hypothesis that the earth has expanded.

In using paleomagnetic results to test one of the hypotheses enumerated above, assumptions about the validity of one or more of the other hypotheses are usually necessary. These assumptions will be emphasized in the following discussion, since failure of the data to satisfy a test set up for a particular hypothesis can be due to failure of the hypothesis or of the assumptions. An important assumption of a different nature, to be understood in the following theoretical discussions, is that each paleomagnetic datum has satisfied the field and laboratory tests necessary to establish that the magnetic direction corresponds to a direction of the earth's field at a known time in the past. Moreover, it will further be assumed that changes due to secular variation have been averaged out so that each datum represents the average magnetic field for at least some 10^4 to 10^5 years. For many paleomagnetic data little evidence useful for evaluating these assumptions is available, and some departure from theory is to be expected.

The first hypothesis, that the earth's magnetic field has been axially symmetrical and possibly also dipolar, and that the average magnetic axis has also coincided with the rotation axis, may be considered conveniently in several steps. First, the consistency of the paleomagnetic data with an ancient magnetic field having axial symmetry may be examined; other techniques may then be used to determine whether the ancient field was also dipolar; and finally, the coincidence of magnetic and rotation axes may be investigated by noting whether the rotation axis, as found from independent evidence, is parallel to the magnetic axis.

If paleomagnetic data are used to test these possibilities, it is necessary to assume that the orientation of the average magnetic symmetry axis has not changed during the period under consideration, and that there have been no relative displacements between the sampling areas being considered.

Two methods are available for testing the hypothesis of axial symmetry. From potential theory, it can be shown that either of the two following conditions is necessary and sufficient to establish the hypothesis, provided only that the ancient field, like the present one, was essentially a potential field of internal origin. The first condition is that there exists a geocentric spherical coordinate system such that the declination equal zero everywhere. This condition is satisfied if great circles defined by the horizontal projections of the magnetic vectors at each sampling site intersect at a common point, which is the symmetry axis. The second alternative condition is that isoclinal lines defined by the paleomagnetic inclination data are concentric circles; the center of the circles is the magnetic symmetry axis. If the ancient geomagnetic field satisfies one of these conditions, it must also satisfy the other under the assumptions stated.

A special technique has been used by Runcorn [185] to determine whether the ancient field was that of an axial multipole. Possible symmetry axes were first found from the horizontal projections of the magnetic vectors, as described above. It was then shown that the paleomagnetically determined inclinations do not fit the theoretical functions of inclination *vs.* latitude for the field of an axial dipole [$n = 1$, $m = 0$ in equation (2.1)], an axial quadrupole ($n = 2$, $m = 0$), or an axial field of higher degree. However, it should be noted that there is no theoretical reason for expecting a single axial multipole of degree $n > 1$ rather than a linear combination of multipoles of different degree, i.e., the general axial symmetrical case discussed above.

The simplest test to determine whether the ancient geomagnetic field was dipolar as well as axially symmetrical consists of calculating virtual geomagnetic poles, as described in Section 5.1.1. The hypothesis is valid if the virtual geomagnetic poles from the sampling areas under consideration coincide.

To test for a coincidence of the magnetic and rotation axes, the orientation of the magnetic symmetry or dipole axis, as found by one of the above methods, may be compared with the orientation of the rotation axis as determined by some independent method.

In general, the area under consideration is the entire earth and if there has been relative displacement between the sampling areas being considered the tests described above will fail—not because the hypothesis is invalid, but because one of the assumptions does not hold. However, if relative displacements have occurred, both the magnetic and rotation axes for a given area will have been displaced and a test for the coincidence of the two axes may still be possible using local rather than worldwide data. Certain characteristics of sedimentary rocks and of assemblages of fossils may be indicative of ancient climates and by inference also of geographic latitude (although it should be noted that some specialists in this field have strong reservations about the validity of such inferences based on data from sampling areas of subcontinental dimensions). Furthermore, the assumption that the geomagnetic field was axially symmetrical requires that inclination be a function only of distance from the symmetry axis. Thus, if we assume some functional relationship between inclination and distance from the symmetry axis, e.g., the dipole relationship [equation (5.4)], a unique magnetic latitude may be calculated for each sampling area [127]. Agreement between the inferred magnetic latitudes and the latitudes based on geologic evidence supports the hypothesis.

Contemporary interest in hypothesis (2), displacement of the earth's axis of rotation, stems largely from re-examinations of the theory of polar wandering by Gold [186], Munk [187, 188], and Munk and MacDonald [189], and also from paleomagnetic results [161]. In addition to the assumption that there be no relative displacements between the areas being considered, it must also be assumed that the hypothesis concerning coincidence of magnetic and rotation axes discussed above is true. Under these assumptions, the hypothesis is valid if magnetic symmetry (or dipole) axes differ from period to period.

Almost all variants of hypothesis (3), "continental drift," envisage rotations and translations of large land masses which retain their original shape and size during displacement. No assumption about the coincidence of magnetic and rotation axes is necessary, but the configurations of the ancient geomagnetic field must be assumed in order to test this hypothesis with paleomagnetic data. Usually the dipole field configuration is assumed, and under this assumption the hypothesis is valid if virtual geomagnetic poles calculated from measurements on rocks of a given age from one land mass differ from contemporaneous data from another land mass. The former relative positions of the two land masses are not determined uniquely by one pair of ancient field directions, however, since the only restriction for a given land mass is that it lie, with proper orientation, on an appropriate isoclinal line. This fundamental ambiguity may also be seen by noting that consistency between results from two land masses is achieved so long as their magnetic poles coincide, thus allowing the land masses to move along geomagnetic latitude (or isoclinal) circles. Some special techniques and examples of reconstructions are given by Irving [190].

In recent discussions of the expanding earth hypothesis by Egyed [191], Carey [192], and Heezen [193], the continents are thought to maintain a constant area and nearly the same shape during expansion, and the expansion is considered as being due to an increase in the area of the oceans. As pointed out by Egyed [191], this hypothesis may be tested by paleomagnetic data from two widely separated localities on the same continent. If the earth has expanded according to this model, the distance between the two points will not have changed, but the geocentric angle between them will have decreased. To test the hypothesis with paleomagnetic data, the ancient field configuration must be assumed. If a dipolar configuration is assumed, then the ancient radius is given by:

$$R_0 = \frac{d}{[\cot^{-1}(\frac{1}{2}\tan I_1) - \cot^{-1}(\frac{1}{2}\tan I_2)]} \tag{5.10}$$

where I_1 and I_2 are the paleomagnetic inclinations at the two sampling sites and d is the linear distance between the ancient latitude circles on which they lie.

A brief summary of these hypotheses, the tests used in comparing them with the paleomagnetic data, and the assumptions made in setting up the tests is given in Table IV. It may again be noted that a failure of any test indicates that either the hypothesis or the assumptions (or both) are not valid.

5.4.1. Hypotheses 1a and 1b: Dipolar Axially Symmetrical Field. All but one of the virtual geomagnetic poles for the late Pleistocene to Recent (0 to ½ m.y.) (Fig. 10) fall within a circle less than 10° in radius centered on the present geographic pole; the one exception is based on a study of sediments, which are known experimentally to be subject to an inclination error. Comparison of these data with virtual geomagnetic poles calculated from current field directions over the surface of the earth (Fig. 8) reveals that the paleomagnetically determined poles are more tightly grouped than the poles calculated from the present geomagnetic field. The reason for this initially surprising result is that every paleomagnetic datum is based on an average of at least several directions of remanent magnetization, each cor-

TABLE IV. Tests and Assumptions Made in Comparing Paleomagnetic Data with Several Hypotheses.

Hypotheses	Tests	Assumptions
1a. Axially symmetrical field	Declination is zero everywhere; or isoclinal lines are concentric circles	Hypotheses 2, 3, and 4 are not valid
1b. Dipolar field	Virtual geomagnetic poles all coincide	Hypotheses 2, 3, and 4 are not valid
1c. Magnetic and rotation axes coincide	Comparison of magnetic axis and rotation axis, as determined by independent method	Hypotheses 2, 3, and 4 are not valid
2. Polar wandering	Ancient magnetic axis displaced with time	Hypothesis 1c is valid
3. Continental drift	Contemporaneous poles from different continents displaced	Hypothesis 1b is valid
4. Expanding earth	Change in length of degree of latitude	Hypothesis 1b is valid

responding to an instantaneous direction of the ancient geomagnetic field. The statistical confidence limits about each virtual geomagnetic pole is different because the poles are based on different numbers of measurements and because the observational errors differ; the magnetic poles farther from the present geographic pole usually have larger confidence limits which, in almost all cases, encircle the geographic pole. Thus the small deviations of these virtual geomagnetic poles from the present geographic pole may not be significant [5]. These paleomagnetic results made from widely separated areas indicate that the geomagnetic field during the past half-million years, when averaged over an interval of time of the order of 10^4 to 10^5 years, has very closely approximated the field of a geocentric dipole [184].

Turning to the Pliocene to early Pleistocene (12 to ½ m.y.) and Oligocene to Miocene (35 to 12 m.y.) maps, the same generalizations appear valid, with the qualification that as we go back in time the groupings of poles are progressively more dispersed. Several possibilities are available for explaining why these older rocks show more dispersion of virtual geomagnetic poles. The hypothesis being tested, that the average geomagnetic field was dipolar, may not be valid. Alternatively, one of the assumptions made in setting up the test may not be justified. The paleomagnetic data from these older rocks may not correspond to average ancient field directions so well as did the data from rocks half a million years old or less. For magnetically unstable rocks, this would result in south magnetic poles (solid symbols) too close to the present pole; for random disturbing effects, the virtual geomagnetic poles will show random displacement and will have larger statistical confidence intervals. Among such virtual geomagnetic poles that probably do not correspond to average ancient field directions are poles 21 and 27 (only one point in time), 25 and 26 (inclination error in sediments), and probably most of the sediments in Fig. 12 (unstable) with the exception of pole 63 of Oligocene age, for which there is evidence of magnetic stability. Alternatively, the assumption that the magnetic axis, when averaged over 10^4 to 10^5 years, has had the same orientation throughout the time interval represented on each map may not be valid. A third possibility is that sampling areas may have been displaced, possibly indicating continental displacements. If the last alternative is correct, then virtual geomagnetic poles from the same continent will be internally consistent but will differ from those of other continents [2]. If the average axis has moved, between-continent and within-continent consistency should be about the same.

The most widely held interpretation of the data from the Oligocene period to the Recent (Figs. 12, 11, 10) is that the average magnetic dipole axis as seen from North America and Europe moved northward along the 120th east meridian a total distance of 10° during this interval of time, coming into coincidence with the geographic pole during the Pleistocene [4, 110, 161]. Virtual geomagnetic poles from other continents are then interpreted as indicating displacements of these continents with respect to North America and Europe. Rotations of Japan, for example, have been suggested by Irving [2] on the basis of pole 16 and by Nagata *et al.* [106] on the basis of pole 46. These two rotations are not progessive, however; the first is clockwise, and the second counter clockwise (Figs. 12, 11).

The alternative possibility that the observed dispersion of poles is due to movements of the average dipole axis appears to us to fit the available data from these two periods at least as well. Superimposed on the tendency for magnetic poles to group around the present rotation axis, there may be a fine structure of displacements of the magnetic axis as large as 15°, even when the magnetic axis is averaged over several million years. The outlines of one such pattern may be dimly discernible in the Miocene. In Fig. 24

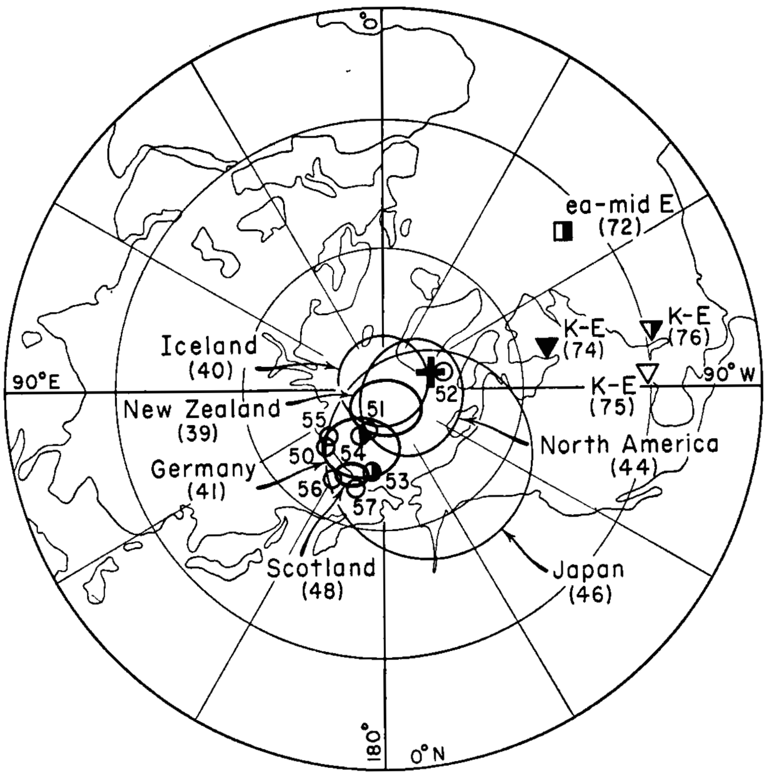


FIG. 24. Summary of Miocene paleomagnetic data, shown as 95% statistical confidence ovals. Also shown are Eocene virtual geomagnetic poles from India (triangles) and North America (square) and poles of doubtful age from Europe (numbered small circles).

are plotted the statistical confidence limits [α_{95} of equation (5.7)] of what, in our judgment, are the six most reliable paleomagnetic studies of relatively well dated Miocene formations. All but two have confidence limits at the 95% probability level (shown as ovals) which include the present geographic pole; the two remaining ovals exclude the present geographic pole but agree with each other. One of the pair (pole 48) is based on Kahn's study of the Skye lavas from Scotland, in which 488 samples from 53 lava flows were used; his laboratory analysis for magnetic stability is one of the most exhaustive ever made [44]. The other, Angenheister's study of the Vogelsberg basalts from Germany (pole 41), is based on 200 samples from 42 lava flows [112]; laboratory tests and excellent agreement between normal and reversed groups of flows point to paleomagnetic reliability. All six of these Miocene measurements are consistent with an average magnetic axis within several degrees of 77°N and 165°E.

Also shown on Fig. 24 are European poles 50, 51, 53, 54, 55, 56, and 57; the ages of the volcanic rocks on which these poles are based is uncertain, and the range of geologic uncertainty is from Eocene to Miocene or younger [115, 116]. Whether the age of these formations is Eocene or younger will later be seen to be a crucial point in assessing the significance of Eocene paleomagnetic results from all over the world (also shown on this map). In most paleomagnetic interpretations, an early Tertiary or Eocene age has been favored for poles 50, 51, 53, 54, and 55 [2, 4]. The directions of magnetization, however, appear to us to be equally consistent with a Miocene age.

Much greater dispersion of virtual geomagnetic poles is apparent on maps for the older periods, and the hypothesis (or its related assumptions) clearly fails when the area under consideration is the entire earth. However, Permian data are now available from sampling sites almost 5000 km apart in western Europe and Siberia (poles 130a, 130b, 130c, 130d, 130e, 130f, Fig. 17). The good agreement between these poles from widely separated, probably undisplaced sampling areas would be improbable if the Permian field had not been dipolar. Although Triassic data from the eastern and western United States are in general agreement [142] the scatter in the virtual geomagnetic poles (Fig. 16) appears to be too great to make a definitive test of whether the field was dipolar or not. No other sets of contemporaneous data from widely separated localities on one continent are available for additional tests of this type.

In summary, the geomagnetic field during the last ½ m.y. has very closely approximated that of a geocentric dipole parallel to the axis of rotation rather than the present geomagnetic pole. Going back to the Oligocene (35 m.y.) the same general conclusions may be drawn, but with a progressively poorer fit. This may be due to a failure of the dipole hypothesis in earlier times, but is more probably due to small amounts of polar wandering or continental drift during this time. The dipole hypothesis, or more probably the assumptions, fail when data from earlier periods and for the entire earth are tested; however, limited Permian data from European and Siberian sampling areas almost 5000 km apart are quite consistent with the dipole hypothesis, but indicate a magnetic axis far removed from the present axis of rotation.

5.4.2. Hypothesis 1c: Dipole Axis Coincident with Rotation Axis. The difficulty in testing this hypothesis is less in determining past positions of the magnetic pole than in determining past positions of the rotation axis. As previously discussed, there is little doubt that during the past half-million years the virtual geomagnetic poles were tightly grouped about the present position of the rotation axis rather than the present geomagnetic pole in Greenland. The probable position of the rotation axis may be inferred from two lines of evidence. The distribution of glacial moraines and other geologic evidence of cold climate during the Pleistocene are not consistent with large displacements of the rotation axis from its present position, but polar wandering less than 10° to 15° would probably not be detectable. Estimates of displacements of the rotation axis from astronomical observations indicate that between 1900 and 1940 the axis moved in the direction of Greenland a distance not exceeding 15 ft [194], corresponding to a rate of 1° per million years. Thus, unless average rates of displacement of the rotation axis during the past half-million years have greatly exceeded the rate during this century, the rotation and magnetic dipole axes coincided very closely. This is certainly one of the most important results obtained from paleomagnetism, since as Hospers [88, 184] and Runcorn [18, 195] have pointed out, it indicates that the present geomagnetic pole 11½° from the geographic pole is a transient phenomenon; the persistent part of the geomagnetic field during the past half-million years has been a dipole component aligned with the rotation axis, as predicted by the dynamo theory.

Turning to the Oligocene to Miocene and Pliocene to early Pleistocene groups of paleomagnetic data, virtual geomagnetic poles are more dispersed and geologic and geophysical evidence for the positions of the rotation axis is less conclusive. Some authorities interpret the geologic evidence as indicating geographic pole displacements of several tens of degrees [196–198], while others find evidence of no polar displacement [199–201]; paleogeographic techniques probably do not have sufficient resolving power to permit a detailed comparison of possible fine structures present in the average positions of geographic and geomagnetic poles.

Stepping back to view these data on a larger scale, we note that the more reliable paleomagnetic data are in agreement with most geologic interpretations in requiring no displacements exceeding 20° during the past 35 million years; moreover, there is no evidence to show that the average magnetic and rotation axes were not in detailed agreement during this time as they were during the past half-million years and as predicted by the dynamo theory.

The position of the rotation axis as reconstructed from geologic evidence is less well known for earlier times. The reconstructions from paleoclimatic evidence by Kreichgauer [196] and by Köppen and Wegener [197] differ from each other by distances usually less than several tens of degrees for periods back to the Carboniferous, but both are in much greater disagreement with paleomagnetically determined virtual geomagnetic poles. On the other hand, Irving [127] and subsequently others [202, 203] have cited a number of sedimentary deposits suggestive of ancient tropical or cold climates which are discordant with their present geographic latitudes but which agree with paleomagnetically determined latitudes. The converse situation has not, to our knowledge, been reported. Additional work in the important field of paleogeography is certainly desirable.

5.4.3. Hypothesis 2: Displacement of Earth's Rotation Axis. If the assumptions for testing this hypothesis are valid, then the paleomagnetic evidence clearly indicates that the rotation axis has been displaced with respect to the earth's crust. It is impossible to look objectively at the grouping of virtual geomagnetic poles for the Carboniferous and Permian periods from North America and Europe, coming as they do from a wide variety of rock types, and come to any conclusion other than that these data represent the general direction of the geomagnetic field as seen from Europe and North America in the latter part of the Paleozoic era. The crucial assumption required to infer displacement of the rotation axis is that it has coincided with the average magnetic axis in the past.

Öpik [204] has argued that even given the dynamo theory for the origin of the geomagnetic field, the rotation and magnetic symmetry axes need not coincide; he suggests that lateral temperature differences in the mantle may give rise to convective motions in the fluid core not symmetrical about the rotation axis. A theoretical analysis of the effects of temperature boundary conditions on the magnetohydrodynamic regime of a rotating fluid sphere has not been made to our knowledge, and Öpik's suggestion must be kept in mind. On the other hand, the position of the rotation axis is better defined during the past half-million years than at any earlier time, and the impressive agreement of the paleomagnetic results from this time interval with the usual form of the dynamo theory makes coincidence of the magnetic and rotational axes in the past appear rather likely.

Details of the paleomagnetic polar wandering curves will be discussed in the next section, since these curves also are of great importance in assessing the validity of the continental drift hypothesis.

5.4.4. Hypothesis 3: "Continental Drift." If virtual geomagnetic poles from rocks magnetized at the same time in different sampling areas do not coincide, and if all the pertinent assumptions are valid, then it follows that the sampling areas have undergone relative displacements since the rocks were magnetized. The maps in Section 5.2.2. show a large number of poles from formations of about the same age that are not coincident, indicating that if all the assumptions are valid many of the sampling areas have undergone immense relative displacements.

A method commonly used to compare paleomagnetic results from different continents is to join successive virtual geomagnetic pole positions from each continent with a smooth curve terminating at the present geographic (and average geomagnetic) pole. Such a path of polar wandering was first drawn for the combined paleomagnetic results from Europe and North America by Creer, Irving, and Runcorn [161], and subsequently separate paths have been drawn for North America [4, 110, 127], Europe [4, 110, 127], Australia [146], Japan [106], and Antarctica [132]. The Precambrian data from North America and Europe have been interpreted in terms of two pairs of curves, curve 1 (Fig. 25) by Creer and associates [4] and curve 2 by Du Bois [205]; Collinson and Runcorn [110] discuss the difference between these two interpretations. In Fig. 25 the numbers adjacent to the polar wandering paths indicate the approximate age, in millions of years, of some of the formations from which paleomagnetic data have been derived.

On the basis of these polar wandering curves, a westward displacement of North America with respect to Europe of 24° to 45° has been proposed by Irving [127], Runcorn [206], and Du Bois [151]. As data from other continents have accumulated, displacements of most of the continents have also been put forward; one of the larger displacements is that proposed for India, amounting to a translation of 4000 to 5000 km and rotation of 24° with respect to Europe during the past 70 million years [100, 120, 125]. Perhaps the best documented case for large-scale continental drift is that suggested by Irving [146] for Australia, where Carboniferous and Permian paleomagnetic results from both sediments and lava flows are internally consistent, but have virtual geomagnetic poles displaced more than 70° from those for North America and Europe.

These interpretations of the paleomagnetic data are certainly among the most challenging geophysical ideas to be put forward in recent years and are worthy of careful consideration by all who are concerned with the origin and history of the continents. However, a comparison of these curves with the virtual geomagnetic poles in Section 5.2.2 shows that while some parts of the curves are supported by considerable amounts of consistent data, others are less well supported by the data now available; for most geologic

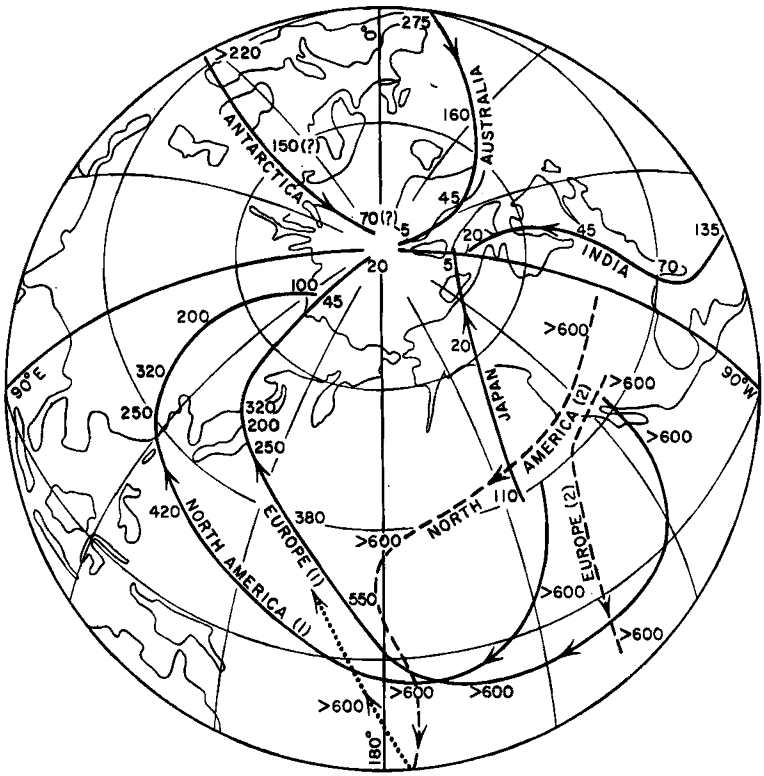


FIG. 25. Paleomagnetic polar wandering curves proposed for Anarctica [132] Japan [106], India [100, 106, 120, 125], Australia [146], and two sets of curves for Europe and North America (1) [4, 127] and (2) [205]. Numbers adjacent to paths indicate approximate age in millions of years of some of the formations from which paleomagnetic data were derived.

periods and most continents, the present state of the science is one requiring additional paleomagnetic study of suitable, well-dated rock formations in order to better delineate the configuration of the geomagnetic field. For only a few geologic periods is the geomagnetic field configuration known with sufficient precision to be useful for geologic applications such as dating rocks.

In concluding this section, several problems of current interest will be described. The first concerns the interpretation of the paleomagnetic results from Europe and North America for the Permian and Carboniferous. The virtual geomagnetic poles for North America appear to be systematically displaced from those for Europe, especially for the Permian period. However, when the Carboniferous and Permian data are considered jointly, a most interesting problem emerges. The two paths of polar wandering going

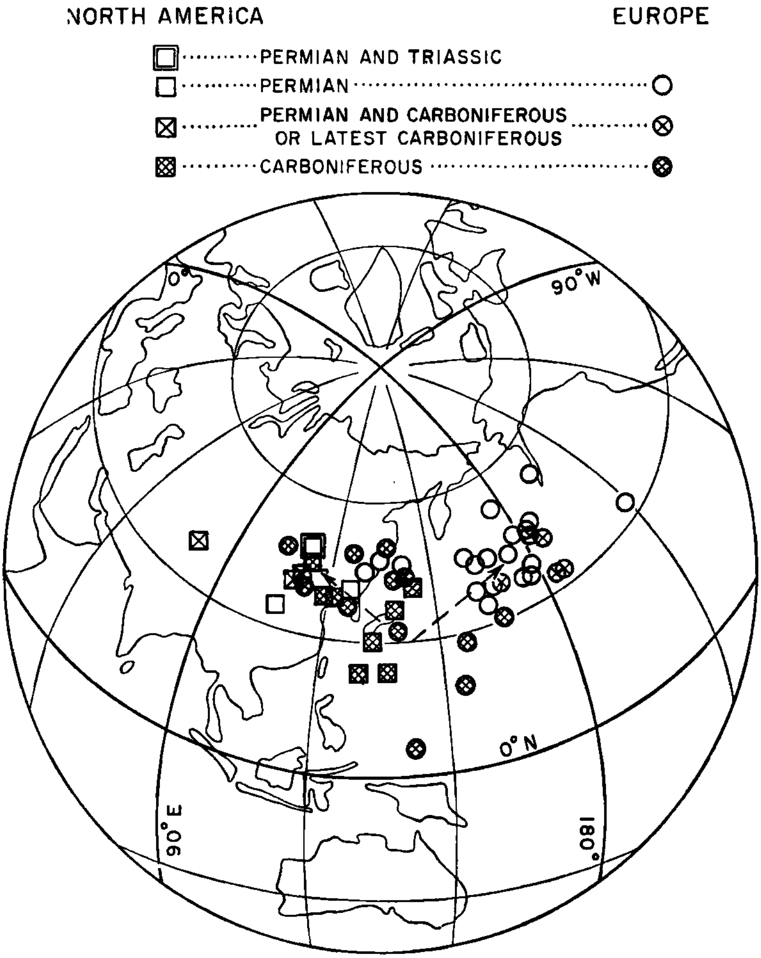


FIG. 26. Permian and Carboniferous virtual geomagnetic poles from North America and Euorpe. Divergence of the two possible polar wandering paths (dashed arrows) corresponds to the two kinks in the curves of Collinson and Runcorn [110].

from Carboniferous to Permian, as shown in Fig. 26 by dashed arrows, may be seen to diverge. These two arrows are nearly parallel to the two kinks in the otherwise smooth polar wandering curves in a very recent paper by Collinson and Runcorn [110], which appeared after the illustrations for this review had been prepared. Collinson and Runcorn suggest that these kinks may be the result of North America and Europe moving apart during the Permian. This explanation, however, does not account for the direction of the movement of the poles, since if North America and Europe had moved apart during the Carboniferous and Permian, the two reconstructed paths of polar wandering would converge, rather than diverge by more than 90° as indicated. As pointed out by Collinson and Runcorn, additional data from rocks of this age, especially igneous rocks from North America, are highly desirable.

A second continental drift problem of current interest concerns the postulated drift of 4000 to 5000 km of India during the Tertiary period. This displacement is based on a comparison of data from the Eocene or Cretaceous Deccan Traps of India having a virtual geomagnetic pole near Florida, with data from western North America and Europe, indicating virtual geomagnetic poles near the present geographic pole (Figs. 13 and 24). Pole 74 is based on younger lava flows from India and pole 68 on lava flows of possible Eocene age or younger from Tasmania; however, other inclination data [207] from Tasmanian basalts of probable early Tertiary age indicate a contemporaneous pole closer to that for Australia (pole 67).

A weak link in this reconstruction appears to us to lie in the argument that at the time the Indian Deccan Traps were magnetized, the field as seen from North America and Europe corresponded to a virtual geomagnetic pole in the vicinity of 75°N and 130°E. The data from Europe usually regarded as Eocene in paleomagnetic interpretations (poles 50, 51, 52, 53, 54, 55) are all from volcanic formations which, as far as the geologic evidence is concerned, may be of any age from Eocene to Miocene or younger [115, 116]. As discussed in Section 5.4.1, their directions of remanent magnetization agree well with Miocene results from elsewhere in the world. The Eocene data from North America are mostly from sedimentary formations; since most of the directions of magnetization are close to the present field, since reversals are completely absent, and since other evidence of stability is also lacking, these data may not be a true indication of the Eocene magnetic field in North America. One well-dated volcanic formation of lower to middle Eocene age from North America (pole 72) has been studied paleomagnetically; indications of paleomagnetic reliability include the observations that the remanent magnetization is demonstrably stable in the laboratory, the magnetizations of normal and reversed groups are consistent, and Graham's fold test is satisfied [64]. Irving [2] has pointed out that the sampling area in western Oregon is in a tectonically active region, and a rotation of the sampling area may have resulted in the present position of pole 72. An alternative explanation is that the geomagnetic field in the lower Eocene was oriented with its dipole axis in the Atlantic Ocean near Florida [5]. Additional sampling in widely separated localities will be necessary to decide between these two interpretations.

A final intriguing result relevant to the problem of continental drift is the distribution of Precambrian virtual geomagnetic poles from all continents. Most of these poles lie within a circle centered about 15°N and 165°W covering one-third the area of a hemisphere. Especially interesting are the two poles from Africa, the radiometric ages of which are 1290 million years and 2000 million years [171, 208] (poles 260 and 261, respectively). Although the difference in their ages is greater than all of the time elapsed since the end of the Precambrian, the distance between these poles is only 18°, suggesting very little polar wandering during this part of the Precambrian. The general grouping of Precambrian poles from different continents does not appear to be entirely random and may be useful in setting an upper limit to the amount of possible continental drift during and since the Precambrian. Random displacements of all sampling areas at rates comparable with that suggested for India for the past 70 million years, for example, would certainly have produced Precambrian virtual geomagnetic poles more randomly dispersed over the earth's surface.

5.4.5. Hypothesis 4: Expanding Earth. In order to make a direct calculation of the earth's radius in the past, data are required from two sampling localities which have not undergone relative displacements and which are at different magnetic latitudes. An ideal geologic period to use is the Permian, during which the geomagnetic field was remarkably steady. The abundant data from rocks of this age from western Europe, together with a recent careful paleomagnetic study by Gusev [148] of 1600 samples from Siberia, now make it possible to establish some limits for changes in the earth's radius during the past 250 million years.

By means of the technique described above, 80 values of R_0 [equation (5.10)] for the Permian period have been found by using each of the 16 European determinations (poles 134 to 144 and 151 to 155) with each of the five Siberian data. The average of the 80 values thus calculated for the earth's radius during Permian time is 6310 km. Although no rigorous statistical analysis of this result is justified because of the widely varying statistical confidence limits of the individual data, the results of a standard statistical analysis will give some measure of the spread of the data: the standard deviation of the individual measurements is 1080 km, and the standard error ($N = 21$) of the mean is 230 km.

Comparing these results with the present earth radius of 6370 km, we may come to the following conclusions. The average Permian magnetic field, as seen from two sampling areas almost 5000 km apart on the Eurasian land mass, shows no significant departure from the dipolar configuration and is perfectly consistent with a Permian earth radius equal to the present one. Although precise statistical confidence limits cannot be assigned to the value found for the Permian earth radius of 6310 km, the accuracy is probably not great enough to confirm or reject expansion at a rate as small as the 0.4 to 0.8 mm/year figure suggested by Egyed [191]. However a total increase as large as 15% in the past 250 million years seems unlikely, and this small expansion is inconsistent with Heezen's recent suggestion [193] that differences in virtual geomagnetic poles from different sampling areas may in general be due to an expansion of the earth.

An alternative interpretation of this result is that one of the assumptions is not valid. For example, if the sampling areas have been displaced so that d [equation (5.10)] has changed proportionally to the change in earth radius, then an earth expansion would be completely masked. Carey's [192] detailed reconstruction for Eurasia in the late Paleozoic time indicates a relative displacement of the European and Siberian sampling areas used in the above analysis, but the implied increase in d does not appear sufficient to mask an increase in the area of the earth as large as the 45% postulated by Carey.

To sum up the interpretation of paleomagnetic data in terms of polar wandering, continental drift, and an expanding earth, there can be little doubt that the magnetic pole has wandered during geologic time; similar wandering of the rotation axis also appears a likely possibility. Moreover, making reasonable assumptions about the earth's field in the past, it is difficult to explain all of the presently available paleomagnetic data without invoking continental drift. However, serious problems remain concerning many details of the polar wandering curves, and those shown in Fig. 25 may be expected to undergo marked changes in the future. In our opinion, future investigations of these details may well alter some of the current conclusions about continental displacements. The hypothesis of an expanding earth is not supported by the paleomagnetic evidence, although present data are not adequate to preclude an expansion at a rate as slow as that suggested by Egyed [191]; it does, however, preclude expansion of an amount sufficient to have caused the departures observed between present field directions and those determined paleomagnetically.

6. Summary

At the conclusion of the section describing the salient characteristics and trends of the geomagnetic field during the period of direct observation, several questions were posed concerning the extension of these properties back to earlier times. Some of these questions can now be answered from the results of paleomagnetic research.

The intensity of the main dipole component of the geomagnetic field appears to have been decreasing since well before the period of direct observation. The Thelliers' careful paleomagnetic measurements of geomagnetic field intensities back to 600 B.C. indicate a decrease at an average rate of about half that found from spherical harmonic analysis during the past century.

When averaged over 10^4 to 10^5 years, the configuration of the earth's field has been much closer to that of a geocentric dipole than it is at present. Moreover, during the past half-million years, the geomagnetic pole has not been at its historic position in north Greenland, but rather has had an average position closely coincident with the present geographic pole. With progressively less precision, the geomagnetic field has had these same general properties during the past 35 million years. However, during this earlier period a fine structure of displacements of the average magnetic axis by amounts up to 15° appears possible, and data now available suggest one such displacement of about 12° in the Miocene.

Reversals of polarity of the geomagnetic field appear very probable, the most recent one having occurred shortly after the beginning of the Pleistocene period. Reversals are abundant in all parts of the stratigraphic column with the exception of the Permian period, from which they are entirely absent.

In addition to these important results, paleomagnetic research has furnished much new evidence of interest to the geological hypotheses of polar wandering and continental drift, and an expanding earth. As this relatively young subject grows, it will undoubtedly yield many more answers of great importance to the earth sciences.

List of Symbols

a	present radius of earth
D	magnetic field declination
d	linear distance between ancient geomagnetic latitudes
F	geomagnetic field intensity
H_c	coercive force of a Weiss domain
H_{cr}	remanent coercive force
$\tilde{\tilde{H}}_d$	intensity of alternating demagnetizing field
H_0	theoretical dipole field intensity at geomagnetic equator
h	magnetic field strength
I	magnetic field inclination
J_s	spontaneous magnetization of Weiss domain
K	Boltzmann's constant
k	best estimate of precision parameter
M	moment of theoretical geocentric dipole
N	number of unit vectors
P	probability
$P_n^m(\theta)$	spherical harmonic function of Schmidt of degree n and order m
p	dipolar geomagnetic colatitude
$\mathcal{P}$	Fisher probability density function
R	vector sum of N unit vectors
R_0	ancient radius of the earth
r	distance from earth's center to any point
T	absolute temperature
T_c	Curie point temperature
t	time

V	geomagnetic potential
v	volume of a single domain grain
α	angular confidence limit
θ	latitude of sampling site
θ'	latitude of virtual geomagnetic pole
$\bar{\theta}$	geographic latitude
κ	Fisher precision parameter
τ	relaxation time of a single domain grain
ϕ	longitude of sampling site
ϕ'	longitude of virtual geomagnetic pole
$\bar{\Phi}$	geographic longitude
ψ	angular displacement of unit vector from mean direction

References

1. Blackett, P. M. S., Clegg, J. A., and Stubbs, P. H. S. (1960). An analysis of rock magnetic data. *Proc. Roy. Soc.* (*London*) **A256**, 291–322.
2. Irving, E. (1959). Palaeomagnetic pole positions: A survey and analysis. *Geophys. J. Roy. Ast. Soc.* **2**, 51–79.
3. Runcorn, S. K. (1959). Rock magnetism. *Science* **129**, 1002–1012.
4. Creer, K. M., Irving, E., and Runcorn, S. K. (1957). Palaeomagnetic investigations in Great Britain VI, Geophysical interpretation of palaeomagnetic directions from Great Britain. *Phil. Transact. Roy. Soc.* (*London*) **A250**, 144–156.
5. Cox, A., and Doell, R. (1960). Review of paleomagnetism. *Bull. Geol. Soc. Am.* **71**, 645–768.
6. Bullard, E. C., Freedman, C., Gellman, H., and Nixon, J. (1950). The westward drift of the earth's magnetic field. *Phil. Transact. Roy. Soc.* (*London*) **A243**, 67–92.
7. Vestine, E. H., La Porte, L., Cooper, C., Lange L., and, Hendrix, W. C. (1947). Description of the earth's main magnetic field and its secular change, 1905-1945. *Carnegie Inst. Wash. Pub.* **578**, 532 pp.
8. Vestine, E. (1953). On variation of the geomagnetic field, fluid motions, and the rate of the earth's rotation. *J. Geophys. Res.* **58**, 127–145.
9. Finch, H. F., and Leaton, E. R. (1957). The earth's main magnetic field—epoch 1955.0. *Monthly No. Roy. Ast. Soc. Geophys. Suppl.* **7**, 314–317.
10. Nagata, T., and Rikitake, T. (1957). Geomagnetic secular variation during period from 1950 to 1955. *J. Geomagnetism and Geoelectricity* (*Kyoto*) **9**, 42–50.
11. Lowes, F. (1955). Secular variation and the nondipole field. *Ann. Géophys.* **11**, 91–94.
12. Jacobs, J. A. (1956). The interior of the earth. *Advances in Geophys.* 378 pp.
13. Elsasser, W. M. (1955). Hydromagnetism I. A review. *Am. J. Phys.* **23**, 590–609.
14. Elsasser, W. M. (1956). Hydromagnetism II. A review. *Am. J. Phys.* **24**, 85–110.
15. Bullard, E., and Gellman, H. (1954). Homogeneous dynamos and terrestrial magnetism. *Phil. Transact. Roy. Soc.* (*London*) **A247**, 213–278.
16. Urey, H. (1952). "The Planets," 245 pp. Yale Univ. Press, New Haven, Connecticut.
17. Verhoogen, J. (1960). The heat balance of the earth's core. *J. Geophys. Res.* **65**, 2530.
18. Runcorn, S. K. (1954). The earth's core. *Transact. Am. Geophys. Un.* **35**, 49–63.
19. Pockels, F. (1898). Ein Versuch die bei Blitzschlägen erreichte maximale Stromstärke zu schätzen. *Meteorol. Z.* **15**, 41–46.

20. Koenigsberger, J. (1932). Über remanenten Magnetismus von Gesteinen. *Beitr. Geophys.* **35**, 204–216.
21. Chevallier, R. (1925). L'aimantation des lavas de l'Etna et l'orientation du champ terrestre en Sicilie du 12^e au 17^e siècle. *Ann. Phys.* [Ser 10] **4**, 5–162.
22. Thellier, É. (1951). Propriétés magnétiques des terres cuites et des roches. *J. Phys. Radium* **12**, 205–218.
23. Thellier, É., and Rimbert, F. (1954). Sur l'analyse d'aimantations fossiles par action de champs magnétiques alternatifs. *Compt. rend. Acad. Sci.* **239**, 1399–1401.
24. Thellier, É., and Rimbert, F. (1955). Sur l'utilisation, en paléomagnétisme, de la désaimantation par champs alternatifs. *Compt. rend. Acad. Sci.* **240**, 1404–1406.
25. Nagata, T., Akasi, K., and Rikitake, T. (1943). The natural remanent magnetism of sedimentary rocks (Preliminary note). *Bull. Earthquake Res. Inst.* (*Tokyo*) **21**, 276–297.
26. Nagata, T., Akimoto, S., and Uyeda, S. (1951). Reverse thermoremanent magnetism. *Proc. Acad. Japan* **27**, 643–645.
27. Nagata, T. (1953). "Rock Magnetism," 225 pp. Marusen, Tokyo, Japan.
28. Nagata, T., Akimoto, S., Uyeda, S., Momose, K., and Asami, E. (1954). Reverse magnetization of rocks and its connection with the geomagnetic field. *J. Geomagnetism and Geoelectricity* (*Kyoto*) **6**, 182–193.
29. Graham, J. W. (1949). The stability and significance of magnetism in sedimentary rocks. *J. Geophys. Res.* **54**, 131–167.
30. Graham, J. W. (1953). Changes of ferromagnetic minerals and their bearing on magnetic properties of rocks. *J. Geophys. Res.* **58**, 243–260.
31. Graham, J. W. (1956). Paleomagnetism and magnetostriction. *J. Geophys. Res.* **61**, 735–739.
32. Balsley, J. R., and Buddington, A. F. (1954). Correlation of reverse remanent magnetism and negative anomalies with certain minerals. *J. Geomagnetism and Geoelectricity* (*Kyoto*) **6**, 176–181.
33. Balsley, J. R., and Buddington, A. F. (1958). Iron-titanium oxide minerals, rocks, and aeromagnetic anomalies of the Adirondack area, New York. *Economic Geol.* **53**, 777–805.
34. Petrova, G. (1956). Three modes of magnetization of rock formations. *Akad. Nauk SSSR Izv. ser. geofiz.* **4**, 431–436 (in Russian).
35. Petrova, G. (1959). On magnetic stability of rocks. *Ann. Géophys.* **15**, 60–66.
36. Uyeda, S. (1958). Thermoremanent magnetism as a medium of palaeomagnetism, with special reference to reverse thermoremanent magnetism. *Jap. J. Geophys.* **2**, 1–123.
37. Akimoto, S., and Kushiro, I. (1960). Natural occurrence of titanomaghemite and its relevance to the unstable magnetization of rocks. *J. Geomagnetism and Geoelectricity* (*Kyoto*) **11**, 94–110.
38. Haigh, G. (1958). The process of magnetization by chemical change. *Phil. Mag.* **3**, 267–286.
39. Kobayashi, K. (1959). Chemical remanent magnetization of ferromagnetic minerals and its application to rock magnetism. *J. Geomagnetism and Geoelectricity* (*Kyoto*) **10**, 99–117.
40. Kawai, N. (1954). Instability of natural remanent magnetism of rocks. *J. Geomagnetism and Geoelectricity* (*Kyoto*) **6**, 208–209.
41. Kawai, N., Ito, H., Yasukawa, K., and Kume, S. (1959). Chemical-pressure remanent magnetism in sedimentary and metamorphic rocks. *Mem. College Sci. Univ. of Kyoto, Ser.* **B26**, 235–239.

42. Momose, K., Kobayashi, K., and Yamada, T. (1959). Palaeomagnetic and geologic researches for the volcanic rocks around Lake Suwa.—Palaeomagnetic researches for the Pliocene volcanic rocks in Central Japan (2). *Bull. Earthquake Res. Inst. (Tokyo)* **37**, 433–481.
43. Schmucker, U. (1959). Gesteinsmagnetische Untersuchungen an Permischen Nahe-Eruptiven. *Geol. Rundschau* **48**, 184–195.
44. Khan, A. M. (1960). The remanent magnetization of the basic Tertiary igneous rocks of Skye, Inverness-shire. *Geophys. J. Roy. Ast. Soc.* **3**, 45–62.
45. Cox, A. (1961). Anomalous remanent magnetization of basalt. *Bull. U.S. Geol. Survey* **1083**, 131–160.
46. Rimbert, F. (1956). Sur la désaimantation par action de champs magnétiques alternatifs, de la magnétite et du sesquioxyde de fer_{α}. *Compt. rend. Acad. Sci.* **242**, 890–893.
47. Néel, L. (1955). Some theoretical aspects of rock-magnetism. *Advances in Phys.* **4**, 191–243.
48. Verhoogen, J. (1959). The origin of thermoremanent magnetization. *J. Geophys. Res.* **64**, 2441–2449.
49. Nagata, T. (1943). The natural remanent magnetism of volcanic rocks and its relation to geomagnetic phenomena. *Bull. Earthquake Res. Inst. (Tokyo)* **21**, 1–196.
50. Johnson, E. A., Murphy, T., and Torreson, O. W. (1948). Pre-history of the earth's magnetic field. *Terr. Magn.* **53**, 349–372.
51. Clegg, J. A., Almond, M., and Stubbs, P. H. S. (1954). The remanent magnetism of some sedimentary rocks in Britain. *Phil. Mag.* **45**, 583–598.
52. King, R. F. (1955). The remanent magnetism of artificially deposited sediments. *Monthly No. Roy. Ast. Soc. Geophys. Suppl.* **7**, 115–134.
53. Griffiths, D., King, R., Rees, A., and Wright, A. (1960). The remanent magnetism of some recent varved sediments. *Proc. Roy. Soc. (London)* **A256**, 359–383.
54. Martinez, J. D., and Howell, L. G. (1956). Palaeomagnetism of chemical sediments. *Nature* **178**, 204–205.
55. Doell, R. R. (1956). Remanent magnetization of the upper Miocene "blue" sandstones of California. *Transact. Am. Geophys. Un.* **37**, 156–167.
56. Thellier, É. (1938). Sur l'aimantation des terres cuites et ses applications géophysiques. *Ann. Inst. Phys. Globe Univ. Paris Bur. Cen. Magn. Terr.* **16**, 157–302.
57. Rimbert, F. (1956). Sur l'action de champs alternatifs sur des roches portant une aimantation rémanente isotherme de viscosité. *Compt. rend. Acad. Sci.* **242**, 2536–2538.
58. Brynjólfsson, A. (1957). Studies of remanent magnetism and viscous magnetism in the basalts of Iceland. *Advances in Phys.* **6**, 247–254.
59. Irving, E. (1961). Palaeomagnetic methods: A discussion of a recent paper by Nairn. *J. Geol.* **69**, 226–231.
60. Graham, J. W., Buddington, A. F., and Balsley, J. R. (1957). Stress-induced magnetizations of some rocks with analyzed magnetic minerals. *J. Geophys. Res.* **62**, 465–474.
61. Powell, D. (1960). Stress-dependent magnetization in some quartz-dolerites. *Nature* **187**, 225.
62. Stott, P. M., and Stacy, F. D. (1959). Magnetostriction and palaeomagnetism of igneous rocks. *Nature* **183**, 384–385.
63. Kern, J. W. (1960). The effects of stress on the remanent magnetization of rocks. Unpublished Ph.D. Thesis, Univ. of California, Berkeley, California, p. 86.

64. Cox, Allan (1957). Remanent magnetization of lower to middle Eocene basalt flows from Oregon, *Nature* **179**, 685–686.
65. Irving, E. (1957). Directions of magnetization in the Carboniferous glacial varves of Australia. *Nature* **180**, 280–281.
66. Creer, K. M., Irving, E., and Nairn, A. E. M. (1959). Palaeomagnetism of the Great Whin sill. *Geophys. J. Roy. Ast. Soc.* **2**, 306–323.
67. Creer, K. M. (1959). A. C. demagnetization of unstable Triassic Keuper marls from S.W. England. *Geophys. J. Roy. Ast. Soc.* **2**, 261–275.
68. Creer, K. M. (1958). Preliminary palaeomagnetic measurements from South America. *Ann. Géophys.* **14**, 373–390.
69. Thellier, É., and Thellier, O. (1959). Sur l'intensité du champ magnétique terrestre dans le passé historique et géologique. *Ann. Géophys.* **15**, 285–376.
70. As, J. A., and Zijderveld, J. D. A. (1958). Magnetic cleaning of rocks in palaeomagnetic research. *Geophys. J. Roy. Ast. Soc.* **1**, 308–319.
71. Minakami, T. (1941). Magnetization of the new lava-flows of Miyakesima Island. *Bull. Earthquake Res. Inst. (Tokyo)* **19**, 612.
72. Kawai, N., Kume, S., and Yasukawa, K. (1956). Exsolution of titanomagnetite and its time effect on rock magnetism. *Proc. Acad. Japan* **32**, 455–458.
73. Balsley, J. R., and Buddington, A. F. (1960). Magnetic susceptibility anisotropy and fabric of some Adirondack granites and orthogneisses. *Am. J. Sci.* **258A**, 6–20.
74. Girdler, R. (1959). Anisotropy of susceptibility of rocks. Unpublished Ph.D. Thesis, Cambridge Univ., Cambridge, England.
75. Howell, L. G., Martinez, J. D., and Statham, E. H. (1958). Some observations on rock magnetism. *Geophysics* **23**, 285–298.
76. Granar, L. (1959). Magnetic measurements on Swedish varved sediments. *Arkiv Geofysik* **3**, 1–40.
77. Blackett, P. M. S. (1952). A negative experiment relating to magnetism and the earth's rotation. *Phil. Transact. Roy. Soc. (London)* **A245**, 303–370.
78. Collinson, D., Creer, K., Irving, E., and Runcorn, S. (1957). Palaeomagnetic investigations in Great Britain I. The measurement of the permanent magnetization of rocks. *Phil. Transact. Roy. Soc. (London)* **A250**, 73–82.
79. Johnson, E. A., and McNish, A. G. (1938). An alternating-current apparatus for measuring small magnetic moments. *Terr. Magn.* **43**, 393–399.
80. Komarov, A. (1956). On the question of the age of the gabbro-peridotite formation in the Urals. *Akad. Nauk SSSR Izv. ser. geol.* **9**, 44–50.
81. Elsasser, W., Ney, E., and Winckler, J. (1956). Cosmic ray intensity and geomagnetism. *Nature* **178**, 1226–1227.
82. Fisher, R. A. (1953). Dispersion on a sphere. *Proc. Roy. Soc. (London)* **A217**, 295–305.
83. Watson, G. S. (1956). Analysis of dispersion on a sphere. *Monthly No. Roy. Ast. Soc. Geophys. Suppl.* **7**, 153–159.
84. Watson, G. S., and Irving, E. (1957). Statistical methods in rock magnetism. *Monthly No. Roy. Ast. Soc. Geophys. Suppl.* **7**, 289–300.
85. Runcorn, S. (1957). The sampling of rocks for palaeomagnetic comparisons between the continents. *Phil. Mag. Suppl.* **6**, 169–176.
86. Kulp, J. (1959). Geological time scale. *Bull. Geol. Soc. Am.* **70**, 1634.
87. Irving, E. (1960). Palaeomagnetic pole positions, Part I. *Geophys. J. Roy. Ast. Soc.* **3**, 96–111.
88. Hospers, J. (1955). Rock magnetism and polar wandering. *J. Geol.* **63**, 59–74.

89. Cook, R. M., and Belshé, J. C. (1958). Archaeomagnetism: A preliminary report on Britain. *Antiquity* **32**, 167–178.
90. Bancroft, A. M. (1951). The magnetic properties of varved clays. M.S. Thesis, Univ. of Birmingham, England.
91. Thellier, É., and Thellier, O. (1951). Sur la direction du champ magnétique terrestre retrouvée sur des parois de fours des époques punique et romaine, à Carthage. *Compt. rend. Acad. Sci.* **233**, 1476–1478.
92. Griffiths, D. H. (1955). The remanent magnetism of varved clays from Sweden. *Monthly No. Roy. Ast. Soc. Geophys. Suppl.* **7**, 103–114.
93. Watanabe, N. (1958). Secular variation in the direction of geomagnetism as the standard scale for geomagnetochronology in Japan. *Nature* **182**, 383–384.
94. Wilson, B. W., and Nanz, R. H. (1960). Measurements of magnetism in sandstone samples. Symposium on Paleontology and Mineralogy Aspects of Polar Wandering and Continental Drift, Atlantic City (abstract), pp. 73–74, Soc. Econ. Paleontologists and Mineralogists, and letter of Aug. 18, 1960 to R. Doell.
95. Kumagai, N., Kawai, N., and Nagata, T. (1950). Recent progress in palaeomagnetism in Japan. *J. Geomagnetism and Geoelectricity (Kyoto)* **2**, 61–65.
96. Roche, A. (1958). Sur les variations de direction du champ magnétique terrestre au cours du Quaternaire. *Compt. rend. Acad. Sci.* **246**, 3364–3366.
97. Doell, R. R (1955). Remanent magnetism in sediments. Unpublished Ph.D. Thesis, Univ. of California, Berkeley, California.
98. Matuyama, M. (1929). On the direction of magnetization of basalt in Japan, Tyôsen, and Manchuria, *Proc. Acad. Japan* **5**, 203–205.
99. Nagata, T., Akimoto, S., Uyeda, S., Shimizu, Y., Ozima, M., Kobayashi, K., and Kuno, H. (1957). Palaeomagnetic studies on a Quaternary volcanic region in Japan. *J. Geomagnetism and Geoelectricity (Kyoto)* **9**, 23–41.
100. Irving, E., and Green, R. (1957). The palaeomagnetism of the Kainozoic basalts of Victoria. *Monthly No. Roy. Ast. Soc. Geophys. Suppl.* **7**, 347–359.
101. Roche, A. (1951). Sur les inversions de l'aimantation rémanente des roches volcaniques dans les monts d'Auvergne. *Compt. rend. Acad. Sci.* **230**, 1132–1134.
102. Sigurgeirsson, Th. (1957). Direction of magnetization in Icelandic basalts. *Advances in Phys.* **6**, 240–246.
103. Khramov, A. N. (1957). Paleomagnetism—the basis of a new method of correlation and subdivision of sedimentary strata. *Doklady Akad. Nauk SSSR ser. geol.* **112**, 849–852 (in Russian) English translation Consultants Bureau, Inc., 1958).
104. Coombs, D. S., and Hatherton, T. (1959). Palaeomagnetic studies of Cenozoic volcanic rocks in New Zealand. *Nature* **184**, 883–884.
105. Hatherton, T. (1954). The magnetic properties of the Whakamaru ignimbrites. *New Zealand J. Sci. and Tech.* **B35**, 421–432.
106. Nagata, T., Akimoto, S., Shimizu, Y., Kobayashi, K., and Kuno, H. (1959). Palaeomagnetic studies on Tertiary and Cretaceous rocks in Japan. *Proc. Acad. Japan* **35**, 378–383.
107. Torreson, O. W., Murphy, T., and Graham, J. W. (1949). Magnetic polarization of sedimentary rocks and the earth's magnetic history. *J. Geophys. Res.* **54**, 111–129.
108. Gus'kova, E. G. (1959). Paleomagnetic studies of sedimentary rocks of southeastern Turkmenia. *Bull. Acad. Sci. USSR, Geophys. Ser.* **3**, 460–464 (English translation by *Am. Geophys. Union*, May 1960, pp. 311–314).
109. Du Bois, P. M. (1959). Late Tertiary geomagnetic field in northwestern Canada. *Nature* **183**, 1617–1618.

110. Collinson, D. W., and Runcorn, S. K. (1960). Polar wandering and continental drift: evidence from paleomagnetic observations in the United States. *Bull. Geol. Soc. Am.* **71**, 915–958.
111. Turnbull, G. (1959). Some palaeomagnetic measurements in Antarctica. *Arctic* **12**, 151–157.
112. Angenheister, G. (1956). Über die Magnetisierung der Basalte des Vogelsberges. *Nach. Ges. Wiss. Göttingen; Math.-phys. Kl.* **9**, 187–204.
113. Nairn, A. E. M. (1960). Paleomagnetic results from Europe. *J. Geol.* **68**, 285–306.
114. Campbell, C. D., and Runcorn, S. K. (1956). Magnetization of the Columbia River basalts in Washington and Northern Oregon. *J. Geophys. Res.* **61**, 449–458.
115. Hospers, J., and Charlesworth, H. A. K. (1954). The natural permanent magnetization of the lower basalts of northern Ireland. *Monthly No. Roy. Ast. Soc Geophys. Suppl.* **7**, 32–43.
116. Wilson, R. L. (1959). Remanent magnetism of late Secondary and early Tertiary British rocks. *Phil. Mag.* **4**, 750–755.
117. Bruckshaw, J. M., and Vincenz, S. A. (1954). The permanent magnetism of the Mull lavas. *Monthly No. Roy. Ast. Soc. Geophys. Suppl.* **6**, 579–589.
118. Vincenz, S. A. (1954). The magnetic properties of some Tertiary intrusives of the Isle of Mull. *Monthly No. Roy. Ast. Soc. Geophys Suppl.* **6**, 590–603.
119. Blundell, D. J. (1957). A palaeomagnetic investigation of the Lundy dyke swarm. *Geol. Mag.* **94**, 187–193.
120. Deutsch, E. R., Radakrishnamurty, C., and Sahasrabudhe, P. W. (1959). Palaeomagnetism of the Deccan Traps. *Ann. Géophys.* **15**, 39–59.
121. Bruckshaw, J. M., and Robertson, E. I. (1949). The magnetic properties of the tholeiite dykes of North England. *Monthly No. Roy. Ast. Soc. Geophys. Suppl.* **5**, 308–320.
122. Roche, A. (1950). Anomalies magnétiques accompagnant les massifs de pépérites de la Limagne d'Auvergne. *Compt. rend. Acad. Sci.* **230**, 1603–1604.
123. Roche, A. (1950). Sur les caractères magnétiques du système éruptif de Gergovie. *Compt. rend. Acad. Sci.* **230**, 113–115.
124. Almond, M., Clegg, J. A., and Jaeger, J. C. (1956). Remanent magnetism of some dolerites, basalts, and volcanic tuffs from Tasmania. *Phil. Mag. Ser. 8*, **1**, 771–782.
125. Clegg, J. A., Deutsch, E. R., and Griffiths, D. H. (1956). Rock magnetism in India. *Phil. Mag. Ser. 8*, **1**, 419–431.
126. Clegg, J. A., Deutsch, E. R., Everitt, C. W. F., and Stubbs, P. H. S. (1957). Some recent palaeomagnetic measurements made at Imperial College, London. *Advances in Phys.* **6**, 219–231.
127. Irving, E. (1956). Palaeomagnetic and palaeoclimatological aspects of polar wandering. *Geofis. Pura Appl.* **33**, 23–41.
128. Runcorn, S. K. (1956). Paleomagnetic survey in Arizona and Utah: Preliminary results. *Bull. Geol. Soc. Am.* **67**, 301–316.
129. Roche, A., and Cattala, L. (1959). Remanent magnetism of the Cretaceous basalts of Madagascar. *Nature* **183**, 1049–1050.
130. Irving, E. (1956). The magnetization of the Mesozoic dolerites of Tasmania. *Papers Proc. Roy. Soc. Tasmania* **90**, 157–168.
131. Clegg, J. A., Radakrishnamurty, C., and Sahasrabudhe, P. W. (1958). Remanent magnetism of the Rajmahal Traps of northeastern India. *Nature* **181**, 830–831.
132. Bull, C., and Irving, E. (1960). Palaeomagnetism in Antarctica. *Nature* **185**, 834–835.

133. Blundell, D. J., and Stephenson, P. J. (1959). Palaeomagnetism of some dolerite intrusions from the Theron Mountains and Whichaway Nunataks, Antarctica. *Nature* **184**, 1860.
134. Hargraves, R. B., and Fischer, A. G., (1959). Remanent magnetism in Jurassic red limestones and radiolarites from the Alps. *Geophys. J. Roy. Ast. Soc.* **2**, 34–41.
135. Nairn, A. E. M. (1956). Relevance of palaeomagnetic studies of Jurassic rocks to continental drift. *Nature* **178**, 935–936.
136. Girdler, R. (1959). A palaeomagnetic study of some lower Jurassic rocks of N.W. Europe. *Geophys. J. Roy. Ast. Soc.* **2**, 353–363.
137. Nairn, A. E. M. (1957). A palaeomagnetic study of Jurassic and Cretaceous sediments. *Monthly No. Roy. Ast. Soc. Geophys. Suppl.* **7**, 308–313.
138. Nairn, A. E. M. (1957). Palaeomagnetic collections from Britain and South Africa illustrating two problems of weathering. *Advances in Phys.* **6**, 162–168.
139. Graham, K. W. T., and Hales, A. L. (1957). Palaeomagnetic measurements on Karroo dolerites. *Advances in Phys.* **6**, 149–161.
140. Nairn, A. E. M. (1960). A palaeomagnetic survey of the Karroo system. *Overseas Geology and Mineral Resources* **7**, 398–410.
141. Creer, K. M., Irving, E., Nairn, A. E. M., and Runcorn, S. K. (1958). Palaeomagnetic results from different continents and their relation to the problem of continental drift. *Ann. Géophys.* **14**, 492–501.
142. Du Bois, P. M., Irving, E., Opdyke, N. D., Runcorn, S. K., and Banks, M. R. (1957). The geomagnetic field in Upper Triassic times in the United States. *Nature* **180**, 1186–1187.
143. Graham, J. W. (1955). Evidence of polar shift since Triassic time. *J. Geophys. Res.* **60**, 329–347.
144. Kintzinger, P. R. (1957). Paleomagnetic survey of Triassic rocks from Arizona. *Bull. Geol. Soc. Am.* **68**, 931–932.
145. Bowker, D. E. (1960). Remanent magnetization of eastern United States Triassic rocks. Unpublished Ph.D. Thesis, Massachusetts Inst. Tech., Cambridge, Massachusetts.
146. Irving, E., and Green, R. (1958). Polar movement relative to Australia. *Geophys. J. Roy. Ast. Soc.* **1**, 64–72.
147. Khramov, A. N. (1958). Paleomagnetic correlation of sedimentary rocks. Leningrad, *Gostoptehizdat*, 218 pp.
148. Gusev, B. V. (1959). Age of alkaline-ultrabasic rocks of Maymecha-Kotuy region according to paleomagnetic data. *Informatsionnyy Byulleten Instituta geologii arktiki, Leningrad* **4**, 30–33 and *Intern. Geol. Rev.* **2**, 327–329 (1960).
149. Nairn, A. E. M. (1957). Observation paléomagnétiques en France: Roches Permiennes. *Bull. Soc. Geol. France Ser. 6* **7**, 721–727.
150. Creer, K. M. (1957). Palaeomagnetic investigations in Great Britain IV, the natural remanent magnetization of certain stable rocks from Great Britain. *Phil. Transact. Roy. Soc. (London)* **A250**, 111–129.
151. Du Bois, P. M. (1957). Comparison of palaeomagnetic results for selected rocks of Great Britain and North America *Advances in Phys.* **6**, 177–186.
152. Roche, A. (1957). Sur l'aimantation des roches volcaniques de l'Estérel. *Compt. rend. Acad. Sci.* **244**, 2952–2954.
153. Rutten, M. G., van Everdingen, R. O., and Zijderveld, J. D. A. (1957). Palaeomagnetism in the Permian of the Oslo graben (Norway) and of the Estérell (France). *Geol. en Mijnbouw* [N. S.] **19**(6), 193–195.
154. Armstrong, D. (1957). Dating of some minor intrusions of Ayrshire. *Nature* **180**, 1277.

155. Runcorn, S. K. (1955). Palaeomagnetism of sediments from the Colorado Plateau. *Nature* **176**, 505–506.
156. Doell, R. R. (1955). Palaeomagnetic study of rocks from the Grand Canyon of the Colorado River. *Nature* **176**, 1167.
157. Belshé, J. C. (1957). Palaeomagnetic investigations of Carboniferous rocks in England and Wales. *Advances in Phys.* **6**, 187–191.
158. Du Bois, P. M. (1959). Palaeomagnetism and rotation of Newfoundland. *Nature* **184**, B.A. 63–64.
159. Howell, L. G., and Martinez, J. D. (1957). Polar movement as indicated by rock magnetism. *Geophysics* **22**, 384–397.
160. Nairn, A. E. M., Frost, D. V., and Light, B. G. (1959). Palaeomagnetism of certain rocks from Newfoundland. *Nature* **183**, 596–597.
161. Creer, K. M., Irving, E., and Runcorn, S. K. (1954). The direction of the geomagnetic field in remote epochs in Great Britain. *J. Geomagnetism and Geoelectricity* (*Kyoto*) **6**, 163–168.
162. Komarov, A. G. (1959). Paleomagnetic measurements on lower Paleozoic basalts from the Ukraine. *Akad. Nauk SSR Izv. ser. geofiz.* **8**, 1219–1225 (in Russian).
163. Chang, Wen-You, and Nairn, A. E. M. (1959). Some paleomagnetic investigations on Chinese rocks. *Nature* **183**, 254.
164. Graham, J. W. (1954). Rock magnetism and the earth's magnetic field during Paleozoic time. *J. Geophys. Res.* **59**, 215–222.
165. Irving, E., and Runcorn, S. K. (1957). Palaeomagnetic investigations in Great Britain II. Analysis of the palaeomagnetism of the Torridonian sandstone series of Northwest Scotland. *Phil. Transact. Roy. Soc.* (*London*) **A250**, 83–99.
166. Harland, W. B., and Bidgood, D. E. T. (1959). Palaeomagnetism in some Norwegian sparagmites and the late Precambrian ice age. *Nature* **184**, 1861–1862.
167. Du Bois, P. M. (1959). Correlation of Keweenawan rocks of Lake Superior district by palaeomagnetic methods. *Proc. Geol. Assoc. Canada* **11**, 115–128.
168. Du Bois, P. M. (1955). Palaeomagnetic measurements of the Keweenawan. *Nature* **176**, 506–507.
169. Hood, P. J. (1958). Paleomagnetic studies of some Precambrian rocks in Ontario. Unpublished Ph.D. Thesis, Univ. of Toronto, Canada, p. 200.
170. Gough, D. I. (1956). A study of the palaeomagnetism of the Pilansberg dykes. *Monthly No. Roy. Ast. Geophys. Suppl.* **7**, 196–213.
171. Gough, D. I., and van Niekerk, C. B. (1959). A study of the palaeomagnetism of the Bushveld gabbro. *Phil. Mag.* **4**, 126–136.
172. Nagata, T., and Shimizu, Y. (1959). Natural remanent magnetization of Precambrian gneiss of Ongul Islands in the Antarctic. *Nature* **184**, 1472–1473.
173. Néel, L. (1951). L'inversion de l'aimantation permanente des roches. *Ann. Géophys.* **7**, 90–102.
174. Nagata, T. (1953). Self-reversal of thermo-remanent magnetization of igneous rocks. *Nature* **172**, 850–852.
175. Gorter, E. W., and Schulkes, J. A. (1953). Reversal of spontaneous magnetization as a function of temperature in LiFeCr spinels. *Phys. Rev.* **90**, 487–488.
176. Verhoogen, J. (1956). Ionic ordering and self-reversal of magnetization in impure magnetites. *J. Geophys. Res.* **61**, 201–209.
177. Hospers, J. (1953–1954). Reversals of the main geomagnetic field. *Kgl. Nederlandse Akad. Wetensch. Proc. Ser. B*, Pt. I, **56**, 467–476, Pt. II, **56**, 477–491, Pt. III, **57**, 112–121.
178. Einarsson, T. (1957). Magneto-geological mapping in Iceland with the use of a compass. *Advances in Phys.* **6**, 232–239.

179. Roche, A. (1956). Sur la date de la dernière inversion du champ magnétique terrestre. *Compt. rend. Acad. Sci.* **243,** 812–814.
180. Cox, Allan. (1959). The remanent magnetization of some Cenozoic rocks. Unpublished Ph.D. Thesis, Univ. of California, Berkeley, California, pp. 193.
181. Roche, A. (1953). Sur l'origine des inversions d'aimantation constantées dans les roches d'Auvergne. *Compt. rend. Acad. Sci.* **236,** 107–109.
182. Einarsson, T., and Sigurgeirsson, T. (1955). Rock magnetism in Iceland, *Nature* **175,** 892.
183. Evernden, J. F., Curtis, G. H., and Kistler, R. (1957). Potassium-argon dating of Pleistocene volcanics. *Quaternia* **4,** 13–17.
184. Hospers, J. (1951). Remanent magnetism of rocks and the history of the geomagnetic field. *Nature* **168,** 1111–1112.
185. Runcorn, S. (1959). On the hypothesis that the mean geomagnetic field for parts of geological time has been that of a geocentric axial multipole. *J. Atmos. Terr. Phys.* **14,** 167–174.
186. Gold, T. (1955). Instability of the earth's axis of rotation. *Nature* **175,** 526–529.
187. Munk, W. H. (1956). Polar wandering—a marathon of errors. *Nature* **177,** 551–554.
188. Munk, W. H. (1958). Remarks concerning the present position of the pole. *Geophysics* **6,** 335–354.
189. Munk, W. H., and MacDonald, G. J. F. (1961). "The Rotation of the Earth." Cambridge Univ. Press, Cambridge, England.
190. Irving, E. (1958). Palaeogeographic reconstruction from palaeomagnetism. *Geophys. J. Roy. Ast. Soc.* **1,** 224–237.
191. Egyed, L. (1960). Some remarks on continental drift. *Geofis. Pura Appl.* **45,** 115–116.
192. Carey, S. (1958). A tectonic approach to continental drift. *In* "Continental drift, a symposium" (S. Carey, Convener), pp. 177–355. Univ. of Tasmania, Hobart, Tasmania.
193. Heezen, B. (1959). Paleomagnetism, continental displacements, and the origin of submarine topography. *In* International Oceanographic Congress Preprints (M. Sears, ed.), pp. 26–28. American Association for Advancement of Science, Washington, D.C.
194. Elsasser, W. M., and Munk, W. (1958). Geomagnetic drift and the rotation of the earth. *In* "Contributions in Geophysics," In Honor of Beno Gutenberg, (H. Benioff *et al.*, eds) pp. 228–236 Pergamon Press, New York.
195. Runcorn, S. K. (1959). On the theory of the geomagnetic secular variation. *Ann. Géophys.* **15,** 87–92.
196. Kreichgauer, P. D. (1902). "Die Äquatorfrage in der Geologie" pp. 394. Missionsdruckerei in Steyl, Kaldenkirchen, Germany.
197. Köppen, W., and Wegener, A. (1924). "Die Klimate der geologischen Vorzeit," pp. 225. Gebrüder Borntraeger, Berlin.
198. Köppen, W. (1940). Die Wanderung des Nordpols seit der Steinkohlenzeit. *Meteorol. Z.* **57,** 106–110.
199. Chaney, R. W. (1940). Tertiary forests and continental history. *Bull. Geol. Soc. Am.* **51,** 469–488.
200. Durham, J. W. (1950). Cenozoic marine climates of the Pacific Coast. *Bull. Geol. Soc. Am.* **61,** 1243–1264.
201. Durham, J. W. (1952). Early Tertiary marine faunas and continental drift. *Am. J. Sci.* **250,** 321–343.

202. Schove, D. J., Nairn, A. E. M., and Opdyke, N. D. (1958). The climatic geography of the Permian. *Geografis. Ann.* **3**, **4**, 216–231.
203. Opdyke, N. D., and Runcorn, S. K. (1960). Wind direction in the western United States in the late Paleozoic. *Bull. Geol. Soc. Am.* **71**, 959–972.
204. Öpik, E. J. (1955). Palaeomagnetism, polar wandering, and continental drift. *J. Irish Ast.* **3**, 234–236.
205. Du Bois, P. M. (1958). Palaeomagnetism and geological correlation. *Ann. Géophys.* **14**, 509–514.
206. Runcorn, S. K. (1956). Palaeomagnetic comparisons between Europe and North America. *Proc. Geol. Assoc. Canada* **8**, 77–85.
207. Green, R., and Irving, E. (1958). The palaeomagnetism of the Cainozoic basalts from Australia. *Proc. Roy. Soc. (Victoria)* **70**, 1–17.
208. Schreiner, G. (1958). Age of a Pilansberg dyke of palaeomagnetic significance. *Nature* **181**, 1330–1331.

This Page Intentionally Left Blank

NUMERICAL PREDICTION OF STORM SURGES

Pierre Welander[1]

Swedish Research Council of the Natural Sciences and International Meteorological Institute, Stockholm, Sweden

Page

[1] Present address: Meteorologic Institute, University of Stockholm, Stockholm, Sweden.

1. Introduction

1.1. *Effects of Storm Surges*

Every year surges created by storms and hurricanes strike certain exposed coastlines, where they cause flooding of low land areas, damage to houses and harbor installations, and even loss of human lives. The height of these surges is generally a few feet. Several times each century a major surge will occur. These major surges, in which the sea level rises 10 to 15 ft or even more, are caused by severe storms that strike under the most unfavorable conditions. In the past, they have caused many catastrophes. Thousands of people have been killed, and material damage has been widespread.

In the shallow Bay of Bengal, two major surges struck in 1864 and 1876 in which 250,000 people were killed, either by the surge or by secondary causes. One surge, flooding the islands of Santa Lucia and Martinique in 1780, resulted in a loss of 15,000 lives. In 1900, another surge hit Galveston. Texas, killing 6000 people. In Europe only a few major surges are known, One struck, however, as recently as 1953 in the North Sea. In Holland, many dikes were breached at that time, and 25,000 km^2 of land was flooded, 2000 people were killed, and 600,000 people were forced from their homes. Figure 1 shows some of the effects of this surge.

Fig. 1 Effects of the 1953 storm surge in Holland.

In September 1959, typhoon Vera struck Japan. Almost 5000 people were killed and 1,600,000 were left homeless. The majority of these casualties were caused by the associated surge.

There are also known cases of major surges occurring in lakes; for example, the surge in Lake Okeechobee that struck in 1928, killing 2000 people.

1.2. The Prediction Problem

These major surges usually occurred unexpectedly. There can be little doubt that the number of casualties would have been considerably lower if the surges could have been predicted, say, 24 hr in advance, thus allowing effective warnings in threatened areas. The prediction must, of course, be accurate enough so that one can distinguish between the dangerous surges and the surges that cause little harm, as people cannot be evacuated from exposed areas for every approaching storm. This may be a delicate point. In the North Sea surge of 1953, for example, where the sea level at certain places rose more than 12 feet, it was the prediction of the last few feet of rise that was critical. A rise of 12 feet could be relatively harmless, while the rise of an additional foot caused a breach in the dike, and catastrophic flooding took place. Is it possible, then, by using our best knowledge of weather and sea and utilizing modern technical means, to make such accurate predictions? At the moment, the answer is probably no. We have, nowadays, an effective network of meteorological stations supplemented by observations from ships and airplanes, and we will probably in the near future have data from weather reconnaissance satellites. This allows us to detect the majority of storms and hurricanes that approach our coasts. However, the motion and the strength of these storms cannot as yet be predicted accurately 24 hr in advance. Further errors will be added in the computation of the sea level elevation from the available meteorological data.

It seems, however, not unlikely that relatively accurate predictions of the storm surges can be made in the future when the methods of numerical weather prediction have been further improved. In fact, for the surges of the extra-tropical storms we may not be so far from the goal today. In this case, the sea level changes are essentially determined by certain area- and time-averaged meteorological conditions, and the prediction of these elements can certainly be made more accurately than the prediction of the local values. For hurricanes which are of much smaller scale the prediction of the local meteorological conditions becomes more critical, and the hurricane surge offers thus more difficult problems.

In this chapter, we will not discuss much of the meteorological prediction problem but confine ourselves to the problem of how to predict the storm surge once the meteorological conditions are given. This problem can be faced with optimism. In fact, already by existing crude methods, it seems possible to predict the storm surge amplitudes from given meteorological conditions with an accuracy of the order of 10–20%. *With the more general numerical and empirical methods to be discussed, it seems possible to make predictions of the storm surge amplitude from given meteorological conditions with an error of only a few per cent.* Although the problem of the meteorological forecast remains, this will be an important step forward in the attempt to make useful storm surge predictions.

1.3. Comparison with the Weather Prediction Problem

It may be of value to indicate at the beginning why the prediction of the storm surge itself can be made so much more accurately than the prediction of the meteorological elements. In a *meteorological prediction*, one has essentially an initial value problem. The initial conditions are mostly given in terms of the pressure and temperature fields. From these one tries to forecast the new fields, say, 24 hr later. The errors in the forecast depend on incomplete or erroneous information about the initial fields, on the defects of the basic models used for prediction, and, if numerical methods are used, on the errors accumulated in finite difference approximations. In general, the model does not include such effects as nonadiabatic heating and friction that govern the long-time behavior of the atmosphere. For this reason any initial error will grow in the course of time. Thus, for reasonably reliable predictions, the forecast time cannot be very long.

For the *prediction of a storm surge* in a given meteorological case, one has to introduce initial values also, but these play a small role because of the presence of the strong forcing functions such as the wind and the atmospheric pressure acting on the sea surface. If these forcing functions are sufficiently familiar over a past time, we can start from a sea at rest. The forcing functions will continually control the development and prevent the growth of the initial errors. Finally, the basic model equations are in themselves both simpler and more accurate than in the meteorological case. The thermodynamic processes in the sea can be safely neglected, and the equations can, with good approximation, be linearized.

1.4. Other Aspects of the Prediction Problem

There is additional interest in predicting smaller surges and sea level changes that may not threaten human lives but which may cause some local flooding and prevent normal ship operation in shallow harbors, channels, etc. For example, the storm surge warning service in Hamburg has made such predictions for a long time for the German North Sea coast, and these have been quite valuable.

Another important aspect of the storm surge problem is the prediction of the frequency of surges within a certain range of amplitudes. The frequency of the most severe surges is, of course, of decisive importance in the design of harbor protections, dikes, etc. This problem is one of the most difficult to handle, because of the uncertainty in extrapolating statistical material from the normal storm surge amplitudes to the range of the most severe surges. According to statistical analysis a surge such as the one occurring in the North Sea in 1953 would occur only once in a thousand years. We cannot, of course, test the accuracy of such a statement at present, but there is some indication that the frequency of the major surges is somewhat higher than that predicted by presently used statistical distributions.

Although the above problem will not be discussed specifically in this chapter, it should be pointed out that a theoretical-numerical model of the storm surges may be of great help also here. Knowing the statistics of the meteorological elements, the statistics of the storm surge amplitudes can be theoretically deduced by using the model equations as a "statistical filter." One can also predict in advance certain categories of dangerous storms, moving in such a way that unusually severe surges are built up (resonance effects etc.).

2. Mechanism of the Storm Surge

2.1. *Definition of the Surge*

First, there is need for a definition of the storm surge that allows separation of other effects which contribute to the sea level changes. An initial natural separation is in *tides* and *transient effects*. The tides are strictly periodic or multiple periodic effects (the word "tide" should only be used in that connection). We have the astronomical tide, with a main semidiurnal period, and the meteorological-climatological tide with a main annual period. A small seismic tide exists, also. The transient effects are those remaining after the tides have been separated out.

There may be cases where strong seismic transients occur (tsunamis), but these are relatively rare. From the transient meteorological effect, we can further separate the short gravity waves with periods of only a few seconds. The remainder is what is usually called the "meteorological sea level effect." Of this effect, one part is caused by evaporation or precipitation and by heating or cooling of the water. These effects are small compared to the direct effects of wind and pressure, at least for the periods in which we are interested, and they are mostly neglected. There may be exceptions, of course. For example, the water level in a lake may be determined essentially by the rainfall and not by wind and pressure.

The term "storm surge" means the effects of wind and atmospheric pressure on the sea level associated with a single storm. To compute the storm surge from a sea level record, the main problem is the elimination of the astronomic tide. This can be done by various methods (see, for example, Doodson and Warburg, *Admiralty Manual of Tides*). Also, the meteorological tides and long-period meteorological transients should be eliminated. In most cases, this is achieved simply by using as a reference the actual sea level averaged over, say, a month. The short gravity waves are already suppressed by the sea level recorder and thus offer no problem; however, there may be some net effects of the gravity waves (nonlinear water transport, etc.) that will be recorded. Such effects cannot be separated from the direct effects of wind and atmospheric pressure, and they should be included in the storm surge.

2.2. The Deep Water Surge

The storm surges actually observed will vary in character depending on the structure and location of the storm. In the open sea the effects are usually small. On small islands in the open sea, the storm surge will seldom be more than a few feet. When a storm approaches such an island, the sea level may first fall somewhat, but when the storm center comes close the water level will rise to a height given approximately by the "law of the inverted barometer," i.e., it rises about 1 cm for every mb drop in pressure. This static law appears to apply also for rapidly moving storms. In cases when islands in the open sea have been struck by large surges, it has almost always been possible to explain the surge by shallow water effects.

For a theoretical understanding of the deep water surge, one may consider the case of a storm that suddenly develops in the open sea. When the pressure falls in the center, the sea surface responds quickly to the changes in the pressure, by the law of the inverted barometer. Normally such an adjustment does not require large transports of water. Further out from the center, the wind effect has to be considered, also. The wind drives the surface water cyclonically around the storm. However, the Coriolis forces due to earth's rotation will deflect the direction of motion of the water to the right (on the Northern Hemisphere), and after some time one finds a surface flow of water out from the storm area. This may explain the lowering of the sea surface that is sometimes observed. The outward flow must ultimately be compensated by a return flow at greater depth. This water will start to flow in radially, but the earth's rotation again causes a deflection to the right of the original direction, so that a cyclonic circulation is also created in deeper layers (Fig. 2). If the sea is homogeneous and the storm is stationary, the motions may penetrate to considerable depth. In actual cases the depth is limited by the motion of the storm and the stratification

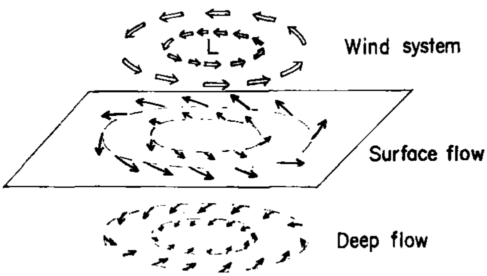


FIG. 2. Water circulation created by a cyclonic wind system.

of the water. In a moving storm adjustment of the sea level to the atmospheric pressure can still take place as long as the speed of the storm is small compared to the speed of the long gravity waves (about 250 meters/sec in the deep sea). The adjustment to the wind field cannot occur so quickly, however, and it seems that the wind effects depend more critically on the speed of the storm. When the storm has passed, it leaves behind a wake of wind-driven currents that may persist for a long time.

2.3. The Shallow Water Surge

The severe storm surges develop over shallow water. In this case the wind effect usually predominates. One can often assume a local steady state. For a vertical column of water there exists an approximate balance of the wind stress acting on the surface, the eventual bottom stress and the vertically integrated pressure gradient (the Coriolis effect can be neglected to a first approximation in this case). The latter force can be expressed as the surface slope multiplied by the gravity, the density, and the depth of the water. If the depth decreases, the maintenance of a balancing integrated pressure gradient requires that the surface slope increase, and the elevations thus become larger. Such an effect of the depth is also observed. The above-mentioned law of balance can be used to compute approximate amplitudes of the surge over shallow water. For a uniform wind stress and uniform depth of the water one has

$$\zeta \sim \kappa \frac{\tau^w F}{g\rho h} \tag{2.1}$$

where ζ is the sea surface elevation, τ^w the wind stress, F the fetch (the distance over which the wind is acting), g the acceleration of gravity, ρ the density of the water, and h the depth. Here, κ is a constant of the order 1, its value depending on the assumed bottom stress. If the bottom stress is zero, the coefficient is 1, if the net flow is zero (implying a reversed bottom current) the coefficient is $3/2$.

The wind stress can be estimated from the wind according to the formula

$$\tau^w = k\rho_a W_a^2 \tag{2.2}$$

where W_a is the wind speed measured at anemometer height, ρ_a the density of the air, and k a constant with a value around 2.5×10^{-3}.

As an example, take the case of the North Sea surge of 1953, where the mean wind speed over the North Sea, at the peak of the storm, was about 35 meters/sec. With $\rho_a = 1.3$ kg/meter^3, $\rho = 10^3$ kg/meter^3, $g = 10$ meters/sec^2, $F = 7 \times 10^5$ meters (700 km), $h = 100$ meters, and $\kappa = 1.5$, one finds $\zeta \sim 4$ meters, which was the order of magnitude observed. The one-dimensional considerations used here can be applied with some success when the scale of the storm is large compared to the scale of the shallow sea. When the scale of the storm is smaller, two-dimensional features will come into play. Generally, the surge will be smaller than predicted by the one-dimensional theory, because of the freedom of horizontal adjustments that exists.

2.4 The Time Effects

The above reasoning has been carried out under the assumption of a steady state. As mentioned, this state is often closely realized for shallow seas. One example where the sea level follows the wind in approximate balance is shown in Fig. 3. A slight time delay can, however, be seen. In other cases the time effects become more important. The maximum elevations are then noticeably delayed relative to the peak of the storm, and

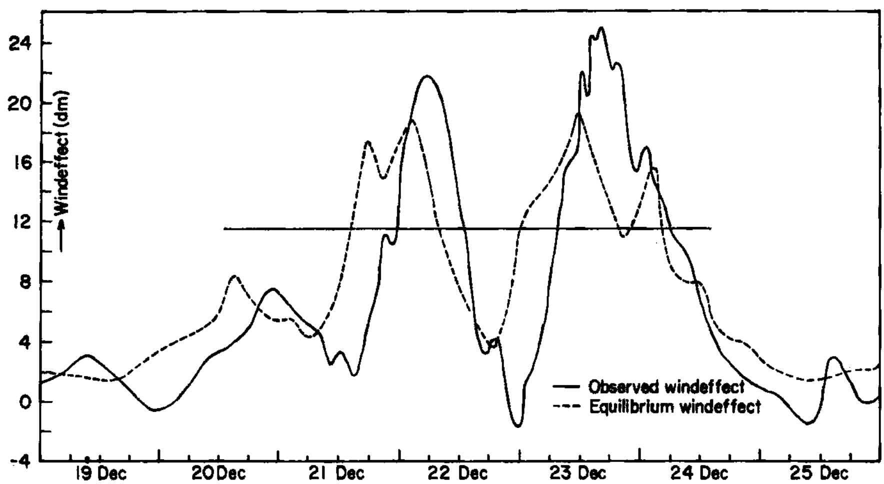


FIG. 3. Comparison between observed wind effect and equilibrium wind effect at Atlantic City, according to A. R. Miller [15].

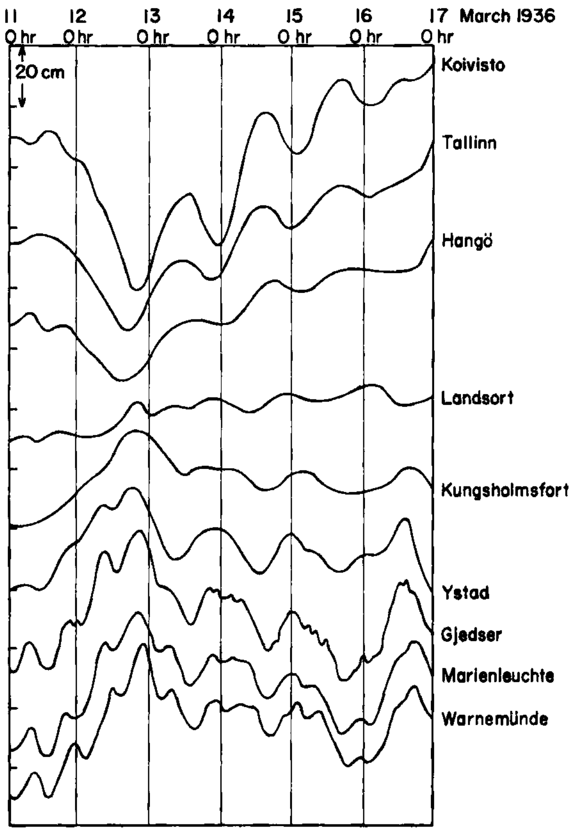


FIG. 4. Seiches in the Baltic Sea, according to G. Neumann [*Arch. Deut. Seewarte* **61**, 4 (1941)].

the main surge may be followed by a series of oscillatory smaller surges. Figure 4 gives an example of strong time effects in the Baltic. Here the whole body of water is in regular oscillations (seiches). Oscillations may also occur over the continental shelf. Examples of such oscillations during the passage of a hurricane are shown in Fig. 5. The periods of the seiches can be computed theoretically with relatively good accuracy, and they can thus be checked. In the case of the oscillations over the shelf shown in Fig. 5, it is possible that these represent seiches in the basin bounded by the shoreline and the continental slope, but one cannot exclude the possibility that they are a series of progressive waves, so-called "edge-waves" that trail with the hurricane.

Figure 5 shows another interesting feature that is sometimes associated with the hurricane surge: the "forerunner" which is represented by the gradual rise of water before the first surge arrives. It is not clear how these

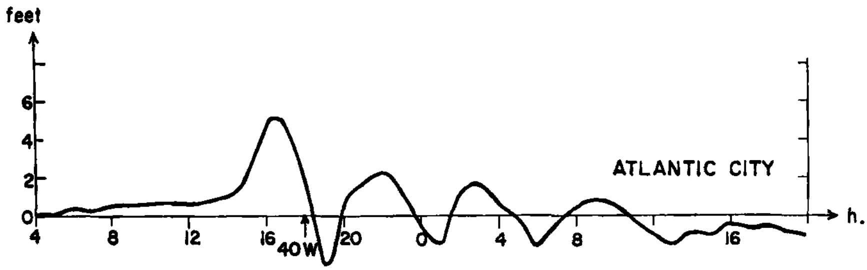


FIG. 5. Oscillations in the sea level of Atlantic City, September 14–15, 1944, according to A. C. Redfield and A. R. Miller [Memorandum on water levels accompanying Atlantic coast hurricanes, Woods Hole Ocean. Inst., 1955].

forerunners occur. It has been suggested that they are caused by nonlinear water transport of long waves or by swell running before the storm. They are certainly not formed by local meteorological effects.

When the storm moves at a speed close to $\sqrt{gh}$, that is the speed of the long gravity waves, it is possible that some resonance effects occur amplifying the surge considerably. In some cases surges have been observed coming in with very steep fronts, as a bore, and it has been suggested they are cases of such a resonance (Fig. 6). However, the question needs further study.

If a deep water surge moves in over a shallow water, one may build up a large shallow water surge even in the absence of strong winds. If the slope

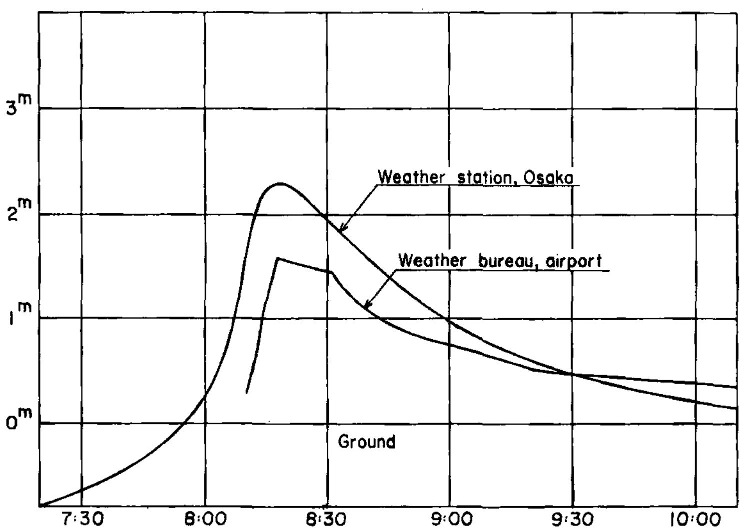


FIG. 6. Sea level variations recorded at Osaka, Japan, indicating a bore-type surge.

of the sea bottom is very steep most of the deep surge will be reflected, but if it is gradual enough the surge proceeds and amplifies as the depth decreases. According to observations, long waves propagating far away from a storm region can also cause surges in shallow waters. Long waves from the Atlantic penetrate, for example, into the North Sea. They seem to swing around Scotland and then travel along the North Sea coast in a counterclockwise direction (Fig. 7). They can be identified with the so-called Kelvin waves, which are also found in the astronomical tide. The counterclockwise rotation of the wave can be shown to result from the effect of Coriolis forces due to earth's rotation.

In many cases the local conditions will strongly influence the surge, and the sea level rises may vary widely between points a few miles apart. A storm blowing straight into a bay may pile up a strong surge in that area, whereas the same storm may affect the sea level only slightly on a straight part of the coast, where the water can escape sideways. An external surge hitting an island may give small effects at the front side, while strong rises occur at points on the other side, because of a local refraction of the surge. The theoretical discussion of such local phenomena is necessarily very difficult, and many of these factors are not yet well understood.

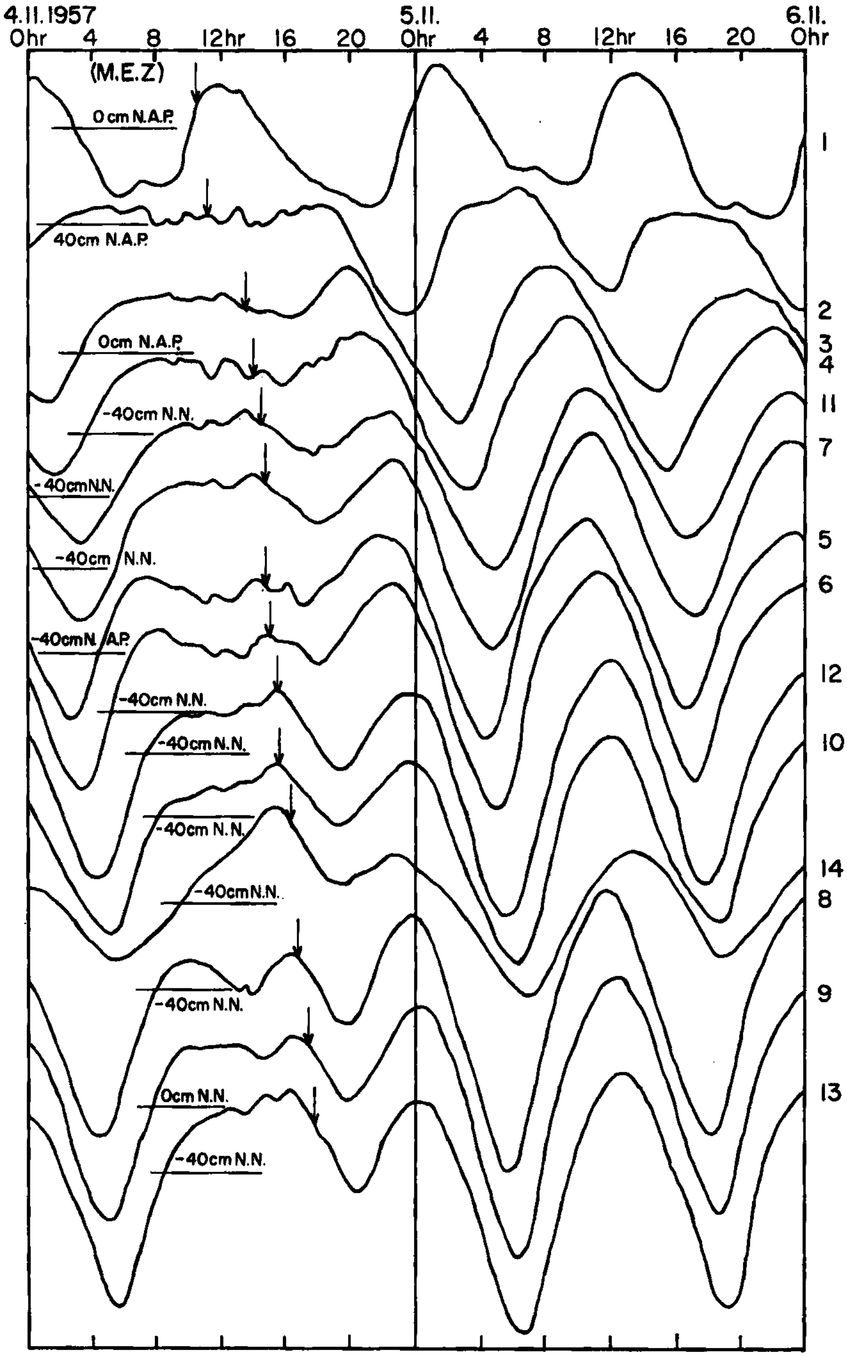
2.5. *Linearity of the Surge*

From the prediction standpoint, one of the most important properties of the storm surge is its approximately linear behavior. Except in very shallow water the surge can be described by linear equations with time-independent coefficients. This permits use of the powerful principle of superposition. The influence from a certain wind stress and air pressure acting at a given point and at a given instant can be simply superimposed on the effects acting at other points and at other times. This principle forms the basis of many of the prediction methods mentioned later.

The nonlinear phenomena that may become noticeable in very shallow water are not yet well understood. These nonlinear effects may be caused by nonlinear friction, nonlinear horizontal accelerations, nonhydrostatic effects, or a combination of these three factors. In particular, the nonlinear effects become a serious complication when one wishes to predict the actual sea level caused by the combined effect of the astronomical tide and the meteorological effect. These effects can often not be linearly superimposed, but more complicated rules of addition have to be used.

3. Some Analytical Studies of Storm Surges

In the past a number of analytical studies of storm surges have been carried out, and it may be of interest to list a few of the more important contributions (Table I, p. 328). This list is not meant to be complete but should serve


4.11.1957
0hr
4
8
12hr
16
20
5.11.
0hr
4
8
12hr
16
20
6.11.
0hr
(M.E.Z)
0 cm N.A.P.
40cm N.A.P.
0cm N.A.P.
-40cm N.N.
-40cmN.N.
-40cm N.N.
-40cmN.A.P.
-40cm N.N.
-40cm N.N.
-40cm N.N.
-40cmN.N.
-40cm N.N.
0cm N.N.
-40cm N.N.
1
2
3
4
11
7
5
6
12
10
14
8
9
13

to demonstrate some types of storm surge problems that have been successfully tackled by analytical methods.

As a general experience one may say that the case of a channel of uniform depth and also the case of a two-dimensional basin of uniform depth and not too complicated boundaries can be handled theoretically. Severe difficulties are encountered in the discussion of models with a variable depth, because the equations will no longer have constant coefficients. This is probably one of the more important theoretical problems that now has to be tackled. Some progress has recently been made, notably by a group working at the Agricultural and Mechanical College of Texas, but certainly much remains to be done.

4. Empirical Methods for Prediction of Storm Surges

4.1. Correlation Methods

The methods used for actual prediction in the past have been based primarily on correlations between the sea level and some meteorological parameter. The simplest method relates the sea level directly to a local wind or to the mean wind over a neighboring area. This method is obviously based on the assumption of a steady state. For example, Fig. 8 shows a diagram of a prediction formula for surges at Hoek van Holland, constructed by Schalkwijk [14]. This gives the amplitude of the surge for different wind directions and wind speeds.

A similar correlation method has been used by Miller [15] to predict the sea level at Atlantic City (Fig. 9). However, Miller added a time lag of 12 hr between wind and sea level, and thus partially took the time effect into consideration. Some of the results obtained by his method are shown in Fig. 10.

A method for prediction of surges at the German North Sea coast has been developed by Tomzcak [16], and it is at present used routinely. This method is based on the wind and pressure data over the German Bay, as well as over some other parts of the North Sea, and also considers certain time lags. The method has given quite accurate results, as can be seen from the example shown in Fig. 11. In Great Britain, several methods have been developed of predicting the North Sea surges. The method obtaining the best result was that by Corcan [17], with later improvements by Rossiter [18]. Corcan used the pressure gradients at two points of the North Sea and the observed meteorological sea level rises at Dunbar, with certain time lags. The prediction is made for Southend. By using the sea level at

---

FIG. 7. Long waves travelling along the eastern North Sea coast on November 4, 1957, according to G. Tomzcak [*Deut. Hydrog. Z.* **11**, 2 (1957)]. The stations run from Hoek van Holland (1) to List/Sylt (14).

TABLE I. Analyses of storm surges.

Investigation (year)	Dimension	Depth	Boundary	Earth's rotation	Friction	Wind stress field	Pressurized field	Nonlinearized effects
Ekman [1], 1923	Two	Uniform	Straight coast or closed basin	Considered	Considered	Uniform, steady	—	—
Proudman-Doodson [2], 1924	One	Uniform	Closed channel	—	Considered	Time variable		—
Eckart [3], 1951	Two	Nonuniform	—	—	—	Free waves		—
Goldsbrough [4], 1952	One	Uniform	Channel open on one end	—	Considered	External surge		—
Schönfeld [5], 1955	One	Uniform	Channel open on one end	—	—	External surge		Considered
Proudman [6], 1955	One	Uniform	Channel open on one side	—	Considered	External surge		Considered
Lauwerier [7], 1955	Two	Uniform	Straight coast	Considered	Considered (prop. to flow)	Time variable	—	—
Kreiss [8], 1956	One	Uniform	Channel open on one side	—	Considered (prop. to flow)	External surge		Considered
Reid [9], 1957	One	Constant slope	Straight coast	—	—	Wind fetch approaching normal to coast	—	—
Kajiura [10], 1956	Two	Uniform	—	—	—	Circular symmetrical moving at constant speed		—
Munk *et al.* [11], 1956	Two	Constant slope	Straight coast	—	—		Circular symmetrical moving along coast at constant speed	—
Crease [12], 1956	Two	Uniform	Semi-infinite barrier	Considered	—	External Surge		—
Kajiura [13], 1959	Two	Uniform	Straight coast shelf	—	—	Circular symmetrical moving at constant speed		—

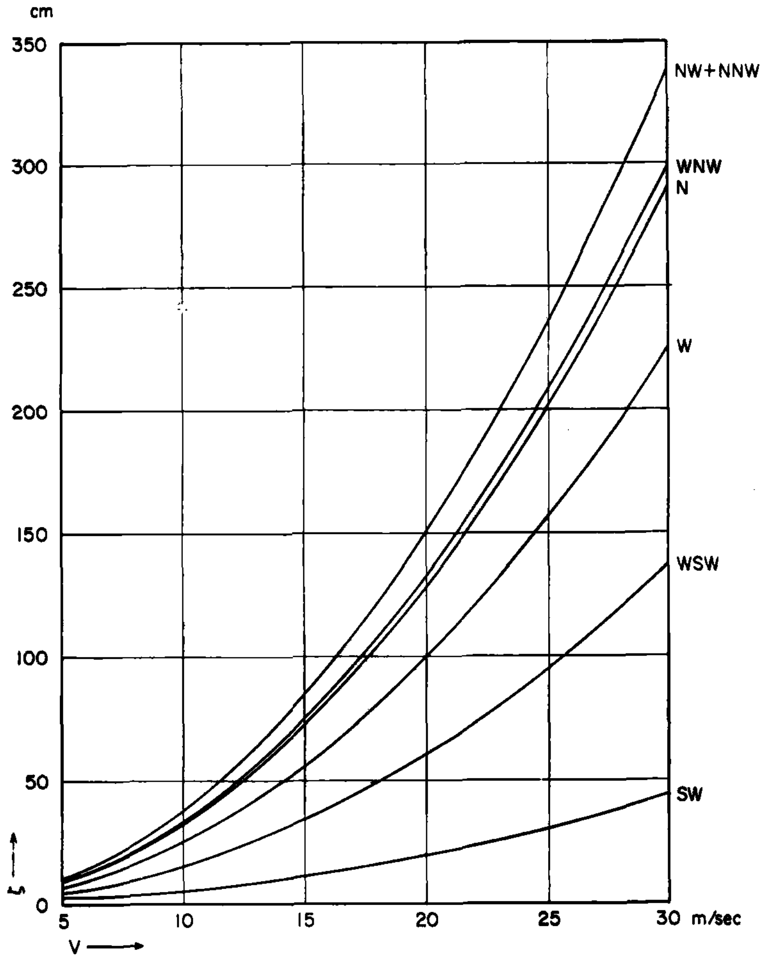


FIG. 8. Diagram for prediction of wind effect at Hoek van Holland, according to W. F. Schalwijk [14].

Dunbar, allowance is made for external surges entering from the Atlantic. The results of this method have been quite satisfactory, as the example in Fig. 12 shows. However, for the North Sea surge of 1953, the result was not so accurate, as can be seen from Fig. 13. This may indicate that the empirical methods based on the use of data from "normal surges" have difficulties when it comes to predicting more severe, abnormal surges. In Holland, Schalkwijk [14] constructed a prediction scheme that is partly theoretical and partly empirical. This has been relatively successful and is used by the storm surge warning service. A similar method has also been discussed by Weenink [19].

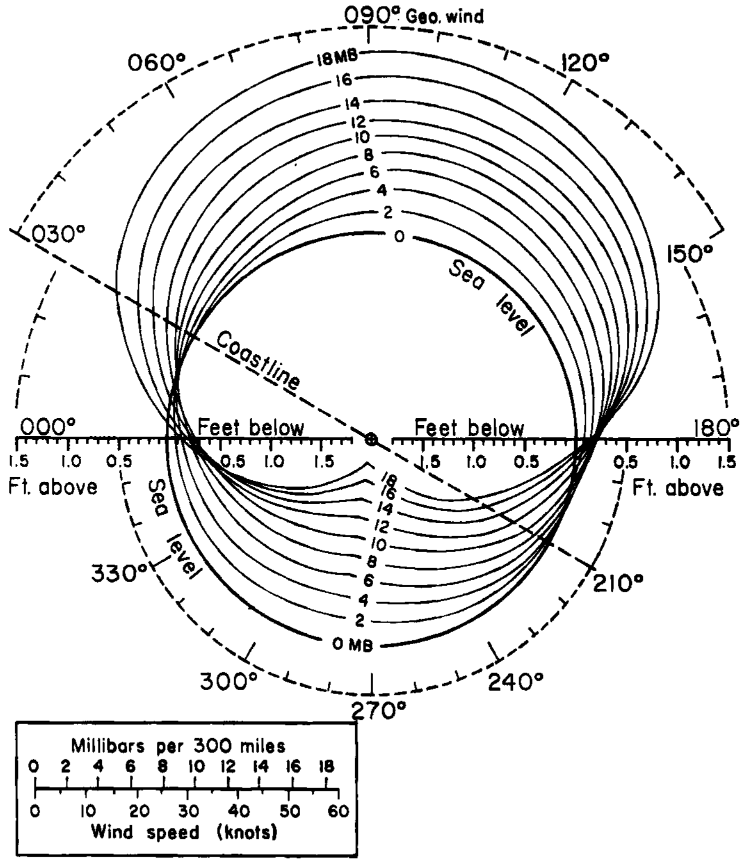


FIG. 9. Nomogram for prediction of wind effect at Atlantic City, according to A. R. Miller [15]. It is used in the following way. Place a straight edge to intercept the center at the angle of the geostrophic wind determined from the weather map. Measure from the center to the isotach representing the maximum pressure difference across a 300 mile distance. The distance when laid out along the scale of the horizontal axis, will give the estimated sea level elevation after 12 hr.

For hurricane surges some empirical prediction schemes have been proposed by Harris [20, 21]. Harris uses a formula in which the key parameter is the central pressure of the hurricane. He also found that the width of the shelf had an important effect. Kajiura [13] also uses the central pressure but includes a "factor of dynamic amplification." Wilson [22], in a very extensive study of the New York Bay area, included sea level data from neighboring points, as in Corcan's method. Wilson's method is essentially based on a theoretical-numerical model by Reid [23]. The coefficients that enter in his formula are determined by a least squares method. Altogether eight forcing functions are used.

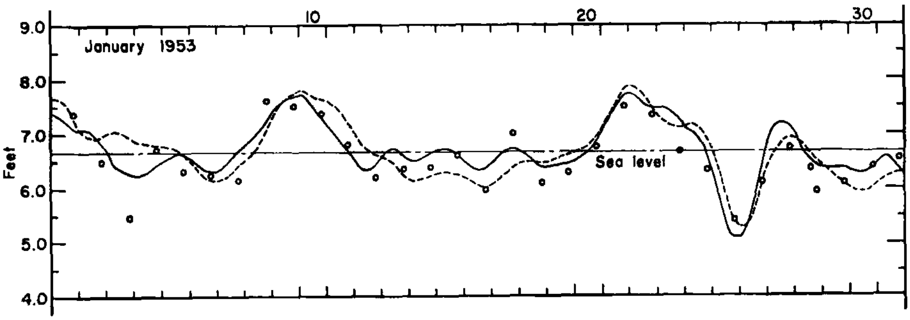


FIG. 10. Prediction of the sea level at Atlantic City, according to A. R. Miller's method. The dashed curve shows the observed sea level, after correcting for the tide; the solid line shows the sea level after further correcting for the barometric effect, and the circles give the predicted values.

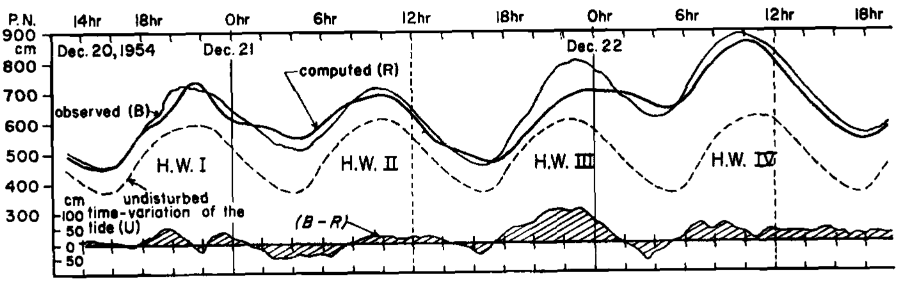


FIG. 11. Prediction of the sea level at the German North Sea Coast, according to G. Tomzcak [16].

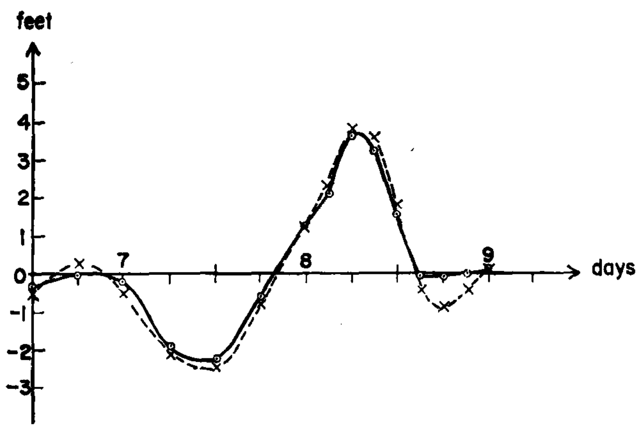


FIG. 12. Prediction of the meteorologic sea level effect at Southend, according to R. H. Corcan [*Phil. Trans. Roy. Soc. London* **A242**, 523 (1950)]. The crosses give observed values; the circles give the predicted values.

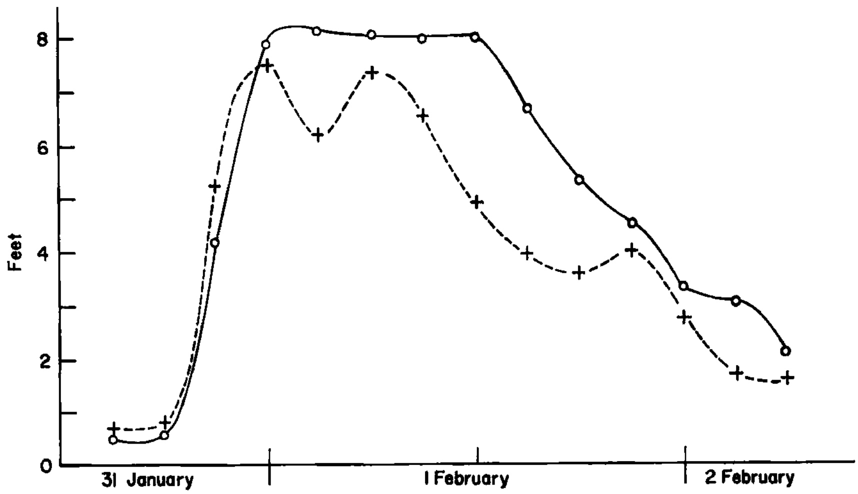


FIG. 13. Prediction of the meteorologic sea level effect at Southend at the storm surge of 1953, according to J. R. Rossiter [*Phil. Trans. Roy. Soc. London* **A246**, 397 (1954)]. The crosses give observed values; the circles give the predicted values.

4.2. The Influence Method

Practically all empirical prediction methods are based on the assumption of a linear surge, allowing for a superposition of solutions. One may then ask for the most general formulation of the prediction problem based on the following two assumptions:

1. The sea level at a certain point and at a certain time is completely determined by the past time history of the wind stress field τ_x^w, τ_y^w and the surface pressure field p^a.
2. The equations giving the sea level elevation are linear (differential or integral) equations with time-independent coefficients.

From the above assumptions it follows that the sea level can be computed, for arbitrary wind stress and pressure fields, if one knows the response of the sea level at the point of interest for three special forcing functions. These forcing functions are obtained as follows. We assume initially no elevations and apply, at a point x, y and acting over a small surrounding area $\Delta x \Delta y$, a wind stress $\tau_x^w = 1$, i.e., a step function. Similarly, we apply step functions in τ_y^w and in p^a. The corresponding responses at the point of interest are proportional to the area $\Delta x \Delta y$, these are called $Z_1 \Delta x \Delta y$, $Z_2 \Delta x \Delta y$, and $Z_3 \Delta x \Delta y$. It is assumed that the forces are applied at the time $t = 0$. Here, Z_1, Z_2, and Z_3 are functions of time and also of the position x, y at which the force is applied (see Fig. 14).

The responses to a time-varying wind stress and pressure acting over the area $\Delta x \Delta y$ can now be obtained by the law of superposition. If the wind

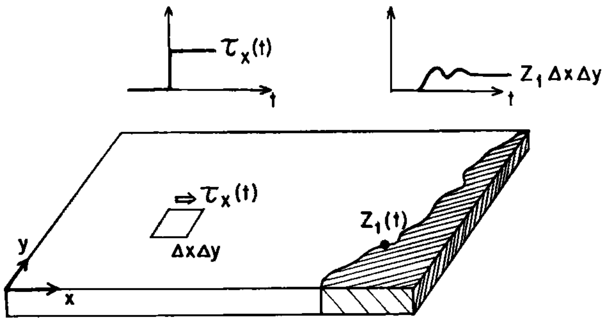


FIG. 14. The response of sea level $Z_1\Delta x\Delta y$ due to a step function in the wind stress τ_x^w acting over the area $\Delta x\Delta y$.

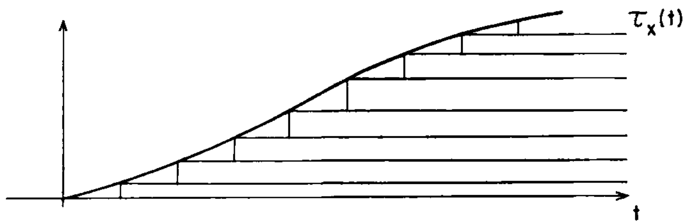


FIG. 15. Building up the general solution by summation of step functions.

stress is $\tau_x^w(t)$, the response can be found by splitting up the time function in steps and summing the responses of these with consideration of the shift of time origin from step to step (Fig. 15). Mathematically, the contribution can be written, in the limit of infinitesimal steps,

$$\int_{T=0}^{\infty} \frac{d\tau_x^w}{dt}(t - T)Z_1(T)dT,$$

where T is the time backward at which the step has been applied. The contribution from the whole area A over which the storm is acting can be obtained by integration over x, y. Similar contributions are obtained from the other two forcing functions. Thus, we arrive at the final formula for the sea level elevation at the selected point.

$$\zeta(t) = \iint_A \int_{T=0}^{\infty} \left\{ \frac{\partial \tau_x^w}{\partial t}(t - T, x, y)Z_1(T, x, y) + \frac{\partial \tau_y^w}{\partial t}(t - T, x, y) \cdot Z_2(T, x, y) + \frac{\partial p^a}{\partial t}(t - T, x, y)Z_3(T, x, y) \right\} dTdxdy \tag{4.1}$$

In practice, it is sufficient to carry out the time integration over a certain finite time, the influence time. Perturbations generated further back in time will not contribute noticeably.

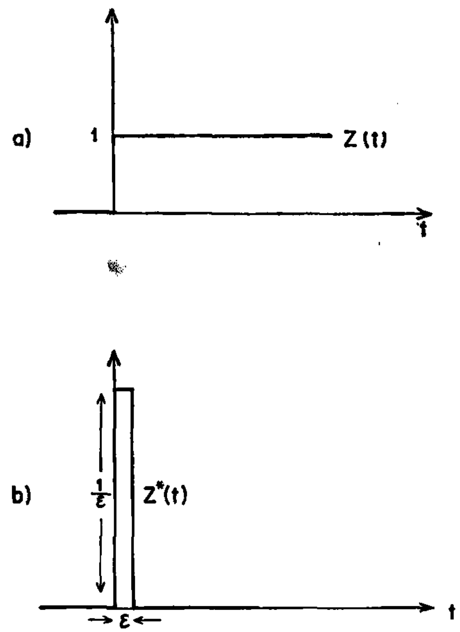


FIG. 16. (*a*) The unit step function. (*b*) The unit impulse function.

The above formula can be used for prediction provided we know the step function responses. The functions can be found for an actual case by an empirical method, as indicated by Welander [24]. In this case past records of the sea level elevation at the point of interest ζ, and of τ_x^w, τ_y^w, and p^a are needed. With these records available, we can consider equation (4.1) as an integral equation for the determination of Z_1, Z_2, and Z_3. In practice, one splits the records in three sets so as to arrive at three simultaneous integral equations, and these are treated by numerical methods. The method is straightforward but requires long and quite accurate records of both the sea level and the meteorological elements. The numerical solution of the integral equations involves so much work that an electronic computer has to be used. However, this calculation needs to be performed only once. When the step functions are found, the prediction is reduced to a simple summation.

For solving the integral equations it has been found practical to go from the step responses to the impulse responses (Fig. 16). If Z is the step response, the impulse response is $Z^* = \partial Z/\partial t$. Integrating partially in equation (4.1) and introducing Z^*, the new integrals become

$$\iint_A \int_{T=0}^{\infty} \tau_x^w(t - T, x, y) Z_1^*(T, x, y) dT dx dy, \quad \text{etc.}$$

For the actual prediction one should, however, use the step responses and not the impulse responses, since the former seem to give better accuracy for the types of storm surges one usually encounters.

One simple illustration of the method is shown in Figs. 17 and 18. Two years of selected sea level records at Ystad, Sweden, were used in constructing the response functions. Only the effects of the area mean stresses $\bar{\tau}_x^w$ and $\bar{\tau}_y^w$ acting over the Baltic were considered in this case. These were computed

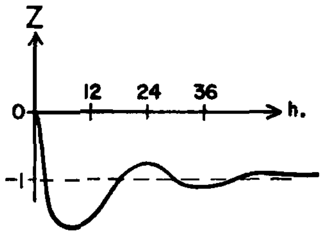


FIG. 17. The step function response Z_2 computed for Ystad, Sweden.

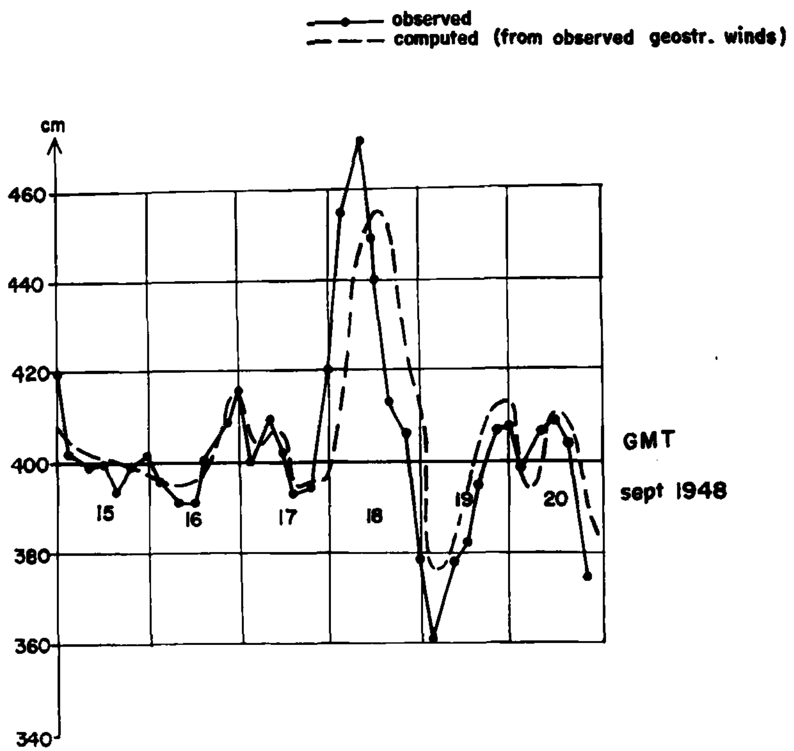


FIG. 18. Prediction of the sea level at Ystad, Sweden by use of the influence method by Welander [24].

from root mean square wind velocities. The available sets of time records $(\bar{\tau}_x^w, \bar{\tau}_y^w, \zeta)$ were then paired two by two. In each pair, $(\bar{\tau}_x^w, \bar{\tau}_y^w, \zeta)$ and $(\bar{\tau}_x^{w'}, \bar{\tau}_y^{w'}, \zeta')$, one set was multiplied by a time function $A(t)$ and the other by a time function $A'(t)$ and added. By the law of superposition, this yielded another possible solution. $A(t)$ and $A'(t)$ were in one case chosen such that the new stress $\bar{\tau}_x^{w''} = A(t)\bar{\tau}_x^w + A'(t)\bar{\tau}_x^{w'}$ disappeared, and in another case they were chosen such that the similar stress $\tau_y^{w''}$ disappeared. In this way one obtained two separate integral equations for Z_1 and Z_2. The impulse responses Z_1^* and Z_2^* were calculated by help of the Swedish computer BESK through direct solution of the finite difference approximation of the integral equations. The actual number of points taken were 40×40. It turned out that the response Z_1^* was relatively small and did not contribute significantly. For Z_2^* a significant response curve was obtained. The corresponding step function response Z_2 is shown in Fig. 17. Some results obtained by using this in making predictions are shown in Fig. 18.

It should be possible to extend the influence method so as to include as forcing functions observed sea level values at other points, as was tried by Wilson [22]. The theory of such methods has not been developed yet, however.

5. Theoretical Basis for a Numerical Prediction

5.1. Use of Numerical Prediction Methods

The method of numerical prediction is based on a numerical integration of the hydrodynamic equations governing the motion of the sea. The observed wind stress components and the surface pressure are again used as forcing functions. This method is also applicable when the forces and the boundary conditions are described by very complicated functions, as is almost always the case in reality. For practical forecasting it is superior to the analytical methods which can only handle certain simple types of surges. The computations may, however, be technically very complicated. In fact, it was not before high-speed electronic computers became available that such a numerical prediction could be made practical. So far, only a few numerical experiments have been made, but in view of the present rapid development of electronic computers and numerical techniques, it is expected that more systematic experiments and perhaps, also, routine forecasting by numerical methods will soon be carried out.

5.2. Basic Equations

It has not been possible to integrate numerically the complete hydrodynamic equations for the sea, but some basic simplifications have to be

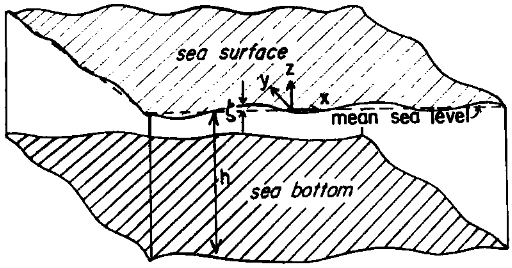


FIG. 19. Coordinates used in the basic equations.

introduced. The water is generally assumed to be incompressible and homogeneous, the Coriolis effect due to the horizontal component of earth rotation is neglected, as is also the effect of the spherical geometry of earth.

These assumptions seem to be very well fulfilled for storm surges that have characteristic periods of at most a few days and for which the region of influence is at most a few thousands of kilometers. One further acceptable simplification is the dropping of the frictional terms, except those depending on the vertical shear. In the deep sea, all frictional effects are very small over the time periods considered, and they may be safely neglected. In very shallow water, a simple type of boundary layer argument shows that the friction caused by vertical shear will strongly dominate other frictional effects. There may be cases where frictional effect produced by strong horizontal shear is important, but these are exceptions.

With the above basic assumptions, the hydrodynamic equations, comprising three equations of motion and the continuity equation, can be easily written.

If x and y are horizontal rectangular coordinates, z a vertical coordinate, counted positive upward (right-hand system), (u, v, w) the corresponding velocity components, p the pressure, ρ_0 the density, g the acceleration of gravity, τ_x, τ_y the two horizontal components of the frictional stress, and f twice the value of the vertical component of earth's rotation, the equations are (Fig. 19):

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z} - fv = -\frac{1}{\rho_0}\frac{\partial p}{\partial x} + \frac{1}{\rho_0}\frac{\partial \tau_x}{\partial z} \tag{5.1}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + w\frac{\partial v}{\partial z} + fu = -\frac{1}{\rho_0}\frac{\partial p}{\partial y} + \frac{1}{\rho_0}\frac{\partial \tau_y}{\partial z} \tag{5.2}$$

$$\frac{\partial w}{\partial t} + u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y} + w\frac{\partial w}{\partial z} = -\frac{1}{\rho_0}\frac{\partial p}{\partial z} - g \tag{5.3}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \tag{5.4}$$

In general, additional equations expressing the turbulent motion of the sea are needed to relate the stresses to the velocity field.

If $z = 0$ is the undisturbed sea surface, $z = \zeta(x, y, t)$ the free surface, $z = -h(x, y)$ the bottom, if τ_x^w, τ_y^w denote the wind stress components, and p_a the surface pressure, the boundary conditions become

$$(5.5) \qquad u = v = w = 0 \quad \text{at} \quad z = -h$$

$$\left.\begin{aligned} &(5.6) \qquad \tau_x = \tau_x^w, \quad \tau_y = \tau_y^w \\ &(5.7) \qquad p = p^a \\ &(5.8) \qquad \frac{\partial \zeta}{\partial t} + u\frac{\partial \zeta}{\partial x} + v\frac{\partial \zeta}{\partial y} = w \end{aligned}\right\} \text{at } z = \zeta$$

The last condition expresses the fact that the free surface is materially following the fluid.

5.3. *Derivation of the Linear Prediction Equations*

For the purpose of a numerical prediction, even the above three-dimensional equations become too complicated and the problem is further simplified by linearizing the equations and integrating them vertically.

The dependent variables are the transport components

$$(5.9) \qquad U = \int_{z=-h}^{\zeta} u dz, \qquad V = \int_{z=-h}^{\zeta} v dz$$

and the elevation ζ. To introduce these variables, we integrate the two first equations of motion and the continuity equation from $z = -h$ to $z = \zeta$, and transform the left sides by the rules for interchanging integration and differentiation. After application of the boundary conditions the equations can be written

$$(5.10) \quad \frac{\partial U}{\partial t} + \frac{\partial}{\partial x}\widetilde{u^2} + \frac{\partial}{\partial y}\widetilde{uv} - fV = -\frac{1}{\rho_0}\int_{z=-h}^{\zeta}\frac{\partial p}{\partial x}dz + \frac{1}{\rho_0}(\tau_x^w - \tau_x^b)$$

$$(5.11) \quad \frac{\partial V}{\partial t} + \frac{\partial}{\partial x}\widetilde{uv} + \frac{\partial}{\partial y}\widetilde{v^2} + fU = -\frac{1}{\rho_0}\int_{z=-h}^{\zeta}\frac{\partial p}{\partial y}dz + \frac{1}{\rho_0}(\tau_y^w - \tau_y^b)$$

$$(5.12) \qquad \frac{\partial \zeta}{\partial t} + \frac{\partial U}{\partial x} + \frac{\partial V}{\partial y} = 0$$

Here, τ_x^b, τ_y^b are the components of the bottom stress. We have further introduced the notation

$$\widetilde{u^2} = \int_{z=-h}^{\zeta} u^2 dz, \quad \text{etc.}$$

It is now assumed that the amplitude of the surge is small compared to the depth and that the horizontal scale of the surge is large compared to the depth. By this assumption one can justify the hydrostatic approximation for the pressure and also justify neglecting the nonlinear acceleration terms. To demonstrate this we use the following arguments developed by Charnock and Crease [25]. The vertical velocity varies from zero at the bottom to a value of the order Z/T at the surface, where Z is the characteristic amplitude of the surge and T the characteristic period. By the continuity equation we expect the ratio of the vertical and the horizontal velocities to be of the order H/L, where H is the characteristic depth and L the characteristic horizontal scale. Thus the horizontal velocities should be of the order $(L/H)\cdot(Z/T)$. We use this result in the equation obtained by eliminating the pressure between equations (5.1) and (5.3):

$$
\begin{array}{cccccc}
\dfrac{\partial^2 u}{\partial t \partial z} + & \dfrac{\partial^2}{\partial x \partial z} u^2 + & \dfrac{\partial^2}{\partial y \partial z} uv + & \dfrac{\partial^2}{\partial z^2} uw - & f \dfrac{\partial v}{\partial z} & \\
1 & \dfrac{Z}{H} & \dfrac{Z}{H} & \dfrac{Z}{H} & fT &
\end{array}
\tag{5.13}
$$

$$
\begin{array}{ccccc}
= \dfrac{\partial^2 w}{\partial t \partial z} + & \dfrac{\partial^2}{\partial x^2} uw + & \dfrac{\partial^2 vw}{\partial x \partial y} & + & \dfrac{\partial^2 w^2}{\partial x \partial z} \\
\left(\dfrac{H}{L}\right)^2 & \dfrac{Z}{H}\left(\dfrac{H}{L}\right)^2 & \dfrac{Z}{H}\left(\dfrac{H}{L}\right)^2 & & \dfrac{Z}{H}\left(\dfrac{H}{L}\right)^2
\end{array}
$$

We find the order of magnitude of the different terms relative to the first term as indicated above. When $(H/L)^2$ is small, neglect of all terms on the right side is justified, assuming that the surge amplitude is at most of the order of the depth. This is equivalent to using the hydrostatic approximation. When Z/H is small, one justifiably further neglects the nonlinear acceleration terms in the horizontal equations of motion, represented on the left side of equation (5.13). With the hydrostatic approximation and neglecting the nonlinear horizontal accelerations the equations simplify considerably. First the pressure terms are evaluated. One finds

$$
\frac{\partial p}{\partial x} = g\rho_0 \frac{\partial \zeta}{\partial x} + \frac{\partial p_a}{\partial x}, \qquad \int_{z=-h}^{\zeta} \frac{\partial p}{\partial x}\, dz \cong g\rho_0 h \frac{\partial \zeta}{\partial x} + h \frac{\partial p^a}{\partial x}, \quad \text{etc.} \tag{5.14}
$$

Consistent with previous assumptions ζ has here been neglected against h. Neglecting the nonlinear acceleration terms, equations (5.10), (5.11), and (5.12) become

$$
\frac{\partial U}{\partial t} - fV = -gh \frac{\partial \zeta}{\partial x} - \frac{h}{\rho_0} \frac{\partial p^a}{\partial x} + \frac{1}{\rho_0} (\tau_x^{\,w} - \tau_x^{\,b}) \tag{5.15}
$$

$$\frac{\partial V}{\partial t} + fU = -gh\frac{\partial \zeta}{\partial y} - \frac{h}{\rho_0}\frac{\partial p^a}{\partial y} + \frac{1}{\rho_0}(\tau_y^w - \tau_y^b) \tag{5.16}$$

$$\frac{\partial \zeta}{\partial t} + \frac{\partial U}{\partial x} + \frac{\partial V}{\partial y} = 0 \tag{5.17}$$

5.4. The Bottom Stress

The above three equations are now expressed in U, V, ζ, the forcing functions, and the bottom stress. To get a closed system of equations the bottom stress must be expressed in terms of the other variables. This problem can be solved when the coefficient of friction is known, as shown by Welander [26]. The local velocity profile (and thus also the bottom stress) can be expressed as certain integrals over the time history of the local surface slope and local forcing functions, and a single integrodifferential equation for ζ can be derived from this. However, this equation is relatively complicated to solve, even on a computer, and it has not so far been used for actual forecasting. One tries instead to use some simpler approximation for the bottom stress. In general, we distinguish between the three following types of surges, which require different laws for the bottom stress:

1. The transient surge;
2. The quasi-steady surge with vertical circulation;
3. The quasi-steady surge with horizontal circulation.

The surge of the first type has a partly developed frictional layer at the bottom. The stress has the direction of the flow (note that the stress is defined as the force that acts on an *underlying* layer of fluid so that the bottom stress is the force acting from the fluid on the bottom) and is about inversely proportional to the period of the surge. For a discussion of this case see, for example, Lamb [27].

The surge of the second type corresponds to the one-dimensional steady surge in a channel. The water circulates in a vertical cell, and the vertically integrated transport is zero. In this case the bottom stress acts in a direction opposite to the flow at the surface, i.e., in general opposite to the acting wind stress. When the coefficient of friction is constant, the magnitude of the bottom stress is found to be half that of the acting wind stress, as shown by Ekman.

The surge of the third type is one in steady state, where the flow has essentially the same direction along the depth, and the water circulates horizontally. The bottom stress is in the direction of the flow, when the depth is small. When the depth becomes comparable to, or larger than, the so-called "Ekman depth" $D = \pi\sqrt{2A/f}$, where A is the coefficient of vertical friction, the conditions become more complicated, however. In this case the effect of the earth's rotation becomes appreciable, and one finds a stress component perpendicular to the transport direction and to

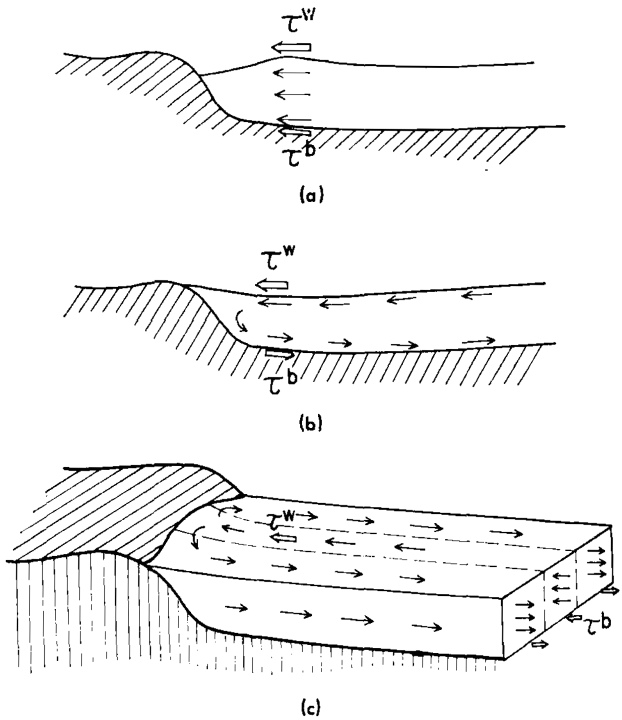


FIG. 20. Circulation and bottom stress in a surge of: (*a*) type 1; (*b*) a surge of type 2; and (*c*) a surge of type 3.

the left of this (to the right on the southern hemisphere) [28]. The Ekman depth is of the order of 100 meters at higher latitudes and increases toward the equator with the decrease of f. The different types of surges discussed are schematically represented in Fig. 20.

Under the assumption of a constant coefficient of friction the bottom stresses will be linearly dependent on the velocities occurring above the friction layer. One may expect that the coefficient of friction increases with the velocities also, so that, say, a quadratic law may be a better approximation. However, the great advantage of using a linear law makes this the preferred alternative.

A fairly general linear law covering the three types of surges mentioned would contain terms proportional to the transport, the acceleration, and the wind stress components. The coefficients should be functions of the depth h. The terms proportional to the transverse component of the transport, to the acceleration, and to the wind stress can formally be eliminated by introducing in the problem a new time, a new elevation, a new gravity, and a new wind stress. An essential new term proportional to the transport component along the flow will, however, remain. The transformed equations take on the form:

$$\frac{\partial U}{\partial t} - fV + rU = -gh\frac{\partial \zeta}{\partial x} - \frac{h}{\rho_0}\frac{\partial p^a}{\partial x} + \frac{1}{\rho_0}\tau_x^w \tag{5.18}$$

$$\frac{\partial V}{\partial t} + fU + rV = -gh\frac{\partial \zeta}{\partial y} - \frac{h}{\rho_0}\frac{\partial p^a}{\partial y} + \frac{1}{\rho_0}\tau_y^w \tag{5.19}$$

$$\frac{\partial \zeta}{\partial t} + \frac{\partial U}{\partial x} + \frac{dV}{\partial y} = 0 \tag{5.20}$$

where r is a certain coefficient of friction. It will, in general, depend on the depth.

5.5. Boundary and Initial Conditions

So far we have not mentioned the boundary and initial conditions. Theoretically the only boundary condition needed in the vertically integrated system is that the normal transport vanish at the coasts:

$$U \cos \varphi + V \sin \varphi = 0 \tag{5.21}$$

where φ is the angle between the x axis and the normal $\hat{n}$ to the coastline (Fig. 21).

The computations cannot be carried out over the entire ocean, hence artificial boundaries in the open sea have to be introduced. What boundary condition should be applied at those is more uncertain. If the water is deep, it seems most reasonable to put $\zeta = 0$, i.e., one assumes no sea level changes. Flows may, however, take place over the boundary. Introduction of a solid boundary is also possible, but if the deep water volume is not large compared to the shallow water volume, this yields erroneous results, as demonstrated in Fig. 22. When the boundary crosses over a narrow sound, one may relate the flow through the sound to the difference in the sea level

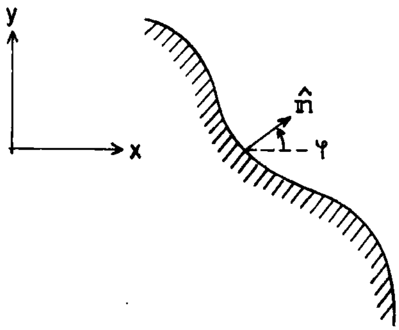


FIG. 21. Definition of $\hat{n}$ and φ.

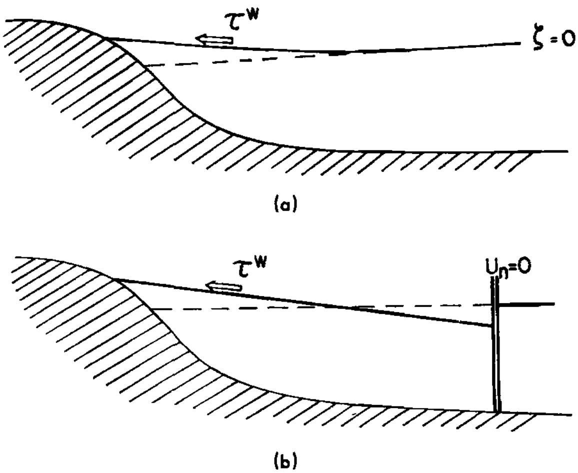


FIG. 22. Elevation set up by an onshore wind: (*a*) with a free outer boundary ($z = 0$); (*b*) with a solid outer boundary ($V\hat{n} = 0$).

on both sides. Such boundary conditions have, for example, been used successfully for the English Channel when predicting surges in the North Sea.

In an actual prediction it becomes, of course, impractical to depict all details of an irregular shoreline and of depths. The numerical prediction scheme discussed here is valid only for computation of the general sea level to a distance from the coast where small-scale irregularities become important. The prediction of the sea level near the shoreline is a local problem, in which the results of the general numerical prediction should enter as boundary values. The boundary conditions in the general prediction may still be taken as a zero transport normal to the main direction of the coast (Fig. 23). One may possibly use a hierarchy of models, in which each mode

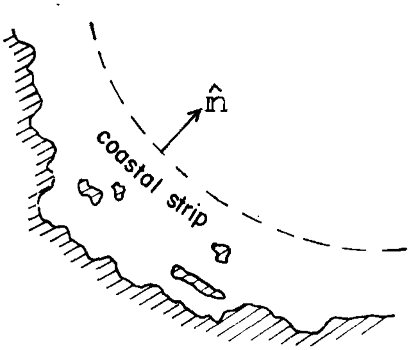


FIG. 23. Division of the sea in an area of general prediction and in a coastal strip.

successively determines the boundary conditions of the next model. As an example, one may think of a general prediction of the sea level changes in the Atlantic. These changes may serve as a boundary condition for a prediction of the sea level changes in, say, the North Sea, and this prediction may, in turn, be used as a boundary condition for a local prediction at special points of the coast.

By splitting the problem into the general surge and the local surge, one also gains mathematical simplification. Leaving out the shoreline strip in the general numerical prediction permits the use, with good approximation, of the linear model developed here, with all its advantages. The nonlinear effects that may become important close to the shoreline are only dealt with in a second stage. Further comments on this particular problem will be given later.

In the actual computations initial values for U, V, and ζ are needed. The usual choice is

$$U = V = \zeta = 0 \tag{5.22}$$

The initial time $t = 0$ is then chosen so far back that the initial conditions do not enter critically, rather the solution is governed only by the forcing functions. That the solution is determined by τ_x^w, τ_y^w and p^a alone, when the integration time is sufficiently long, can be proved as follows. Assume, for example, the system equations (5.18)–(5.20). If we take the difference between two solutions governed by the same τ_x^w, τ_y^w, and p^a, the difference transport ΔU, ΔV, and the difference height $\Delta\zeta$ satisfy the equations

$$\frac{\partial}{\partial t}\Delta U - f\Delta V + r\Delta U + gh\frac{\partial}{\partial x}\Delta\zeta = 0 \tag{5.23}$$

$$\frac{\partial}{\partial t}\Delta V + f\Delta U + r\Delta V + gh\frac{\partial}{\partial y}\Delta\zeta = 0 \tag{5.24}$$

$$\frac{\partial}{\partial t}\Delta\zeta + \frac{\partial}{\partial x}\Delta U + \frac{\partial}{\partial y}\Delta V = 0 \tag{5.25}$$

Multiplying equation (5.23) by ΔU, equation (5.24) by ΔV, adding and integrating over a region, at the boundaries of which the normal transport component U_n or ζ disappears, one finds the equation

$$\frac{\partial E}{\partial t} = -rK \tag{5.26}$$

where E and K are the energy quantities

$$E = \tfrac{1}{2}\iiint [(\Delta U)^2 + (\Delta V)^2 + gh(\Delta\zeta)^2]\,dxdy$$

$$K = \tfrac{1}{2}\iiint [(\Delta U)^2 + (\Delta V)^2]\,dxdy$$

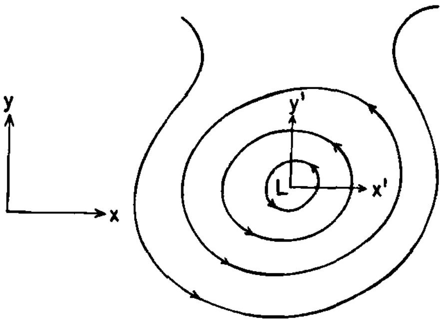


FIG. 24. Coordinate system x',y' moving with the storm.

Since both E and K are positive, one sees that E continually decreases, until $\Delta U = \Delta V = 0$. Then $\Delta\zeta$ must also vanish. For the area mean value of $\Delta\zeta$ will, by definition, vanish, and if $\Delta\zeta$ did not vanish at every point, there would be pressure gradients that in view of equations (5.23)–(5.25) required nonzero values of ΔU and ΔV.

5.6. Moving Coordinate System

In certain cases it may be advantageous to use other forms of the fundamental equations. If one considers small-scale storms such as hurricanes, it is not practical to use a fixed coordinate system. As one requires a high resolution there would be need for a very fine grid covering the whole area swept by the storm. It is therefore better to use a coordinate system moving with the storm (Fig. 24). For a hurricane that is of more or less circular symmetric form one may think that cylindric coordinates would be best suited. Because of difficulties encountered in such a coordinate system near the origin, and because the surge will not generally have a circular symmetric shape, Cartesian coordinates seem preferable, however. The system of equations corresponding to equations (5.15)–(5.17) is obtained by the transformation to new coordinates:

$$x' = x - \int_0^t u_0(t)\,dt$$
$$y' = y - \int_0^t v_0(t)\,dt \tag{5.27}$$

where u_0, v_0 give the velocity of the storm center. In the new system the transport components are

$$U' = U - h(x', y', t)u_0(t)$$
$$V' = V - h(x', y', t)v_0(t) \tag{5.28}$$

where h is the depth, which is also a function of time in the new coordinates.

The gradient operator will not be changed by the transformation. We find the form corresponding to equations (5.15)–(5.17):

$$\frac{\partial U'}{\partial t} - fV' = -gh\frac{\partial \zeta'}{\partial x'} - \frac{h}{\rho_0}\frac{\partial p^a}{\partial x'} + \frac{1}{\rho_0}(\tau_{x'}^w - \tau_{x'}^b) - \frac{\partial}{\partial t}(hu_0) + fhv_0 \tag{5.29}$$

$$\frac{\partial V'}{\partial t} + fU' = -gh\frac{\partial \zeta'}{\partial y'} - \frac{h}{\rho_0}\frac{\partial p^a}{\partial y'} + \frac{1}{\rho_0}(\tau_{y'}^w - \tau_{y'}^b) - \frac{\partial}{\partial t}(hv_0) - fhu_0 \tag{5.30}$$

$$\frac{\partial \zeta'}{\partial t} + \frac{\partial U'}{\partial x'} + \frac{\partial V'}{\partial y'} = -\frac{\partial h'}{\partial t} \tag{5.31}$$

The boundary condition is now

$$U' \cos\varphi + V' \sin\varphi = -hu_0 \cos\varphi - hv_0 \sin\varphi \tag{5.32}$$

at the coast, φ is again the angle between x axis and the outward normal. The position of the coastline is also a function of time in the new system.

As initial conditions of the new problem we have finally

$$U' = -hu_0, \quad V' = -hv_0, \quad \zeta' = 0 \quad \text{at} \quad t = 0 \tag{5.33}$$

5.7. *Near-Shore Effects*

The shallow strip next to the shoreline offers special problems. The solution in this strip should be determined by the elevation at the outer edge, obtained by a more general prediction model, and by the local wind and pressure. In many cases an empirical method such as the influence method seems to be advantageous, but it may also be worthwhile to attempt a numerical integration method. It seems likely that a steady state model will work satisfactorily when the strip has a width of only a few miles and the adjustment takes place rapidly. The surge should then be determined by the instantaneous values of the boundary height, the local wind, and the pressure. Difficulties arise, however, when it comes to computing the bottom stress, which is generally very important in this region, and the nonlinear acceleration terms. The main question is whether the surge corresponds to type 2 or type 3; i.e., whether the surge gives essentially a vertical circulation or a horizontal circulation. A uniform onshore wind blowing over a strip of uniform depth will give rise to a vertical circulation. If the wind is nonuniform the water moves onshore in regions of maximum winds and returns in regions of weaker winds, thereby producing a horizontal

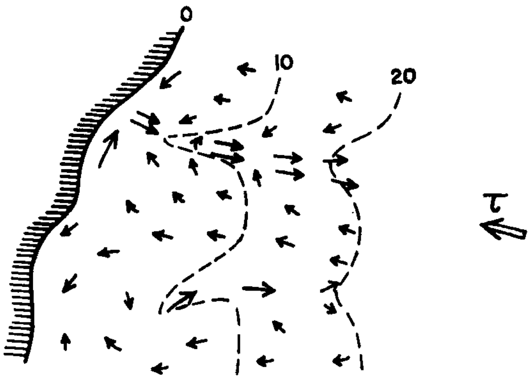


FIG. 25. Backflow in deeper channels set up by an onshore wind.

circulation. Even more important are the variations in the depth along the shore. Water that is elevated in the more shallow regions will escape toward the sides and flow back into the sea where the depth is larger. Concentrated return flows occur where deep channels cross over the strip (Fig. 25).

For a hurricane moving in over a continental shelf with no depth variation along the coast a theoretical investigation by Kajiura [13] has shown that the vertical circulation dominates when the scale of the storm is several times the width of the shelf, but horizontal circulation plays a noticeable role \ hen the storm has a scale comparable to this width.

The question of the type of surge is important in all estimates of the nonlinear accelerations. In principle, the nonlinear accelerations can be included when the velocity profile is known. Assume, for example, that we can write $u = \bar{u}(x, y)F(z/h)$ where $F(z/h) = F(z^*)$ is the vertical profile, normalized so that

$$\int_{-1}^{0} F(z^*)\, dz^* = 1.$$

One has then

$$U = \int_{-h}^{\zeta} \bar{u}F\, dz \cong h \int_{-1}^{0} \bar{u}F\, dz^* = h\bar{u} \tag{5.34}$$

and a nonlinear acceleration term such as $\partial/\partial x\, \widetilde{u^2}$ in equation (5.10) becomes

$$\begin{aligned} \frac{\partial}{\partial x}\widetilde{u^2} &= \frac{\partial}{\partial x}\int_{-h}^{\zeta} \bar{u}^2F^2\, dz \cong \frac{\partial}{\partial x}\left\{h\bar{u}^2 \int_{-1}^{0} F^2\, dz^*\right\} \\ &= \lambda \frac{\partial}{\partial x}(h\bar{u}^2) = \lambda \frac{\partial}{\partial x}\left(\frac{U^2}{h}\right) \end{aligned} \tag{5.35}$$

where

$$\lambda = \int_{-h}^{0} F^2 \, dz^* \tag{5.36}$$

is a constant.

In this case we can express the horizontal nonlinear accelerations in terms of the transports and get a closed system of equation to determine U, V, and ζ. We can now see what happens for surges of the types 2 and 3. For a surge of type 3 the horizontal velocity is expected to have about the same direction and magnitude along the vertical. In this case $F(z^*) \sim 1$ and thus $\lambda \sim 1$. The nonlinear term $(\partial/\partial x)(\widetilde{u^2})$ can then be well approximated by $(\partial/\partial x)(U^2/h)$. For a surge of type 2, on the other hand, U is close to zero, while $\bar{u}$ may be large. In this case $F(z^*)$ will take on large values and λ will be a large number.

However, the problem of how to include the nonlinear terms in an actual prediction scheme has not yet been satisfactorily solved. In general the velocity profile is unknown and to compute it in each step of time integration requires an amount of work that exceeds the capacity of even a large electronic computer.

6. Numerical Methods

6.1 Representation

In order to solve the differential equations numerically they have to be broken down so that the field variables are represented by discrete numbers. One may use a finite difference representation or a representation by suitable continuous functions (for example, a truncated Fourier Series). When the prediction model is linear it is preferable to use as such functions the natural eigenfunctions corresponding to the region. Determining these eigenfunctions may be troublesome for regions of complicated shape, but if a large number of predictions are to be made for the same region, this method is certainly advantageous. Once the eigenfunctions are known, the only remaining work will be the development of the forcing functions in terms of these functions and a simple summation procedure. If the computation of the eigenfunctions cannot be made, a more direct method corresponding to the influence method described earlier can be used. We will return to this later.

6.2. Finite Difference Methods

Most numerical prediction experiments made in the past have been based on a step by step integration. In this case one uses a grid, generally with quadratic meshes. The variables are the values of U, V, and ζ at the grid points. The differential equations are approximated by difference

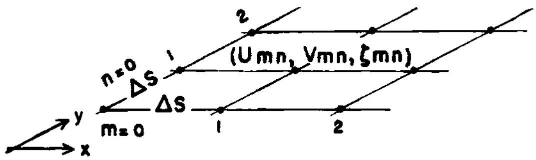


FIG. 26. Numbering of the gridpoints.

equations, and the time integration can then be made step-wise for each grid point. One may think that the finite difference formulas should be chosen so as to approximate the differential equations as closely as possible; however, it is not certain that a finite difference formula approximating the differential equations most closely in each step gives the smallest final error. In many cases difference equations that are of a higher order than the differential equations, and that are therefore expected to be more accurate, have solutions that amplify initial errors and make the whole solution diverge after a few time-steps.

6.3. Finite Difference Form of the Prediction Equation

Assume a grid with a mesh width Δs in the x and y directions (Fig. 26). At the grid points assume that U, V, and ζ are known at some initial time t. Writing coordinates for the grid points $x = x_0 + m\Delta s$, $y = y_0 + n\Delta s$, the values of U, V, and ζ at the grid points are denoted by $U_{mn}(t)$, $V_{mn}(t)$, and $\zeta_{mn}(t)$. At each grid point it is also necessary to know the depth h_{mn}, the atmospheric pressure $p^a_{mn}(t)$, and the wind stress components $\tau^w_{x\,mn}(t)$, $\tau^w_{y\,mn}(t)$. By estimating the space derivatives occurring in equations (5.15)–(5.17) by central differences, and using a forward time difference with a step Δt to approximate the derivatives $\partial U/\partial t$, $\partial V/\partial t$, and $\partial\zeta/\partial t$, we arrive at the system

$$\begin{aligned}\frac{U_{mn}(t+\Delta t) - U_{mn}(t)}{\Delta t} - fV_{mn}(t) &= -gh_{mn}\frac{\zeta_{m+1,n}(t) - \zeta_{m-1,n}(t)}{2\Delta s} \\ &\quad - \frac{h_{mn}}{\rho_0}\frac{p^a_{m+1,n}(t) - p^a_{m-1,n}(t)}{2\Delta s} + \frac{1}{\rho_0}[\tau^w_{x_{mn}}(t) - \tau^b_{x_{mn}}(t)]\end{aligned} \tag{6.1}$$

$$\begin{aligned}\frac{V_{mn}(t+\Delta t) - V_{mn}(t)}{\Delta t} + fU_{mn}(t) &= -gh_{mn}\frac{\zeta_{m,n+1}(t) - \zeta_{m,n-1}(t)}{2\Delta s} \\ &\quad - \frac{h_{mn}}{\rho_0}\frac{p^a_{m,n+1}(t) - p^a_{m,n-1}(t)}{2\Delta s} + \frac{1}{\rho_0}[\tau^w_{y_{mn}}(t) - \tau^b_{y_{mn}}]\end{aligned} \tag{6.2}$$

$$\begin{aligned}&\frac{\zeta_{mn}(t+\Delta t) - \zeta_m(t)}{\Delta t} + \frac{U_{m+1,n}(t) - U_{m-1,n}(t)}{2\Delta s} \\ &\qquad + \frac{V_{m,n+1}(t) - V_{m,n-1}(t)}{2\Delta s} = 0\end{aligned} \tag{6.3}$$

These equations can be solved for U_{mn}, V_{mn}, ζ_{mn} at the time $t + \Delta t$, and knowing the forcing functions at this time a new step can then be computed.

As central differences are exact to a higher order in Δt than the forward differences, one may prefer to use central differences also in the time direction.

The corresponding system is

$$\frac{U_{mn}(t+\Delta t) - U_{mn}(t-\Delta t)}{2\Delta t} - fV_{mn}(t)$$

$$(6.4) \qquad = -h_{mn}\frac{\zeta_{m+1,n}(t) - \zeta_{m-1,n}(t)}{2\Delta s} - \frac{h_{mn}}{\rho_0}\frac{p^a_{m+1,n}(t) - p^a_{m-1,n}(t)}{2\Delta s} + \frac{1}{\rho_0}[\tau^w_{x_{mn}}(t) - \tau^b_{x_{mn}}(t)]$$

$$\frac{V_{mn}(t+\Delta t) - V_{mn}(t-\Delta t)}{2\Delta t} + fU_{mn}(t)$$

$$(6.5) \qquad = -h_{mn}\frac{\zeta_{m,n+1}(t) - \zeta_{m,n-1}(t)}{2\Delta s} - \frac{h_{mn}}{\rho_0}\frac{p^a_{m,n+1}(t) - p^a_{m,n-1}(t)}{2\Delta s} + \frac{1}{\rho_0}[\tau^w_{y_{mn}}(t) - \tau^b_{y_{mn}}(t)]$$

$$(6.6) \qquad \frac{\zeta_{mn}(t+\Delta t) - \zeta_{mn}(t-\Delta t)}{2\Delta t} + \frac{U_{m+1,n}(t) - U_{m-1,n}(t)}{2\Delta s} + \frac{V_{m,n+1}(t) - V_{m,n-1}(t)}{2\Delta s} = 0$$

In the special case of no bottom stress, $\tau_x{}^b = \tau_y{}^b = 0$, one may arrange the computation as shown in Fig. 27. One has one set of ζ-points and one set of U,V points, that together form the grid. A ζ-point is occupied only every second time-step, in the way shown in the figure, while the U and the V values alternate at a U,V point. This arrangement is possible because we only need $U_{mn}(t - \Delta t)$, $V_{mn}(t)$ (the Coriolis force), and $\zeta_{m+1,n}(t)$, $\zeta_{m-1,n}(t)$ (the pressure gradient) to compute $U_{mn}(t + \Delta t)$. Similarly, we only need $V_{mn}(t - \Delta t)$, $U_{mn}(t)$ and $\zeta_{m,n+1}(t)$, $\zeta_{m,n-1}(t)$ to compute $V_{mn}(t + \Delta t)$. Computing $\zeta_{mn}(t + \Delta t)$, finally, only requires the knowledge of $U_{m+1,n}(t)$, $U_{m-1,n}(t)$, $V_{m,n+1}(t)$, $V_{m,n-1}(t)$, and $\zeta_{mn}(t - \Delta t)$.

If one includes a bottom stress that depends on the transport, as in the system equations (5.18)–(5.20), both U and V are needed at every U, V point. The grid will then have the appearance shown in Fig. 28.

6.4. Smoothing Technique

In the case shown in Fig. 27 one works with two overlapping subgrids connected with each other only through the Coriolis term. If the Coriolis

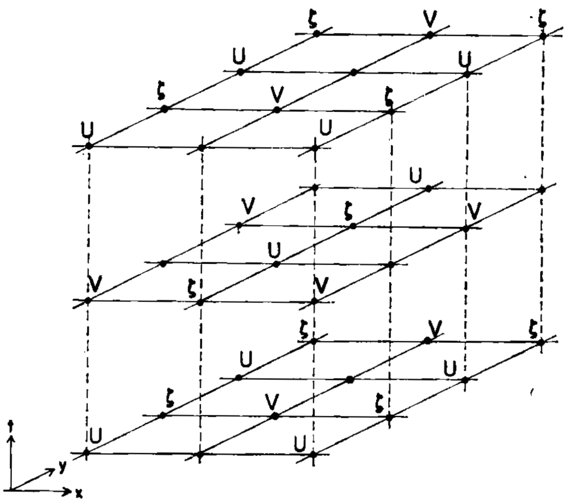


FIG. 27. Arrangement of the gridpoints for consecutive time steps in the case of no bottom stress.

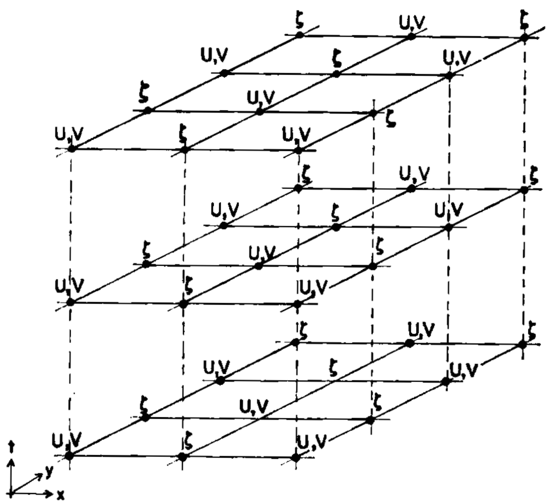


FIG. 28. Arrangement of the gridpoints for consecutive timesteps in the case of a bottom stress depending on the transport.

force were zero, the subgrids would be independent of each other, and it is sufficient to use one of them. Using both grids would not increase the accuracy but instead cause extra complications, because the errors would build up independently in the two grids and cause oscillations of the values at neighboring points belonging to the different subgrids. Because the Coriolis force is a relatively small effect, the coupling between the two subgrids will never be strong, and spurious waves with a wavelength equal to the mesh width will appear. To avoid this, one may connect the subgrids by means of a smoothing operator. The usual smoothing operator is given by the formula

$$A_{mn}^{s} = \alpha A_{mn} + \frac{1-\alpha}{4}\left[A_{m+1,n} + A_{m-1,n} + A_{m,n+1} + A_{m,n-1}\right]$$

where A_{mn}^{s} is the smoothed value, and α is a suitable constant, $0 \leq \alpha < 1$. In this case, one simply adds a certain proportion of the mean value computed from the four surrounding points. Such smoothing is also useful when working with the nonlinear model, in which small-scale noise is constantly generated.

6.5. Boundary Conditions

At the boundaries some complications may arise. At a coast the normal component of the transport should vanish. One example of how this can be done is seen in Fig. 29. The coast is assumed to run in the direction of the y axis. In this case $U = 0$ at the boundary, and all other values can then be computed, proceeding by the system shown in Fig. 27 or Fig. 28. If the coast runs in another direction, one may divide it into sections running parallel to the x or y axis (Fig. 30). An open boundary, where ζ is to be prescribed, can be treated similarly, as shown in Fig. 31.

6.6. Numerical Stability

This raises the question of numerical stability. We consider first the case where one has no Coriolis force and bottom stress, and where the depth is uniform. Our system is then equivalent to the ordinary equation for long gravity waves. The stability of the corresponding finite difference equation

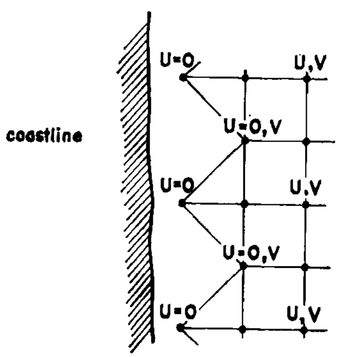


FIG. 29. Arrangement of the boundary points at a straight coast.

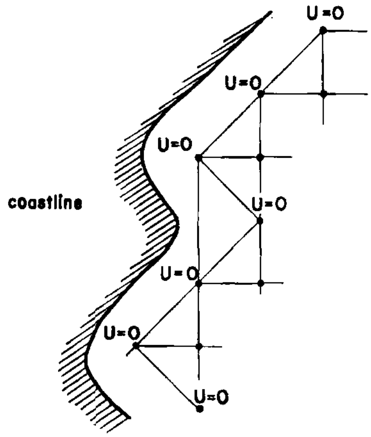


FIG. 30. Arrangement of the boundary points at an irregular coast.

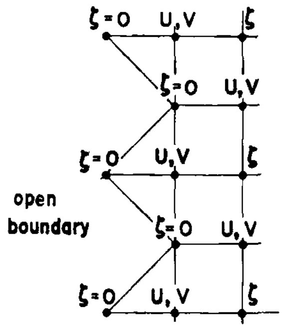


FIG. 31. Arrangement of the boundary points at an open straight boundary.

has been studied by Platzmann [29] and others. For centered time differences, Platzmann shows that the condition

$$\sqrt{2gh}\,\Delta t/\Delta s < 1 \tag{6.7}$$

has to be fulfilled. This is the so-called "Courant-Friedrichs-Lewy criterion." Practically one should use such small time-steps that the above parameter is well below one. If one uses forward differences the finite difference solution becomes unstable, regardless of what small steps are taken.

Introducing the Coriolis force and the bottom stress, the conditions are different, and one can in this case switch over to forward time differences. Further improvement is obtained by making use of the so-called "implicit method" discussed by Lax and Richtmyer [30]. By use of the implicit method all the terms needed for the computation of the time differences are evaluated at the new time $t + \Delta t$. As these values are unknown from the beginning, iterations at each time-step are needed, starting from some assumed values. The method allows large time-steps, but on the other hand more complicated computations arise in each time-step. A combination of the explicit and implicit method has been used for storm surge computations by Fischer [31]. Fischer evaluated the terms in the equations of motion at the time t, put the terms in the continuity equation at time $t + \Delta t$, and arrived at the following system:

$$\begin{aligned}(6.8)\qquad \frac{U_{mn}(t + \Delta t) - U_{mn}(t)}{\Delta t} - fV_{mn}(t) &= -gh_{mn}\frac{\zeta_{m+1,n}(t) - \zeta_{m-1,n}(t)}{2\Delta s}\\ &- \frac{h_{mn}}{\rho_0}\frac{p^a_{m+1,n}(t) - p^a_{m-1,n}(t)}{2\Delta s} + \frac{1}{\rho_0}[\tau^w_{x\,mn}(t) - \tau^b_{x\,mn}(t)]\end{aligned}$$

$$\begin{aligned}(6.9)\qquad \frac{V_{mn}(t + \Delta t) - V_{mn}(t)}{\Delta t} + fU_{mn}(t) &= -gh_{mn}\frac{\zeta_{m,n+1}(t) - \zeta_{m,n-1}(t)}{2\Delta s}\\ &- \frac{h_{mn}}{\rho_0}\frac{p^a_{m,n+1}(t) - p^a_{m,n-1}(t)}{2\Delta s} + \frac{1}{\rho_0}[\tau^w_{y\,mn}(t) - \tau^b_{y\,mn}(t)]\end{aligned}$$

$$\begin{aligned}(6.10)\qquad \frac{\zeta_{mn}(t + \Delta t) - \zeta_{mn}(t)}{\Delta t} &+ \frac{U_{m+1,n}(t + \Delta t) - U_{m-1,n}(t + \Delta t)}{2\Delta s}\\ &+ \frac{V_{m,n+1}(t + \Delta t) - V_{m,n-1}(t + \Delta t)}{2\Delta s} = 0\end{aligned}$$

For the case where the bottom stress is proportional to the transport [see equations (5.18)–(5.20)] Fischer showed that two conditions have to be statisfied to obtain stability:

$$(6.11)\qquad \sqrt{\frac{gh}{2}}\,\frac{\Delta t}{\Delta s} < 1$$

$$(6.12)\qquad f^2\frac{\Delta t}{r} < 1 \qquad (f \neq 0)$$

With a strong friction and small Coriolis effect, equation (6.12) does not limit the time-step, while this may be the case when the friction is small or the Coriolis effect strong. Fischer made practical experiments with the system equations (6.8)–(6.10) for different values of Δt, r and the parameter α that enters in the smoothing operator given in equation (6.7). Some results of these experiments are shown in Fig. 32.

Stability problems similar to the ones discussed here have been encountered earlier in numerical weather prediction. In this case it has been found possible to get around the Courant-Friedrichs-Lewy criterion and use quite large time-steps of integration by working with a vorticity equation rather than the primitive equation and by applying the quasi-geostrophic

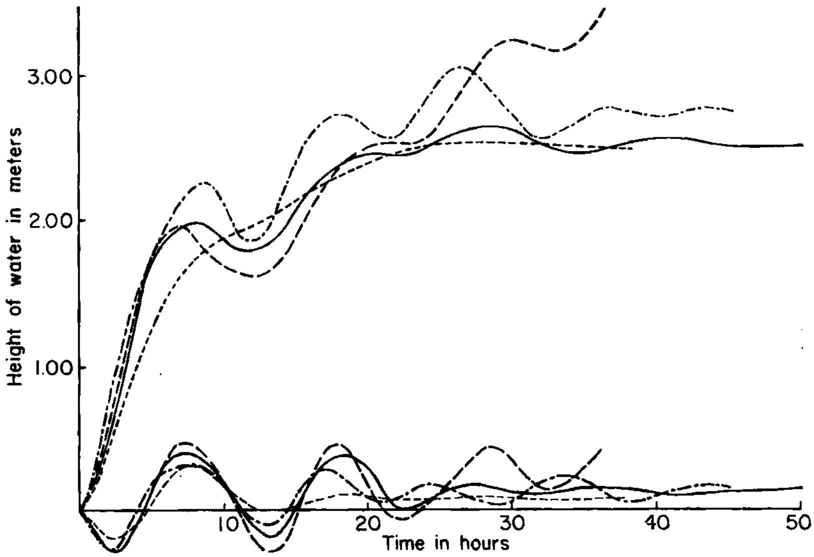


FIG. 32. Variations in the prediction curves depending on the choice of timestep, mesh width and strength of the smoothing applied according to G. Fischer [31].

Solid line $\Delta t = 20$ min, $\Delta s = 74$ km, $\alpha = \frac{7}{8}$
Long-dashed line $\Delta t = 20$ min, $\Delta s = 74$ km, $g = 1$
Short-dashed line $\Delta t = 20$ min, $\Delta s = 74$ km, $\alpha = 0$
Dashed-dotted line $\Delta t = 10$ min, $\Delta s = 37$ km, $\alpha = \frac{7}{8}$

The upper set of curves represents the approach to a steady state in the case of a wind that suddenly starts to blow uniformly over the basin. The lower set of curves represents the decay of an original perturbation in the case of no wind.

assumption. Thus, the long gravity waves are essentially eliminated from the problem. In our case such approximations cannot be made. In the storm surges the long gravity waves are essential and it is rather the quasi-geostrophic motions that can be neglected. Nothing can then be gained by using a vorticity equation, and the primitive equations should be retained.

7. The Meteorological Problem

7.1. The Meteorological Data

Even if we have a perfect model of the storm surge, a limit to the accuracy of the prediction is set by the errors in the meteorological data fed into the model. For prediction times more than a few hours the meteorological errors are always significant. In particular, large errors can be expected when one deals with hurricanes and other small-scale storms, where relatively small variations in the predicted storm track can considerably change the surge at a given point. Errors occur, however, not only in the meteorological prediction but also in the observed meteorological data that enter the storm surge computation. The preparation of these data will thus require great care.

The final meteorological data which enter into the prediction are the two wind stress components and the surface pressure. One needs data for a past period based on actual observations, and data for the prediction period based on a subjective or objective weather forecast. The surface pressure is one of the meteorological elements that is directly measured and predicted. The wind stress cannot be directly measured, but it has to be computed from wind data. The computations of the wind stress for past times should as much as possible be based on measured surface winds. At the coast one may have many observations, while generally only few observations are made over the open sea. One must then use the analyzed 1000-mb geostrophic winds. For the prediction period one is generally limited entirely to geostrophic winds.

7.2. Computation of Surface Winds

From the measured or predicted geostrophic 1000-mb winds one should first try to estimate the surface winds. So far no good theory exists on which such a computation can be based. The simplest method is to use the geostrophic winds themselves, multiplied by a constant factor. A further step is the use of the cyclostrophic winds. In hurricanes and other intense storms, this certainly gives important corrections. It is also important to correct for the friction. As a first approximation one may simply turn the wind vector by a fixed angle in a direction toward the low pressure side.

One may consider more complicated computations where factors such as stability of the air and roughness of the surface enter in. For example, theoretical studies of this problem have been made by Rossby and Montgomery [32]. However, little can be gained at present by complicating the problem in this way. Probably, in the case of strong storms, stability plays a small role, and there are indications that the roughness parameter is relatively constant.

7.3. Wind Stress Formula

With the surface winds measured or estimated one can then proceed to the computation of the stress at the sea surface. There is some doubt about the correct relation between the wind speed and the wind stress at light winds, but at strong winds there seems to be agreement that the formula

$$\tau^{w} = 2.6 \times 10^{-3}\, \rho_a W_a^{2} \tag{7.1}$$

gives satisfactory results. Here ρ_a is the air density and W_a the wind speed.

Results of some different direct and indirect determinations of the wind stress are shown in Fig. 33.

Since W_a enters quadratically in equation (7.1) the strongest winds will get a larger weight. This means, for example, that a gusty wind gives a

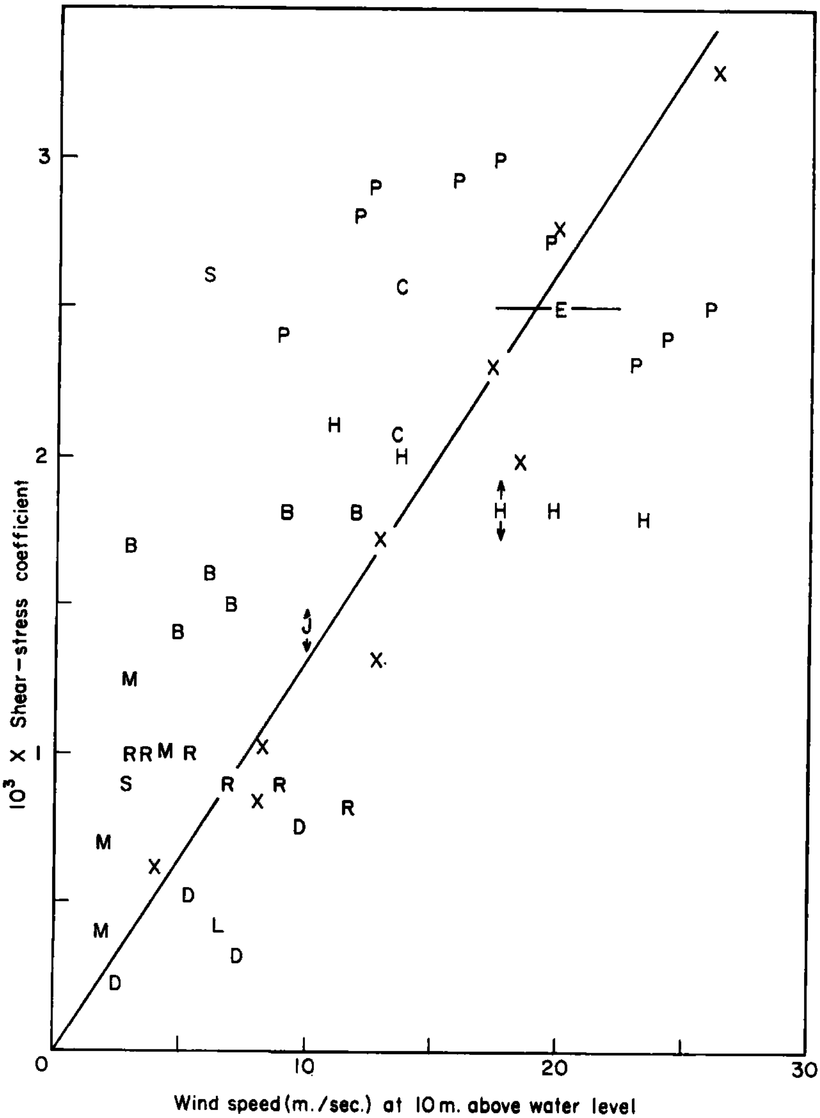


FIG. 33. Different determinations of the wind stress constant according to J. R. Francis [*Proc. Roy. Soc.* (*London*) **A206**, (1951)]. (X = wind tunnel results; S = Shovlejkin; P = Palmen; E = Ekman; M = Montgomery; R = Roll; B = Bruch; H = Hela; J = N. K. Johnson; D = Durst; C = Corcan; L = Sutcliffe.)

higher stress than a constant wind with the same mean speed. If one uses wind values that are averaged over certain areas, one will generally get too small stress for the same reason. The averages should be computed as root mean square averages and not as linear averages.

8. Experiments in Numerical Prediction

8.1. Hand Computations

Numerical integration by hand of the two-dimensional basic equations has been carried out by Kivisild [33]. Kivisild studied surges in Lake Okeechobee, Florida. He used the linearized transport equations with a bottom stress proportional to the flow. The lake was divided into 13 equilateral triangles, and the terms in the equations were estimated by a combined graphical and finite difference technique. In the cases studied, his numerical method functioned quite well. Figure 34 shows Lake Okeechobee with the gauge stations and the grid system introduced by Kivisild. The side of his triangles was 60,000 ft. The time-step of integration was chosen to be 1 hr. Kivisild used a wind stress formula based on the logarithmic law with a turbulent length $m = 0.2$ cm. He found later that the value $m = 0.15$ would have given the best agreement. The computations were based on observed winds. Figure 35 compares the predicted and observed elevations for the hurricane of August 26–27, 1949. The mean error in the prediction of the peak surge amplitudes is about 30 %. The arrival of the peak is generally predicted too early by 1–2 hr. It seems likely that the discrepancies can be explained to a large extent by nonlinear effects. The mean depth of the lake is approximately 7 ft, and noticeable nonlinear effects may be expected already when the surge is of the order of 1 ft.

Numerical integration by hand has also been performed by Doodson [34] for a one-dimensional surge set up in a uniform channel by external variations in the sea level. In this case the velocities can be assumed to be independent of depth. Doodson included the first-order nonlinear terms proportional to ζ/h, and also considered a bottom stress proportional to the velocity square. His finite difference method was based on centered time differences introduced in the vertically integrated equation of motion and the equation of continuity. The purpose of his investigation was not so much the prediction of the surge itself as the study of the nonlinear effects occurring when the astronomical tide and the surge interact.

A numerical method for studies of free or forced surges in a nonuniform channel has been described by Reid [35]. Reid did not use a step by step integration but an iterative procedure. The bottom friction is neglected in his case. It has been possible to compare the numerical results with analytic solutions. The numerical method proposed seems to be completely stable.

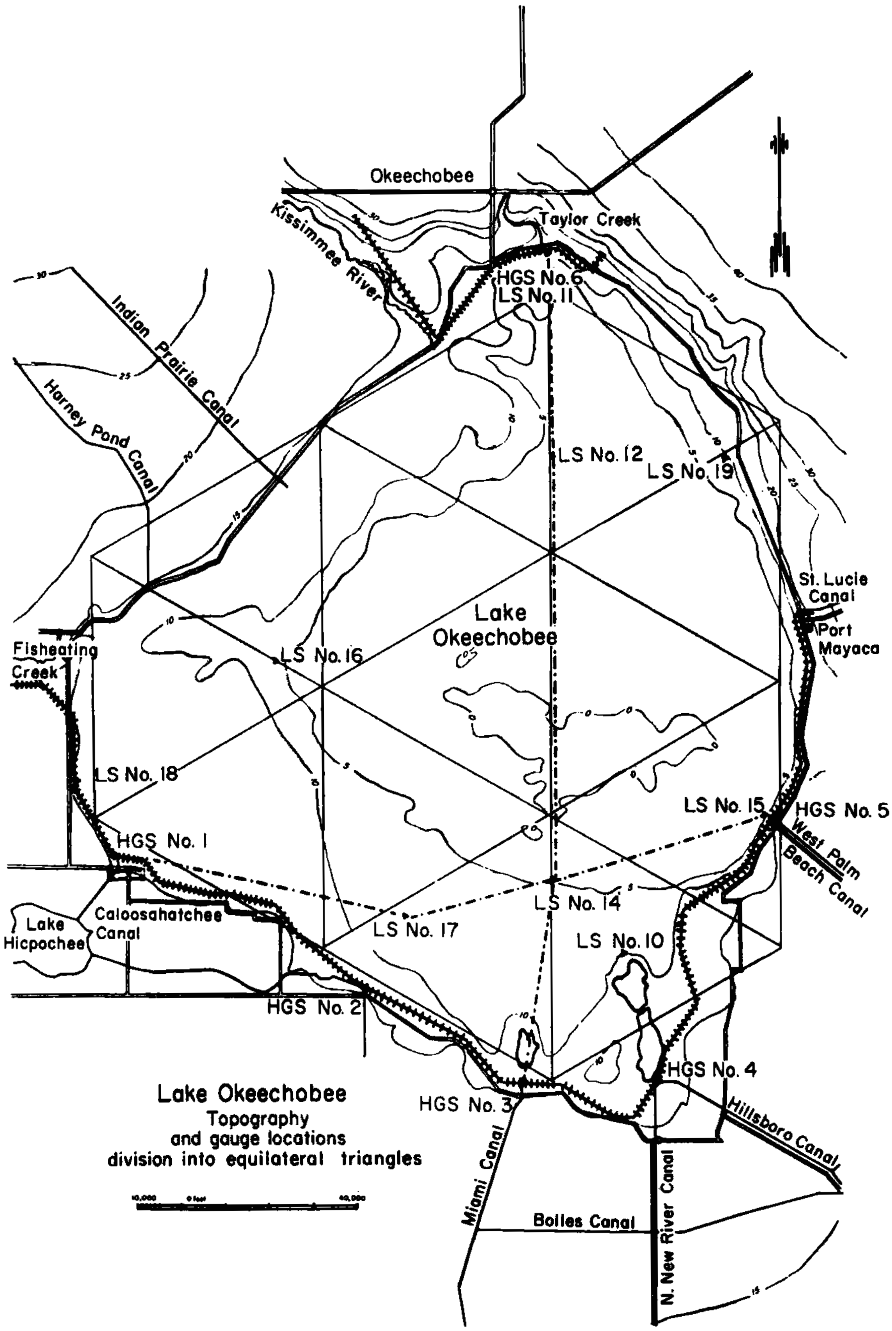


FIG. 34. The grid system used by H. Kivisild [33] in his computation for Lake Okeechobee. The gauges at which the water level was measured are indicated by circles.

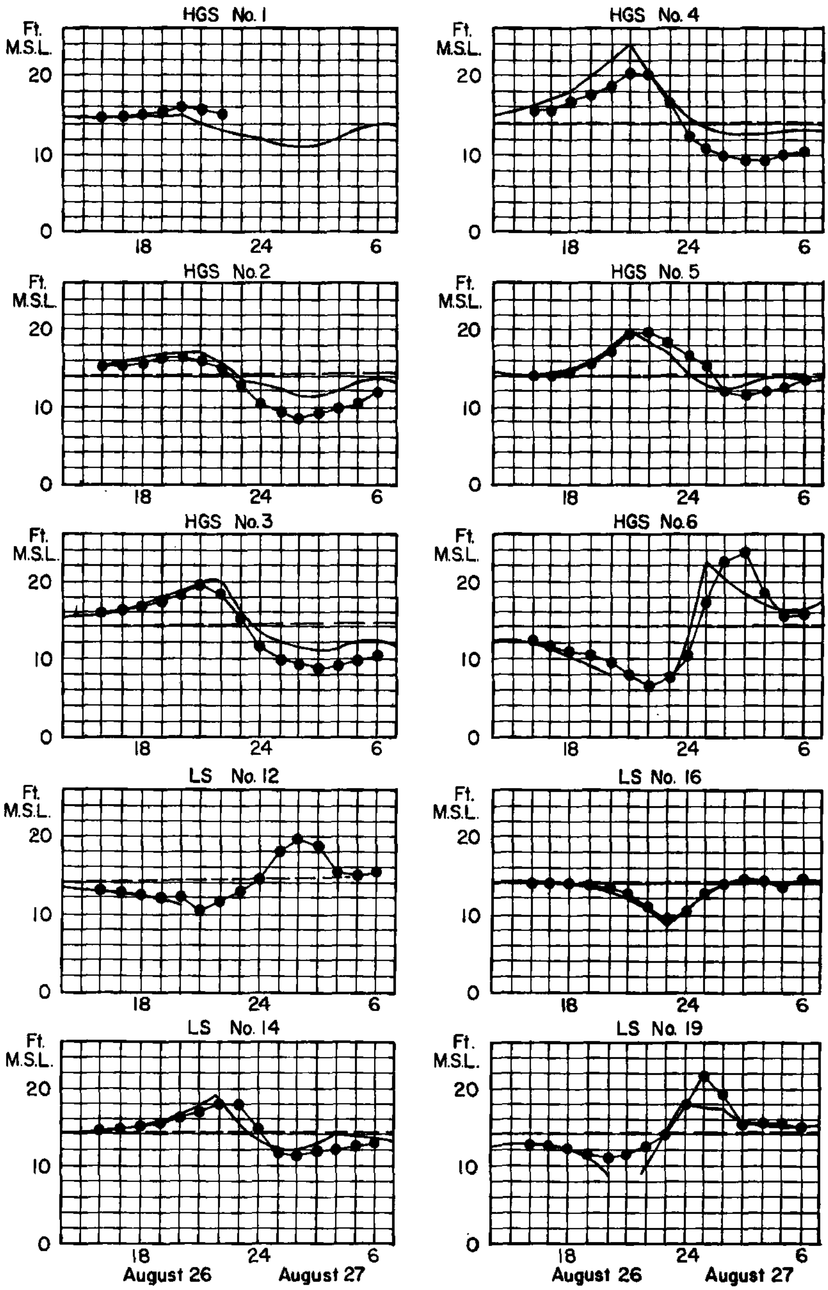


FIG. 35. Prediction of sea level in Lake Okeechobee for the hurricane of August 26–27, 1949, according to H. Kivisild [33]. The observations are given by the solid curve; the prediction by the dotted curve.

8.2. Machine Computations of Surges in the North Sea

The first numerical integration of the two-dimensional prediction equations using an electronic computer that gave the speed of computation necessary for practical predictions was carried out by Hansen [35]. Hansen used the system of differential equations.

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + suW + g\frac{\partial \zeta}{\partial x} = \frac{\tau_x^{\ w}}{h\rho_0} \tag{8.1}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + svW + g\frac{\partial \zeta}{\partial y} = \frac{\tau_y^{\ w}}{h\rho_0} \tag{8.2}$$

$$\frac{\partial \zeta}{\partial t} + \frac{\partial}{\partial x}(hu) + \frac{\partial}{\partial y}(hv) = 0 \tag{8.3}$$

where u and v are the horizontal velocity components and $W = \sqrt{u^2 + v^2}$.[1] It has been assumed that the water velocity is uniform over the vertical and that the bottom stress is proportional to the square of the velocity. The surface winds were computed from the geostrophic winds by applying a reduction factor of 0.7, and the wind stress was obtained from the formula

$$\tau^w = 3.2 \times 10^{-6}\, W_a^{\ 2} \quad \text{(cgs units)} \tag{8.4}$$

where W_a is the wind speed. The stress was assumed to be directed along the geostrophic wind. The coefficient of bottom friction was $s = 2.6 \times 10^{-3}$ (cgs units). Hansen carried out a numerical experiment with these equations for the case of the North Sea surge from January 31 to February 1, 1953. The grid he used is seen in Fig. 36. The boundary conditions were zero normal flow at all sides except at the open northern boundary, where the height was made equal to zero.

His finite difference scheme was based on the convectional centered time differences. The computations, carried out on the computer BESK in Stockholm, gave some promising results. However, spurious oscillations were observed, because of instability of the finite difference system. A comparison between the computed and observed surge at some points along the North Sea coast is given in Fig. 36. The mean error in the prediction of the peak surge is in the range 15–20 %. In the western and southern North Sea the prediction is reasonably good, while large errors are found for the stations on the east side. This may be partly explained by the effects of the artificial coast introduced at the eastern open side.

Further work on numerical prediction of surges has been carried out by

[1] In Hansen's, as well as in several other investigators' formulas, the factor ρ_0^{-1} before the stress does not occur. With the usual definition of the stress as a force per unit area, the above equations are the more correct ones.

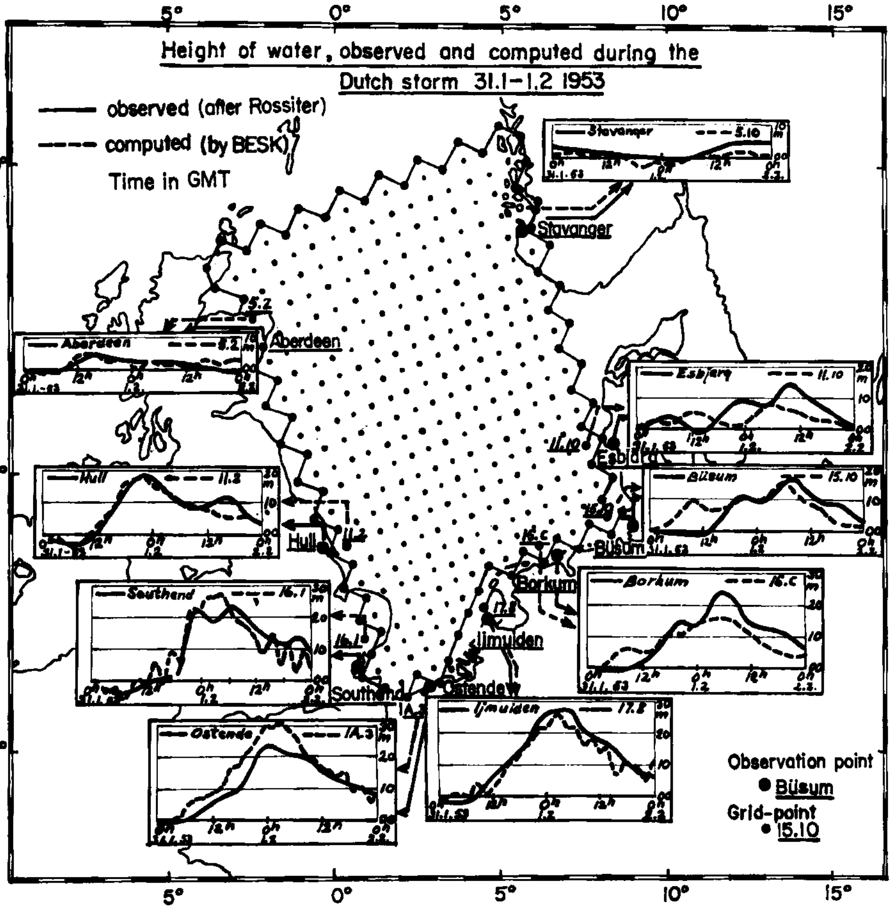


FIG. 36. The grid system used for the North Sea by W. Hansen [35] in his computation for the 1953 storm surge, and the prediction at different points along the coast.

a group at the International Meteorological Institute in Stockholm. After studies of the stability problem, the so-called "half-implicit method" described earlier was adopted. By this method and by applying a weak smoothing to the fields, the instability observed in Hansen's case could be eliminated. As the nonlinear terms were not found essential, the linear transport equations were used, including a linear bottom stress term. The bottom stress coefficient (earlier denoted r) was given the value $2.5 \times 10^{-2}h^{-2}$ (cgs units), this corresponds to a turbulent coefficient of friction $A = 200$ gm-cm^{-1} sec^{-1} in Ekman's steady state model. As in Hansen's computations only the wind effect was considered. It was verified that the pressure contribution was very small. The grid size used was $\Delta s = 76$ km.

The wind stress constant had to be lowered from the value 3.2×10^{-6} to 1.9×10^{-6} in order to get a good agreement in the case of the 1953 surge. This may, partly at least, be an effect of the smoothing term. The wind data were fed every 3 hr into the computer, and an automatic interpolation between these values was made.

Some steady state computations by Fischer [31] may be of interest and are shown in Fig. 37. They give the elevation and transport in the North Sea for two cases of a uniform, constant wind. In this case the Skagerrak Sea is also included. By the law of superposition, the two computations allow the construction of the solution to a uniform, constant wind of any direction and any magnitude. For example, one can compute the maximum elevation at any point in the North Sea and the wind direction for which this maximum elevation is attained, assuming a wind of given speed. The computation carried out for a wind of 20 meters/sec is given in Fig. 38.

The transports shown in Fig. 37 need commenting upon. One sees that a horizontal circulation is set up in both cases computed (surge of type 3). These circulations arise if one assumes that the wind drives the water over the more shallow parts of the North Sea, while compensating pressure flows take place where the depths are large. Some effects of the Coriolis forces are also noticeable. This can be seen, for example, in the isolevels that show the turning to the right predicted by Ekman's theory [26].

8.3. Machine Prediction of Surges in the Baltic Sea

In the Baltic some numerical predictions of surges have been carried out by Swansson [36] and Uusitalo [37].

Swansson used a one-dimensional linearized model, where the Baltic was considered as a nonuniform canal. The bottom stress was assumed proportional to the flow, and the pressure effects as well as wind effects were considered. Swansson used forward time differences and a smoothing term. His division of the Baltic in sections is shown in Fig. 39 and some of the results are given in Fig. 40.

Uusitalo used the linearized two-dimensional equations with a bottom stress term that varies inversely proportional to the depth. He made use of an implicit method of integration with forward time differences. The terms are then evaluated at the time $t + \Delta t/2$, by forming the mean of their values at the times t and $t + \Delta t$. In this case it was necessary to iterate. The mesh width in the grid was 44 km, the time-step 10 min. The wind stress constant was the same as in Fischer's case. Two principal cases were studied in the Northern Baltic, this part of the sea being separated from the rest of the Baltic by the shallow Ålands Sea. The sea was assumed to be at rest initially, and a wind was then suddenly introduced, blowing constantly along the Baltic (Fig. 41). In a second experiment the wind direc-

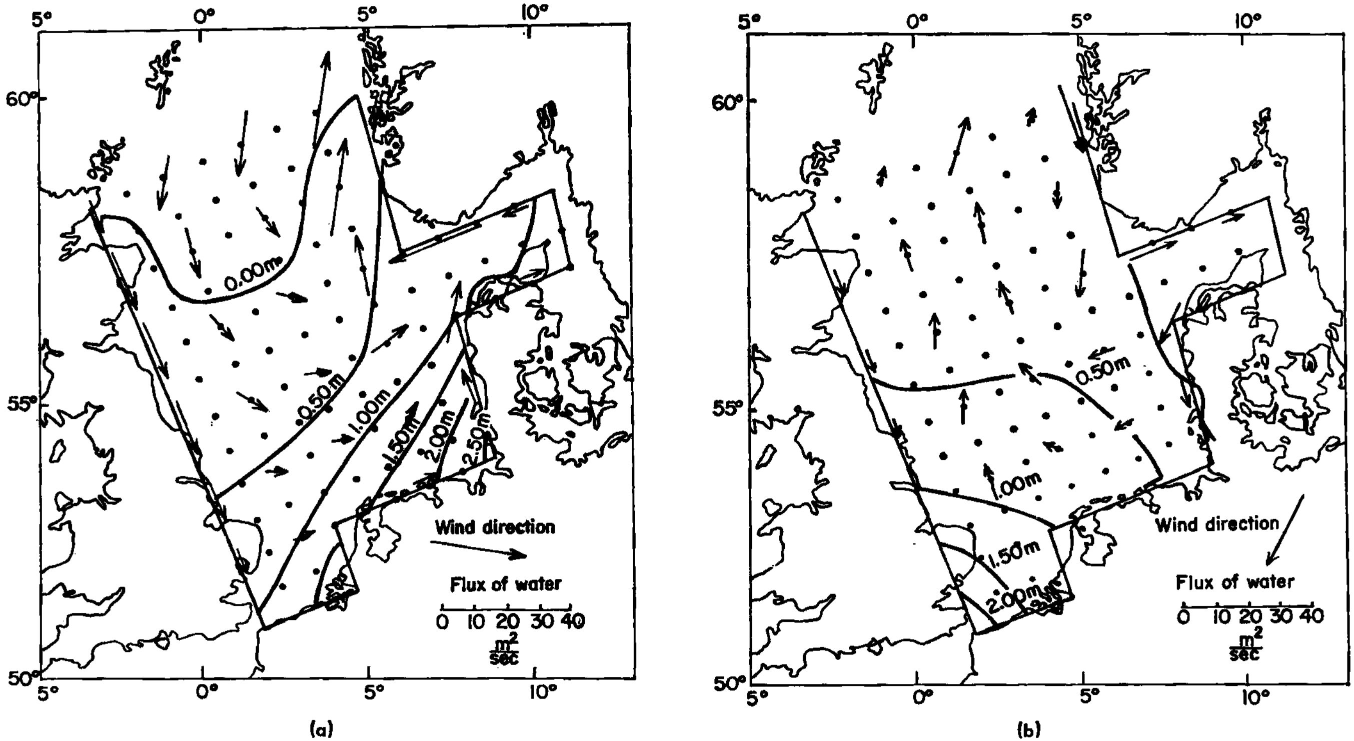


FIG. 37. Elevations and transports in the North Sea set up by a constant uniform wind of 20 meters/sec, according to G. Fischer [31]. The wind has the direction indicated in the lower right-hand side of the picture.

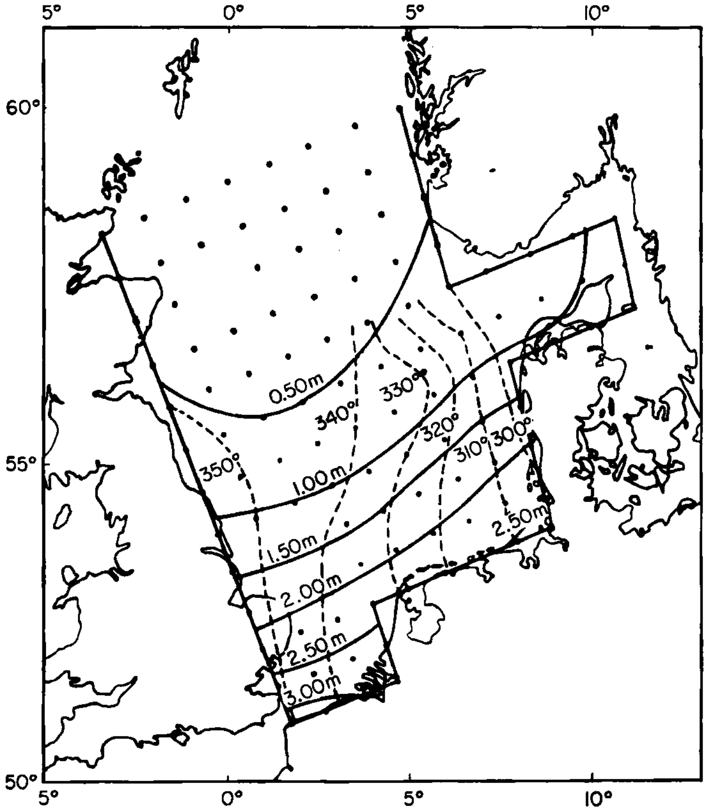


FIG. 38. Maximum elevations at different points in the North Sea set up by a constant wind of 20 meters/sec, and the wind direction at which this occurs, according to G. Fischer [31]. The figures give this wind direction measured clockwise from the northerly direction.

tion was changed 90° (Fig. 42). In the first case a relatively strong circulation persisted after the transient had died out, the strongest currents occurring on the eastern side (surge of type 3). However, in the second case the transport became small, indicating a vertical circulation (surge of type 2). Oscillations were found in the southern part of the sea (the Sea of Bothnia), with a main period of 5.4 hr. These seem to correspond to the transverse seiches that should have a period of about 5.3 hr. In the northern part (Bay of Bothnia) all oscillations were damped out quickly. By combining these two computations, in the way described earlier, the general response to a uniform but time-variable wind field over the Northern Baltic could be computed.

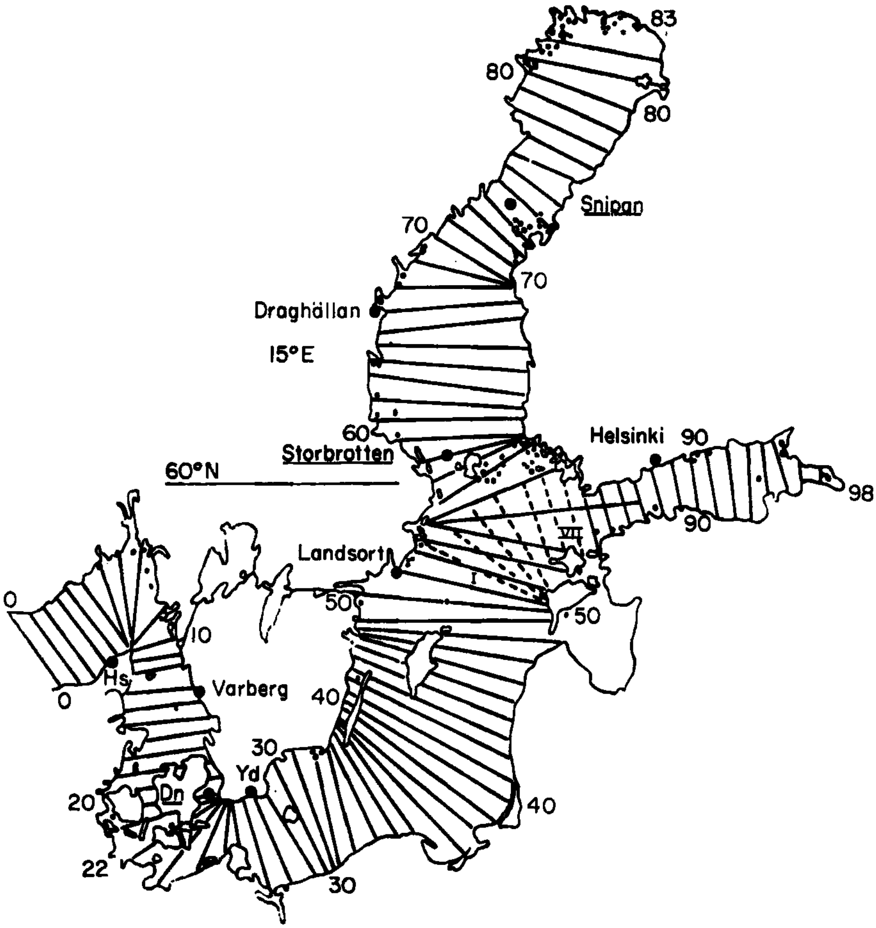


FIG. 39. Division of the Baltic Sea in section, according to A. Swansson [36].

8.4. Machine Computation of a Surge in Lake Michigan

An important contribution to the field has been given by Platzmann [38] who made a study of a surge created in Lake Michigan on June 26, 1954, when an intense squall line passed the southern end of the lake. Platzmann used the linearized transport equations (5.15)–(5.17) but neglected the Coriolis force and the bottom stress. In his finite difference formula, Platzmann used centered time differences according to the scheme shown in Fig. 43. For the special problem studied, he neglected, as a first approximation, the wind stress and considered only the effect of the atmospheric pressure variations. The southern end of Lake Michigan, with the bound-

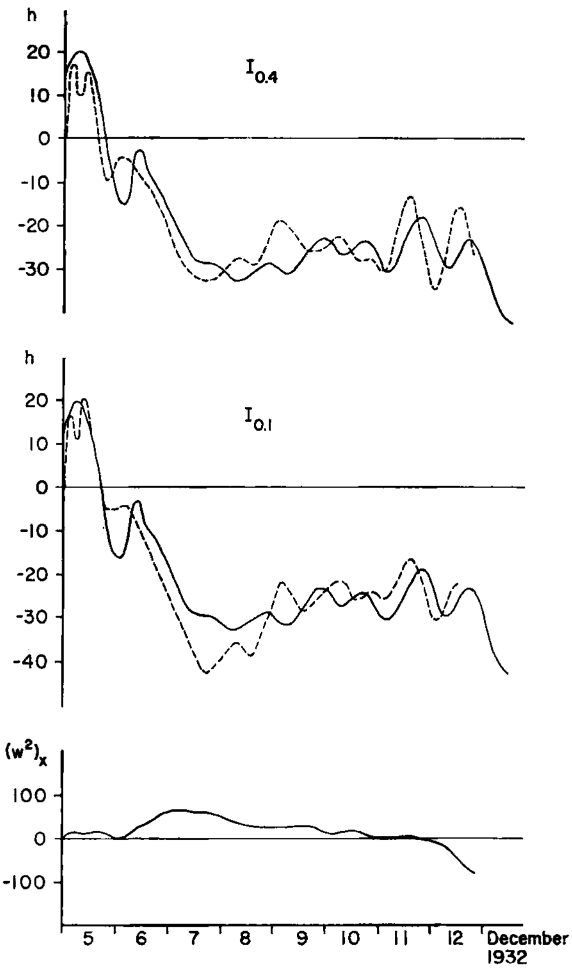
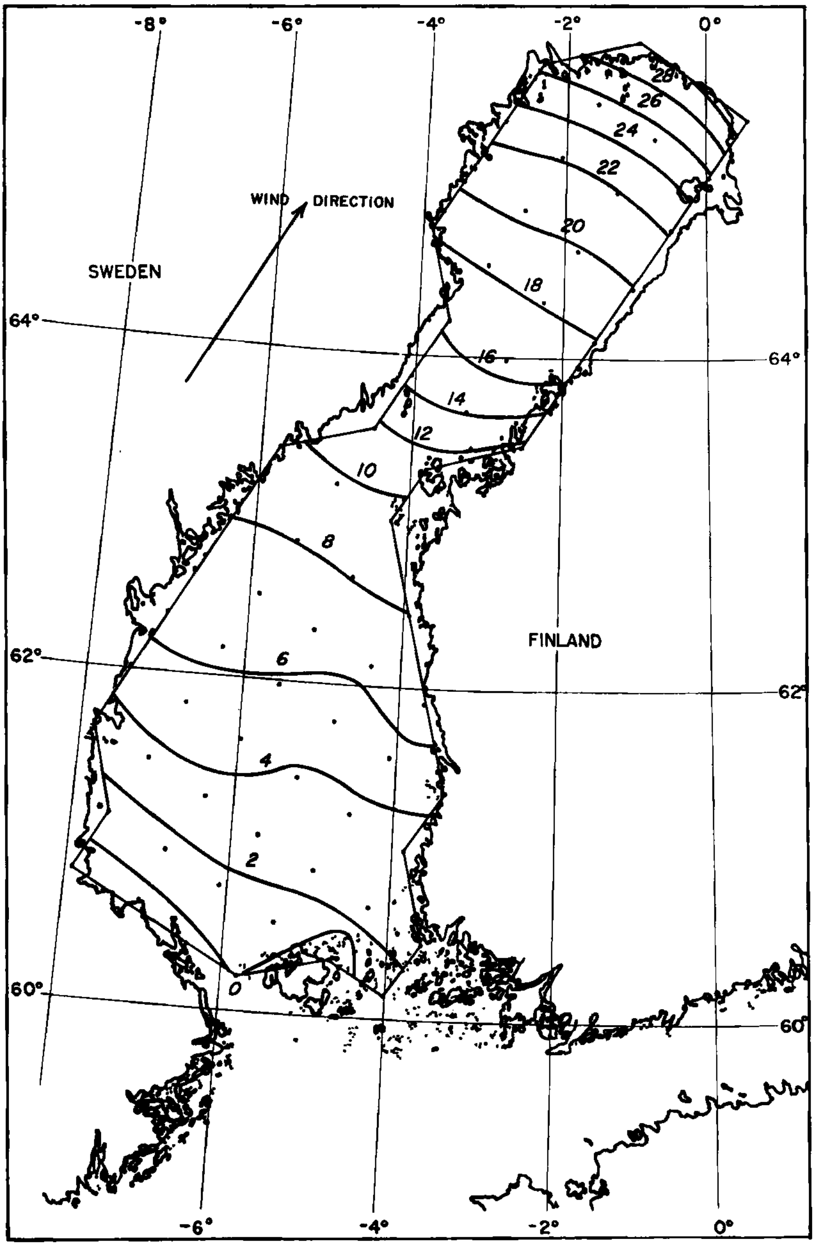


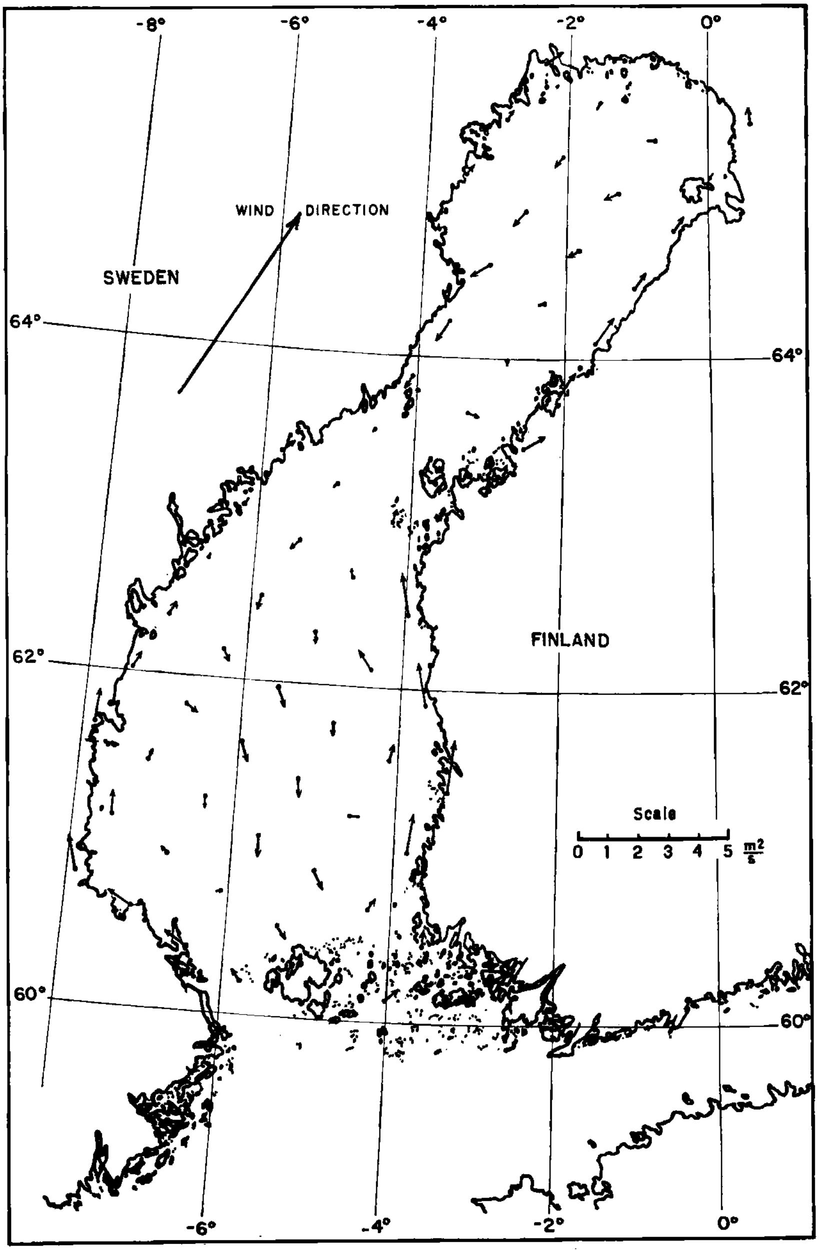
FIG. 40. Prediction of water levels at Varberg, Sweden, according to A. Swansson [36]. The solid lines represent observed values, the dashed lines computed values. The bottom curve gives the x-component of the wind velocity squared. The two cases shown differ by the values of the minimum cross-section area in the entrance to the Baltic Sea (0.4 km² and 0.1 km², respectively).

aries of the grid introduced, is seen in Fig. 44. The mesh width was about 2 nautical miles and the time-step 1 min. The total number of grid points was 6101. The squall line was represented in the simplest possible way. It was assumed to be bounded by two parallel lines, between which the pressure changed linearly. The squall line was further assumed to move with a constant speed and without change of shape. The computation was carried out on an IBM computer, type 704.



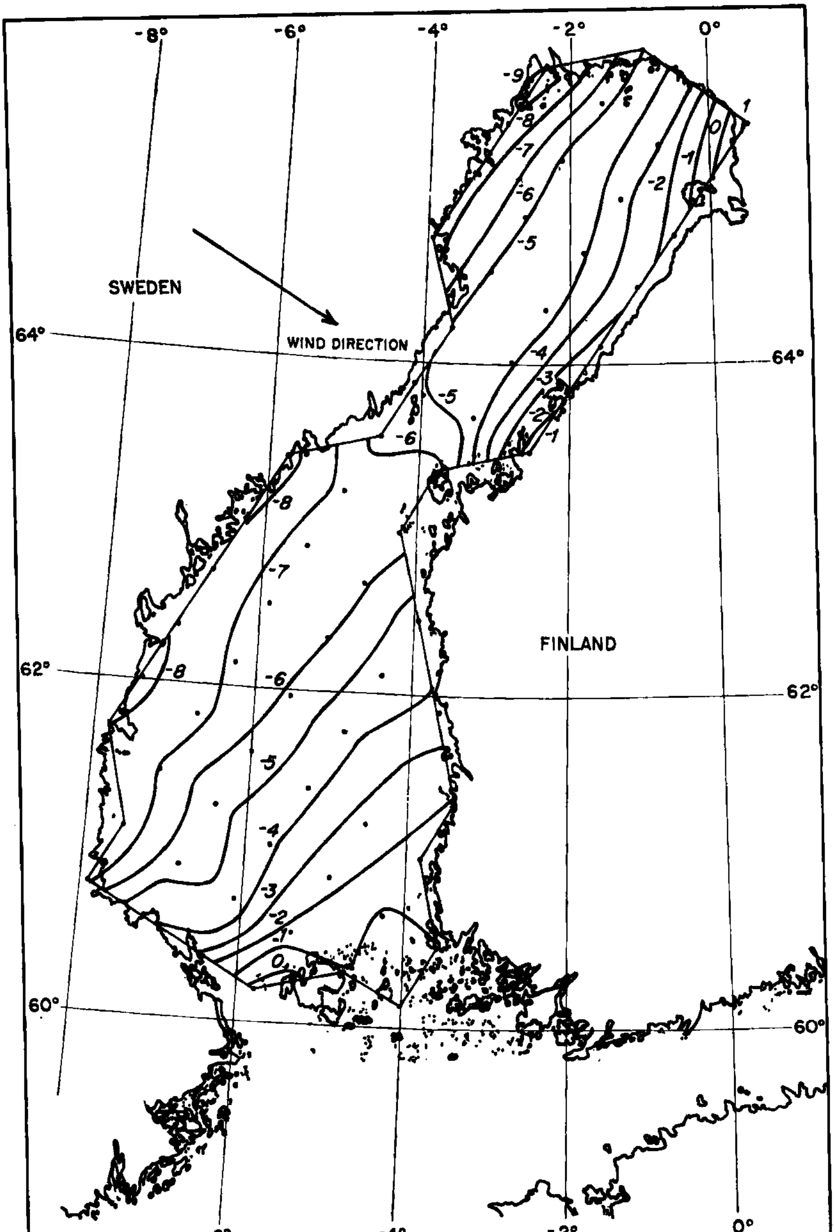
-8°
-6°
-4°
-2°
0°
WIND DIRECTION
SWEDEN
FINLAND
64°
62°
60°
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
-6°
-4°
-2°
0°

(a)

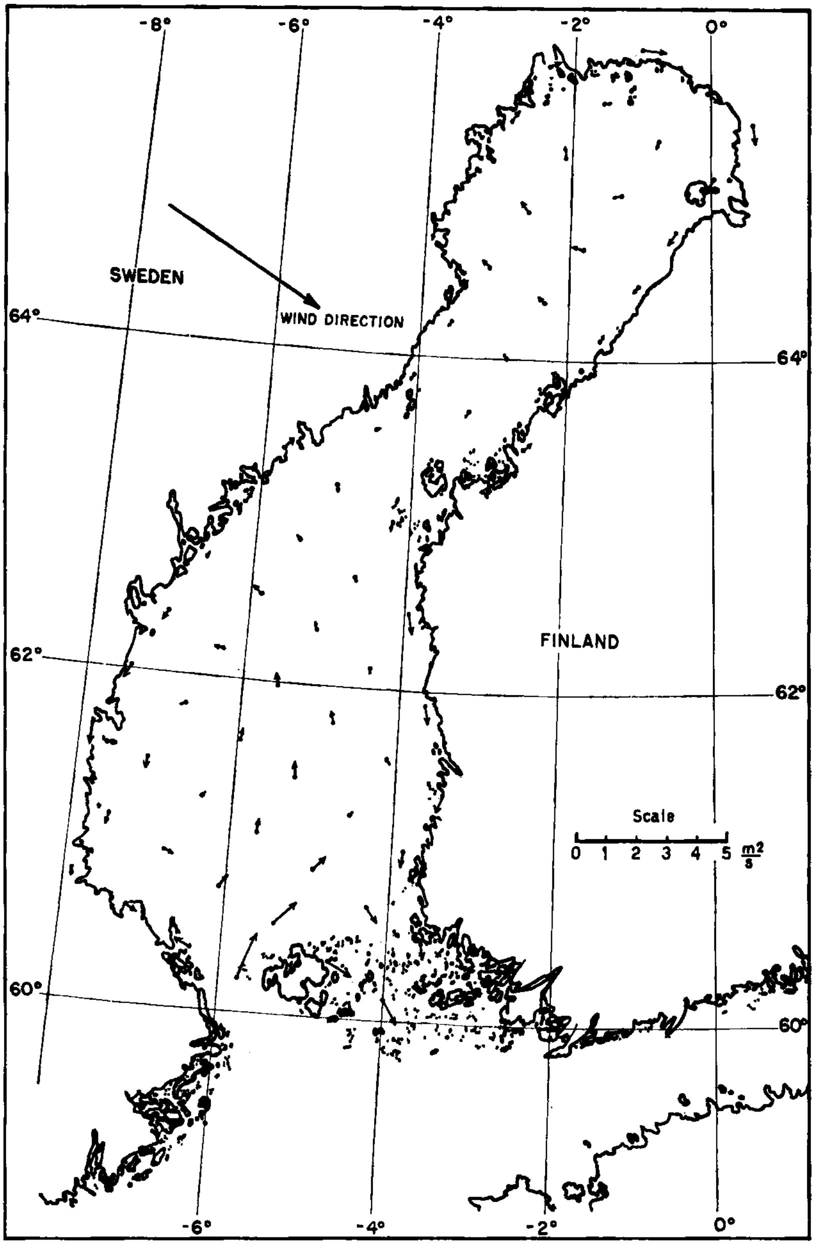


(b)

FIG. 41. Sea level elevations in cm (*a*) and transports (*b*—on facing page) set up in the Baltic Sea by a uniform northwesterly wind of 10 meters/sec. The wind has been blowing for 28 hours. From a computation by S. Uusitalo [37].



(a)



(b)

FIG. 42. Sea level elevations (*a*) and transports (*b*—on facing page) set up in the Baltic Sea by a uniform southwesterly wind of 10 meters/sec. The wind has been blowing for 12 hours. From a computation by S. Uusitalo [37].

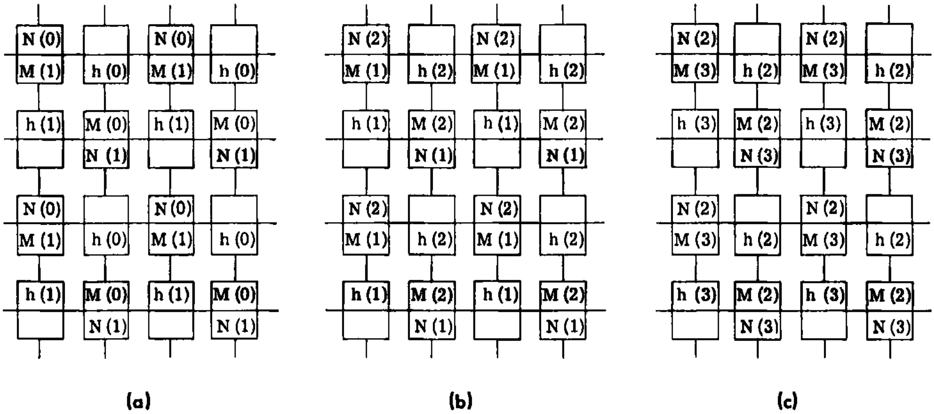


FIG. 43. Arrangement of gridpoints in the prediction scheme by G. W. Platzmann [38]. M and N stand for the transport components; h for the elevation. Each small box represents a half cell in the computer. The depths are stored in the empty half cells. The figure shows three successive time steps (a, b, c).

Some results of his computation are shown in Fig. 45 and 46. The computed and measured lake level and the record of the atmospheric pressure at Wilson Avenue Crib are shown in Fig. 47. The predicted time of arrival of the peak surge about 2 hr after the passage of the squall line was in agreement with the observations. However, the amplitude of the predicted peak surge was too low by some 40 %. This can probably be explained by the fact that the wind effect was not taken into consideration. The model also had difficulties in correctly reproducing the later oscillations. Platzmann suggested that this essentially was a result of an insufficient resolving power of the grid and of meteorological noise.

8.5. Further Research

To improve the present numerical sea level predictions, future work should concentrate on the following points.

1. Development of a numerical weather forecast scheme that can provide the necessary meteorological elements in the best way.
2. Testing the present wind stress formulas under different conditions and applying necessary corrections.
3. Development of a technique which permits the use of isolated observations of wind, pressure, and sea level.
4. Theoretical and experimental studies of bottom stress and nonlinear effects, and inclusion of these findings in the prediction.
5. Study of other near-shore effects and the problem of the prediction of the local surge.
6. Systematic comparison of an advanced statistical method, such as the influence method, and a numerical prediction method.

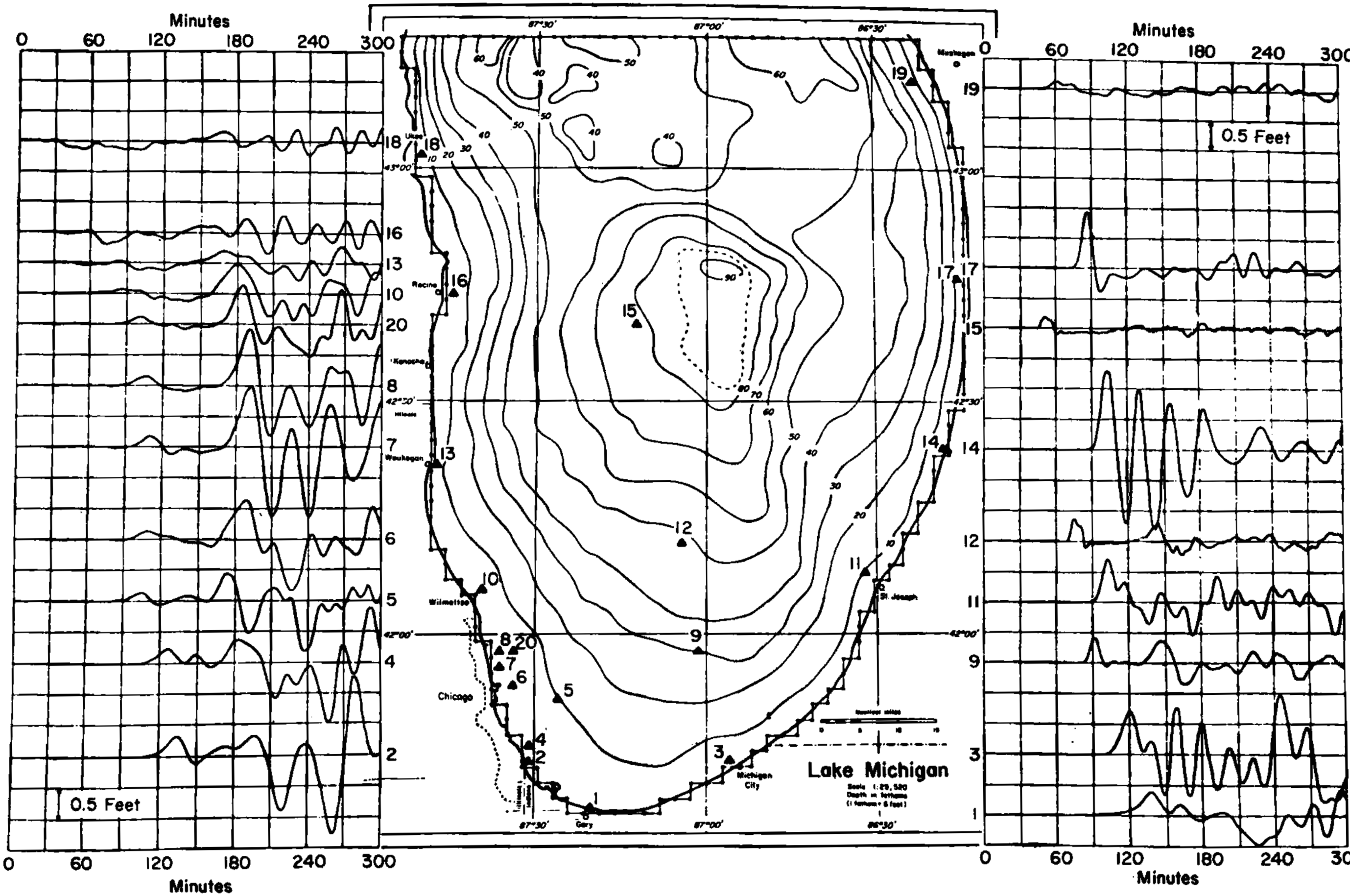


FIG. 44. Southern end of Lake Michigan with the boundaries of the grid used by G. W. Platzmann [38].

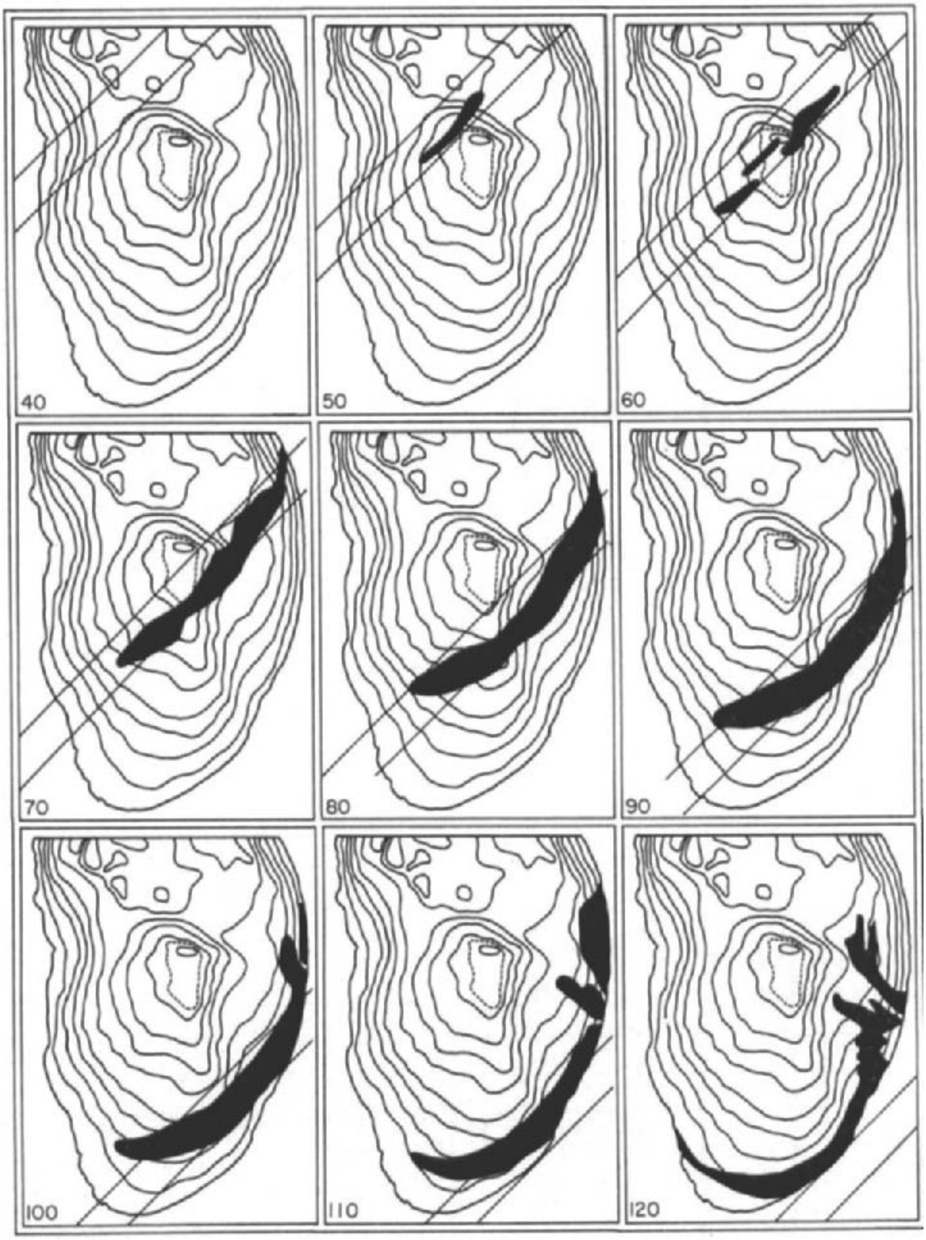


Fig. 45. Prediction of the water level in Lake Michigan for the surge of June 26, 1954, according to G. W. Platzmann [38]. The squall line moves at a speed of 54 knots. The numbers in the lower left corners give the time in minutes since the squall line entered over the lake. Shaded areas show positive (black) or negative (gray) stages in excess of 0.2 feet. The figure shows the development of the surge during the first two hours.

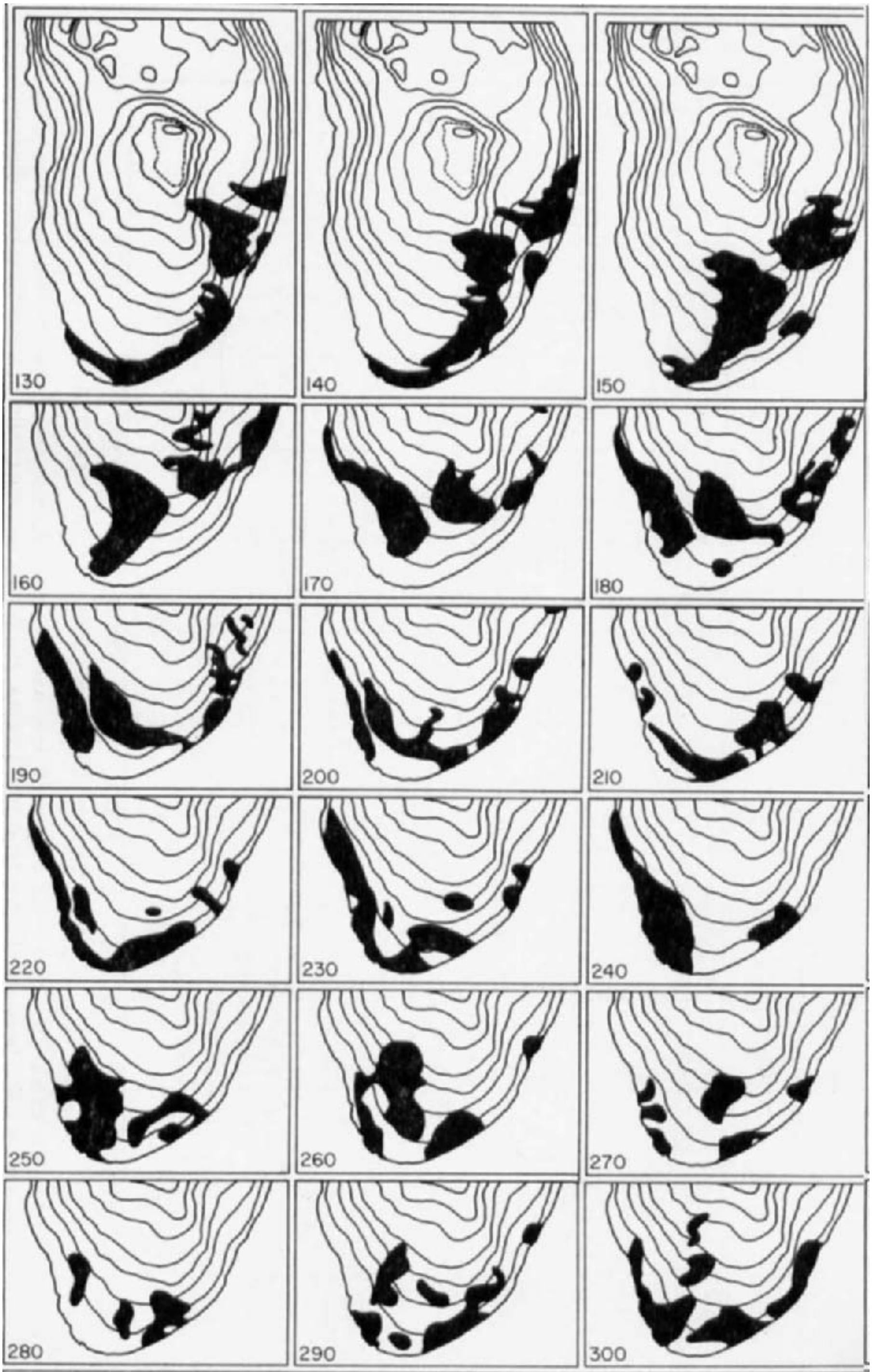


FIG. 46. The continued development during 3 hours of the surge shown in Fig. 45.

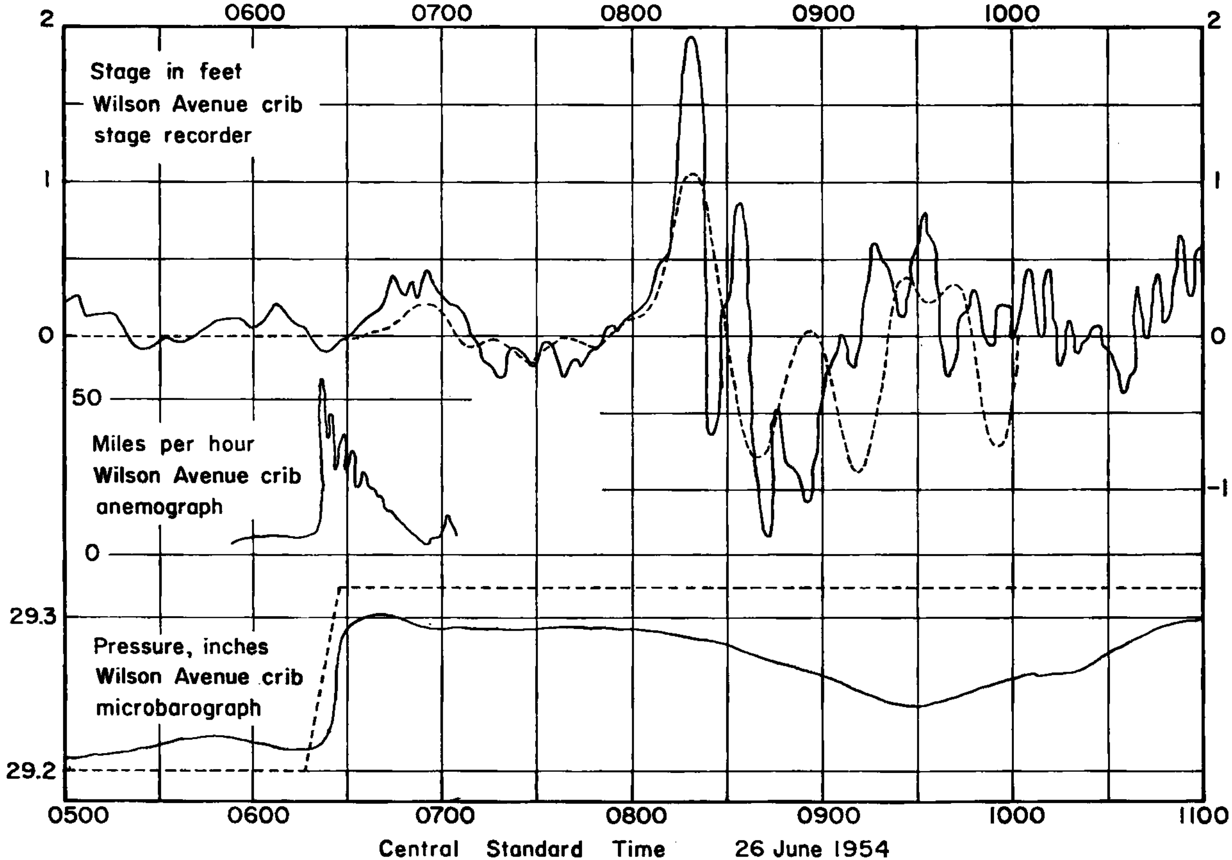


FIG. 47. Prediction of the water level at Wilson Avenue Crib for the surge of June 26, 1954, according to G. W. Platzmann [38]. The solid line gives the observed values; the dashed line, the predicted values for a squall line moving at 54 knots.

7. Study of the stability of linear and nonlinear numerical models.
8. Study of interaction of the astronomical tide and the meteorological effects, and development of a scheme for predicting the total sea level height.
9. Further study of the application of numerical prediction to surges producted by small-scale intense storms such as hurricanes.

The results obtained from the first numerical experiments have been promising, and with more systematic research it should be possible to make rapid progress.

The problem is sufficiently interesting from a scientific viewpoint to motivate further studies by meteorologists and oceanographers, and it seems that much can be learned about the atmosphere-sea coupling from numerical prediction experiments. When better knowledge of the dynamics of the real atmosphere-sea system, as well as a better knowledge of the models used in the numerical experiments have been obtained, we should also be able to develop an accurate prediction scheme, which can be used by a warning service. The solution of this problem may be an important contribution in our attempts to decrease the loss of human lives and the material damage caused by storm surge action.

List of Symbols

ζ	sea level elevation
τ^w	wind stress
$\tau_x{}^w$, $\tau_y{}^w$	wind stress components
τ_x , τ_y	stress components in the sea
$\tau_x{}^b$, $\tau_y{}^b$	bottom stress components
h	depth
F	fetch
g	acceleration of gravity
ρ, ρ_0	density of sea water
ρ_a	density of air
k, κ	constants
W^a	wind speed at anemometer height
p	pressure
p^a	atmospheric pressure at sea level
Z	unit step response function or characteristic amplitude of a surge
Z^*	unit impulse response function
T	integration time or characteristic period of a surge
f	Coriolis parameter
x, y, z	coordinates
x', y'	coordinates relative to storm center
u, v, w	velocity components
u_0 , v_0	velocity components of storm center
W	horizontal velocity
U, V	horizontal transport components
$\widetilde{u^2}$, etc.	vertically integrated value of u^2, etc.
$\bar{u}$, etc.	mean vertical value of u, etc.
H	characteristic depth
L	characteristic horizontal scale
D	Ekman depth
A	coefficient of vertical friction
E, K	energy integrals
φ	angle between x axis and outward normal of the coast
$F(z/h)$	vertical profile
λ, α	constants
m, n, s	grid point numbers
$A^s{}_{mn}$	smoothed value of A_{nm}
r, s	frictional coefficients

References

1. Ekman, V. W. (1923). Ueber Horizontalcirkulation bei winderzeugten Meeresströmungen. *Arkiv. Mat. Astr. Fysik* 17, 26.
2. Proudman, J., and Doodson, A. T. (1924). Time-relations in meteorological effects on sea level. *Proc. London Math. Soc.* [2], 24, 2.
3. Eckart, C. (1951). Surface waves on water of variable depth. Marine Phys. Lab. Scripps Inst. Oceanogr. Wave Report 100, Ref. 51-12.
4. Goldsbrough, G. R. (1952). A theory of North Sea storm surges. *M.N.R. Astr. Soc. Geophys. Suppl.* 6, 365–371.
5. Schönfeld, J. C. (1955). Discussion of tidal waves in a canal with variable cross-section. *Proc. 6th Gen. Meeting Intern. Assoc. Hydraulic Research* Vol. 1.
6. Proudman, J. (1955). The propagation of tide and surge in an estuary. *Proc. Roy. Soc. (London)* A231, 8–24.
7. Lauwerier, H. A. (1955). The motion of a shallow sea under influence of a non-stationary wind. Math. Centre Report T.W. 31, Amsterdam.
8. Kreiss, H. (1957). Some remarks about non-linear oscillations in tidal channels. *Tellus* 9, 1.
9. Reid, R. O. (1956). Dynamic storm-tide potential. Technical Report 127-1. Ref. 56-3T. Dept. Ocean. Met., Texas A. & M. College, College Station, Texas.
10. Kajiura, K. (1956). A forced wave caused by atmospheric disturbances in deep water. Technical Report 133-1, Ref. 56-26T. Dept. Ocean. Met., Texas A. & M. College, College Station, Texas.
11. Munk, W., Snodgrass, F., and Carrier, G. (1956). Edge waves on the continental shelf. *Science* 123.
12. Crease, J. (1956). Long waves on a rotating earth in the presence of a semi-infinite boundary. *J. Fluid Mechanics* 1.
13. Kajiura, K. (1959). A theoretical and empirical study of storm induced water level anomalies. Technical Report. Ref. 59-23F. Dept. Ocean. Met., Texas A. & M. College, College Station, Texas.
14. Schalkwijk, W. F. (1947). A contribution to the study of storm surges on the Dutch Coast. *Mededeel. en Verhandel. K. Nederl. Meterol. Inst., Ser. B* 1, 7.
15. Miller, A. R. (1957). The effect of steady winds on sea level at Atlantic City. *Meteorol. Monographs* 2, 10.
16. Tomzcak, G. (1954). The storm surge warning service of the German Hydrographic Institute in Hamburg. *Deut. Hydrog. Z.* 7, 1–2.
17. Corcan, R. H. (1948). Storm surges in the North Sea. H. O. Misc. 15072 USA.
18. Rossiter, J. R. (1959). Research on methods of forecasting storm surges on the east and south coasts of Great Britain. *Quart. J. Roy. Meteorol. Soc.* 85, 262.
19. Weenik, M. P. H. (1958). A theory and method of calculation of wind effects on sea levels in a partly enclosed sea, with special application to the south coast of the North Sea. *Proc. Koninkl. Nederl. Akad. Wetensch.* 73.
20. Harris, D. L. (1957). The hurricane surge. *Proc. 6th Conf. Coastal Eng. Council Wave Research, Berkeley, California.*
21. Harris, D. L. (1959). An interim hurricane surge forecasting guide. Natl. Hurricane Research Project Report No. 32, U.S. Weather Bureau.
22. Wilson, B. W. (1959). The prediction of hurricane storm tides in New York Bay. Technical Report 65-2. Ref. 5S-1F. Texas A. & M. College, College Station, Texas.
23. Reid, R. O. (1957). Forced and free surges in a narrow basin of variable depth

and width: A numerical approach. Techn. Report. Ref. 57-25T. Texas A. & M. College, College Station, Texas.
24. Welander, P. (1956). On the wind action on shallow seas with particular reference to the problem of numerical sea level prediction. Report Inst. Meteorol., University of Stockholm.
25. Charnock, H., and Crease, J. (1957). North Sea surges. *Sci. Progr.* **45,** 494–511.
26. Welander, P. (1957). Wind action on a shallow sea: some generalizations of Ekman's theory. *Tellus* **9,** 1.
27. Lamb, H. (1953). "Hydrodynamics" (6th ed.), p. 590. Cambridge Univ. Press, London and New York.
28. Ekman, V. W. (1905). On the influence of the earth's rotation on ocean currents. *Arkiv. Mat. Astr. Fysik* **2,** 4.
29. Platzmann, G. W. (1953). The computational stability of meteorological prediction equations. Dept. Meteorol. Report, Univ. of Chicago, Illinois.
30. Lax, P. D., and Richtmyer, R. D. (1956). Survey of the stability of finite difference equations. *Communs. Pure and Appl. Math.* **9,** 2.
31. Fischer, G. (1959). Ein numerisches Verfahren zur Errechnung von Windstau und Gezeiten in Randmeeren. *Tellus* **11,** 1.
32. Rossby, C. G., and Montgomery, R. B. (1935). The layer of frictional influence in wind and ocean currents. *Papers Phys. Ocean. Meteor.* **3,** 3.
33. Kivisild, H. (1954). Wind effect on shallow bodies of water with special reference to Lake Okeechobee. Thesis, Roy. Tech. Univ. of Stockholm.
34. Doodson, A. T. (1956). Tides and storm surges in a long uniform gulf. *Proc. Roy. Soc. (London)* **A237,** 325–343.
35. Hansen, W. (1956). Theorie zur Errechnung des Wasserstandes und der Strömungen in Randmeeren nebst Anwendungen. *Tellus* **8,** 3.
36. Swansson, A. (1958). Some computations of water heights and currents in the Baltic. *Tellus* **11,** 2.
37. Uusitalo, S. (1960). The numerical calculation of wind effect on sea level elevations. *Tellus* **12,** 4.
38. Platzmann, G. W. (1958). A numerical computation of the surge of June 26, 1954 on Lake Michigan. *Geophysics* **6,** 3–4.

This Page Intentionally Left Blank

AUTHOR INDEX

Numbers in parentheses are reference numbers and indicate that an author's work is referred to although his name is not cited in the text. Numbers in italic show the page on which the complete reference is listed.

S

T

U

V

SUBJECT INDEX