TEXT FLY WITHIN
THE BOOK ONLY

UNIVERSAL
LIBRARY
OU_160356
UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 551/L26A Accession No. 10387

Author Landsberg, H.E. jed.

Title Advances in geophysics. v.1. 1953

This book should be returned on or before the date last marked below.

ADVANCES IN GEOPHYSICS

VOLUME I

Advances in

GEOPHYSICS

EDITED BY

H. E. LANDSBERG

Geophysics Research Directorate,
Air Force Cambridge Research Center

EDITORIAL ADVISORY COMMITTEE

Bernhard Haurwitz
Walter D. Lambert
James B. Macelwane, S.J.
Roger Revelle

VOLUME 1

1952

ACADEMIC PRESS INC., PUBLISHERS
NEW YORK, N. Y.

COPYRIGHT 1952, BY
ACADEMIC PRESS INC.
125 East 23rd Street, New York 10, N.Y.

All Rights Reserved

No part of this book may be reproduced in any form, by photostat, microfilm, or any other means, without written permission from the publishers.

Library of Congress Catalog Card No. 52-12266

PRINTED IN THE UNITED STATES OF AMERICA

LIST OF CONTRIBUTORS

JAMES R. BAILSLEY, *Geological Survey, U. S. Department of the Interior, Washington, D. C.*

JOHN C. BELLAMY, *Cook Research Laboratories, Chicago, Illinois*

BERT BOLIN, *University of Stockholm, Sweden*

ARNOLD COURT, *Statistical Laboratory, University of California, Berkeley, California*

N. C. GERSON, *Geophysics Research Directorate, Air Force Cambridge Research Center, Cambridge, Massachusetts*

D. W. PRITCHARD, *Chesapeake Bay Institute, The Johns Hopkins University, Baltimore, Maryland*

FRED L. WHIPPLE, *Department of Astronomy, Harvard University, Cambridge, Massachusetts*

GEORGE PRIOR WOOLLARD, *University of Wisconsin, Madison, Wisconsin*

Foreword

New knowledge in the geophysical sciences is accumulating rapidly. Since the start of World War II the research effort in geophysics has increased by several orders of magnitude. This trend has been fostered by commercial, industrial, and military interests. The journals and monographic series publishing results in the various subfields of geophysics have also multiplied in number and scope. In recent years there has further been a tendency toward greater specialization.

While many workers in geophysics are busily engaged to advance the frontiers of this science dealing with our planet, the earth, little has been done toward integration of the findings. Textbooks appear only at intervals of several years—and then usually cover just one topic, such as meteorology or oceanography. The handbooks, now in existence, have become very dated by the rapid progress. The cross fertilization possible between closely allied fields has been retarded by the widely scattered literature and by the aggravating tendency of using obscure and limited means of communication. Duplicated progress reports, distributed in a haphazard fashion and in limited number, especially on some government-sponsored projects, hardly deserve to be called *publications*.

For these reasons it seems that the time is ripe for a series of monographic treatises that summarize, from time to time, the advances that have been made in geophysics. Other rapidly expanding fields of science have found this method advantageous. It serves many useful purposes. One is, of course, stock-taking. Another is an attempt at keeping investigators in closely allied fields aware of progress in each other's specialties. Finally, there is the value of critical reviews, which will disclose the gaps and serve as stimuli for further work.

In this first volume of *Advances in Geophysics*, several such review articles appear. Some of them summarize established knowledge, others point out new work which is needed to fill in the lacunae. Two of the papers deal with problems of evaluating geophysical data. The rate at which such data have accumulated is staggering. This concerns all fields of geophysics; they are all plagued by vast archives of raw material. Evaluation and interpretation is imperative and the old-fashioned approaches are manifestly inadequate.

We have further selected papers in fields where major progress has been made in the recent past. Most of these should be of interest to geophysicists at large. They cover the general circulation in the atmosphere; the action of the ocean at the land-sea interface in estuaries; the gravity field of the earth; the airborne magnetic survey methods. Two other papers concern themselves with the very high atmosphere. Both of them point to the fact that other sciences can contribute materially to geophysics. One covers the specific results obtained by astronomical techniques through the survey of meteorites. The other points to the great opportunities still open in solving the many puzzles of suprastratospheric regions. This should be both a challenge and an invitation to the physicists.

Not all specialties of geophysics find coverage in this first collection of *Advances*. Hydrology, Seismology, Volcanology, and Tectonophysics will have to await further volumes in this series. It is not that they have no advances to report but considerations of space set a limit to the present effort. We hope that it will be received kindly by our colleagues whose future collaboration in this endeavor we bespeak.

H. E. LANDSBERG

Cambridge, Massachusetts
September 1952

Contents

LIST OF CONTRIBUTORS v

FOREWORD vii

Automatic Processing of Geophysical Data

BY JOHN C. BELLAMY, *Cook Research Laboratories, Chicago, Illinois*

1. Introduction 2
2. Description of Present Techniques 3
3. Evaluation of Present Techniques 18
4. Unitary Records 23
5. Automatic Processing of Unitary Records 33
6. Summary 42

List of Symbols 42

References 43

Some New Statistical Techniques in Geophysics

BY ARNOLD COURT, *Statistical Laboratory, University of California, Berkeley, California*

1. Introduction 45
2. Extremes 53
3. Circular Distributions 75

List of Symbols 82

References 83

Studies of the General Circulation of the Atmosphere

BY BERT BOLIN, *University of Stockholm, Sweden*

1. Introduction 87
2. The Mean State of the Motion of the Atmosphere and Its Seasonal Variations 89
3. Basic Physical Principles Governing the General Circulation of the Atmosphere 91
4. The Momentum Balance in the Atmosphere 95
5. Some Basic Principles for the Energy Balance in the Atmosphere 102
6. Fluctuations in the Circulation of the Atmosphere. The Index Cycle . . . 103
7. Principal Aspects of the Approach to a Theory for the General Circulation of the Atmosphere 107
8. The Barotropic Model 109
9. The Baroclinic Model 113
10. Effects of the Non-uniformity of the Surface of the Earth 114

List of Symbols 116
References 116

Exploration of the Upper Atmosphere by Meteoritic Techniques

By Fred L. Whipple, *Department of Astronomy, Harvard University, Cambridge, Massachusetts*

1. Introduction 119
2. Techniques of Observation and Astronomical Results 122
3. Theory of the Meteoric Process 133
4. Results Concerning the Upper Atmosphere 139
5. Circulation in the Upper Atmosphere 147
List of Symbols 151
References 151

Unsolved Problems in Physics of the High Atmosphere

By N. C. Gerson, *Geophysics Research Directorate, Air Force Cambridge Research Center, Cambridge, Massachusetts*

1. Introduction 156
2. The Terrestrial Atmosphere 158
3. Static Properties and Processes of the High Atmosphere 179
4. Dynamics of the Ionosphere and Mesosphere 212
5. Conclusions 230
Acknowledgments 234
General References 234
References 235

Estuarine Hydrography

By D. W. Pritchard, *Chesapeake Bay Institute, The Johns Hopkins University, Baltimore, Maryland*

1. Introduction 243
2. Classification of Estuaries 244
3. The Physical Structure and Circulation Pattern in Coastal Plain Estuaries . 247
4. The Physical Structure and Circulation Pattern in Fiord Estuaries 253
5. The Bar-Built Estuaries 254
6. Theoretical Studies of the Dynamics of Estuarine Circulation 256
7. The Flushing of Tidal Estuaries 268
List of Symbols 278
References 279

The Earth's Gravitational Field and Its Exploitation

By George Prior Woollard, *University of Wisconsin, Madison, Wisconsin*

1. Introduction 281
2. The Exploitation of Gravity 286
3. The Exploitation of Gravity Measurements in Geodetic Studies 293
4. Geologic Uses of Gravity Data 301

Appendix 305
List of Symbols 310
References 310

Aeromagnetic Surveying

By James R. Balsley, *Geological Survey, U. S. Department of the Interior, Washington, D. C.*

1. Introduction 314
2. Basic Instrument 314
3. Associated Equipment 322
4. Field Survey Technique 323
5. Office Compilation of Field Data 326
6. Interpretation of Results 329
7. Results of Aeromagnetic Surveys 329
8. Advantages and Limitations 342
9. Applicability 344
List of Symbols 344
References 345

Author Index 351

Subject Index 358

ERRATUM

Page 100, line 2 of the legend for Figure 5 should read:

"CGS-units $\times 10^{29}$"

"ADVANCES IN GEOPHYSICS," VOL. 1

Automatic Processing of Geophysical Data

JOHN C. BELLAMY

Cook Research Laboratories, Chicago, Illinois

CONTENTS

		Page
1.	Introduction	2
2.	Description of Present Techniques	3
2.1.	General Remarks	3
2.2.	Measured Conditions	3
2.3.	Observation	3
2.4.	Observational Record	4
2.5.	Evaluation	5
2.6.	Original Records	5
2.7.	Transcription to Perforated Tapes	6
2.8.	Perforated Tape Records	6
2.9.	Teletype Transmission of Data	6
2.10.	Teletype Page Printed Records	7
2.11.	Plotting of Data	7
2.12.	Weather Station Charts	7
2.13.	Contour and Isotherm Analysis	8
2.14.	Contour and Isotherm Maps	8
2.15.	Facsimile Transmission	9
2.16.	Facsimile Charts	9
2.17.	Transcription to Punched Cards	9
2.18.	Punched Card Records	9
2.19.	Processing of Punched Cards	10
2.20.	Tabulations from Punched Cards	11
2.21.	Transcription to High Speed Digital Inputs	11
2.22.	High Speed Digital Input Records	11
2.23.	Large Scale Digital Computations	12
2.24.	Large Scale Digital Output Records	12
2.25.	Transcriptions to Analogue Inputs	13
2.26.	Analogue Input Records	14
2.27.	Analogue Computations	15
2.28.	Analogue Computer Output Records	17
2.29.	Microfilming	17
2.30.	Microfilm Records	17
2.31.	Final Analysis	17
2.32.	Final Records	18
3.	Evaluation of Present Techniques	18
3.1.	Manual Operations	18
3.2.	Graphs	19

Page
3.3. Printed Tables ... 20
3.4. Perforated Tapes ... 20
3.5. Maps ... 21
3.6. Punched Cards ... 21
3.7. High Speed Digital Input Records ... 22
3.8. Analogue Input Records ... 22
3.9. Microfilm Records ... 23
4. Unitary Records ... 23
4.1. Requirements of a Universal Record ... 23
4.2. Numerical Notations ... 25
4.3. Unitary Strip Chart Records ... 27
4.4. Unitary Vectorial Records ... 27
4.5. Isometric Geographical Records ... 30
4.6. Representation of Singular Values ... 31
4.7. Space Requirements of Unitary Records ... 32
5. Automatic Processing of Unitary Records ... 33
5.1. Analogue Recording ... 33
5.2. Digital Recording ... 33
5.3. Tabular Recording ... 34
5.4. Analogue Playback ... 34
5.5. Digital Playback ... 36
5.6. Tabular Computations ... 36
5.7. Choice of Parameters ... 38
5.8. Selection Operations ... 41
6. Summary ... 42
List of Symbols ... 42
References ... 43

1. Introduction

The need for processing large amounts of observational data in most fields of science has been growing very rapidly in recent years. This growth can be attributed to the fact that those problems involving essentially steady states of but few interdependent parameters have been solved in large part, leaving for study those problems involving rapidly changing conditions of several interdependent parameters. Investigations of these latter problems have produced vast amounts of observational data by the accumulation of observations, and by the introduction of instruments and recorders capable of high frequency response. This is especially true of the geophysical sciences in which the experimental conditions cannot be controlled in order to limit the rate and range of variations or the number of effective parameters. It has become quite evident that conventional methods of manually processing these great masses of data are wholly inadequate, and that rapid and convenient means of automatic processing must be used for both basic research and routine use such as is required for weather forecasting.

Although most recent instrumentation developments have con-

centrated on obtaining large numbers of more automatic, more accurate and faster observing instruments, considerable attention has also been given to the development of automatic data processing equipment, especially automatic data communication and computing equipment. However, there are a few operations in present data processing systems which have not yet been developed sufficiently to permit the full utilization of the capabilities of the newer types of observing instruments, communication systems or automatic computers. It is the plan of this article to describe in some detail the characteristics of presently available equipments in terms of how they are or can be used in a complete system of data processing so that the weak links in the data processing chain become obvious. In addition, the basic concepts of some present developments of means of strengthening these weak links are to be described. It is hoped that these descriptions will aid geophysicists in planning and executing expanded research and operational programs in the future as well as stimulate the development of automatic data processing equipment which will better serve the needs of the geophysical sciences.

2. Description of Present Techniques

2.1. General Remarks

In order to maintain and clarify over-all system relationships, the following descriptions of presently available techniques are organized with respect to the way they are used in the processing of one particular type of geophysical data, namely the observations made with radiosondes of upper air pressure, temperature and moisture conditions. The relationships among the various processes and records are indicated in Fig. 1.

Although most of the following descriptions refer specifically to radiosonde observations, essentially the same type of detailed conditions and operations occur with most other types of geophysical measurements. Thus these descriptions and subsequent conclusions apply quite generally to the processing of almost any kind of geophysical data.

2.2. Measured Conditions

The geophysical conditions which are measured in this particular example are the pressure, temperature and moisture content of the air along essentially vertical lines above the various radiosonde observation stations at the time of the observations.

2.3. Observation

In the radiosondes used in the United States, rising balloon-borne pressure, temperature, and moisture sensing elements cause changes of

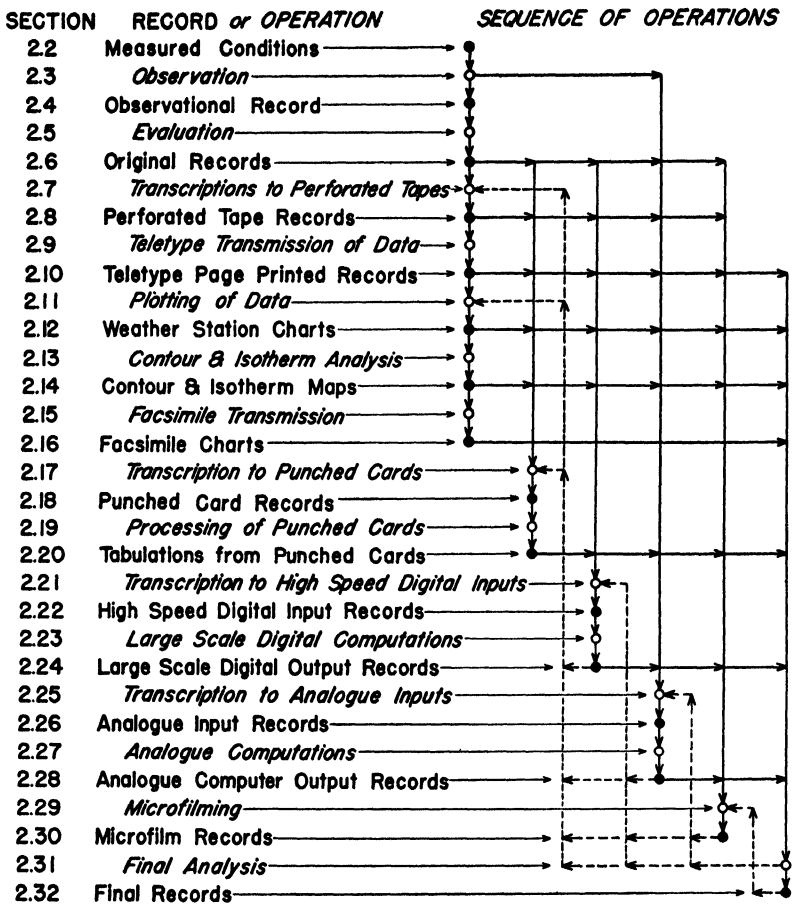


FIG. 1. Operational diagram.

the audio frequency of modulation of radio signals. The radio signals are received at a ground station and demodulated into electrical signals of varying audio frequencies which are automatically measured and recorded. The pressure sensing element is used to switch the frequency determining temperature and moisture sensing elements, along with fixed calibration elements, into the active circuit in a predetermined sequence at definite (calibrated) values of pressure.

2.4. Observational Record

The record of the received audio frequencies consists of the position of a broken ink line with respect to frequency and time scales preprinted on a paper strip chart.

2.5. Evaluation

The values of the temperature and humidity at various values of pressure are determined by manual reading of the frequency record and reference to calibration charts of switching points vs. pressure and of frequency vs. temperature and humidity. Errors due to oscillator drift and changing registry of the inking pen with respect to the preprinted grid are reduced by comparison to the reference frequency signals which have been periodically received. The results of these evaluations are written on the original records, see Section 2.6.1.

Since the indication of pressure values is in terms of the switching between various input sensing elements, and since there is no identification of the various types of signals other than continuity and expected values, considerable human selection and judgment is exercised in this evaluation.

In order to eliminate excessive evaluation times, only those points in the record (significant levels) are evaluated at which marked changes of the slope of the frequency vs. time curves for either temperature or moisture are evident.

The values of the temperature and moisture are then plotted with pencil on thermodynamic diagrams, see Section 2.6.2, and successive points are connected with straight lines. These graphs are used to check the consistency of the evaluations and to determine graphically mean temperatures between various pressure surfaces for manual numerical hydrostatic calculations of the height of prescribed standard (mandatory) pressure surfaces. The results of these calculations are then manually entered on both types of original records (2.6.1 and 2.6.2). In addition, the interpolated values of temperature and moisture at the mandatory pressure surfaces are read from the graphs of 2.6.2 and entered in the record of 2.6.1.

2.6. Original Records

The original records of the evaluations of the radiosonde observations are made in the following forms:

2.6.1. Raob Computation Data Record. This record consists of tabulations in Arabic decimal numerals, of the values of the temperature, moisture and height of all mandatory and significant levels. These forms also include the values of the frequencies read at the significant levels, as well as a simplified or modified written record of the message which is to be transmitted over communication systems as described in the following sections.

2.6.2. Graphical Thermodynamic Evaluation Diagram. This record consists of graphs of the temperature, moisture and geodynamic height

as functions of pressure in which the points plotted for each significant level are connected with straight lines.

2.7. Transcription to Perforated Tapes

The original tabulations (see 2.6.1), are manually rearranged and recorded on the Raob computation data record in the form in which it is to be transmitted over communication links. This arrangement consists of selecting the "more significant" of the significant levels, and arranging the values into a predetermined sequence of groups of five digits each. The first two groups of five digits are used to identify the station and time at which the observation was made. Usually the values for about twenty significant and mandatory pressure levels are included in this message.

Sometimes this data is transmitted manually with Morse code either over land lines or radio links. Usually, however, the written encoded form of the recording is transcribed into perforated tape records by manually typing with a teletype perforator.

2.8. Perforated Tape Records

This type of record of the observations consists of holes punched in a paper tape with a hole in a given position representing a unit of a binary digit and the absence of a hole in that position representing a zero of the corresponding binary digit (see Fig. 2a3 and Section 4.2). Five binary digit positions (hence $2^5 = 32$ possible combinations of holes) are usually used across the tape to represent decimal digits, alphabet letters or machine instructions such as space, line shift, upper or lower case, etc. By using upper and lower case shifts, $2 \times 2^5 - 2 = 62$ different characters or additional instructions can be represented. The normal size of the tape is 0.7 inch wide with 0.1 inch along its length allotted for each character or machine instruction.

2.9. Teletype Transmission of Data

The perforated tape record is manually introduced to a teletype transmitter which then automatically transforms the entire record into a time sequence of "on" or "off" electrical conditions corresponding respectively to values of the binary digits zero or one. The usual rate of sending this information on land lines is six characters per second with an occasional speed-up to ten characters per second.

Usually the transmissions from the various observing stations are collected at distribution centers where duplicate perforated tape records of the observations are automatically made with teletype reperforators.

These tapes are then manually arranged into desired sequences and automatically retransmitted to various weather analysis and forecasting offices. At each of these receiving points teletype page printers automatically produce the following type of record.

2.10. Teletype Page Printed Records

This record consists of Arabic decimal numerals (or alphabetical characters) printed on a strip of paper 8½ inches wide in the same coded form as originally set up in the process described in Section 2.7. The records from the various observing stations are recorded in sequence, with the station identification being given in the first five digit group of the report from a given station.

The same type of record is also obtained by manual typing by a radio or telegraph operator when Morse code transmission is used.

2.11. Plotting of Data

The values reported and recorded in Section 2.10 are manually selected and transcribed with either pen or pencil into the following forms:

2.12. Weather Station Charts

2.12.1. Constant Pressure Maps. The values of the geodynamic height, temperature and dew point observed at each observation station at any particular mandatory level are written in Arabic decimal numerals at predetermined positions with respect to the positions of the observation stations on a map. One such map is normally plotted for each conventional level, subject to the plotting manpower, time, and desires of the individual forecaster. Normally, wind observations for the same time and height are also plotted on these maps.

2.12.2. Thermodynamic Diagrams. The values of the temperature and dew point at each transmitted pressure level are plotted with pencil on a graphical reference grid of, essentially, temperature vs. pressure. These diagrams are similar to those of Section 2.6.2 with the exception that the ones considered here usually are overprinted with many related thermodynamic reference grids to facilitate thermodynamic evaluation of the soundings. One such plot on a separate sheet of paper is normally made for each of the observations the forecaster desires, with each sounding identified with respect to time and the observation station with Arabic decimal numerals.

2.12.3. Consolidated Thermodynamic Charts. For the special purposes of operation 2.15, several small scale simplified thermodynamic diagrams

are plotted on one sheet of paper. Normally sheets 12 inches by 19 inches, each containing 16 soundings, are used.

2.13. Contour and Isotherm Analysis

The primary use of radiosonde observations in present weather forecasting practice is an interpretation or analysis of the observed (actually calculated) heights of various mandatory pressure surfaces as plotted in constant pressure maps (Section 2.12.1). This analysis consists of manually drawing contour lines of the geodynamic height of the mandatory pressure surfaces at the time of the observations. Normally, this analysis includes considerations of the relationship between winds and horizontal pressure gradients in which the direction and horizontal space of the contour lines are indicated by the direction of the wind.

In similar fashion, isotherms of temperature and of dew point (hence lines of constant vapor pressure or constant mixing ratio) on the mandatory pressure surfaces are normally drawn by interpolation between the observed temperature values as plotted in 2.12.1. In addition, isotherms of the mean temperature (hence the vertical thickness) between two mandatory pressure surfaces are sometimes drawn. Although this can be accomplished by numerical calculation and plotting of the difference in height between the two pressure surfaces, it is usually accomplished by means of a graphical subtraction technique (called the intersection method) using superimposed contour charts for the two pressure surfaces involved.

2.14. Contour and Isotherm Maps

The results of the above analyses are recorded on maps with pencil or ink lines representing the contours and isotherms of temperature and dew point on the various mandatory constant pressure surfaces and the mean isotherms between such pressure surfaces. Two general forms of such maps are drawn, as follows:

2.14.1. Direct Analysis. In this case, the various isolines are usually drawn directly on the same maps on which the observations were plotted in 2.12.1. Distinctions between the various isolines are obtained by using various colored lines, and values corresponding to the lines are entered as Arabic decimal numerals.

2.14.2. Analysis for Facsimile Transmission. Copies of the maps of 2.14.1 are sometimes made in which the distinction between the various isolines is obtained by the types (solid, dashed, dotted, etc.) of lines used and on which a selected few of the original observations of 2.11 are copied. These maps are used as inputs into the type of data transmission described below.

2.15. Facsimile Transmission

In this type of communication the charts in 2.12.3 and 2.14.2 are automatically scanned point by point with a photoelectric cell which produces an electrical signal dependent upon the light reflection from the surface of the chart at the point being observed. These electrical signals are amplified and used to modulate carrier currents for land line or radio transmission, and, after demodulation of the received signal, to control the intensity of current passing through an electro-sensitive paper.

The electrode through which this current passes scans the area of the electro-sensitive paper in synchronism with the scanning of the original chart. Thus, the markings on the original chart are reproduced on the electro-sensitive paper. Normally a chart 12 inches by 19 inches with a resolution to approximately 0.010 inch (or essentially an on-off signal at about 2,280,000 points) is transmitted by this means in about twenty minutes.

2.16. Facsimile Charts

Facsimile charts or records in their usual present form consist of a silver gray sheet of metallized paper upon which black lines have been formed by the passage of current through the paper. The radiosonde data is represented by black lines in the coding as described in Sections 2.12.3 and 2.14.2.

2.17. Transcription to Punched Cards

In order to perform automatic evaluations of radiosonde observations, primarily for research and climatological purposes, many of the records are transcribed into punched cards. This operation involves the manual reading, selection, and encoding of the data as recorded in 2.6.1 or as interpreted at points other than observing stations in 2.14.1, and the manual operation of a key operated card punch called a key punch. This punch can be operated at speeds up to the order of six characters or decimal digits per second and sometimes can be operated in parallel with an electrical typewriter which produces a typewritten copy of the data that has been punched on the card.

2.18. Punched Card Records

Punched cards in common use are $7\frac{3}{8}$-inches long, $3\frac{1}{4}$-inches wide and 0.0065 inch thick and contain holes which are punched in predetermined positions to represent decimal digits or alphabetical characters. The International Business Machines cards contain 80 columns of 12 rows of punch positions with decimal digits represented by punching of the 10 lower positions; alphabetical characters are represented by combinations

of two punches, one in one of the three top rows and the other in one of the lower nine rows. (See Figs. 2a9 and 3a.) The Remington Rand card consists of two 45 column by 6 row punch positions, one above the other. The decimal digits and alphabetical characters are represented by various combinations of from one to three punches in the six possible positions in any given column.

2.19. Processing of Punched Cards [1]

Automatic processing of data recorded on punched cards starts with automatic reading of the values (and sometimes instructions as to what to do) punched in the cards. This reading is usually accomplished with brushes which sense the entry in each column by the time of completion of a current path through the holes with respect to a sequence of electrical pulses. Each such pulse corresponds to the passage of each possible punch position past the sensing position. In some types of punch card machines, the sensing of the hole positions and operations are entirely mechanical, rather than electromechanical.

Following is a listof the basic types of punch card processing machines, with a brief description of the type and order of magnitude of the speed of operations performed by them.

2.19.1. Verifier. This is usually a key punch modified to test the accuracy of the punches in the card and can be operated at speeds up to six columns per second.

2.19.2. Sorter. Sorters are used for selecting and sorting cards with respect to values punched in any one column at a rate of about six cards per second per pass.

2.19.3. Interpreter. Interpreters are used for printing on the card in alphabetical characters or Arabic numerals the values corresponding to the punched holes at a rate of 1.2 seconds per card.

2.19.4. Reproducer. This machine is used for reproducing (in the same or different columns) the values contained in one set of cards on another set of cards at a rate of two cards per second.

2.19.5. Collator. Collators are used for selecting, matching and merging cards from two groups at rates up to four mergers per second.

2.19.6. Calculating Punch. Calculating punches are used for the operations of addition or subtraction (three per second), multiplication (3.6 seconds per eight decimal digits) and division (9 seconds per eight decimal digits) and punching the results in the card.

2.19.7. Tabulator. The tabulator selects and prints, as described in Section 2.20, any selected values on the cards, or the algebraic sums of the values of any given columns of a set of cards at a rate of 2.5 cards (lines of table) per second.

2.20. Tabulations from Punched Cards

These records consist of decimal digits (or alphabetical characters) automatically printed in tabular form on either a continuous strip or a sheet of paper. The International Business Machine tabulations include 88 characters per line, while the Remington Rand tabulations include 100 characters per line. These records are normally used as the output records of the selections and computations which can be performed with the punch card machinery, and are usually used as the inputs for additional manual procedures as indicated in Section 2.31.

2.21. Transcription to High Speed Digital Inputs

The purpose of this operation is to obtain a record which is suited to the existing large scale digital computers to be described in Section 2.23. Although the records of 2.8 and 2.18 can be used as inputs to some large scale digital computers, many such computers have calculating speeds much in excess of that obtainable from punched paper records. In order to speed up the inputs to such machines the photographic and magnetic inputs described in the next section have been, or are being, developed.

Machines are presently available for making either photographic or magnetic input records by operating typewriter-like keyboards with manual reading of records such as described in Sections 2.4, 2.6, 2.10, 2.12, 2.14, 2.20, or 2.30; or by automatic playback of the punched paper records such as described in Sections 2.8 or 2.18. In these transcription machines signals of the on-off type usually are generated by the keyboard or punched paper playback device. These signals are then used to control recording lamps or passed through magnetic recording heads placed in proximity to moving photographic films or magnetic tapes or drums. The speed of such transfers is thus usually of the order of that required for manual typing of the data.

2.22. High Speed Digital Input Records [1, 2]

Presently available records for use as inputs to large scale digital computers consist of punched paper, photographic films, or magnetic tapes or drums. In all types the data is usually coded in terms of binary numbers although sometimes direct decimal notations are used. Punched paper tapes usually use hole space allotments of the order of 0.100 inch square, while both photographic films and magnetic tapes use a minimum of the order of 0.010 by 0.020 inch for corresponding spot sizes.

2.23. Large Scale Digital Computations [1, 2]

In recent years there has been a very intensive development of large scale (and, in general, high speed) digital computers. These devices perform essentially simple electromechanical (with relays) or electronic (vacuum tube) operations of comparison, selection, addition, subtraction, multiplication and division at speeds ranging from those indicated in Section 2.19 up to more than 1000 multiplications per second. In general these machines have been designed so that instructions for sequencing the simple operations are introduced into them in essentially the same form as the input data. By virtue of their very rapid speeds of simple operations, they can perform complicated operations made up of many simple operations in relatively short times. In some machines, internal memory (or short time, erasable recordings) are used so that several sequential operations can be performed for each data input. Most machines of this type have been designed for extreme accuracy using from ten to nineteen decimal digits. Thus they are ideally suited for producing tables of the solutions of mathematical equations.

The input devices to these computers consist of equipment for reading the records on punched paper, photographic films or magnetic tapes or drums and producing on-off electrical signals in various wires in the machines. The smaller size digit space and ease of rapid photoelectric or electromagnetic detection of signals obtained from the photographic or magnetic tape records permits feeding data into the machines at rates of the order of 1,000 to 10,000 transverse lines per second. This is to be compared with maximum rates of the order of 25 lines per second with perforated paper tapes.

Some of the large scale digital computers can automatically produce output records of the same form (2.22) as is used as their inputs, but this type of record usually must be transcribed into a more readable form for manual interpretation and additional processing. The usual final output devices are thus the automatic tabulators similar to the teletype page printer, 2.9, or the punch card tabulators, 2.19.7. A new type of output device called a numeroscope [2], which electronically writes Arabic decimal digits on cathode ray oscilloscopes has recently been developed. These representations can then be recorded on photographic film at speeds comparable to those obtained with the computers.

2.24. Large Scale Digital Output Records

As indicated above, the usual output record of the large scale digital computers consists of a tabular record of the same form as described in 2.10 or 2.20. The output record formed by the numeroscope is essentially in the form of a microfilm copy of such tabular records.

2.25. Transcriptions to Analogue Inputs

Although analogue computer techniques have not been applied, to the author's knowledge, to the automatic processing of radiosonde data, presently available devices for such processing are mentioned here so that all known types of automatic data processing equipment are included in this list.

By analogue techniques are meant those in which the value of an observation is represented by the magnitude of some other physical quantity such as an electrical voltage, a frequency, a shaft position, a length, etc. The types of devices for producing records which can be conveniently played back to produce analogue signals for introduction into analogue computers are described in this section. The analogue signals generated by these transcription devices often could be introduced directly into the analogue computers. This process, however, would usually be much slower than the capabilities of the analogue computers.

With radiosonde observations the most applicable analogue recording would be the direct recording of the varying audio frequency electrical signals described in Section 2.3. Any high quality sound recorder, using photographic films, magnetic tapes, or wax disks, could be used for this application. Usually, however, it is desirable to record simultaneously a fixed reference frequency on a parallel track or channel so that the effect of any variations of drive speed can be corrected. This technique of analogue recording on magnetic tapes in terms of variable audio frequency signals with parallel fixed frequency reference channels has been used extensively for the direct recording of many types of observational data at the Cook Research Laboratories.

Slightly modified sound recorders can also be used to transcribe the various other types of records into the desired analogue input records. Experimental models of modified adding machine keyboards have been built to produce voltage signals proportional to the depressed numbers. Using these signals to modulate the inputs to a sound recorder then effects a transfer from tabular or graphical types of records into the desired analogue form. Similarly manual, semiautomatic, and automatic curve following devices are available which can convert the position of a stylus or tracking photocell into voltage signals proportional to graphical data. These signals can then be recorded with the sound recorders to complete the transcription of graphical data. Finally, equipment used for demodulating pulse-code types of communication signals (in which rapidly sampled intensity values of sound waves are transmitted in terms of digital binary codes) could be modified for use for automatically transferring the signals obtained by automatic playback of the records described

in Sections 2.8, 2.18, or 2.22 into the proper signals for introduction into sound recorders.

In the general case of recording geophysical data in analogue form, the great advantage of sound track types of recording is the very high speed of response (up to at least 10,000 cps) which can be achieved with respect to other types of recordings.

2.26. Analogue Input Records

Records similar to sound tracks on photographic films or magnetic tapes seem to be most adaptable for automatic introduction of essentially continuous data into electronic analogue computers. With either recording medium the minimum practical wavelength resolution in the direction of travel is of the order of 0.001 inch and the minimum single channel track width is of the order of 0.020 inch. Either variable area or variable density modulation can be used with the photographic records, but usually only variable intensity modulation is used with magnetic tape records. In both recording mediums either direct amplitude modulation, usually with accuracy limits of the order of 1 to 5% of full scale, or frequency or phase modulation, with accuracy limitations from 0.1 to 1%, can be used. In principle, there is no restriction on the width of either kind of recording medium so that the possible number of parallel channels are essentially unlimited. Usually, however, 1 inch wide magnetic tapes and 35 mm. wide photographic films are used.

Although in most respects the characteristics of photographic film and magnetic tape records are comparable, several important differences exist. In the playback process the voltage generated in magnetic reading heads is proportional to the rate of change of magnetic flux, rather than the actual value of the flux. Thus, the output signal is essentially the time derivative of the input signal used to make the record, in contrast to the direct reproduction of the recorded signals with photographic films. This effect seriously restricts the usefulness of amplitude modulated magnetic tape records. In addition, it limits the slowest speed at which a frequency modulated magnetic tape record can be played back, since, for a given wave length and intensity of record, the output voltage is directly proportional to the playback speed. Magnetic tape records have the additional disadvantage that they are not directly visible for manual inspection, in contrast to variable area photographic film records. On the other hand, however, the elimination of protection against unwanted exposure, the elimination of chemical processing, and the possibility of easy erasing and re-use make magnetic tapes more suitable than photographic films for this type of record in many applications.

2.27. Analogue Computations [1, 3, 4, 5]

In recent years there has been an intensive development of electronic or electromechanical analogue computers of comparable magnitude to the development of large scale digital computers. Usually either the magnitude of electrical voltages or the linear or angular positions of shafts are used in analogue computers to represent the values of the data. Brief descriptions of the characteristics of these types of computation, with comparisons to the characteristics of large scale digital computers, are given in this section.

Very high speed, but usually low accuracy (up to about 1%) computations are conveniently accomplished by introducing continuously varying voltages corresponding to the input values into various networks of resistors, capacitors, inductors, transformers, rectifiers, vacuum tubes, etc. These networks can be chosen to modify the input voltages corresponding to many types of computations such as addition, subtraction, multiplication, division, differentiation, integration, etc. In addition, certain types of differential equations can be solved directly in this way by choosing the electrical components in the network to represent the coefficients in the analogous electrical differential equation. As an example of an electrical analogue computation, a Fourier analysis of the time variation of some geophysical parameter can easily be made with electrical filters at rates which would correspond to inputs into digital computers for the same operation in the order of 100,000 inputs per second, and the digital computer would be required to perform a large number of its basic arithmetic operations for each input.

Slower, but somewhat more accurate (up to about 0.1%) computations of practically all kinds can be carried out with computers of the shaft position analogue type. Particular problems are usually solved by using applicable mechanical linkages such as gear trains, differentials, ball-disk-cylinder mechanisms, lever and cam arrangements, etc., or variable electrical components such as variable resistors, capacitors, inductors or transformers. For the more accurate computations of this type the variable electrical components are usually used to provide a desired relationship between shaft positions and a voltage generated in the component. Self-balancing servo loops are then used to drive the shafts of such components until the voltages generated in them are equal to the input voltages or to the voltages generated in similar types of input components. In general, these types of computers are adaptable to the solution of many more kinds of problems of greater complexity than are the direct variable voltage analogue computers. They can quickly and easily solve many problems which would require extremely complicated

programming and long solution times with even the fastest large scale digital computers.

Of course, many analogue computers are combinations of the voltage and shaft position types. Such computers are ideally suited to problems such as linear differential equations with variable coefficients in which variable electrical components are used as analogues of the coefficients and the voltage is used as an analogue of the dependent variable. Both speeds and accuracies of such computers are relatively low, but their versatility make them very useful for many problems.

From these characteristics of analogue computers it is seen that whenever calculations with accuracies greater than about 0.1% of full scale are desired digital computers are used. Also digital computers usually are more adaptable to problems requiring but a few simple arithmetic operations on each input, even though desired accuracies are about only 1%. For such problems digital speeds are considerably greater than those obtained with most shaft position analogue computers, and usually the stabilization circuits required for even 1% operation of direct voltage analogue computers are relatively complicated.

Although input signals for analogue computers can be derived directly from the original observing instruments or from other types of records, as described in Section 2.25, the most adaptable type of analogue input record is that described in Section 2.26. These sound track types of records are played back by drawing the photographic film or magnetic tape past photocell or magnetic head pickups, thus producing voltage signals which represent the input data. If amplitude modulation has been used on the records these signals can be used directly as the input signals to the computer. If, however, frequency modulated records are used, the frequency of the playback signals usually must be measured, and usually corrected with respect to the reference frequency, before introduction into the computers. In general, the relatively low accuracy and high speed obtained with amplitude modulated signals suits them to direct voltage analogue types of computers, while the higher accuracy and lower frequency response of frequency modulated signals suits them for inputs to shaft position types of computers.

The final recording of the outputs of analogue computers is accomplished most rapidly (up to about 10,000 cps) with cathode-ray-oscilloscope-camera combinations. Slightly slower (up to about 1,000 cps) recordings can be made with photographic galvanometers, and somewhat slower (up to about 100 cps) by the movement of pens on paper or of styli on heat sensitive or electro-chemical recording papers. Recordings made with pens or styli, helices with tapper bars and typewriter ribbons, etc., in which recording elements are moved by the output shaft of shaft

position analogue computers or by shafts of self-balancing potentiometers for voltage analogue computers are relatively very slow (up to perhaps 10 cps).

2.28. Analogue Computer Output Records

2.28.1. Graphs. The records of the results obtained with analogue recorders are usually in the form of graphs superimposed on a preprinted reference scale grid. Accuracies and readability are usually adjusted to be compatible, ranging from 5% of full scale with the faster recorders to 0.1% of full scale with the slower recorders.

2.28.2. Intensity Modulated Areas. One form of record which can be produced conveniently only with analogue recorders consists of intensity modulation of photographic or electrochemical markings throughout the area of the recording paper. Although the percentage accuracy (perhaps up to 5%) obtained in the intensity modulation is somewhat low, the versatility, easy visualization and space saving for representations of variables which are functions of two independent parameters make this type of output record very suitable for many types of geophysical problems.

2.28.3. Tables. Some analogue computers are provided with digital counting wheels on the output shafts which produce tabular records in the form of Arabic numerals.

2.29. Microfilming

In order to obtain duplicate copies, to conserve space for storage and shipment, and in some cases to obtain more rapid and convenient access to desired records, most of the basic tabular or graphical records such as those described in Sections 2.6.1, 2.6.2, 2.20 and 2.28 are recorded photographically with standard 35 mm. copying cameras. Selection, arrangement and handling of the original records is performed manually.

2.30. Microfilm Records

The size of individual microfilm records, which are direct photographic copies of the original records, are usually of the order of one inch. The spaces for Arabic decimal numerals on records 2.6.1, etc., are reduced to the order of 0.010 by 0.015 inch, and the required resolutions of the microfilm copies of graphs such as 2.6.2 are reduced to the order of 0.001 inches.

2.31. Final Analysis

The final processing of the observational data beyond that already described almost always consists of human interpretation. Sometimes

considerable additional manual plotting, calculation, etc. are required. As an example, the weather chart records such as described in Sections 2.12.1, 2.12.2, 2.14.1 and 2.16 are studied in relation to other weather charts by weather forecasters or research workers and the conclusions are recorded in some form such as a written weather forecast. Usually the data contained in the tabular records such as those of Sections 2.10, 2.20, 2.24 or 2.30 are manually plotted in many different ways in order to present them to a research analyst in more readily visualized graphical forms. Such graphical plots, or the tabular records of Section 2.20 or 2.24 themselves for cases such as climatological summaries, are then studied, rearranged, etc., and the results recorded either as numerical tables or summary graphs. Many times the process of final recording includes setting type, photoreproduction of graphs, etc., and printing many copies with printing presses.

2.32. Final Records

The form of almost any of the previously described records can and sometimes is used for the final records of the results of the analysis. The size of records produced by printing presses are such that original 15 inch graphs are usually reduced to the order of 5 inches and the smallest Arabic decimal numeral sizes are of the order of $\frac{1}{8}$ and $\frac{1}{16}$ inch.

3. Evaluation of Present Techniques

3.1. Manual Operations

The above descriptions clearly indicate that although great advances have been made in recent years in the automatic handling or processing of observational data, many manual operations remain which seriously limit the overall efficiency of the processing operations. Most of the slower operations consist almost entirely of the manual transcription of records from one form to another which is more suitable for some kind of automatic processing, or which is more readily visualized by the analyst of the data. It is somewhat surprising that the relatively simple and routine operation of transcription between various forms of records is still performed manually, while relatively complicated operations such as communication and involved calculations can be performed automatically by machines.

Obviously all manual operations can never be eliminated from the processing system, especially in research work, since human judgement and interpretation must be used, at least in the final analysis (see Section 2.31), and in the control of the other processes. However, the time and effort required for such manual interpretations and control represents

a very small fraction of that now required for transcriptions from one form of record to another. Thus, if these transcriptions could be accomplished automatically, or better yet, eliminated entirely, the overall efficiency of the processing of data would be increased tremendously. In fact, such developments are required if the full capabilities of the newer types of automatic processing equipment for overall speed, versatility, and efficiency are to be realized.

The various kinds of records listed in Section 2 are evaluated here with respect to their adaptability as a universal type of record for all operations such that the transcription processes could be eliminated.

3.2. Graphs

The graphical forms of records consist of ink or pencil lines drawn on an under-grid of coordinate lines. They are used to provide easy manual visualization, efficiency of recording space for continuous or successive discreet observations, and sometimes easy graphical calculations.

For clarity the under-grid lines are usually quite widely spaced so that manual interpolation is required to obtain accurate numerical values. When high accuracies (say 0.1% of full scale) are required, quite large scales and careful interpolation must be used.

The use of under-grid lines does not permit the recording of a large number of observations (for example, 100 radiosonde observations) on a single sheet of paper. The possibilities of color distinction between many different graphs on the same under-grid are severely limited, and overlapping of the grid lines of adjacent graphs is usually highly undesirable. This shortcoming is demonstrated by the consolidated thermodynamic charts, 2.12.3, in which severe restrictions of accuracy, clarity and relative geographic positions have been imposed in order to represent but 16 out of approximately 100 radiosonde soundings on a single sheet of paper.

Usual graphical forms are not suitable for the recording of singular values such as the date, place, time, etc. of a particular observation. For this purpose they are wasteful of recording space, inconvenient for accurate manual reading, and not adaptable to automatic reading.

Color distinctions between the under-grid and the plotted curve provides for the use of automatic curve followers for producing input signals for digital or analogue computers. However, such automatic curve followers are neither rapid nor convenient and seriously limit the operational speed of electrical analogue computers and most large scale digital computers.

Although graphs can serve as both the input and output records of facsimile communication equipment, this is usually a relatively inefficient

form of communication. In this form of communication an on-off signal is sent for every point of the area of the record, rather than just the values of the observation. This results in a requirement for much greater band width, or transmission time, unless a large portion of the record is covered with useful data representations. In addition, it is not convenient to introduce any automatic selection or computation of the data into the facsimile operations, resulting in a loss of overall efficiency with respect to other forms of communication.

3.3. Printed Tables

Records in which Arabic numerals or alphabetical characters are used to represent the values of observations are included in this classification. They are usually used to provide reliably accurate visual reading. They are especially applicable to records of a few discrete values, or singular values such as identification, etc., in which cases a saving of recording space is usually obtained.

In general Arabic numerals are not useful as inputs for any form of automatic reading. Special forms of characters could be designed which could be read automatically, but it is probable that such a process would be quite complex, and that the space saving feature would in large part be lost. In any event signals from such equipment would usually not be appropriate for introduction into voltage analogue computers.

The fact that so many graphical records are used in the processes listed in Section 2 demonstrates the fact that tables of Arabic numerals does not permit convenient visual interpretation of the records. This is especially true of continuous, or successive discrete, values of a single parameter. Since most geophysical observations are largely of these forms, this characteristic is very disadvantageous.

3.4. Perforated Tapes

Perforated tape records are primarily useful as input records for automatic digital computers or communication links. They are, of course, capable of maintaining any desired accuracy.

The fact that many teletype tapes are overprinted with Arabic numerals or alphabetical characters indicates that direct visual interpretation is more difficult than in the case of printed tables, which themselves are none too convenient for this purpose. In addition, the area required for punched paper tape records is usually of the order of five times that required for Arabic numeral records.

Although punched tapes can be used as inputs to analogue computers, the required reading equipment is relatively complicated, switching tran-

sients are usually difficult to eliminate, and practical input speeds are slower than the capabilities of voltage analogue computers.

3.5. Maps

Two types of records are considered here; the plots of Arabic numerals at positions on maps, space cross-sections, etc., corresponding to the positions of the observations (2.12.1); and maps or cross-sections on which isolines of a given parameter have been drawn (2.14). These records are usually used to aid direct visualization and interpretation of the observations.

That the first of these forms does not provide in itself the desired aid to visualization is indicated by the fact that almost invariably isolines of the plotted values are drawn. Although in some instances the isolines represent the results of an interpretation or computation involving several different parameters, they are usually just another form of record of the plotted observations.

Although the prime purpose of these records is to provide a picture of the spatial distribution of the conditions represented by the observations, experience has indicated that this requirement is incompletely satisfied when three dimensional distributions, even without time variations, are required. This is a result of the necessity of using too many charts on too many sheets of paper to permit manual viewing and interpretation without an excessive amount of study and memory on the part of the analyst. As a corollary to this characteristic, the space efficiency of these records is usually quite poor. In some cases, however, such as with "sea level" weather maps on which many observations of many different parameters are plotted, quite good recording space efficiency is obtained.

Obviously neither of these forms of records is satisfactory as an input record for either digital or analogue computers. Although they can be transmitted automatically by facsimile, as indicated in Section 3.2 this is not usually an efficient form of processing.

3.6. Punched Cards

Punched card records are appropriate primarily as input records for automatic digital computations. They are most useful for operations of the logical type such as sorting, collating, etc., with respect to singular values or a few successive discrete values. There is no restriction on the accuracy which can be maintained with this form of record.

That punched cards are not suitable for direct visual interpretation is demonstrated by the fact that interpreting machines are used to print Arabic numerals or alphabetical characters on the cards, that tabulators are used extensively to transcribe the records to tabular form, and that,

as indicated in Section 3.3, even this tabular form is not usually suitable for the visual interpretation.

The space efficiency of punched cards is usually very low; of the order of magnitude of ten times as much recording area is required than for Arabic numeral records. This seriously limits their application to the recording of continuous or multitudinous discrete observations.

Although punched cards can be used as input records for large scale digital computers, they are in general not capable of the desired speed of operation. Similarly, although punched cards can be used, with some difficulty, as inputs to analogue computers, this process is in general very slow with respect to the capabilities of the analogue computers.

3.7. High Speed Digital Input Records

High speed digital input records, as described in Section 2.12, consist of digital notations, usually in binary form, on photographic films or magnetic tapes or drums. They permit very high speed reading of the digits and provide for the maintenance of any desired accuracy.

In general these records conserve recording space, which is usually a direct corollary of the fact that reading rates are high. However they do require of the order of four times the recording area of Arabic numerals on microfilm.

Direct manual interpretation of these records is more difficult than with any of the records previously mentioned. Magnetic tape recordings are not usually visible, and photographic film records of this type require optical magnification before they can compare even to punched paper tape records.

In common with the other types of digital records, these records have but limited applicability as inputs to analogue computers.

3.8. Analogue Input Records

Analogue input records (Section 2.26) use either variable area or variable density markings, with amplitude, phase or frequency modulation, on either photographic film or magnetic tape. They are primarily useful as records of continuous or multitudinous successive discrete observations to be used as very high speed inputs to analogue computers.

The use of intensity or amplitude modulation permits very efficient utilization of storage space for successive observations, but such records are limited to accuracies of less than 1% of full scale. Frequency or phase modulation permits the extension of maintainable accuracies to perhaps 0.1% of full scale, but this increases the required size of the record to be comparable to that of high speed digital input records of comparable accuracy capabilities.

Of the various forms of records of this general type, only amplitude modulated, variable area records on photographic films are easily interpreted visually. Even in this case optical magnification and registry of superimposed scales or grids are usually required.

None of the analogue records are readily adaptable to the recording of singular values since their accuracy requirements are usually beyond the scope of such records. Similarly, although equipment is available for changing analogue signals into digital signals, the accuracy restrictions limits the applicability of analogue records as inputs to digital computers.

3.9. Microfilm Records

Since microfilm records are but photographic reductions of graphical and tabular records, their characteristics are similar to those previously described. The saving of recording space is largely offset by the increased difficulty of visual interpretation which requires optical enlargement.

4. Unitary Records

4.1. Requirements of a Universal Record

It is obvious from the discussions in Section 3 that none of the types of records now in common use are adaptable to both universal automatic processing and convenient direct visualization. The various desirable characteristics of records discussed in Section 3 are summarized here to clearly outline the requirements of a universal type of record.

4.1.1. Visualization. Probably the most important characteristic of a universal record for geophysical observations, or results of reductions or calculations, is that it can be interpreted visually both easily and accurately.

4.1.2. Automatic Processing. The universal record should be capable of simple, convenient and very rapid automatic reading and recording in conjunction with communication systems, digital computers, and analogue computers.

4.1.3. Adaptability to Types of Data. The universal record should be capable of efficient representation of continuous, successive discrete, and singular values. The accuracy requirements of these types of data usually varies from about 1% of full scale for most continuous data to perhaps 0.1% of full scale for successive discrete values to much higher accuracies for singular identifications such as date, time, latitude, longitude, etc.

4.1.4. Space Utilization. The record should require a minimum of recording area, consistent with the desired or required reading resolution.

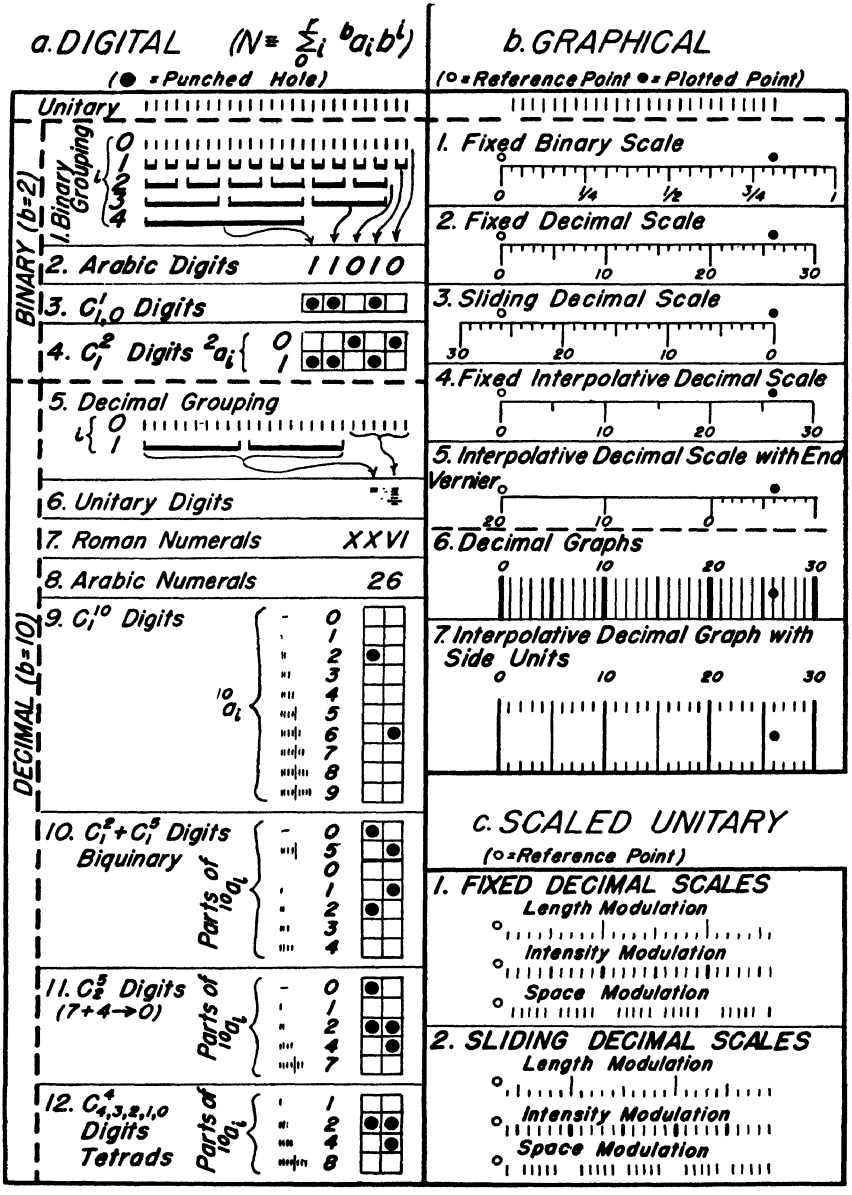


FIG. 2. Numerical notations.

4.2. Numerical Notations

Thus one of the major requirements of a universal type of record is that it be capable of convenient manual and automatic interpretation in both digital and graphical (analogue) form. That such a record is possible and practical is illustrated in Fig. 2, which is described here.

4.2.1. Digital Notations. In books on the theory of numbers it is proved that any, and all, integral numbers, N, can be uniquely determined by specifying the digits, ${}^{b}a_i$, in the equation given in Fig. 2a, where b is any desired integral number (greater than 1), and where the individual digits, ${}^{b}a_i$, can range from 0 to, but not including, b.

The fact that it is stated that the validity of this equation can be proved indicates that there is some other concept, and probable notation, of a number, N, which is more basic or axiomatic than the digital form of notation. Apparently this more axiomatic notation is the unitary notation, illustrated in Fig. 2, in which one mark, or signal, is used to represent each unit in the number N. This is the type of notation which would be used, for example, in making one mark on a piece of paper for each person that passed a given place if it were desired to count those persons.

The derivation of the common methods of digital notation (hence the "proof" of the equation for N) is illustrated in Fig. 2a. It is seen that the process of forming a digital number consists of counting off, or grouping, the units of the unitary representation into groups of b units each; grouping these groups into larger groups each containing b subgroups, etc. The number of units, or subgroups, remaining after each grouping process are called the digits, ${}^{b}a_i$, of the number. These digits are usually recorded side by side, with increasing order of grouping increasing toward the left. For example, as illustrated in Fig. 2a1, the binary ($b = 2$) notation is obtained by pairing off the units of the number, pairing these pairs, etc., as far as possible. The number (0 or 1) of remaining pairs left after each pairing process is then recorded as shown in Fig. 2a2 to obtain the Arabic numeral binary digital notation. The binary notation of Fig. 2a3, in which a remainder is represented by a punched hole and no remainder by no hole in the corresponding digital position, illustrates the notation sometimes used with perforated paper tapes. For purposes of the processing described in Section 5.8, it is sometimes desirable to provide separate positions for representing both the 0 and 1 values, as illustrated in Fig. 2a4.

In a similar fashion, decimal digital representations are obtained by marking off groups of ten units, etc., as illustrated in Fig. 2a5. The decimal digits can then be represented by the unitary decimal digital notation of Fig. 2a6. It is noteworthy that this type of representation is

used for dial telephone signals. The derivation of the Roman numerals of Fig. 2a7 is obvious, as are the definitions (Fig. 2a9) of the Arabic numerals used in Fig. 2a8. It is interesting to note that the Arabic numerals, with the exception of 7 and 9, consist essentially of unitary marks which have been curved and/or connected together for ease of writing.

The remaining notations illustrated in Fig. 2a are those commonly used for automatic reading. The ten positional code of Fig. 2a9 is used on IBM punched cards; the seven positional code of Fig. 2a10 is essentially that used with the abacus and some large scale digital computers. The notation of Fig. 2a11 represents the minimum number of positions which can be used in this type of coding for representing decimal digits for the rapid selection processes described in Section 5.8. The binary representation (of the form of Fig. 2a3) of decimal digits shown in Fig. 2a12 is used for the inputs of many large scale digital computers.

4.2.2. Graphical Notations. Graphical notations are essentially length analogue notations in which values are represented by the distance between a reference point or line and a mark representing the desired value. In the case of bar graphs the values are represented directly by the length of the line.

Graphical lengths or distances are measured with auxiliary length scales, or rulers, or with background grid scales preprinted on the graph paper. The more accurate rulers or grid scales are constructed with marks for each unit of distance corresponding to the desired resolution, or minimum required difference between successive values, as illustrated in Figs. 2b1, 2b2, 2b3, and 2b6. Frequently, in order to increase the ease of reading, only the marks or lines at every other, every fifth, or every tenth unit of desired resolution are included, as shown in Figs. 2b4 and 2b7. Visual interpolation is then required for estimating the positions of the omitted values. That this is in general undesirable when reliable accuracy is desired is indicated by the fact that many rulers and graphs are constructed with the unitary marks included at the ends of the scales, as illustrated in Figs. 2b5 and 2b7.

It is noteworthy that the construction of the grid scales, or rulers, is similar to the process of grouping in the derivation of digital representations from the unitary notations. These groupings are started either from the reference position (fixed scales, as in Figs. 2b1, 2b2, 2b4, 2b6 or 2b7) or from the plotted positions (sliding scales, as in Figs. 2b3 and 2b5).

4.2.3. Scaled Unitary Notations. It is clear from these illustrations that both graphical and digital notations are derived from unitary notations in much the same manner. Conversely, it is also clear that the scaled unitary, or unitary digital, notations shown in Fig. 2c are essen-

tially both graphical and digital in character. Their graphical character is obtained by specifying that the distance between successive unitary marks be substantially constant so that the end mark occurs at a distance from the reference position which is proportional to the value represented. Their digital character is obtained by modulating the unitary marks, or spaces, in order to provide the grouping characteristics of digital notations.

4.3. Unitary Strip Chart Records

Some adaptions of unitary numerical notations to recordings of either successive discrete values or continuous values are illustrated in Fig. 3. Any of these forms of records could be used, as described in detail in Section 5, as the input or output records for this class of data in any of the operations listed in Fig. 1, but some of them are more adaptable to certain operations than to others. For example, the unmodulated discrete records of Fig. 3a are very suitable for records of, say, mercurial barometer observations for any kind of automatic playback, but they would seldom be used since they are not suitable for direct visual digital interpretation. Similarly the space modulated records of Figs. 3c and 3e are more suitable for intensity analogue playback than are the intensity modulated records of Figs. 3b and 3d, although the error due to the different intensities of the lines in the latter would not usually be objectionable for this operation. On the other hand, the intensity modulated forms (Figs. 3b and 3d) are more suitable for visual graphical interpretation or shaft position analogue playback than are the space modulated forms (Figs. 3c and especially 3e) since in the latter the strictly uniform unitary spacings are slightly compromised by the modulation.

The choice between sliding or fixed scales of modulation is made on the basis of the character of the data to be recorded. The sliding scale records (Figs. 3b and 3c) are more useful for slowly varying data with high accuracy requirements since the accuracy of counting (either manual or automatic) is independent of the alignment of the transverse counting line. The fixed scale records (Figs. 3d and 3e) are more useful for rapidly varying data with lower accuracy requirements since the alignment of the transverse counting line is not important, but the loss of resolution inherent in sliding scale records of the type shown when the data varies rapidly is eliminated.

4.4. Unitary Vectorial Records

The adaptation of scaled unitary records to several interrelated observations, such as the cartesian components of positional or velocity vectors, is illustrated in Fig. 4.

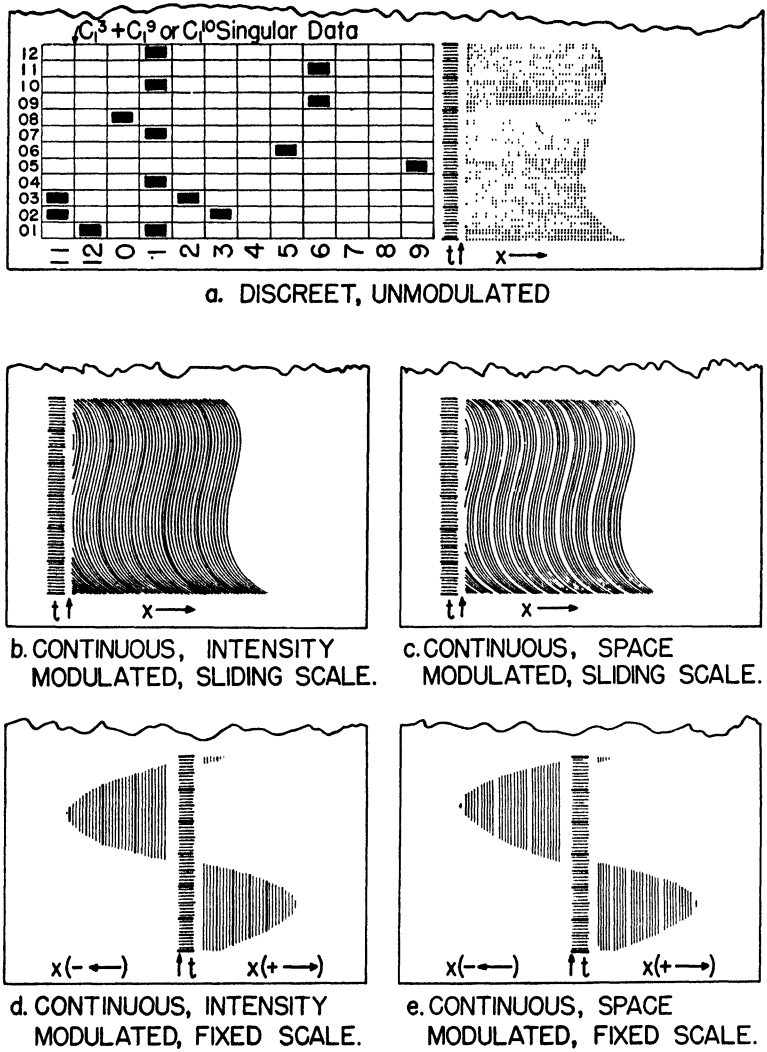


FIG. 3. Unitary strip chart records.

Figure 4a illustrates a possible record of a single airplane observation of the wind velocity components ($u = -21$ knots and $v = +20$ knots) and of the relative height (D of $+120$ feet) of the flight level pressure surface (8,200 feet of pressure altitude as read at the bottom of the z_p scale) which was made at a latitude of 42°07′N and a longitude of 87°53′W. Although a large recording space has been used, a direct visual representation superior to any other form of record known to the author for this type of data has been obtained. In addition this record

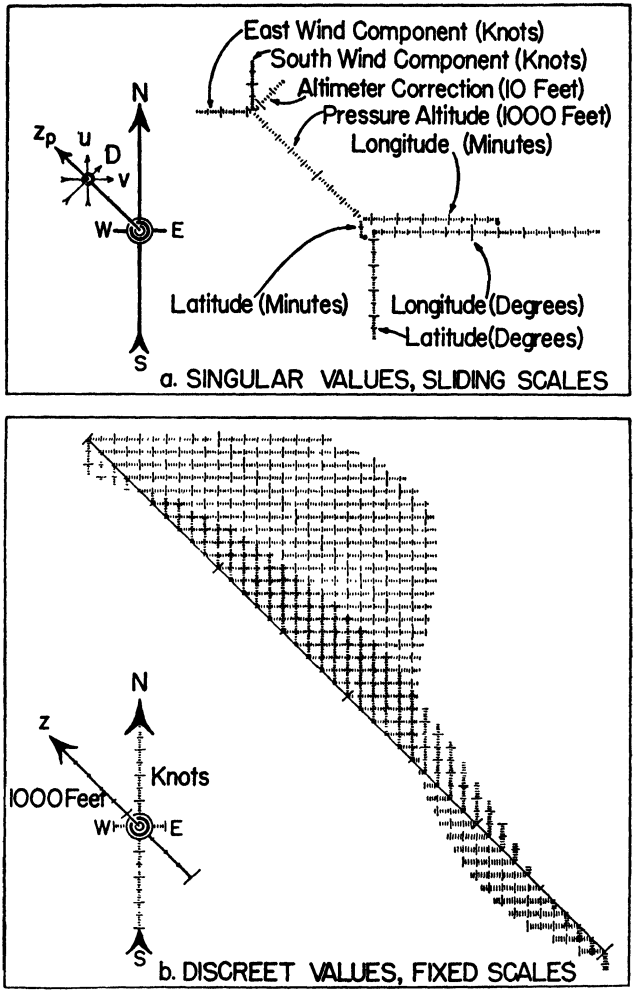


FIG. 4. Unitary vectorial records.

is amenable to automatic recording and playback in either analogue or digital form as described in Section 5.

Figure 4b illustrates a method of recording discrete successive values of a two dimensional vector as a function of a third variable. This particular example could represent an upper air wind observation in which the values of the N-S and E-W wind components are represented (in knots with a fixed modulation scale) at each thousand feet of altitude. The components are plotted in terms of the direction toward which the wind is blowing; hence the wind at ground level is blowing from the

NE with a northerly component of 7 knots and an easterly component of 10 knots. The northerly component decreases with height until the wind is due east (23 knots) at 5,000 feet. The clockwise shift of the wind continues, becoming southerly (about 25 knots) between 14 and 15 thousand feet, and this component increases to a maximum of 31 knots at about 20,000 feet. From this point up the clockwise shift continues due to a rapid increase of the westerly component, although the southerly component is decreasing and reaches zero at 36,000 feet where the maximum west wind component of 107 knots occurs. The wind at the top of the sounding (40,000 feet) is from the WNW and its magnitude is now decreasing.

This form of record not only provides convenient manual interpretation, both in terms of cartesian components and direction and speed; it is also capable of automatic recording and playback, as described in Section 5.4.

4.5. Isometric Geographical Records

Another form of record in which scaled unitary notations provide convenient visual interpretations is illustrated in Fig. 5. This is a record of the temperature (in terms of the parameter S described in more detail in Section 5.7) as a function of pressure (pressure altitude) as observed over the section of the United States included in the region of this sectional map at one particular observational time. The values of S are recorded at each of the significant and mandatory levels (see Sections 2.5 and 2.7) for each of the radiosonde observations. The origins ($S = 0$, $z_p = 0$) of the individual graphs are located at the geographic location of the observing stations. Since the pressure height scales are parallel, and sliding scales are used for the S values, the various values of S at any given pressure surface are read at the correct relative geographical positions with respect to each other.

Experience [6] with this type of record has demonstrated that it provides a more easily visualized three-dimensional representation than any other known form of record. In this interpretation the pressure altitude scales are visualized as sticking up into the air as, say, telephone poles with cross arms of lengths proportional to the values of S placed at the appropriate vertical positions. More common graphical notations using under-grid lines are not adaptable to this form of record since the grids of adjacent stations overlap and result in a hopeless confusion of lines unless an excessively large map is used. The relatively infrequent overlapping of recording lines (such as in the region of 43°N, 80°W), even with quite close spacings of stations, is usually not very troublesome with respect to manual interpretation of the records

of the form shown in Fig. 5. The overlapping of the fields of possible ranges of the records of this type, though, does complicate somewhat completely automatic playback of such records. Thus this form of record is primarily adaptable for manual interpretation.

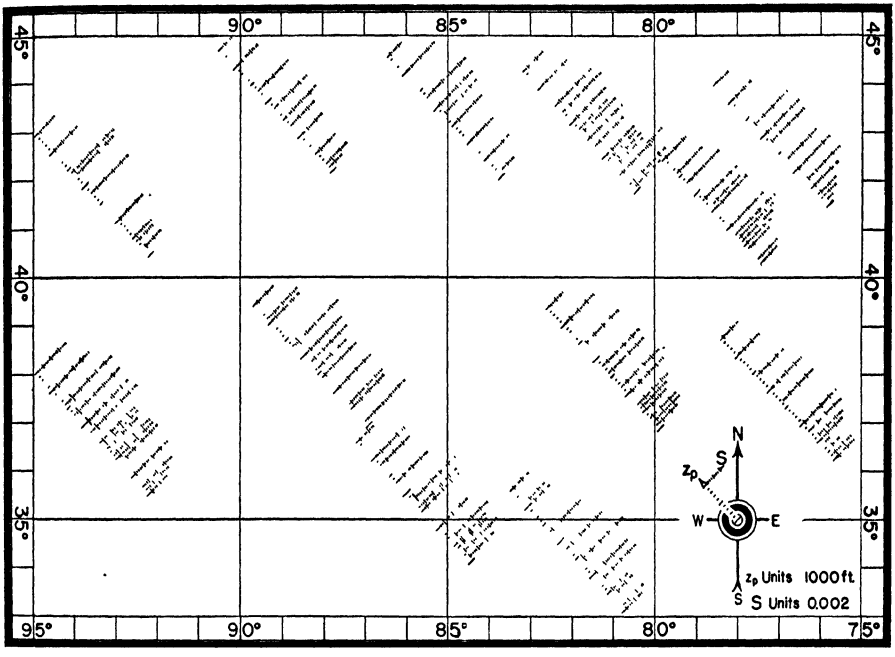


FIG. 5. Isometric geographic positional records.

4.6. Representation of Singular Values

Singular values, such as the identification of the observing station, etc., often can be represented by scaled unitary notations. For example, the singular values represented in Fig. 4a provide more convenient visual interpretation than most other numerical notations. It is also capable of automatic analogue or digital playback.

Unitary decimal digits (Figs. 2a6 and 2a9) are capable of convenient automatic playback, in contrast to Arabic numerals, and require less recording space (See Section 4.7) than the other forms of decimal digit notations illustrated in Fig. 2a. Since they can be recorded conveniently either manually or automatically, and since, with practice, they will probably be at least as convenient as Arabic numerals for direct visual interpretation, these unitary decimal digits should find wide application for the recording of singular data.

The positional digital representations such as illustrated in Figs. 2a4,

2a9, 2a10 and 2a11 are very convenient for rapid selection operations such as described in Section 5.8. The appearance of such a notation for auxiliary data on a strip chart form of record is illustrated in Fig. 3a. This illustration is essentially a representation of a portion of an IBM card in which a $C_1^3 + C_1^9$ notation has been used in the first three columns for alphabetical characters, and a C_1^{10} notation has been used in the rest of the columns for decimal digits.

4.7. Space Requirements of Unitary Records

As mentioned in Section 2, experience [2] has shown that the minimum practical spot size for records such as represented in Figs. 2a3, 2a4, 2a9, 2a10, 2a11, and 2a12 are of the order of 0.010 by 0.020 inch on either photographic film or magnetic tape. On the other hand the minimum practical wavelength on these recording mediums for analogue types of data is of the order of 0.001 inch. This discrepancy is attributable to the synchronization requirements of the digital types of records in which the existence or non-existence of a mark or hole in a predetermined position must be determined. Since it is not necessary to position each of the unitary marks with respect to such predetermined positions, the space required between successive unitary marks is similar to the wave length resolution of analogue records, or 0.001 inch. This fact has been demonstrated experimentally recently at the Cook Research Laboratories.

This same factor of approximately 10 to 1 in required relative mark separations is also apparent from the fact that perforated tapes and punched cards use from 0.1 to 0.25 inch per spot, whereas easily readable unitary resolutions of graphs are of the order of 0.010 to 0.020 inches. It is also apparent in the illustrations of Fig. 2a where the unitary decimal digits, of ten unitary spaces each, occupy approximately the same space as one punched hole digital position or Arabic numeral.

Thus, for records of the form illustrated in Fig. 3, approximately 100 unitary marks can be placed in the space required for one positional digit. Hence, even with respect to the most efficient (Fig. 2a2) positional digital code, direct scaled unitary notations are more saving of recording space for all numbers up to approximately 1000, which requires the use of 10 binary digits. Thus direct unitary digital records are the most efficient form of notation (of those discussed here, at least) for recording most continuous or successive discrete values since the accuracy of such observations seldom exceeds 0.1% of full scale.

For singular values, however, the same registration requirements exist for adjacent entries of either unitary or positional digital notations. Hence in such cases only about 10 unitary marks can be placed in the area of a single digital position. Binary notations then become more saving of space for numbers greater than about 64 (6 binary digits). If,

however, unitary decimal digits are used, each such digit requires an area approximately double that of a positional digit (this provides registration at the ends of the unitary digit). Thus a space saving of approximately 3.3 to 2 with respect to binary digits can be effected, regardless of the full scale of the number involved. In addition, the use of unitary decimal digits provides for convenient automatic playback of records which require a space but little larger than that required by Arabic decimal numerals.

Scaled unitary records usually require less recording space than more common types of graphical records. As illustrated in Fig. 5, the fact that only the required range for a particular observation need be used permits closer spacings of adjacent unitary records. The elimination of the registry requirement between the recording means and the preprinted under-grid of graphical records permits the use of smaller unitary spacings. In addition, since only the last few (at most 4) of the closely spaced unitary marks need be counted manually for accurate values from scaled unitary records, the use of much closer unitary spacings are practical than with auxiliary scales or graphical under-grids. The elimination of the closely spaced lines beyond the plotted, or end, point is largely responsible for this effect.

5. Automatic Processing of Unitary Records

5.1. Analogue Recording

Almost any photographic recorder can be used, with slight modification, for the production of unitary strip chart records of the form illustrated in Fig. 3. For example, an optically recording galvanometer can be modified by replacing the recording point of light with a transverse line of light, the full scale end of which is adjusted to coincide with the position of the usual recording point. If then a grating of opaque and transparent lines corresponding to the unitary digital modulations and opaque to the left of the zero position is placed in contact with the recording surface of the film, records such as illustrated as the positive portions of Figs. 3d and 3e are produced by the motion of the film and galvanometer mirror. Discreet records such as illustrated in Figs. 3a, 4 or 5 can be recorded with this equipment by flashing the light source at appropriate times. If desired, the lamp-slit-mirror light system can be replaced by a cathode ray tube in which the amplitude of a high frequency transverse sweep of the recording spot is controlled to be proportional to the value to be recorded.

5.2. Digital Recording

The recording of numerical values stored in counters at the output of digital computers, digital communication links, etc. can be recorded in

any form of fixed scale unitary record by sweeping a contact unitary recording screen such as mentioned above with a spot of light using either a rotating mirror or cathode ray tube. The number of unitary marks crossed, and hence recorded, are observed with a photoelectric tube, and counted in an electronic counter. When this count reaches the value to be recorded, the sweep is stopped, the light source is switched off, or the light source is deflected off the recording position.

5.3. Tabular Recording

A somewhat more versatile recording technique utilizes unitary recording tables as illustrated in Fig. 6. When used for analogue recording with optical galvanometers, a transverse recording slit is placed next to the recording film. The image of this slit is then focused and deflected with the galvanometer mirror at the desired recording position on the recording table, which is uniformly illuminated. Any desired form of unitary record can then be produced by appropriate construction of the unitary recording tables, two forms of which are shown in Fig. 6.

These same unitary recording tables are readily adaptable to the recording of digital values. In this process the reference, x, scale is continuously illuminated, and the image of the recording slit, or recording position, is swept across the recording table with a rotating mirror. The counts generated by this sweep on the reference, x, scale are observed with a photocell, and counted in electronic counters. When the desired count is reached the desired unitary number is recorded by illuminating the recording, y, scale with a very short duration flash of light.

Similar recording tables can be used for producing the desired records mechanically using a helix-tapper-bar-typewriter ribbon mechanism, or electrochemically with a helix and bar electrode. In either case, the unitary recording tables of Fig. 6 represent the development of the cylinders on which the helices are placed, and the recording positions represent the instantaneous relative position of the tapper bar or bar electrode. The helices can then be continuously positioned by an analogue drive for analogue inputs, or continuously rotated for digital inputs, in which case the tapper bar or recording bar is triggered to produce the record when the desired recording position is reached, as determined by counting the passage of the reference, x, scale marks.

5.4. Analogue Playback

Voltage analogue playback of the records is accomplished (illustrated in Fig. 3) by viewing the records with a photocell focused through a slit on the desired transverse line of the record in the same fashion as with variable area sound on film records. The amount of light received by the

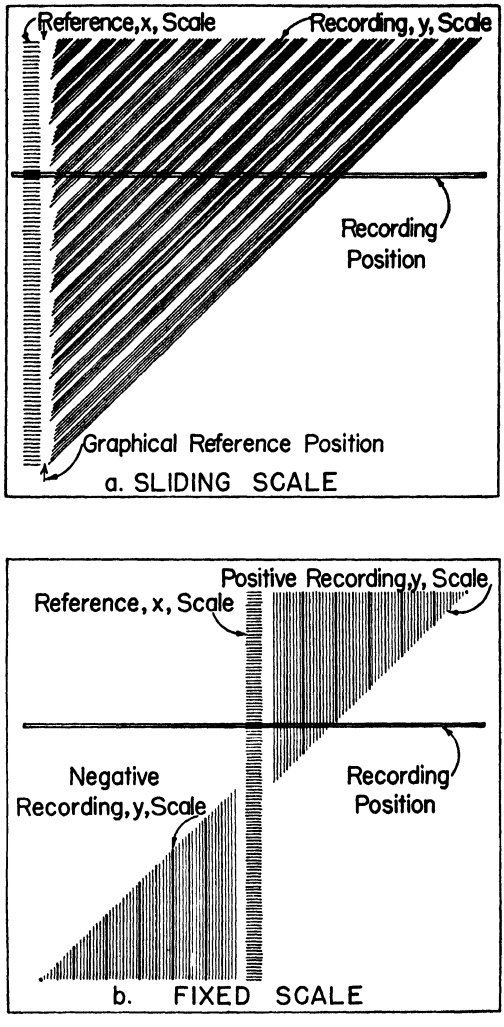


FIG. 6. Unitary recording tables.

photocell is then dependent upon the number of unitary marks that have been recorded in that position. The most convenient and accurate conditions for such playback are obtained with a record which is opaque except for transparent unitary marks with either space modulation or no modulation, in which case the amount of transmitted light at any transverse position is directly proportional to the number of unitary marks. The voltage analogue playback can, however, be obtained with sufficient accuracy for many purposes by using reflected light from opaque records even of the intensity modulated form.

More accurate, shaft position, analogue playback can be obtained from these records with automatic, photocell controlled, curve followers. This is relatively convenient, since it is, in general, easier to detect and follow the edge of a broad lighted area than the single data line characteristic of more common graphical records. The complication introduced by the undergrid lines of common graphs has also been eliminated.

The most accurate form of analogue playback, which might be required for some shaft position types of computers and could be used on records such as shown in Figs. 4 and 5, actually consists of digital playback as described in the next section. The digital counts are either converted into an electrical analogue signal or used directly as the input signals for the control of the input shaft position. This positioning procedure eliminates inaccuracies due to shifting registry on either recording or playback or due to any change of dimensions of the recording medium or recording or playback apparatus.

5.5. Digital Playback

Digital playback of any of the forms of unitary records is accomplished by scanning the desired recorded position with a spot of light and counting, with a photocell and electronic counter of the desired digital base, the number of unitary marks in that position. The scanning light spot can be conveniently generated with either a rotating mirror or a cathode ray tube with essentially the same techniques as used in television or facsimile.

Multiple records such as illustrated in Fig. 4 are conveniently played back in digital form by using a short, very narrow line of light for the scanning "spot." Since the output of light upon crossing a very narrow unitary mark is very sensitive to the relative orientation of the scanning line and the unitary mark, the various overlapping, but differently oriented, unitary marks are easily distinguished and counted separately.

5.6. Tabular Computations

Unitary types of records are very convenient for many types of numerical calculations in which extreme accuracies (greater than 0.1% of full scale) are not required. For such calculations the desired relationships are expressed in tables (such as a logarithm table, a sine table, etc.) using unitary numerical notations. Since, as illustrated in the following examples, a very small space is required for tables of this type, and since such tables are easily constructed for a great variety of functions, this technique offers great promise for very rapid processing of observational data.

Calculations of this type for functions of a single variable can con-

veniently be performed directly in the recording process with the unitary recording tables illustrated in Fig. 6. In this process any desired function, $y = y(x)$, can be represented either by appropriate construction of the recording, y, scale, or by alterations of both scales from the linear forms shown in Fig. 6. In addition to performing desired arithmetic computations, these tables are very useful for automatically including calibrations of analogue recorders or sensing elements directly in the recording process.

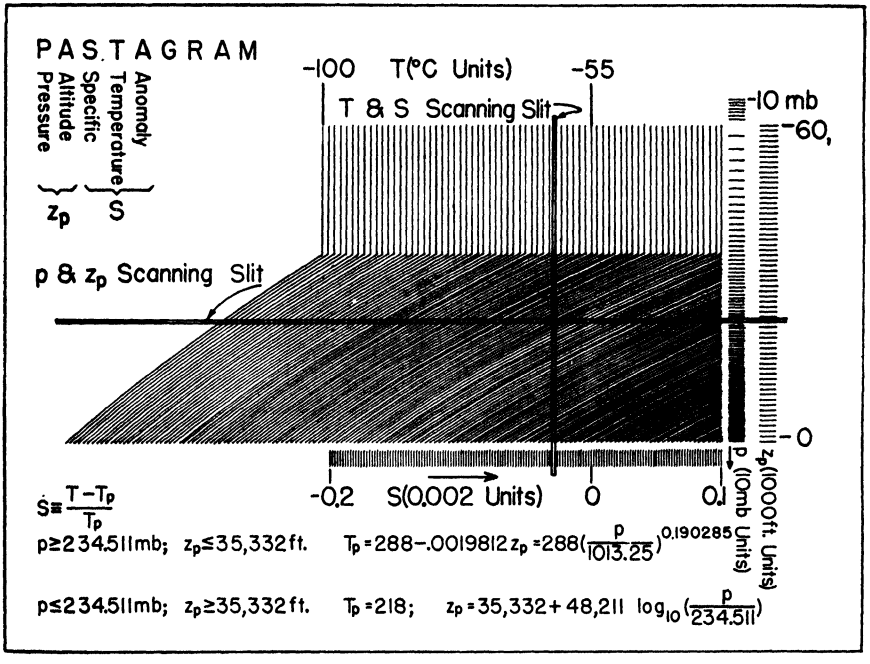


FIG. 7. Double entry unitary table.

An extension of this concept to the calculation of functions of two variables is illustrated in Fig. 7, which represents a table for determining the parameters S and z_p from the parameters T and p which are now commonly used for describing radiosonde observations (see Section 5.7). The unitary table has been constructed as a graph of the parameters p and T with respect to linear orthogonal coordinates of z_p and S. This graph differs from common graphs only in that the basic coordinate grid lines (z_p and S) are not continued through the body of the graph, and that a line is entered for each of the units of desired resolution or accuracy for each of the parameters.

Assuming that the input values of p and T have been stored in

electronic counters, the calculation begins by causing the p and z_p scanning slit to move across the graph. Photocells and electronic counters are used to keep track of the position of this slit by observing the counts generated by the p and z_p scales. When this slit reaches the desired count of p, the count stored in the z_p scale counter is the desired output value of z_p. At this time the T and S scanning slit is caused to sweep across the graph, with photocells and counters reading the S values from the S scale and the T values from that portion of the graph exposed by the p and z_p scanning slit. The desired output value of S is then retained in the S counter when the T counter reaches the input value of T.

Since the scans can be accomplished with rotating mirrors at extremely high rates, and since very intense illumination of the table can be used, the speed of such operations is limited by the maximum reliable counting rate (presently of the order of 1,000,000 counts per second of electronic counters). Thus, assuming that full scales of 1000 are used, the procedure outlined above would in general require one second to be completed, since in the main the time between successive p counts could not be shorter than the time required for a full S sweep. If, however, the unitary marks under the p and z_p scanning slit and on the S scale were to be transferred to a phosphorescent memory screen (by an intense short flash of light at the time of reaching the desired p count), the p count operation could proceed at maximum speed. With this technique, then, in the order of 2 milliseconds would be required for the computation or conversion of parameters. This speed of calculation, coupled with the ease of producing the required table for almost any desired functional relationship, makes this technique outstanding with respect to other known automatic computation techniques.

5.7. Choice of Parameters [7]

At the outset it sometimes appears as though the characteristic that direct unitary records conserve recording space only if the observations require not more than 1000 units for full scale limits their usefulness for many geophysical problems. As an outstanding example of such a situation is the common practice of expressing the pressure height relationship calculated from radiosonde observations by giving the height, z, in say, feet above mean sea level, of the various pressure surfaces described in terms of the pressure parameter, p, expressed in millibars. Since relative accuracies of the order of 10 feet, from heights of zero to the order of 100,000 feet, are required, it appears that a full scale of 10,000 ten-foot units are required, and that direct unitary representations are not applicable

It should be noticed, however, that the use of z and p are not at all

applicable or convenient for the description of this relationship. In the first place, the use of such parameters implies that it is necessary to have the capability of describing a situation in which the height of perhaps the 10 mb pressure surface is at sea level, which of course is absurd when naturally occurring conditions are being described. In addition, the direct use of such parameters practically precludes the possibility of drawing graphs of the relationship in order to obtain direct visual interpretation. This results from two effects; first, the graphs are necessarily very large; and second, on the very large graphs the relatively very small variations which occur, and which are the only conditions of importance, are not readily discernible.

This last effect indicates a procedure which not only makes it possible to provide representations which are conveniently interpreted visually, but also provides for savings of recording space, for savings of calculation time and complexity, and for savings of communication time or band width. Since the height of a given pressure surface never varies very much from some usual or average value, it is apparent that all of the above savings can be accomplished if the height of any given pressure surface is described with respect to some standard height which is approximately the usual height of that pressure surface. In other words, the height of pressure surfaces are described in terms of some quantity, D, defined [7] by the relationship

$$D = z - z_p \tag{1}$$

where z_p is an arbitrary reference height for any given pressure surface.

Meteorological experience indicates that the height of any given pressure surface, hence the range of D, naturally varies over a maximum range of about 5000 feet. Hence, if units of 10 feet of D are used, a full scale of but 500 units are required and the use of unitary representations provides a saving of recording space. In addition, of course, relatively small graphs of continuous variations, and hence convenient manual interpretations are now possible.

The use of the parameter D, in conjunction with the pressure parameter, p, is not, however, convenient, since complete manual interpretation of observations expressed in this form requires that the value of z_p for any given value of p be memorized. This is obviously impossible if a continuous representation along a vertical line of the pressure-height relationship is desired. But it is equally obvious that this difficulty is eliminated by using the parameter z_p itself, rather than p, for describing the value of the pressure. That this procedure is convenient and practicable is demonstrated by the fact that it is already in common use in terms of the pressure altitude, z_p, used for the calibration and interpreta-

tion of pressure measurements made with aircraft pressure altimeters. This procedure is interpreted physically in terms of the concept of using a standard column of air, rather than a standard column of mercury, as a barometer with which to describe the pressure in terms of heights, of or in, the barometer at which given pressures occur.

In order to maintain convenient and complete personal understanding of related parameters, the process of calculating derived parameters from observed parameters (such as the height of pressure surfaces from pressure and temperature observations) should be made as simple and straightforward as possible. This means, for example, that the choice of the standard heights, z_p, of given pressure surfaces should be defined in terms of a simple mathematical definition of a corresponding standard temperature-pressure relationship, such as given for T_p (z_p) in Fig. 7, in which T_p is the temperature corresponding to a given pressure (or pressure altitude) in the standard atmosphere. The desired simple and straightforward relationship between observed and derived parameters is then obtained by choosing the parameter with which to describe temperature measurements to be the specific temperature anomaly, S. With this choice the hydrostatic equation, with which practically all heights of pressure surfaces is derived, takes on the particularly simple form [7]

$$\frac{dD}{dz_p} = S \tag{2}$$

The ease of calculation (and hence also direct fundamental personal understanding) with a linear equation of this form with respect to the usual logarithmic form such as

$$\frac{dz}{d \ln p} = \frac{RT}{g} \tag{3}$$

in which accuracies of the order of 0.01% of z, instead of 0.5% of D, are required, is obvious.

The use of anomalies, such as S, as the parameter for describing upper air temperature observations, usually provides the same sort of reduction of required percentage accuracies as does the use of D with respect to z, with the attendant saving of recording space, calculation time and complexity, etc. In addition, the use of anomalies rather than "absolute" values usually provides more convenient direct interpretation of graphs, especially of the form illustrated in Fig. 5. This results primarily from the fact that a convenient reference line for comparison of various observations is automatically provided by the independent parameter scale (z_p scale in Fig. 5) which is essentially a plot of the standard relationship from which the anomalies are determined. Since the standard

relationship, or standard atmosphere in the example of Fig. 5, is chosen to represent approximately average conditions, both hot or cold conditions and stable or unstable conditions (vertical temperature gradients) are depicted to best advantage.

5.8. Selection Operations

Although, as indicated in Sections 4.6 and 4.7, the positional digital notations such as illustrated in Figs. 2a4, 2a9, 2a10, and 2a11, usually require more recording space and are less easily interpreted visually than Arabic numerals or unitary decimal digits, they are more convenient for some types of automatic processing. The process of automatic selection of records with respect to particular desired values, usually of singular identifying values, is an outstanding example of this characteristic. For example, if it were desired to select all punched cards for which the values in two given columns were as shown in Fig. 2a9, a complementary matching card would be punched with a hole in every position of the two columns except the desired positions corresponding to 2 and 6. If then a light source were placed behind the cards, and each data card were passed in turn over the complementary matching card, the desired match, or selection, would be indicated when, and only when, no light passed through the two cards at the instant they were in alignment. Obviously, this same condition obtains with this procedure for any desired number of columns or entries, and but one light source and photocell are required. The only limitation on the number of entries which can be simultaneously examined in this way is determined by the signal-to-noise ratio available in the photocell circuit and provided by the opacity of the record.

Since predetermined registration is required for this process, it is obvious that unitary decimal digital notations, or Arabic numeral notations, are not suitable for it. Similarly, it is apparent that the notations of Figs. 3a3 or 3a12 are not capable of this type of operation. A little study of the characteristics of this technique indicates that a positional notation of the form $C'_m{}^n$ is adaptable to its utilization. It is thus seen that the minimum number of code positions for decimal digits of this form is 5, obtained with $C_2{}^5$ as illustrated in Fig. 2a11. Similarly, if alphabetical characters are to be represented in this fashion, the minimum number of code positions is 7, with a code of the form $C_3{}^7$. This code, with its 35 possible combinations, could be used for both alphabetical characters and decimal digits, if, say, the letter O and zero were given the same combination of marks.

The convenience and simplicity of this particular automatic selection technique with respect to those in which multiple reading positions and devices, and usually auxiliary memory and matching circuits or mecha-

nisms, are required is obvious. In addition, since in this type of coding a fixed number (n) of marks, or active signals, are used, relatively simple self-checking circuits can be used in computers which eliminate many possibilities of error in automatic digital calculations [1, 2].

6. Summary

It is apparent from the preceding discussion that development of observing, communication, and computing equipment in recent years has produced a capability of very rapid and versatile automatic measurement and processing of geophysical data. This capability has not been realized, however, since many transcriptions from one form of record to another, either manually or at essentially manual rates, are now required. The greatest shortcoming is the inability of the automatic equipment to produce output records which are capable of convenient and efficient mental assimilation and interpretation.

It is also apparent that the development and use of records using unitary numerical notations can eliminate most of the slow and inefficient steps in processing observational data, as well as provide superior records for manual interpretation and assimilation. This capability is summarized in the following list of properties of unitary digital records.

1. In general, they require a minimum of recording space.
2. They are convenient forms of input and output records for any type of digital computer.
3. They are convenient forms of input and output records for all kinds of analogue computers.
4. They are convenient forms of input and output records for any kind of automatic communication.
5. They provide a capability of very rapid and convenient tabular computation.
6. They can be conveniently read and interpreted manually in graphical, isometric, and digital forms.

These characteristics indicate that unitary digital notations should be used for practically all records of observational data, with the possible exception of the $C_n{}^m$ types of positional digital notations for automatic selection processes and perhaps some Arabic numeral and alphabetical character notations for singular record identifications.

List of Symbols

$^b a_i$ the value of the ith digit with respect to the base b
$C_n{}^m$ combination of m things taken n at a time
D altimeter correction ($D = z - z_p$)
g acceleration of gravity

N a general integral number
p pressure (Expressed in force per unit area)
R gas constant for air
S specific temperature anomaly (or south)
t independent parameter (time)
T temperature
T_p standard temperature at a given pressure
u, v W-E and S-N wind components
x parameter dependent on t or independent of y
y parameter dependent on x
z height above mean sea level
z_p pressure altitude (Pressure expressed as feet of a standard air column)
E east
W west
Σ Summation

References

1. Berkely, E. C. (1949). Giant Brains, or Machines That Think. Wiley, New York, 270 pp.
2. Aiken, H. H., and others. (1948). Proceedings of a Symposium on Large Scale Digital Calculating Machinery. Harvard Univ. Press, Cambridge, Mass., 302 pp.
3. Frost, S. (1948). Compact analogue computer. *Electronics* **21,** No. 7, 116–122.
4. Ragazzini, J. R , Randall, R. H., and Russell, F. A. (1947). Analysis of problems in dynamics by electronic circuits *Proc. Inst. Radio Engrs.* **35,** No. 5, 444-452.
5. Svoboda, A. (1948). Computing Mechanisms and Linkages. Radiation Laboratory Series. Vol. 27, McGraw-Hill, New York. 359 pp.
6. Cook Research Laboratories, Chicago. (1949). Graphical Representations of U. S. Weather Observations. Red Bank, N J., Watson Laboratories, Contract No. W28-099ac394, 114 pp.
7. Bellamy, J. C. (1945). The use of pressure altitude and altimeter settings in meteorology. *J. Meteorol.* **2,** No. 1, pp. 1 79.

Some New Statistical Techniques in Geophysics

ARNOLD COURT

Statistical Laboratory, University of California, Berkeley, California

CONTENTS

	Page
1. Introduction	45
1.1. General	45
1.2. Methods	46
1.3. Functions	48
1.4. Graphics	49
1.5. Components	50
1.6. Separation	52
2. Extremes	53
2.1. Intervals	53
2.2. Frequency	55
2.3. Probability	56
2.4. Risks	58
2.5. Theory	60
2.6. Description	61
2.7. Parameters	64
2.8. Computations	67
2.9. Evaluations	69
2.10. Applications	71
2.11. Conclusion	74
3. Circular Distributions	75
3.1. Requirement	75
3.2. Description	76
3.3. Procedure	78
3.4. Graphing	79
3.5. Limitations	81
List of Symbols and Notation	82
References	83

1. INTRODUCTION

1.1. General

Statistical theories and methods are being applied increasingly in all fields of science, especially in geophysics. Until the 1930s, the physical sciences generally used only the rudimentary methods of statistics, preferring, for example, the Gaussian probable error to the analytically stronger and more versatile standard error (or deviation). Statistical

theorems and methods developed in the preceding half-century were employed much more in the biological and social sciences than in the physical.

In the last few decades, however, the physical sciences have adopted a more modern statistical outlook. Geophysics in particular has made rapid strides in adopting statistical practices, and many techniques have been developed for the special requirements of its various component sciences. Some of these techniques are described in detail in this article, in order to acquaint a large circle of geophysicists with their potentialities.

A preliminary discussion of some fundamental aspects of statistics which often are overlooked in geophysical applications, and an explanation of a rediscovered simple method of estimating two normal components from a bimodal distribution are given in this section. The article is largely devoted, in Section 2, to a discussion of the likelihood of occurrence (return period) of extreme values, and the most recent method for estimating them, the theory of extreme values. The final section mentions briefly an even newer development, the statistics of circular variables, still in the descriptive stage.

Applications, interpretation, and limitations of the techniques, rather than underlying theory and proof, are stressed. The only statistical knowledge presumed of the reader is that of a first course in statistics: least squares computations, characteristics of the normal distribution, and simple correlation.

Symbols and notation used in this article are listed at the end of this article. Most of the symbols are those used in the various original papers, but some of the notation is novel, since statistics has developed so rapidly that its notation and symbolism have not yet been fully standardized. In recent years, the overbar ($\bar{x}$) has been accepted to designate the mean; in this article, in addition, the tilde ($\tilde{x}$) denotes the median and the circumflex ($\hat{x}$) the mode. Grave and acute accents ($\grave{x}$ and $\acute{x}$) indicate the largest and smallest values, respectively.

The classical statistical methods of geophysics have been presented recently in great detail by Conrad and Pollak [1]. Some more modern statistical concepts, however, are not included there, and may be overlooked by geophysicists. In the following paragraphs certain aspects are discussed which may make more accurate the application of statistics to geophysics.

1.2. Methods

Statistical methods are of two general categories: *descriptive* and *analytical.* Both depend in large part on the theory of probability which, in the words of Laplace, is merely common sense reduced to figures.

Descriptive methods are those which compress many figures into a few to represent them adequately for the purpose at hand; these methods are largely those formerly known as the *calculus of observations*. They involve few assumptions about the nature of the original figures, and consider the figures as such, and not as samples. Descriptive methods permit computation of means, modes, medians, and of variances and higher moments, as well as of correlations between two or more variables.

Analytical methods use the descriptive techniques to determine how well the observations agree with the theoretical model which they are assumed to follow. From the character of the model, in turn, and the descriptive results, the analytical procedures can indicate the accuracy of generalizations from the data, and of comparisons with other observations.

Emphasis, in most elementary courses in statistics, on the analytical aspects has obscured, for many geophysicists, both the limitations and the utility of the purely descriptive methods of the calculus of observations. Whereas description takes the data as they are, analysis considers them only as a *sample* of a population or universe. This parent population, in turn, is *assumed* to have certain characteristics, whose numerical values are estimated from the description of the sample.

Establishing that the sample does in fact have the attributes of the parent population is therefore essential to any analysis, yet in many cases this correspondence is not established at all. For example, the standard deviation can be computed for any set of figures as a valid measure of the amount of dispersion, but only if the figures are shown to follow a "normal" distribution can it be assumed that two-thirds of them fall within one standard deviation from the mean.

Descriptive methods alone may suffice for many geophysical applications—more so than in the biological and social sciences—where a mass of data is to be reduced to a few characteristic figures (means, modes, variances), without any inferences about the parent population or any detailed comparisons with other sets of observations. But statistical *analysis* of geophysical data must start with a clear expression of the population of which the data are considered to be a sample, and establishment that the sample is indeed drawn from such a population.

For many sets of geophysical data, "It is clear that one cannot define a population out of which the given sample was drawn at random." [2] Most geophysical data concern measurements of a variable which is continuous in both time and space, and may be relatively uniform over certain ranges of one or both. A single reading of air temperature, or magnetic intensity, or sea-swell length, may be considered as a sample of conditions at the spot of observation during a short interval of time,

such as a few minutes; or it may be a sample of conditions over a small area, a few inches to a few miles in radius.

That is, an instantaneous reading is "chosen at random" from all possible similar readings which could have been made at any of an infinite number of other times during the interval, or at an infinite number of places in the vicinity. But when the element is averaged in time it is no longer a sample with respect to time: the parent population of a series of mean daily values (temperature, magnetic activity, or sea-swell length) is composed of all possible values for the vicinity, each averaged in time.

Furthermore, while the individual reading or mean daily value may be a *random* sample from an infinite population, a series of such readings is not a random sample, but a *stratified* sample: one from each of several distinguishable divisions (e.g., days) or strata of the population. Consequently, many analytical procedures, particularly tests of significance, are not strictly applicable to such data.

1.3. Functions

The extensive computations required for statistical description or analysis are laborious if done by hand, but can be done rapidly on modern computing machines. Recent improvements in such machines, in fact, have permitted great simplifications in the routine computations, in that involved calculations can be done more rapidly than simpler calculations which require additional manipulation. Unfortunately, these advances are rarely reflected in elementary textbooks, which describe methods applicable to manual computation, perhaps aided by an adding machine.

For example, combination of observations into classes is desirable when a large mass of data is to be summarized manually, but imposes some loss in accuracy as the price for convenience. With modern machines, individual observations can be squared and the results added in less time than is required to select class limits, assemble data into classes, and perform the computations. Consequently, the classic rules as to the number and size of classes no longer are very important.

Quantitative data or measurements, however, already are grouped by classes, defined by the unit of measurement even though the variable measured is itself continuous. Any further grouping usually is inadvisable.

Likewise, although the standard deviation is defined basically as the square root of the mean of the squares of all *deviations* from the mean (root mean square), in practice it is obtained most readily by the "variable squared" method: the square root of the difference between the mean of the squares of all the original observations and the square of the mean of the observations. Individual departures from the mean need not be computed at all.

Statistical analysis involves the comparison of observed data with a theoretical model, expressed mathematically in either of two ways, one the integral of the other.

A *frequency distribution* represents a set of data, observed or theoretical, of finite size; when all frequencies are reduced to percentages of the total sample size, the result is a *probability distribution.* In either form, this function, denoted by $f(x)$, represents the *density* of the distribution of frequency or probability, and when plotted on cartesian paper it yields a characteristic curve—"bell-shaped" for the normal curve.

The *area* under such a curve represents cumulative frequencies or probabilities; consequently, the *cumulative probability* function is the integral of the probability density distribution or function: $F(x) = \int_{-\infty}^{x} f(t)dt$. The graphing of such an integral, if computed from one end of the distribution to the other, yields an *ogive,* or cumulative frequency or probability graph; in hydrology, a time-frequency ogive has been called a "duration curve." On cartesian paper the cumulative probability ogive of a normal distribution is S-shaped or "sigmoid"; special "probability paper" (Section 1.4) transforms this curve into a straight line.

Each form of frequency function, the density distribution and its integral, has separate uses. In general, the density distribution is used to graph the theoretical function for comparison with a graph of observed values, while its integral, the cumulative probability function, is used for numerical comparison of the agreement between theory and observations, and for discussion and conclusions after correspondence is established. While a density distribution curve can be approximated from area values, and theoretical ordinates can be compared numerically with observed frequencies, such procedures are not as correct as the proper use of the two functions.

1.4. Graphics

For any cumulative probability function, whose ogive plotted on cartesian paper is a sinuous curve, a special graph paper can be designed on which the ogive becomes a straight line. Such papers were first designed by engineers, and they are used chiefly in that field. "Though mathematicians look with disfavor on the use of graphical methods in the evaluation of statistical parameters, engineers find them very convenient and time saving, especially if the accuracy required is not too great." [3]

Graph paper for the normal probability paper was designed and introduced in 1914 by Hazen [4] without comment, and explained in 1916 by his coworker Whipple [5], who also presented a logarithmic normal paper previously suggested by Hazen; revision of this paper has recently been

proposed by Kottler [6]. As soon as the statistical theory of extreme values (Sections 2.5 *et seq.*) was introduced into the United States, Powell [7] designed a probability paper for its function (Section 2.8). Other probability papers include the "Probit" and "Logit" graphs of Berkson and Gumbel's new "Range" paper [8].

The chief virtue of any probability paper is that a set of data which plots along a straight line on it can be assumed to be drawn from a population whose distribution is that on which the paper is based. A further advantage is that such a straight line, whether drawn by inspection or fitted mathematically, can be used to obtain estimates of other values, such as the expected frequency of a given value or the value with a given probability of occurrence.

However, probability paper cannot be used alone to determine how well data follow the assumed distribution, e.g., to test for "normality," because a straight line cannot be fitted to plotted points by inspection: the paper is not linear, and slight departures from a straight line are magnified at both ends. "A Log-Probability Chart should be used only to represent an exact ogive by a straight line but not to judge how the data fit it. It is impossible to achieve any reliable judgment by mere inspection of such a graph." [6]

To plot a set of values on any probability paper they must be arranged in order of magnitude and their cumulative rank established. The smallest value is No. 1, the next-smallest No. 2, etc.; if the smallest occurs twice, it has Nos. 1 and 2, and the next-smallest is No. 3, etc. Alternatively, the largest value may be No. 1.

However, there has been little agreement on how to plot these cumulative ranks on probability paper. If the ranks are divided by the number of observations, N, then the last one is unity, which is at infinity on the graph paper. Compromises have been suggested, by which either $\frac{1}{2}$ or 1 is subtracted from the rank before division by N, or division by $2N$; these either omit an observation or distort the original data [9].

Certain theoretical considerations indicate advantages in dividing each cumulative rank by $N + 1$ for plotting; in addition to providing more realistic frequency values, this method permits all observations to be plotted on graph paper. This procedure is used in analysis by the theory of extreme values (Section 2.8).

1.5. Components

Typical of the subordination of descriptive methods to analytical procedures is the neglect of a very simple and useful technique for estimating two normal components in any frequency distribution. Many measurements, in geophysics as well as other sciences, involve varia-

bles which are not uniform but include subvariables of different basic characteristics.

For example, Landsberg [10] has shown that observed thermal gradients in the earth's crust fall into two groups, possibly for sedimentary and metamorphic rocks, respectively. Similarly, in middle latitudes the tropopause may be either high and cold (tropical) or low and not so cold (polar), so that a frequency distribution of daily tropopause height determinations has two definite modes.

A general method for finding two normal components in any distribution, assuming nothing about them except their existence, was presented by Pearson [11] in the first of his famous "Contributions to the Mathematical Theory of Evolution" before the Royal Society on November 16, 1893. It requires solution of a complete ninth degree equation involving the first five moments of the given distribution.

Pearson applied this method not only to markedly skewed distributions, in which the presence of two components is indicated strongly, but to some which are quite symmetrical (although not normal) to find components with identical means but differing standard deviations. His general method applies even when one of the components is negative, i.e., the given distribution is the difference between normal ones.

To Edgeworth's [12] suggestion for simplifying assumptions, Pearson [13] retorted that the "process is not so laborious that it need be discarded for rough methods of approximation based upon dropping the fundamental nonic and guessing suitable solutions." However, Charlier [14] considered the general solution "a very laborious operation," and developed simple solutions for two special cases: (1) where means are assumed for the two components and (2) where the variances of the two components are assumed to be equal.

Charlier's development, published in English in a journal of the University of Lund (Sweden), attracted little attention, and no mention of it appears in his later textbook nor does it seem to have been used by anyone else. Of the two methods, the first, involving assumption of the means of the two components, is far simpler than the second, which requires computation of the fourth moment of the given distribution and solution of a cubic equation.

However, Charlier devoted little space to the first method and expanded on the second, terming it the "abridged method for dissecting frequency curves." Since the cubic equation involved is actually one step in the general solution, "hence it is no loss of time to begin with this approximate method." He felt that assuming equal variances for two components "is of a more general character" than assuming values for their means: "Especially in biology it is a fairly probable supposition

that two types found together in nature possess *nearly* equal standard deviations. We may then use this method to separate the two components."

He admitted that "this abridged method is applicable only when there are *a priori* reasons for the assumption that the two components have nearly equal standard deviations. There are many problems where no such reasons exist," such as those involving several sets of errors to a reading, each set being of a different type and magnitude.

In geophysics, equal variances may be present in some cases, but in general the first method, of assumed means, is the most applicable. Both methods, and one further simplification, are presented in the next paragraph, without the theoretical basis or development and in more condensed and modern notation [15].

1.6. Separation

In obviously bimodal distributions, and many unimodal ones with pronounced "humps" or "shelves," means M_1 and M_2 for two supposed components may be apparent. Their departures from the mean M of the given distribution,

$$M - M_1 = m_1 \qquad \text{and} \qquad M_2 - M = m_2 \tag{1.1}$$

give the variances of the two components:

$$\begin{cases} \sigma_1^2 = \sigma^2 - 2m_1m_2/3 - (m_1^2/3 + \nu_3/3m_2) \\ \sigma_2^2 = \sigma^2 - 2m_1m_2/3 - (m_2^2/3 - \nu_3/3m_1) \end{cases} \tag{1.2}$$

where σ^2 and ν_3 are the variance and third moment of the given distribution. The total areas or frequencies of each component depend only on the assumed means:

$$N_1 = Nm_2/(m_1 + m_2) \qquad \text{and} \qquad N_2 = Nm_1/(m_1 + m_2) \tag{1.3}$$

Finally, from a table of the normal frequency distribution ordinates (Section 1.5), $\phi(t)$, the ordinates of each component at any distance (in t units) from the mean may be found, since

$$y_1 = (N_1/\sigma_1)\phi(t) \qquad \text{and} \qquad y_2 = (N_2/\sigma_2)\phi(t) \tag{1.4}$$

The larger component always corresponds to the smaller departure from the mean, which in turn is m_1 if ν_3 is positive, m_2 if negative. Should impossible means be assumed for the two components, σ_1^2 or σ_2^2 will be negative, indicating no real solution.

However, the method of assumed means does not give a unique solution: usually trial of several pairs of means is required to find one set yielding two components which, added together, closely approximate the

given distribution. The best pair of means generally has maximum ordinates agreeing well with the observed values, due regard being given to the contribution each component makes to the other's peak.

Such agreement can be made as close as desired by assuming values of the maximum ordinates $\grave{y}_1$ and $\grave{y}_2$ in addition to the means M_1 and M_2. Then

$$\sigma_1 = N_1/\sqrt{2\pi}\ \grave{y}_1 \qquad \text{and} \qquad \sigma_2 = N_2/\sqrt{2\pi}\ \grave{y}_2 \tag{1.5}$$

In effect, this short cut to Charlier's procedure replaces the standard deviation and skewness of the original distribution by a subjective evaluation which may be more effective for some distributions, but is not of as general applicability in finding two normal components.

Assuming the two presumed components to have equal variances, instead of assuming values for their means, led Charlier to a cubic equation involving the difference between the variances of the given distribution and the assumed components:

$$z^3 + \tfrac{1}{2}(\nu_4 - 3\sigma^4)z + \tfrac{1}{2}\nu_3^2 = 0 \tag{1.6}$$

where $z = \sigma_1^2 - \sigma^2$ and ν_4 is the fourth moment of the distribution. The discriminant of this cubic,

$$C^2 = (\sigma^{12}/216)\ (13.5\alpha_3^4 + E^3) \tag{1.7}$$

where $\alpha_3 = \nu_3/\sigma^3$ is the skewness and $E = (\nu_4/\sigma^4) - 3$ the excess, almost always is positive, indicating only one real root:

$$z = 0.4082\sigma^2\ (\sqrt[3]{-3.6742\alpha_3^2 + \gamma} - \sqrt[3]{+3.6742\alpha_3^2 + \gamma}) \tag{1.8}$$

where $\gamma = \sqrt{13.5\alpha_3^4 + E^3}$.

Except for almost symmetrical and very flat-topped distributions, γ is positive, so that z will be negative, and $\sigma_1^2 < \sigma^2$. But if $-z > \sigma^2$, then σ_1^2 is negative, and there is no actual solution, indicating that the assumption of equal variances is unwarranted. If the assumption is justified, and σ_1 is real, the means are:

$$\begin{cases} M_1 = M - m_1 = M - (\nu_3/6) - \sqrt{(\frac{1}{4}\nu_3)^2 - z} \\ M_2 = M + m_2 = M - (\nu_3/6) + \sqrt{(\frac{1}{4}\nu_3)^2 - z} \end{cases} \tag{1.9}$$

The areas N_1 and N_2 of the two components are found from equation 1.3 as before.

2. Extremes

2.1. Intervals

Extremes of any distribution of observations are of interest because they afford a rough indication of the range of the variable: extremes which

have occurred may be expected to occur again. In geophysics, extremes are of greater importance than in many other sciences, because many questions of engineering design hinge on the most extreme value to be expected. Dams must be constructed to withstand the maximum flood anticipated in the lifetime of the structure, skyscrapers must be designed with the most severe earthquake in mind, chimneys should be able to endure the strongest wind, communications circuits should operate during the most severe magnetic and electrical disturbances, and piers must be located and constructed to withstand the heaviest anticipated surf.

In all such problems, specified calculated risks may be taken if the likelihood of occurrence of these extremes can be estimated within known limits of accuracy. The basis for such estimates of risk, and methods for their calculation, are explained first in this section. Then follows a discussion of the most recent method of estimating the most extreme value to be expected in a given period, the statistical theory of extreme values.

By definition:

> An event which happens H times in N trials has a *relative frequency* of occurrence of H/N, and an *apparent return period* of $T = N/H$.

The apparent return period, or reciprocal of the relative frequency, is therefore the *average* interval between recurrences of the event in the particular series of trials. Despite the rigor of this definition, it has not been fully appreciated, and there even have been some attempts to prove it.

Distinctions have been drawn, in hydrology, between two kinds of return periods: the "exceedance interval" and "recurrence interval," respectively the average periods between exceedances and recurrences of an event. These distinctions may be justified in dealing with discrete variables, such as number of points on a throw of two dice, but they grow meaningless for continuous variables as the unit of measurement becomes smaller. The distinction is part of the earlier empirical approach to the problems, which has been superseded by the recent advances outlined in this article.

Events for which relative frequencies and return periods are estimated are defined in one of two ways: by time or by magnitude. Events defined by time are the largest (or smallest) individual values during a given interval, such as a month, year, or solar cycle. Events defined by magnitude are those values which exceed some predetermined base, such as a temperature of 100°F or an earthquake intensity of 6.0; the time unit is usually much smaller than that used for the first type.

In particular, most hydrologic analyses use the relative frequency and apparent return periods of *annual* floods (maximum stream discharge), ignoring the second-highest floods of each year although some of them may be greater than the largest floods of other years. To rectify this apparent fault, other analyses use all floods exceeding the base value ("partial-duration series"), so that "the recurrence interval is the average interval between floods of a given size regardless of their relationship to the year or any other period of time." [16] It is less than the recurrence interval computed on the annual basis, although "for large floods the two approach numerical equality."

2.2. Frequency

If the occurrence or recurrence of an event depends on so many independent factors that it may be considered to follow the laws of chance, its relative frequency usually is assumed to be the same as the *probability of occurrence* in any one trial. This equivalence, which appears intuitively sound to the engineer, is questioned by the mathematician, and has encountered much statistical discussion.

It is the subject of an early theorem, acclaimed as one of the foundations of probability theory, proven by James Bernoulli in his *Ars conjectandi* (published posthumously in 1713):

> As the number of trials increases, the probability approaches unity that the relative frequency of occurrence will differ by less than any desired amount from the true probability of occurrence.

This theorem does not say that the relative frequency itself approaches the true probability as a limit, although Rietz [17] proposed such a statement as the basic definition of probability, from which Bernoulli's theorem would be an immediate consequence. In recent years these fundamental assumptions of probability theory have been the subject of renewed discussion [18].

In most geophysical problems, the true probability is unknown and must be inferred from the relative frequency. "Bernoulli, himself, in establishing his theory, had in mind the approximate evaluation of unknown probabilities from repeated experiments," Uspensky [19] pointed out, quoting Bernoulli as saying: "If somebody for many preceding years had observed the weather and noticed how many times it was fair or rainy, . . . by these very observations he would be able to discover the ratio of cases which in the future might favor the occurrence or failure of the same event under similar circumstances."

While the relative frequency based on very many occurrences provides a reasonable estimate of the true probability of occurrence, the

relative frequency in a few occurrences is not at all reliable. Probability estimates usually are made in terms of two limits which are expected, with some given degree of confidence, to include the true value; for a given relative frequency, the greater the degree of confidence, the wider the interval in which the true probability is estimated to be. The limits of the estimate converge sharply as the number of trials on which it is based increases; this is shown by Table I, for 95% confidence, based on a diagram by Clopper and Pearson [20], which has been reproduced widely [21]; a similar table is presented by Snedecor [22] without explanation.

TABLE I. Limits of estimate of true probability with 95% confidence from relative frequency based on samples of varying size.

	Number of Trials = Sample Size					
Rel. freq.	10	20	30	50	100	1000
.00	.00 to .31	.00 to .17	.00 to .12	.00 to .07	.00 to .04	.00 to .01
.10	.00 to .46	.01 to .32	.03 to .27	.05 to .22	.07 to .17	.08 to .12
.20	.02 to .57	.05 to .44	.07 to .39	.10 to .34	.12 to .30	.17 to .22
.30	.06 to .66	.12 to .55	.15 to .50	.18 to .45	.21 to .40	.27 to .33
.40	.11 to .75	.18 to .64	.22 to .60	.26 to .55	.30 to .50	.37 to .43
.50	.17 to .82	.27 to .73	.31 to .69	.35 to .65	.40 to .60	.47 to .53

Table I shows, for example, that a relative frequency of 0.20 based on 10 trials (2 occurrences in 10 years) may arise from true probabilities anywhere between 0.02 and 0.57. For the same relative frequency observed in 50 trials the corresponding limits are 0.10 to 0.34. Based on 1000 trials the limits are only 0.17 to 0.22. Estimates of the true probabilities based on the rather small samples used in geophysics have very wide confidence intervals—so wide as to vitiate many computations based on them.

Probably the most valuable contribution of the theory of extreme values, discussed in detail later in this section, is that it provides an estimate of the true probability of occurrence of extreme values based, not on one extreme alone, but on all the values. An estimated relative frequency or return period obtained by this method, as outlined in Section 2.9, is the closest obtainable approximation to the true probability or return period.

2.3. Probability

Return periods, observed or estimated, are used extensively in various branches of geophysics, especially in hydrology for flood analysis. Nevertheless, the significance of the return period is not well known,

although it can be developed as a corollary of the oldest problem in the theory of probability. In this problem, 300 years ago, Pascal found that while the probability of a double six on any one throw of two dice is $\frac{1}{36}$ and its return period is therefore 36 throws, there is better than a 50–50 chance of obtaining at least one double six in only 25 throws.

In general, the probability that an event x_T, whose probability of occurrence in a single trial is $p = 1 - q$ and whose return period is therefore $\bar{T} = 1/p$, will *not occur in any of N trials* is (notation as in List of Symbols, page 82):

$$(2.1) \qquad P(\grave{x}_N < x_T) = q^N = (1 - p)^N = (1 - 1/\bar{T})^N$$

Consequently, the probability of *at least one occurrence in N trials* is:

$$(2.2) \qquad P(\grave{x}_N \geq x_T) = 1 - q^N = 1 - (1 - 1/\bar{T})^N$$

In Pascal's dice problem, $p = \frac{1}{36}$, and for $P(\grave{x}_N \geq x_T) = P(\grave{x}_N < x_T) = \frac{1}{2}$, $N = \log(\frac{1}{2})/\log(\frac{35}{36}) = 24.6$.

Similarly, the probability of *occurrence for the first time on the Nth trial* is the compound probability of non-occurrence in $N - 1$ trials and of occurrence in one trial:

$$(2.3) \quad P[(\grave{x}_{N-1} < x_T)(x_N \geq x_T)] = pq^{N-1} = p^{N-1} - p^N = (\bar{T} - 1)^{N-1}/\bar{T}^N$$

This probability is greatest on the first trial, and decreases with each successive trial because the probability of occurrence on the preceding trials increases. In Pascal's dice problem, the probability of a double six for the first time on the Nth trial (equation 2.3) decreases, while that for a double six in at least one of N trials (equation 2.2) increases, as follows:

N:	1	2	3	4	5	10	15	20	25	30	36
$P(x_N \geq x_T)$:	.028	.027	.026	.026	.025	.022	.019	.016	.014	.012	.010
$P(\grave{x}_N \geq x_T)$:	.028	.055	.081	.107	.132	.246	.345	.431	.506	.471	.638

A fourth relationship, extensively used in some probability problems, but rarely of direct interest in geophysics, gives the probability of *exactly H occurrences in N trials:*

$$(2.4) \qquad P[(\grave{x}_N \geq x_T) = H] = [N!/H!(N - H)!]p^H q^{N-H}$$
$$= [N!/H!(N - H)!](\bar{T} - 1)^{N-H}/\bar{T}^N$$

The factorial terms are the binomial coefficient, usually written $\binom{N}{H}$ but formerly written as ${}_NC_H$ or C_H^N; they represent the number of combinations of N objects taken H at a time. For no occurrences, $H = 0$ and the coefficient becomes unity, so equation 2.4 reduces to equation 2.1; for exactly one occurrence, $H = 1$ and the coefficient becomes simply N, so equation 2.4 is N times equation 2.3: the probability of exactly one

occurrence in N trials is N times as great as the probability of occurrence for the first time on the Nth trial.

The significance of these equations, especially equations 2.1 and 2.3, becomes clearer if the number of trials N is expressed as a fraction of the true return period $\bar{T}$ by the substitution $N = \bar{T}/r$, where r is any positive number. This substitution permits the evaluation of the equations as $\bar{T}$ increases without limit, since by the definition of e, the base of natural logarithms, the limit of $(1 - a/\bar{T})^{\bar{T}}$ as $\bar{T}$ increases is e^{-a}. Thus the probability that an event x_T, whose return period is $\bar{T}$, *will not occur* within $N = \bar{T}/r$ trials, is (from equation 2.1),

$$P(\dot{x}_N < x_T) = [(\bar{T} - 1)/\bar{T}]^{\bar{T}/r} \xrightarrow[\bar{T} \to \infty]{} e^{-1/r} \tag{2.5}$$

Likewise, the probability that x_T will *occur for the first time* on the $N = \bar{T}/k$ trial is (from equation 2.3),

$$P[(\dot{x}_{N-1} < x_T)(x_N \geq x_T)] = (\bar{T} - 1)^{(\bar{T}/r)-1}/\bar{T}^{\bar{T}/r} \xrightarrow[\bar{T} \to \infty]{} 0 \tag{2.6}$$

2.4. Risks

These equations illuminate the nature of the intervals between recurrences of x_T in a very long series of trials, of which the average interval $\bar{T}$ is by definition the return period. The median $\tilde{T}$ is the period with a 50% probability of at least one occurrence (Pascal's original problem), $P(\dot{x}_N \geq x_T) = 1 - e^{-1/r} = \frac{1}{2}$. As $\bar{T}$ increases, $1/r$ approaches log 2 $= 0.69315$, so that the median is a little more than $\frac{2}{3}$ of the average, i.e., $\tilde{T} \doteq 0.7\bar{T}$. The mode, $\hat{T}$, or most frequent interval between recurrences, is always 0: there is more chance that an extreme value will recur on the next trial following an occurrence (interval 0) than that it will recur for the first time on any specific trial thereafter, but this probability for any specific trial approaches 0 as $\bar{T}$ increases without limit.

When $r = 1$, that is $N = \bar{T}$, the probability by equations 2.4 and 2.5 for various occurrences of an event x_T during a very long period equalling its average return period $\bar{T}$ approach:

0 occurrences	$1/e = 0.36788$
1 occurrence	$1/e = 0.36788$
2 or more occurrences	$= 0.26424$
	1.00000

Consequently, the probability that the event x_T will occur at least once in an infinitely long series is 0.63212, not much less than the value 0.638 given above for occurrences of a double six in 36 throws of two dice. Actually, the limiting values can be used for practical purposes whenever $\bar{T}$ exceeds 10 or 15, as shown in Table II.

Practical application of these findings can be made readily in terms of calculated risks. The probability (equations 2.1 and 2.5) that an event x_T, whose return period is $\bar{T}$, *will not occur* in any of $N = \bar{T}/r$ trials is also the probability that in each of these trials the variable x will be less than the value x_T. This in turn may be considered as the confidence that a structure, designed to withstand a maximum event

TABLE II. Factor r by which desired lifetime N must be multiplied to obtain design return period T_d for various calculated risks U (equation 2.8).

Calculated risk, U		.632	.500	.400	.333	.300	.250	.200	.100	.050
Desired life, N	2	1.27	1.71	2.22	2.73	3.06	3.73	4.74	9.75	19.76
	10	1.05	1.49	2.01	2.52	2.85	3.52	4.52	9.52	19.57
	∞	1.00	1.44	1.96	2.47	2.80	3.45	4.48	9.49	19.50

whose return period is $\bar{T}$, *will not fail* in a shorter period $\bar{T}/r$. Thus the confidence is 50% that a bridge designed to withstand a 100-year flood, but which will fail in the slightly larger 101-year flood, will not be washed out in less than about 70 years; the confidence that it will not be washed out in 100 years is only 37 percent—the risk of such failure is consequently 63 percent.

Conversely, for any desired lifetime $N = \bar{T}/r$, and a *calculated risk* of failure U within a lesser interval, the *design return period* T_d can be determined by substituting for N in equation 2.2 and solving for r:

$$(2.7) \quad U = P(\dot{x}_r \geq x_T) = 1 - (1 - 1/T_d)^{T_d/r} \xrightarrow[T_d \to \infty]{} 1 - e^{-1/r}$$

$$(2.8) \quad r = \log (1 - 1/T_d)^{T_d}/\log (1 - U) \xrightarrow[T_d \to \infty]{} -1/\log (1 - U)$$

Values of r are given in Table II for various calculated risks U and for lifetimes N of 2 and 10 (trials, e.g. years) as calculated from the exact first portion of equation 2.8, as well as the limiting values from the second part. These limiting values are approached so rapidly that they may be used with sufficient accuracy for any desired lifetimes greater than 10 or 15. This table indicates, for example, that a tower which is to last 50 years, with a risk of only 10% of failure due to strong winds before that time, should be designed for the strongest wind expected in $T_d = 50 \times 9.49 = 475$ years.

Tables II and I show different aspects of the same fundamental fact: that the intervals between recurrences of an event are variable. This fact, though known intuitively and demonstrable as a corollary of a problem solved more than 300 years ago, has not been used extensively in numerical estimates. One of the few investigations of the problem, by Thomas [23], used a different version of equation 2.4 (for the proba-

bility of exactly H occurrences in N trials), considered as a general expression of which others such as equations 2.1, 2.2, and 2.3 are special cases. By this more indirect method, conclusions analagous to those presented here were reached, and the resulting tables are reproduced in a recent textbook [24].

2.5. Theory

Use of Table II implies accurate estimation of the magnitude of x_d, the "design extreme" whose return period T_d is obtained from the table for the desired lifetime N and calculated risk U. Such estimation, however, is subject to the limitations of Table I as long as it is based on only the observed relative frequency of the extreme in question. Improvement in the estimate can be achieved only by increasing the size of the sample from which the relative frequency is determined, or by weighting or correcting the estimate in some way.

The most obvious weighting procedure is to consider all the observed extremes instead of only the extremes equalling or exceeding the required value. In effect, this process increases the sample size synthetically, and thus narrows the confidence limits of the estimate. The various empirical weighting procedures proposed in the last few decades have been replaced in recent years by a newer method, with theoretical foundations: the statistical theory of extreme values.

From foundations laid during the previous 15 years, the statistical distribution of the extreme values in a sample was developed during the 1930s by Dr. Emil J. Gumbel [25]. (The fundamentals of the theory are summarized by Kendall [26].) After applying the theory to such widely diverse things as the ages of the oldest inhabitants of each region and the intervals between radioactive emissions, Gumbel adapted it to flood analysis and introduced it in this form [27] shortly after coming to the United States in 1940.

The theory attracted widespread interest, and was adapted by others [7, 28] for hydrological computations, and applied to breaking strength [29] problems, the determination of gust loads on aircraft [30], and to climatic evaluations [31]; additional refinements were made by Gumbel [32].

The theory applies to the largest (or smallest) values in each of N independent sets of n independent observations each, drawn from the same population. This parent population must be distributed according to some exponential law (as is the normal distribution), so that it is unlimited but tends to zero as the variable increases or decreases; the distribution also must possess all moments.

While based on these premises, in practice the theory may be applied

to many cases in which some of the conditions are met only approximately; in particular, it may be used for extremes of distributions which are limited at either end, as long as the limits are well beyond the region of observation. Temperature has a definite lower limit (absolute zero) and possibly an upper limit, but since these are far removed from the values observed on earth, extremes of air temperature (or water, or rocks) may be analyzed by the theory. Similarly, rainfall amounts and flood stages can be analyzed if the smallest values in each set are still well above zero: the highest flood stage of each year in a perennial river can be analyzed, but not the highest stage in a dry wash which may have no water at all for several years in a row.

The fundamental theorem of the theory of extreme values is:

> In a set of N independent extremes $x_1, x_2, x_3, \ldots x_N$, each being the extreme of one of N sets of n observations each of an unlimited, exponentially-distributed variable, as both N and n grow large the cumulative probability that any one of these N extremes will be less (greater, for smallest values) than any chosen quantity, x, approaches the double exponential expression

$$q(x) = \Phi(x) = \exp\left[-e^{\mp a(x-\hat{x})}\right] \tag{2.9}$$

In the exponent, the $-$ sign applies for largest extremes, the $+$ sign for smallest extremes; "exp" is another way of writing "e to the power": $\exp(x) = e^x$. This expression gives the probability of nonoccurrence $q(x)$ of the event x in a single trial, and thus affords a way of determining the probability of occurrence $p = 1 - q = 1 - \Phi(x)$ used in Sections 2.3 and 2.4. Consequently, the return period of extremes equal to or exceeding x is

$$\bar{T}_x = 1/[1 - \Phi(x)] \tag{2.10}$$

Introduction of the expression for $\Phi(x)$, equation 2.9, into equation 2.10 yields a most unwieldy expression, so that in practice the probability of non-occurrence, $\Phi(x)$, is obtained first, and then the return period is found.

2.6. Description

The manner in which this probability of non-occurrence, $\Phi(x)$, varies with x is shown by differentiation:

$$\Phi'(x) = a \cdot e^{\mp a(x-\hat{x})}\,\Phi(x) \tag{2.11}$$

Further differentiation shows that the density of probability (Section 1.5) is a maximum at $x = \hat{x}$, i.e., that $\hat{x}$ is theoretically the most frequent value (mode) of the set of extremes being considered. Graph-

ing reveals the density function (equation 2.11) to be a generally bell-shaped curve, roughly similar to the normal curve but skewed markedly (to the right for largest values, to the left for smallest values), so that the mean is different from the mode (it is greater for largest values, less for smallest values).

The skewness of the density of probability curve shows that there is a greater likelihood of very great extremes than of very small ones, i.e., than of extremes which are closest to the mean of the parent values. Although derivation of the theory of extreme values is far beyond the scope of this article, some intuitive basis for it can be mentioned.

In any fair-sized sample drawn from a normal distribution, or from one of the same general unimodal, unlimited type, it is almost certain that there will be at least one value as much as one standard deviation greater than the mean. On the other hand, since the distribution from which the extremes are drawn has no limits, a few such samples will contain values greater than the mean by more than three standard deviations. Consequently, when the extremes of each of many such samples are considered as a group, they are found to range from around one standard deviation above the mean of the original distribution up to a few very large values, but to be concentrated close to the lower end of this range.

The skewness of the density distribution of the extreme value function is shown in Fig. 1, which also illustrates the relation between a set of extremes and the observations from which it is drawn. The large histogram, to which a normal curve has been fitted, shows the frequency of occurrence of the highest temperature of each summer day (June-July-August) at Washington, D. C., during 74 years—a total of 6,808 daily observations [33].

In the lower right a solid histogram shows the frequency of occurrence of the highest temperature in each of the 74 summers, with an extreme value probability density curve fitted to it. Since the daily values are by 5°F class intervals, the scale for the annual values has been multiplied by 5 to make the two curves comparable.

One moral of Fig. 1 is that even a small set of extreme values must represent a relatively large number of actual observations, since each value in the set of N extremes is itself the extreme of a large number, n, of readings: here $N = 74$, $n = 92$, since this example involves the extremes of each of 74 sets of observations each containing 92 observations. The theory of extreme values assumes both N and n to be large, and in general it should not be applied if either is less than 20, and preferably 30 or even 50.

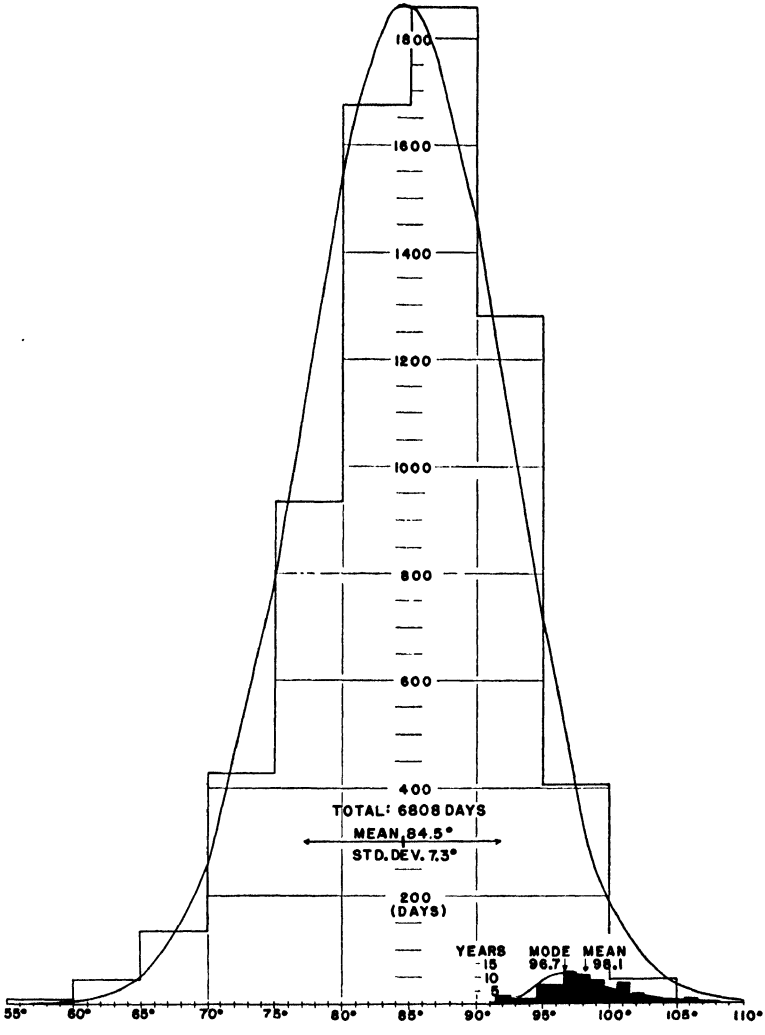


FIG. 1. Frequencies of highest temperatures of each summer day and year. Washington, D. C., 74 years (1872–1945), June-July-August.

The highest daily temperatures in Fig. 1 are not fitted too well by a normal curve—they are skewed somewhat to the right, but not as much as would be required if they were independent values and thus subject to the theory of extreme values. Incidentally, the analysis of the extremes applies only to summer: in five of the 74 years, the highest temperature of the year came outside the three summer months, once in May and four times in September.

2.7. Parameters

In the density distribution (equation 2.11) of extreme values, the inflection points, where curvature changes from convex upward around the mode to concave in the tails, are at $x = \hat{x} = \pm 0.9624/a$; in the normal curve, the inflection points are at $\pm\sigma$. Thus $1/a$ is somewhat analagous to σ, in that it indicates the degree of dispersion of the various extremes about their mode; consequently, "a" itself is a measure of concentration about the mode.

This measure of concentration, a, and theoretical mode, $\hat{x}$, of any set of extremes depend in theory on the density distribution $f(x)$ of the entire set of values and on its integral, the cumulative probability function $F(x)$:

$$a = n\, f(x) \qquad \text{and} \qquad F(\hat{x}) = 1 - 1/n \tag{2.12}$$

Since these theoretical definitions require knowledge of the density distribution of the population from which the set of extreme values has been drawn, and in general the only knowledge of this population is derivable from the sample, these definitions cannot be used in practice. Instead, these two values are estimated by the theory of least squares from the data of the sample (as explained in Section 2.8), using two theoretical quantities:

$$a = \sigma_N/s_x \qquad \text{and} \qquad \hat{x} = \bar{x} \mp s_x(\bar{y}_N/\sigma_N) \tag{2.13}$$

Here $\bar{x}$ is the mean and s_x the standard deviation of the set of extremes, while the mean $\bar{y}_N$ and standard deviation σ_N of a theoretical variate depend only on the sample size N, and thus can be tabulated for ready use. Table III gives their values for every integer of N from 15 to 100, and for selected greater sample sizes; linear interpolation is adequate when $N > 100$ since as N increases both quantities approach limiting values asymptotically. Table III was computed by Dr. Gumbel [31].

Because the double exponential form of the basic equation (2.9) imposes difficulties in computation and analysis, it is reduced to linear form by taking the double ("iterated natural") logarithm of both sides; a new variate, $y = -\log[-\log \Phi(x)]$, is called the *reduced variate:*

$$y = \pm a\,(x - \hat{x}) \tag{2.14}$$

Solved for x, this equation becomes

$$x = \hat{x} \pm y/a \tag{2.15}$$

With the definitions of Eq. 2.13 introduced, this expression becomes

$$x = \bar{x} \pm (s_x/\sigma_N)(y - \bar{y}_N) \tag{2.16}$$

TABLE III. Reduced means and standard deviations of reduced extremes.

Sample size N	Reduced mean $\bar{y}_N$	Std. dev. σ_N	Sample size N	Reduced mean $\bar{y}_N$	Std. dev. σ_N	Sample size N	Reduced mean $\bar{y}_N$	Std. dev. σ_N
15	.5128	1.0206	50	.5485	1.1607	85	.5578	1.1973
16	.5157	1.0316	51	.5489	1.1623	86	.5580	1.1980
17	.5181	1.0411	52	.5493	1.1638	87	.5581	1.1987
18	.5202	1.0493	53	.5497	1.1658	88	.5583	1.1994
19	.5220	1.0565	54	.5501	1.1667	89	.5585	1.2001
20	.5236	1.0628	55	.5504	1.1681	90	.5586	1.2007
21	.5252	1.0696	56	.5508	1.1696	91	.5587	1.2013
22	.5268	1.0754	57	.5511	1.1708	92	.5589	1.2020
23	.4283	1.0811	58	.5515	1.1721	93	.5591	1.2026
24	.5296	1.0864	59	.5518	1.1734	94	.5592	1.2032
25	.5309	1.0915	60	.4421	1.1747	95	.5593	1.2038
26	.5320	1.0961	61	.5524	1.1759	96	.5595	1.2044
27	.4332	1.1004	62	.5527	1.1770	97	.5596	1.2049
28	.5343	1.1047	63	.5530	1.1782	98	.5598	1.2055
29	.5353	1.1086	64	.5533	1.1793	99	.5599	1.2060
30	.5362	1.1124	65	.5535	1.1803	100	.5600	1.20649
31	.5371	1.1159	66	.5538	1.1814			
32	.5380	1.1193	67	.5540	1.1824	150	.5646	1.22534
33	.5388	1.1226	68	.5543	1.1834			
34	.5396	1.1255	69	.5545	1.1844	200	.5672	1.23598
35	.5402	1.1285	70	.5548	1.1854	250	.5688	1.24292
36	.5410	1.1313	71	.5550	1.1863			
37	.5418	1.1339	72	.5552	1.1873	300	.5699	1.24786
38	.5424	1.1363	73	.5555	1.1881			
39	.5430	1.1388	74	.5557	1.1890	400	.5714	1.25450
40	.5436	1.1413	75	.5559	1.1898			
41	.5442	1.1436	76	.5561	1.1906	500	.5724	1.25880
42	.5448	1.1458	77	.5563	1.1915			
43	.5453	1.1480	78	.5565	1.1923			
44	.5458	1.1499	79	.5567	1.1930	750	.5738	1.26506
45	.5463	1.1519	80	.5569	1.1938			
46	.5458	1.1538	81	.5570	1.1945	1000	.5745	1.26851
47	.5473	1.1557	82	.5572	1.1953			
48	.5477	1.1574	83	5574	1.1969	Inf.	.5772	1.28255
49	.5481	1.1590	84	.5576	1.1967			

where, as before, the upper sign is used for extremes of maximums, the lower for those of minimums. This equation gives the *expected extreme* for any set of N extremes, that is, the extreme value for which the true return period $\bar{T}$ corresponds to the probability given by y.

In this form, the results of application of the theory of extreme values to a set of extremes can be compared with results given by earlier, more

empirical formulas. A "general formula for hydrologic frequency analysis," applicable to all analyses of the probabilities or return periods of extreme values, has recently been proposed by Chow [34]. With notation altered to conform to the remainder of this article, it is:

$$x = \bar{x} + Ks_x \tag{2.17}$$

where x is the departure of an individual observation (flood) from the mean $\bar{x}$ of the series; s_x is the standard deviation of x (i.e., of the series), and K is a "frequency factor . . . which depends upon the law of occurrence" of the particular event.

The only difference between various methods, each of which assumes a different law of occurrence, is in their definition of K, computation of which in some cases is quite laborious and requires extensive tables. By dividing equation 2.17 by $\bar{x}$, Chow obtained an expression for the "y-mean ratio" (in his notation y is used where x is used here) in terms of K and the coefficient of variation:

$$x/\bar{x} = 1 + K(s_x/\bar{x}) \tag{2.18}$$

This form he considered more useful than the first (2.17) in comparing various formulas.

From equation 2.16, the "frequency factor" for the theory of extreme values is:

$$K = \pm(y - \bar{y}_N)/\sigma_N \tag{2.19}$$

Since y is the double logarithm ("iterated natural logarithm") of the probability, and $\bar{y}_N$ and σ_N depend only on the sample size, K can be tabulated readily, as in Table IV. With the values in this table, the

TABLE IV. Values of $K = \pm(y_T - \bar{y}_N)/\sigma_N$ for various probabilities $\Phi(x)$ and various sample sizes N.

	$\Phi(x) = 1 - 1/T$						
N	0.999	0.990	0.980	0.960	0.950	0.900	0.800
15	6.265	4.005	3.321	2.631	2.310	1.703	0.967
20	6.006	3.836	3.179	2.517	2.302	1.625	0.919
25	5.842	3.728	3.088	2.444	2.235	1.575	0.888
30	5.727	3.653	3.026	2.393	2.188	1.541	0.866
40	5.576	3.554	2.943	2.326	2.126	1.495	0.838
50	5.478	3.491	2.889	2.283	2.086	1.466	0.820
70	5.359	3.413	2.824	2.230	2.038	1.430	0.797
100	5.261	3.349	2.770	2.187	1.998	1.401	0.779
200	5.130	3.263	2.698	2.129	1.944	1.362	0.755

expected extreme x whose probability of not being equalled or exceeded (equation 2.9) is $\Phi(x) = \exp(-e^{-y})$, and therefore whose return period is (equation 2.10) $T_x = 1/[1 - \Phi(x)]$, can be computed if the mean $\bar{x}$ and standard deviation σ_x of N extremes are available. For example, the extreme expected to occur (on the average over a long period) once in 100 years [$\Phi(x) = 0.990$] is 3.65 s_x greater than the mean of $N = 30$ extremes.

Conversely, the expected return period $\bar{T}_x$ corresponding to any given extreme value x can be obtained from equations 2.10, 2.13, and 2.14, but the resulting expression is cumbersome, and the determination is easier by the methods outlined in Section 2.10.

2.8. Computations

Certain computations based on the theory of extreme values can be made directly from a set of extremes (obtaining the mean $\bar{x}$ and standard deviation s_x) by the use of Tables III or IV, and equations 2.16 or 2.17. For complete analysis of a set of extremes, however, and in particular to determine how well the set follows the theory, it is more convenient to graph the data, using a special extreme probability paper.

On this paper, one of the coordinates is linear, for the observed extremes (denoted by x), while the other is double logarithmic, for $\Phi(x)$ which is (equation 2.9) a double exponential expression. In the original version of this paper [7, 27], the double-logarithmic coordinate was the abscissa; in a revised version [31] the coordinates are reversed so that the observed values, denoted by x, are plotted along the abscissa as is customary, and the double-logarithmic scale is the ordinate. To facilitate plotting and analysis, there are two other ordinate scales: at the left a linear scale for the reduced variate y, and at the right a quasilogarithmic scale for the return period T.

Extreme probability paper is identical in function and use to other probability papers (Section 1.4), and observations are plotted on it by rank and magnitude. Each extreme is plotted at an abscissa corresponding to its value and at an ordinate, on the double-logarithmic scale, corresponding to its cumulative rank divided by $N + 1$. All such points are then connected by short straight lines, producing a zig-zag line which should, if the entire set follows the theory of extreme values, approximate a straight line.

This straight line is simply equation 2.14 or 2.15, which was fitted to the observations by a method of least squares: the estimates of a and $\hat{x}$ (equation 2.13) actually minimize the sum of the diagonal distances from the line to each plotted point representing one of the observed extremes. Ordinary least squares procedure minimizes the sums of either the

horizontal or vertical departures, but this method provides a best fit, independent of whether x or y is considered as the independent variable.

This "line of expected extremes" is expressed customarily by equation 2.15, since in practice specific values of x are determined for various probabilities as represented by y, such as 0 and 5. This procedure, however, implies no dependence of x on y: they are mutually dependent.

To indicate how well the line fits the observations, a confidence band can be drawn on both sides of it. Generally, the limits of this band are chosen so that there is a probability of 0.68268 (corresponding to $\pm\sigma$ of the normal distribution) that the extreme corresponding to any frequency $\Phi(x)$ will fall within the band. For frequencies from 0.15 to 0.85, the width of this band is obtained by dividing a certain theoretical value, here called h, by $a\sqrt{N}$, so that the limits of the band (sometimes called control curves) are, by equation 2.16,

$$x = \bar{x} \pm Ks_x \pm h/a\sqrt{N} \tag{2.20}$$

where the first double sign is + for largest values, − for smallest values, and the second gives, respectively, the upper and lower limits of the band. Values of h for various frequencies are:

Freq. $\Phi(x)$:	.150	.200	.300	.400	.500	.600	.700	.800	.850
h:	1.255	1.243	1.268	1.337	1.443	1.598	1.835	2.241	2.585

For frequencies greater than 0.85, the width of the 0.68269 confidence band is calculated for the largest and next-to-largest extremes:

$$\Delta_{x,N} = \pm 1.1407/a \qquad \Delta_{x,N-1} = \pm 0.7592/a[(N-1)/N] \tag{2.21}$$

On either side of the line of expected extremes, intervals as obtained by dividing the tabular values above by $a\sqrt{N}$ are plotted at the corresponding frequencies; the values computed from equation 2.21 are laid off similarly at the frequencies of the largest and next-to-largest observed values, but symmetrically about the line and not about the points representing those observed extremes. Two lines are drawn connecting the points so plotted, forming a characteristic horn-shaped figure; technically, the two lines should be drawn smoothly, with a french curve, but in practice short straight lines are adequate. For frequencies greater than that of the largest observed extreme, the confidence band is extended parallel to the line of extremes at the same width as for the largest value.

Figure 2 shows, for the same data represented by the solid histogram of Fig. 1, the zig-zag plot of the 74 observed extremes, their "line of expected extremes," and the confidence band centered on this line. The scales and grid of Fig. 2 are skeletonized from extreme probability graph paper. Since the ordinate of this paper is doubly logarithmic, most of the

observations are concentrated in the lower part of the diagram: the median (frequency .500 or return period 2) is less than a third of the way up the figure. Because the largest and next-to-largest values in this

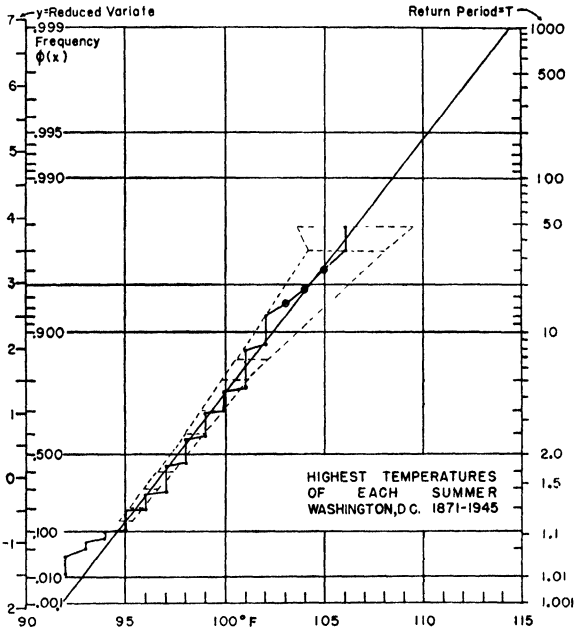


FIG. 2. Highest temperatures of each summer at Washington, D. C. (1871–1945) plotted on extreme probability graph and fitted by line of expected extremes, with confidence band added.

particular example are equal in value (a not uncommon occurrence in some sets of extremes), the confidence band broadens markedly for the last value. In Figure 2 the confidence band has not been extended past the largest observed value, as may be done.

2.9. Evaluations

If about two-thirds of the observed extremes as plotted on the extreme probability paper fall within the confidence band, the extremes may be considered to be represented adequately by the theory of extreme values. Usually the largest few values will show the greatest departures from the line, but unless one of them is well outside the confidence band it is not subject to serious question.

The probability p_Δ that the greatest extreme x_N of the sample will depart, by an amount equal to or less than Δ (its actual departure), from

its expected value x_T as given by the line of expected extremes (or by equations 2.16 or 2.17) is

$$p_\Delta = \exp(-e^{-a\Delta}) - \exp(-e^{a\Delta}) \tag{2.22}$$

Values of $a\Delta$, the "relative departure," for various probabilities are:

Probability, p_Δ:	0.0100	0.1000	0.3000	0.5000	0.6827	0.7500	0.9000
Rel. departure, $a\Delta$:	0.0136	0.1342	0.4200	0.7429	1.1407	1.2940	2.2511

When the actual departure Δ of the largest extreme from its expected value is multiplied by "a" (equation 2.13), this table permits estimation of the probability that the largest extreme of the given set could have such a departure.

Another method of determining the reliability of the largest extreme, if it deviates markedly from the expected value, is to omit it from an entire new computation of $\bar{x}$, s_x, and the line of expected extremes, and then determine its relative departure from the new line for evaluation by the above table.

When the most extreme value of a set of extremes is very different from its expected value, which is based on it and all the others in the set, it may be so as the result of chance: there is always a probability of 0.01 that the 100-year value will occur on the next trial (i.e. year). But such a departure warrants investigation of the original data for possible errors in observation, recording, or transcription.

When the two or three most extreme values depart markedly from the expected values, or when many of the observations plot outside the confidence band, the observations simply may not follow the theory of extreme values, for any of several reasons:

a. The set of extremes in question may not be independent.

b. The individual extremes may not be comparable, i.e., may not be extremes of samples from the same population. For example, annual wind extremes at a weather station where the anemometer height or exposure has changed markedly through the years do not follow the theory; nor do maximum annual river stages (heights) if the channel width increases irregularly with the height.

c. The original population, from which independent samples are presumed to have been drawn with each sample yielding a separate extreme, may not be unimodal and unlimited. Maximum relative humidity values would not follow the theory (except in very arid areas) because the upper limit (100%) is within the range of the observations.

Lack of correspondence between observation and theory does not discredit the theory: it merely shows that the theory of extreme values cannot be used to analyze the observations. Thus, unless it has been

established that the variable in question does fall within the scope of the theory, a complete analysis, using a confidence band on extreme probability paper, is desirable before any conclusions are drawn.

2.10. Applications

Most practical applications of the theory of extreme values, in geophysics as elsewhere, are concerned primarily with return periods. The information desired usually is either the return period of some specified extreme value, or else the converse, the greatest extreme to be expected within some specified period. Either of these questions can be answered satisfactorily, together with the confidence limits of the answers.

As demonstrated in Sections 2.3 and 2.4, the return period $\bar{T}$ is the *average* of all the intervals between recurrences of an event in a long series, but half of the intervals will be less than about .7 of this average, and the most probable interval is zero. The probability that an event will not occur until the end of its return period is only 0.37, which is also the probability that it will occur exactly one time before the end of the period.

Confidence limits of the return period also can be expressed in another way. Instead of a single value, the return period can be indicated by the interval within which there is a given probability P_T that the extreme x_T (whose return period is T) will occur. The limits of this interval are bT and T/b, where $e^{-1/b} - e^{-b} = P_T$. This gives, for various values of P_T:

P_T:	.100	.300	.500	.68269	.750	.900	.95450
b:	1.146	1.522	2.105	3.129	3.909	9.503	21.485
$1/b$:	.873	.657	.475	.319	.256	.105	.0465

Thus the probability is .68 that the extreme value x_T will occur for the first time in at least $.32\bar{T}_x$ and in no more than $3.13\bar{T}_x$.

The first of the two questions concerning extremes, that of the return period $\bar{T}_x$ for a specified extreme value x, is difficult to answer directly. Combination of equations 2.10, 2.13, and 2.14 gives

$$\bar{T}_x = 1/[1 - \exp\{-\exp[\bar{y}_N \pm (x - \bar{x})(\sigma_N/s_x)]\}] \tag{2.23}$$

Fortunately, as x increases, this converges toward

$$\bar{T}_x \xrightarrow[T_x \to \infty]{} \exp[\bar{y}_N \pm (x - \bar{x})(\sigma_N/s_x)] = e^y \tag{2.24}$$

In both these equations, the + applies to largest values, the − to smallest. Thus, with the mean $\bar{x}$ and standard deviation s_x of the set of extremes, and the values of $\bar{y}_N$ and σ_N in Table III, $\bar{T}_x$ can be calculated. Usually it is simpler, however, to obtain it graphically: it is read on the return period scale, at the right of the extreme probability paper, opposite the point of intersection of the line of expected extremes with the desired

value of x, as given on the abscissa scale at the bottom of the sheet. (Equations 2.23 and 2.24 indicate the nature of the relationship between the return period scale on the right side of the extreme probability paper, the frequency scale in the body of the paper, and the reduced variate scale along the left side; all three scales are indicated in Fig. 2.)

The second question concerning extremes, that of the probable extreme with a given return period $\bar{T}_x$, is much simpler: it is answered by equation 2.16 and Table III, or equation 2.17 and Table IV, using $\bar{x}$ and s_x in either case. Or the probable extreme can be read directly on the extreme probability paper: it is the abscissa at which the line of expected extremes intersects the appropriate return period line.

Once the expected extremes, x_1 and x_2, for any two return periods, T_1 and T_2, are determined, the expected extremes x_T for any other return period T_x can be determined:

$$x_T = x_1 + [x_2 - x_1][(y_T - y_1)/(y_2 - y_1)] \tag{2.25}$$

In this equation, the last fraction involving only the reduced variates (y) depends only on the lengths of the two periods T_1 and T_2, and is called Z_T. For two convenient periods of 10 and 100 trials (years), values of

TABLE V. Factor (Z_T) by which difference between 100-year and 10-year extremes must be multiplied to give excess over 10-year value of extreme to be expected in T years.

T	Z_T	T	Z_T	T	Z_T
15	.18018	60	.78118	140	1.14399
20	.30634	70	.84717	150	1.17306
25	.40352	80	.90451	200	1.29607
30	.48257	90	.95497	300	1.46941
35	.54924	100	1.00000	400	1.59159
40	.60682	110	1.04080	500	1.68666
45	.65759	120	1.07812	750	1.85937
50	.70287	130	1.11230	1000	1.98186

Z_T for various other return periods T_x are given in Table V, which can be used to determine the expected extreme for those periods:

$$x_T = x_{10} + Z_T(x_{100} - x_{10}) \tag{2.26}$$

Most of the computations discussed in this and preceding Sections are arranged in logical order on a "Worksheet 2," reproduced as Fig. 3. "Worksheet 1," printed on the reverse of the original of this form, provides space for arranging the extremes in order, computing their mean and standard deviation, and their cumulative frequencies and plotting

PROBABILITIES OF EXTREMES - Worksheet 2

EXAMPLE

I. Mean and Standard Deviation (First line taken from Worksheet 1, on back):

N = 408 Σ(xp) = 13,364 Σ(x²p) = 457,252

√N = 20.199 Mean. $\bar{x}'$ = 32.7549 $\bar{x}^2$ = 1,120.7157

Arbitrary Mean: x_o = 0.0 $(\bar{x}')^2$ = 1,072.8835

True Mean $\bar{x}$ = 32.7549 $(s_x)^2$ = 47.8322

N/(N-1) = 1.0025 Standard Deviation: s_x = 6.9159

II. Parameters (First line taken from Table 1):

σ_n = 1.25484 $\bar{y}_n$ = .5715

1/a = s_x/σ_n = 5.5114 $\bar{y}_n$ (1/a) = 3.1498

1/(a√N) = (1/a)/√N = .27286 u = $\bar{x} + \bar{y}_n(1/a)$ = 29.6051 = mode

NOTE. Upper sign used for maxima, lower sign for minima

III. Line of Expected Extremes

x = u ± (1/a)y = 29.6051 ⊕ ⊖ 5.5114 y

y:	-2.00	0.00	3.00	5.00	2.25	4.60
y(1/a):	-11.0228	0.00	16.5342	27.5570	12.4027	25.3524
x:	18.5823	u = 29.6051	46.1393	57.1621	x_{10} 42.0078	x_{100} 54.9575

NOTE: Values x_{10} and x_{100} are for return periods of 10 and 100

IV. Half-width of 0.68269 Confidence Band, $\sigma_{x,m} = \sigma_{y,m}\sqrt{N}/(a\sqrt{N}) = (\sigma_{y,m}\sqrt{N})\left[(1/a)/\sqrt{N}\right]$:

Φ(x):	.150	.200	.300	.400	.500	.600	.700	.800	.850
$\sigma_{y,m}\sqrt{N}$:	1.255	1.243	1.268	1.337	1.443	1.598	1.835	2.241	2.585
$\sigma_{x,m}$:	.342	.339	.346	.365	.394	.436	.501	.611	.705

For largest value, $\Delta_{x,n} = 1.141\ (1/a)$ = 6.288

For next-to-largest value, $\Delta_{x,n-1} = .759\left[N/(N-1)\right](1/a)$ = 4.191

V. Expected Extreme, in T periods (years, etc) $x_T = x_{10} + Z_T(x_{100} - x_{10})$: $x_{100} - x_{10}$ = 12.95

T	Z_T	$Z_T(x_{100}-x_{10})$	x_T	T	Z_T	$Z_T(x_{100}-x_{10})$	x_T	T	Z_T	$Z_T(x_{100}-x_{10})$	x_T
15	.180			60	.781			140	1.144		
20	.306			70	.847			150	1.173		
25	.404			80	.905			200	1.296		
30	.483	6.25	48.26	90	.955			300	1.469		
35	.549			100	1.000			400	1.592		
40	.607			110	1.041			500	1.687		
45	.658			120	1.078			750	1.859		
50	.703			130	1.112			1000	1.990		

Place: DES MOINES, IOWA

Data: HIGHEST WIND SPEED (FASTEST MILE) IN EACH MONTH, JAN. 1912 to DEC 1945 FROM MANUSCRIPT TABULATION IN U. S WEATHER BUREAU

Computer: U.G. Date: 2/5/51

Environmental Protection Section, Research & Development Branch, Military Planning Division, Office of The Quartermaster General

For Evaluating the Probability of Extreme Values by the Method Developed by Dr. E. J. Gumbel.

FIG. 3. Example of computation form for evaluating extremes by the methods discussed in Sections 2.7 to 2.10. Only those portions of the worksheet necessary to answer a particular question need be used. (Taken from [31].)

positions. These two worksheets, and the form of the extreme probability paper used with them, were developed from Gumbel's original work by the Climatology Unit, Environmental Protection Section, Research and Development Branch, Office of The Quartermaster General; they are discussed in a report of this Unit [31], from which Fig. 3 is taken. This example involves winds, rather than the temperatures of Figs. 1 and 2, to present a different application of the theory and method.

2.11. Conclusion

This Section has shown that a combination of classic probability theory and the very recent theory of extreme values permits accurate analysis and evaluation of the extremes of many geophysical elements. The highest temperature, strongest wind, severest earthquake, greatest magnetic disturbance, or worst flood which has occurred or been exceeded only 5 times in 50 years has a relative frequency of 5 in 50 or 0.10, but the best estimate, with 95% confidence, is that its true probability is somewhere between 0.05 and 0.22. Thus its return period is not necessarily 10 years, but is somewhere between 4.5 and 20 years. When all the 50 observations are considered, instead of only the 5 which have equalled or exceeded the value in question, then the theory of extreme values provides a reasonably accurate method of estimating the return period—or the expected extreme for any given return period.

Even after the return period is established, however, the chances are two out of three that the value in question will occur within a shorter interval, and are also two out of three that it will occur in at least 0.32 and no more than 3.13 times the return period. For engineering and similar applications, the *design return period* T_d can be determined (Section 2.4) for any desired lifetime N and calculated risk U of failure in less than T_d:

$$T_d \doteq -N/\log (1 - U) = rN \tag{2.27}$$

Table II provides a simple way of determining T_d for most risks U actually used.

Once this design return period is established, the expected extreme corresponding to it (x_T) can be obtained by the theory of extreme values. This is done most simply by equation 2.17 ($x_T = \bar{x} + Ks_x$) and Table IV, for K; this requires only the mean $\bar{x}$ and standard deviation s_x of N extremes, provided that extremes of the type in question are known to follow the theory reasonably well.

Fundamentally, the theory of extreme values involves the development on strictly theoretical grounds of a function (equation 2.9) for the probability that a given extreme value will not be equalled or exceeded

by any one of a very large set of extreme values, obtained as specified. Observed extremes are then fitted to this function by an ingenious least squares procedure, involving in addition several approximations based on the assumption that the sample of observed extremes is so large that limiting (asymptotic) values can be used.

This procedure is essentially similar to that used for the "normal" distribution, and many other statistical and mathematical "laws," in which observed data are fitted to a theoretical function. As is often the case in many other fields, the theoretical function has been found to apply to samples which depart markedly from the original premises (small in number, not wholly independent, not unlimited, etc.). In some cases, however, other samples which apparently should follow the theory equally well do not do so, for some reason which may not be apparent.

Hitherto, many distributions of extreme values, falling within the scope of the theory of extreme values, have been analyzed by other methods. Chief of these has been the logarithmic normal distribution; that is, the logarithms of the individual extremes have been considered to be normally distributed (Section 1.4). Some of the earlier hydrologic analyses used a logarithmic transformation, and more recently the breaking strengths and analogous properties (e.g., water penetrability) have been evaluated by using logarithms.

As yet, no simple method has been proposed to determine whether an actual set of observed data are fitted better by one theoretical function than another. Familiarity with the logarithmic normal procedure, and the complexity of the extreme value theory in its earlier stages, has caused many investigators to prefer the former. It is hoped that the exposition of the theory of extreme values in this section will enable geophysicists and others to determine for themselves whether the newer theory cannot be used to greater advantage in analyzing any problem involving extremes.

3. Circular Distributions

3.1. Requirement

Circular variables are those which vary continuously through all angles of a circle, in contrast to the more familiar linear variables, which may have no limits or be limited on one or both ends. More so than any other science to which statistics is applied, geophysics has many problems involving circular variables: many elements (e.g., winds, tidal forces, magnetic fields) vary around the compass, and almost all geophysical elements vary continuously with time through a day, a lunar month, a solar (27-day) cycle, or a year. Hitherto, such data have been analyzed

either as though they were linear, or as trigonometric functions, especially through the use of Fourier series, in which several sine or cosine terms of different amplitudes and periods are added to approximate the original data.

As long as a circular variable does not extend completely around the circle, it can be linearized for statistical analysis without great error. Ocean swells reaching a beach have a total variation in direction of about half a circle, and all days of snowfall in temperate latitudes occur in about half a year. In such cases, statistical analysis based on the normal distribution, or any other linear distribution, is adequate: it may be considered that the variable has no limits on either side. However, when all directions, hours, or months are represented in the distribution of the variable, the linear approach cannot be justified: there is no more logic in considering the day to start and end with midnight than at noon or 7 A.M., and changes in the limits can affect any analysis seriously.

Approximation of a circular variable by a Fourier series avoids the difficulty of artificial limits, but introduces another artificiality: the periods of the various terms usually have no physical basis. What, for example, is the significance of a half-yearly term in a series approximating the annual course of air temperature or geomagnetic intensity? At best, comparison of two circular variables by Fourier series can indicate the phase retardation, i.e., the amount by which the peak of the curve lags behind some point, such as the solstices for temperature. Furthermore, Fourier analysis cannot be applied readily to spatial variables, i.e., those involving directions such as wind.

3.2. Description

During the last year, a *circular normal probability function* has been described by Gumbel [35, 36]; when developed it will permit circular variables to be analyzed in the same way that linear variables now are discussed with the aid of the linear "normal curve." The circular normal distribution has the same theoretical basis as the linear normal one: it assumes a large number of random "errors," or departures from the mean, with the frequency of the departures varying inversely with their magnitude.

A crude experiment illustrating the theory of the circular normal distribution is provided by a tiltable roulette wheel. When horizontal, the frequencies of the numbers on which the ball alights is uniform around the wheel. The more it is tilted, the more the frequencies concentrate toward the numbers at the bottom, regardless of their value. When the wheel is inclined by 30° or 40°, the distribution is confined to the two or three numbers at the bottom.

In the equation of the circular normal distribution, the degree of concentration of the variable at one time or direction is indicated by a parameter, denoted by k. This parameter is 0 for a uniform circular distribution, and has no upper limit, although values of k exceeding 3 indicate so great a concentration within a narrow sector that the distribution may be considered as linear. Thus k, a measure of concentration around the mean, is in many ways analogous to the reciprocal of the standard deviation σ of the linear normal distribution, since σ is a measure of dispersion around the mean; k in analogous to "a" in the theory of extreme values (Section 2.7).

The density of probability of the circular normal distribution is:

$$\Phi(\alpha, k) = \frac{1}{I_0(k)} e^{k \cos \alpha} \tag{3.1}$$

where α is the angle measured from the mean, and the denominator involves an incomplete Bessel function of the first kind of zero order for a pure imaginary argument, and has real values.

This function is completely specified by the two parameters, α, the angular departure from the mean, and k, the measure of concentration about the mean. In turn, k may be estimated by the method of maximum likelihood from the observations themselves: it is uniquely determined by the length of the *vector mean* $\bar{\alpha}$ of the data (Table VI). The vector mean is simply the vector sum of the data divided by the total number of *units*, not observations.

TABLE VI. Values of the parameter k of the circular normal distribution corresponding to lengths of the vector mean, $\bar{\alpha}$.

$\bar{\alpha}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.000	.020	.040	.060	.080	.100	.120	.140	.160	.181
.1	.201	.221	.242	.262	.283	.303	.324	.345	.366	.387
.2	.408	.430	.451	.473	.495	.516	.539	.561	.584	.606
.3	.629	.652	.676	.700	.724	.748	.772	.797	.823	.848
.4	.874	.900	.927	.954	.982	1.010	1.039	1.068	1.098	1.128
.5	1.159	1.191	1.223	1.257	1.291	1.326	1.362	1.398	1.436	1.475
.6	1.516	1.557	1.600	1.645	1.691	1.739	1.790	1.842	1.896	1.954
.7	2.014	2.077	2.144	2.214	2.289	2.369	2.455	2.547	2.646	2.754
.8	2.871	3.000	3.143	3.301	3.479	3.680	3.911	4.177		

For observations which have magnitude as well as direction (such as wind speed by directions or flood stages by dates), the division is by the total number of units (miles per hour, or feet) rather than by the total

number of observations; there is no distinction for data which are merely frequencies of occurrence (such as number of hours of wind from each direction or number of people dying per month).

3.3. Procedure

In fitting a circular normal curve to observed data, the first step is to compute the *resultant direction* (time or date is considered as a direction) and *length,* which together form the *vector mean.* Basically, two methods for such computation are available: graphical and trigonometric. Each has several variants.

In the graphical method, vectors representing all the observations of each class are added, on plain or ruled paper or on a circular plotting board. Magnitude of the resultant vector, from the start of the first to the end of the last, is measured with a scale, and its direction determined by a protractor. Alternatively, the vectors may be plotted on a polar diagram and their components parallel to two perpendicular axes measured by a scale. From the algebraic sums of each component, the resultant is found as in the first method.

In the trigonometric method, components of each vector are obtained by multiplying it by the appropriate sine and cosine values; after addition, the two components are then used to determine the direction of the resultant by a tangent formula, and its magnitude either from a sine or cosine relation or from the root mean square.

From the vector mean, the proper value of k is found from Table VI, and the equation of the function may be written directly. Or, the observed and theoretical frequencies for each class interval (sector) may be compared, numerically or graphically.

For a numerical comparison of theoretical and observed frequencies, the observations must be regrouped into sectors so that one will be centered on the resultant direction. For example, if the resultant of a series of monthly observations turns out to be 86° (1 Jan. being 0° and 360°), the data originally grouped as 0–30°, 30–60°, 60–90°, etc., must be grouped into the following sectors: 11–41°, 41–71°, 71–101°, etc. The number of observations falling within these new classes can then be compared with the theoretical expectancies, as obtained from the appropriate area table, and multiplied by the number of observations.

In the present stage of development of the circular normal theory, such numerical comparison is not very practical or fruitful. Unless the original data were reported to much greater accuracy than the classes used (such as directions to the nearest degree or time to the nearest minute or day of the year), no basis is as yet available for regrouping them into the new classes based on the resultant. Only in case the resultant hap-

pens to fall close to the center of one of the original classes (sectors) can most observational series be compared numerically with the expected frequencies. Furthermore, as yet no criterion has been developed for the goodness of fit of observations to theory (such as is provided by the chi-square test in linear normal theory, or the confidence band in the theory of extreme values).

3.4. Graphing

Comparison of observations and theory can be made most readily and satisfactorily by graphing both the data and the theoretical density of probability. From a carefully drawn graph of the probability density, the expected frequency for each of the original classes (sectors) can be estimated for comparison with the observed frequencies. This eliminates the need to regroup the data for comparison with the probability values given in the area tables.

Such estimates will be most accurate, and any graphical representation or comparison of circular variables more meaningful, if equivalent polar paper is used instead of the customary polar coordinate graph paper. On equivalent polar paper, concentric circles are drawn at distances from the center corresponding to the *square root* of the indicated numbers, instead of the numbers themselves as on the customary paper. Thus, on equivalent polar paper, for each sector the *area is directly proportional to the frequency* which it represents.

The same results may be obtained on conventional polar coordinate paper by using the square roots of the observed and theoretical frequencies. Since equivalent polar paper is not generally available, Table VII gives the square roots, rather than the actual values, of the radius vectors for unit-area circular normal distributions with various k values. This table is condensed from a more extensive one [36], which itself required a complex computational procedure. Table VII gives values for 10° intervals, but satisfactory curves can be plotted by using ordinates at intervals of 20° or 30°.

To obtain a curve for comparison with one plotted from the square roots of n observations grouped into w equal sectors (including any with no observations), the tabular values must be multiplied by $\sqrt{n/w}$; when observations are expressed as percentages, no multiplication of the tabular values is required. In either case, however, square roots of the observed values must be used, until equivalent polar paper becomes available.

Although the principle of equivalent polar paper is obvious, it does not seem to have been applied to any great extent in geophysics, or in graphic presentation generally. Yet a sector is a truer representation of observations which may have fallen anywhere within it than the conven-

tional radius bar or vector, centered in the sector with length directly proportional to the number of observations or their sum or mean.

A similar use of square roots in polar graphing was proposed by Leighly [37] almost a quarter-century ago, but was little used; he did not suggest square-root graph paper. The term "equivalent" for the square-root paper developed for use in graphing the circular normal distribution was suggested by Leighly as the internationally-understood term implying areal equivalence.

3.5. Limitations

As previously indicated, circular normal probability theory is a completely new branch of statistics, and as yet has not been developed to the point of general utility. So far, the basic function has been established, and tables computed for the probability function itself (areas of sectors) and the density of probability (radius vectors), Table VII.

Perhaps the most significant aspect of the development at present is the finding that the vector mean of a unimodal distribution of a circular variable uniquely characterizes the degree of concentration of the variable about the angular mean or resultant direction. This vector mean, translated into the parameter k of the distribution function, affords an index of the degree of concentration. Thus, values of k can be computed for various distributions for comparative purposes: the relative concentrations of winds at various places, or at different hours or months in the same place, can be compared.

Such comparison can be only qualitative, however, since no relation comparable to the "t-test" has yet been developed. This is hardly surprising: although the linear normal distribution was developed more than a century ago, the t-test is barely 40 years old. While the circular normal function, with its cosine exponent and incomplete Bessel function, is far more complicated than the linear normal one, its development can proceed rapidly because of analogies with the linear case.

Another serious limitation on the use of the circular normal theory at present is that it applies properly only to unimodal distributions. Many circular distributions in geophysics, however, are bimodal or trimodal. Depending on the period of time covered, wind distributions may show several peaks, flood crests on some rivers come either in early spring (snow melt) or early summer (heavy rains), and so on. Until a method of separating such distributions into two normal components (as can be done for the linear case, as explained in Sections 1.5 and 1.6) bimodal distributions must be regrouped into broader classes (sectors) to form a unimodal distribution for comparison with the circular normal curve.

Whether indices of skewness and kurtosis can be developed for the

TABLE VII. Radius vectors (ordinates) of the circular normal probability function. Tabular values are square roots of distances from center (pole) at indicated angles α from mean (resultant) to a curve of unit total area. To obtain curve for comparison with observed distribution of n observations divided among w sectors and plotted according to the square roots of their observed frequencies by sectors, tabular values must be multiplied by $\sqrt{n/w}$. When equivalent polar paper (square root) is used, tabular values should be squared.

α = angle from mean or resultant

k	Mean	±10°	±20°	±30°	±40°	±50°	±60°	±70°	±80°	±90°	±100°	±110°	±120°	±130°	±140°	±150°	±160°	±170°	180°
0.2	1.100	1.098	1.093	1.085	1.074	1.061	1.046	1.030	1.012	0.995	0.978	0.962	0.946	0.933	0.922	0.912	0.906	0.902	0.900
0.4	1.197	1.194	1.183	1.166	1.143	1.115	1.084	1.050	1.015	0.980	0.947	0.916	0.887	0.862	0.841	0.824	0.812	0.805	0.803
0.6	1.292	1.286	1.269	1.241	1.204	1.160	1.112	1.060	1.008	0.957	0.908	0.864	0.824	0.789	0.760	0.738	0.722	0.712	0.709
0.8	1.381	1.373	1.348	1.309	1.258	1.197	1.131	1.062	0.992	0.926	0.864	0.808	0.758	0.716	0.682	0.655	0.636	0.624	0.621
1.0	1.465	1.454	1.422	1.370	1.304	1.226	1.141	1.054	0.969	0.889	0.815	0.749	0.692	0.644	0.606	0.576	0.556	0.543	0.539
1.2	1.543	1.529	1.489	1.424	1.341	1.246	1.143	1.040	0.940	0.847	0.763	0.690	0.628	0.576	0.535	0.504	0.482	0.469	0.465
1.4	1.616	1.599	1.549	1.471	1.372	1.258	1.139	1.019	0.906	0.802	0.711	0.632	0.565	0.512	0.469	0.438	0.416	0.403	0.398
1.6	1.682	1.662	1.603	1.511	1.395	1.264	1.128	0.994	0.869	0.756	0.658	0.575	0.507	0.452	0.410	0.378	0.356	0.344	0.340
1.8	1.744	1.720	1.652	1.546	1.413	1.264	1.112	0.965	0.829	0.709	0.606	0.521	0.452	0.398	0.356	0.325	0.304	0.292	0.288
2.0	1.800	1.773	1.695	1.575	1.425	1.260	1.092	0.932	0.788	0.662	0.557	0.470	0.402	0.348	0.308	0.279	0.259	0.247	0.244
2.2	1.853	1.822	1.734	1.599	1.432	1.251	1.069	0.898	0.747	0.617	0.509	0.423	0.356	0.304	0.266	0.238	0.219	0.209	0.205
2.4	1.901	1.867	1.769	1.619	1.436	1.238	1.043	0.863	0.705	0.573	0.465	0.380	0.314	0.265	0.228	0.203	0.185	0.176	0.172
2.6	1.947	1.908	1.800	1.635	1.436	1.223	1.016	0.828	0.665	0.530	0.423	0.340	0.277	0.230	0.196	0.172	0.156	0.147	0.145
2.8	1.989	1.947	1.828	1.649	1.433	1.206	0.988	0.792	0.625	0.490	0.385	0.304	0.244	0.199	0.168	0.146	0.132	0.124	0.121
3.0	2.029	1.983	1.853	1.659	1.428	1.187	0.958	0.756	0.587	0.453	0.349	0.271	0.214	0.173	0.143	0.123	0.111	0.103	0.101
3.2	2.066	2.016	1.876	1.667	1.421	1.167	0.928	0.721	0.551	0.417	0.316	0.241	0.187	0.149	0.122	0.104	0.093	0.086	0.084
3.4	2.102	2.048	1.897	1.673	1.412	1.145	0.898	0.687	0.516	0.384	0.286	0.215	0.164	0.129	0.104	0.088	0.078	0.072	0.070
3.6	2.135	2.078	1.916	1.678	1.401	1.123	0.868	0.653	0.482	0.353	0.258	0.191	0.144	0.111	0.089	0.074	0.065	0.060	0.058
3.8	2.167	2.106	1.933	1.680	1.390	1.099	0.838	0.621	0.451	0.324	0.233	0.169	0.125	0.096	0.076	0.063	0.054	0.050	0.048
4.0	2.198	2.132	1.948	1.681	1.377	1.076	0.809	0.590	0.421	0.297	0.210	0.150	0.109	0.082	0.064	0.053	0.045	0.042	0.040

circular case remains to be seen, as well as many other applications analogous to those of the linear normal curve. Obviously, there are opportunities for many people to develop this new branch of statistics, of such potential value to geophysics.

At present the circular normal theory affords only (1) a measure of the concentration of a circular variable about its resultant and (2) a normal function to which the observations can be compared qualitatively. Yet its development is of benefit to geophysics simply by pointing out that "average" times or dates should be computed vectorially, as are "average" directions, and that circular variables cannot be analyzed adequately by the linear methods of classical statistics.

List of Symbols and Notation

(Section in which first usage is made shown in parentheses)

- a parameter of distribution of extreme values (2.5)
- b factor defining interval of occurrence of extreme value (2.10)
- E excess of distribution $= \nu_4/\sigma^4 - 3$ (1.6)
- e base of natural logarithms $= 2.7182818284$ (2.4)
- $F(x)$ cumulative probability function (1.3)
- $f(x)$ probability density function $= F'(x)$ (1.3)
- H number of occurrences of an event (2.1)
- h factor defining confidence band of extreme values (2.8)
- $I_0(k)$ Bessel function of first kind of zero order for pure imaginary argument (3.2)
- K frequency factor in frequency analyses (2.7)
- k parameter of circular normal distribution (3.2)
- M mean of a bimodal distribution; M_1, M_2 means of components (1.6)
- m_1, m_2 departures of component means from common mean (1.6)
- N size of sample: number of observations in bimodal distribution (1.6); number of trials (2.1); number of observed extremes (2.5); desired lifetime (2.4)
- n number of values in each sample from which extreme is taken (2.5)
- p probability of occurrence (2.3)
- q probability of non-occurrence $= 1 - p$ (2.3)
- r ratio of return period to number of trials $= T/N$ (2.3); of design return period to desired lifetime $= T_d/N$ (2.4)
- s_x standard deviation of x (2.7)
- T return period of an event $= 1/p$ (2.1); T_d = design return period (2.4)
- t normalized deviate of a variable $= (\bar{x} - x)/\sigma$ (1.6)
- w number of sectors of circular distribution (3.4)
- x a variable; an extreme value (2.5)
- y ordinate (1.6); reduced variate of extreme value function (2.7); $\bar{y}_N$ = theoretical mean (2.7)
- Z_T factor to obtain extreme expected in T years (2.10)
- z difference between variances of bimodal distribution and of components $= \sigma_1^2 - \sigma^2$ (1.6)
- α angle of circular distribution measured from mean $\bar{\alpha}$ (3.2)
- α_3 ν_3/σ^3 = skewness (1.6)
- Δ departure of extreme value from expected (2.8)

ν_3, ν_4 third and fourth moments (1.6)
π 3.1415926535 (1.6)
ϕ function: of t (1.6)
Φ of extreme value x (2.5); of circular distribution (3.2)
σ standard deviation (1.6); σ_N = theoretical standard deviation (2.7)

Notation

x a variable (2.3)
x_N Nth value of x (2.3)
$\bar{x}$ mean of all values of x; $\bar{x}_N$ = mean of N values of x (2.7)
$\tilde{x}$ median of all values of x (2.3)
$\hat{x}$ mode of all values of x (2.4)
$\grave{x}$ largest of all values of x; $\grave{x}_N$ = largest among N values (2.3)
$\acute{x}$ smallest of all values of x; $\acute{x}_N$ = smallest of N values (2.3)
x' first derivative of x (1.3)
$x!$ factorial $x = 1 \cdot 2 \cdot 3 \cdot 4 \cdot \ldots x$ (2.3)
$\doteq$ approximately equal (2.4)
$\rightarrow$ approaches as a limit (2.3)
log natural logarithm (2.4)

References

1. Conrad, V., and Pollak, L. W. (1950). Methods in Climatology. Harvard Univ. Press, Cambridge, Mass., 459 pp.
2. Panofsky, H. A. (1949). Significance of meteorological correlation coefficients. *Bull. Am. Meteorol. Soc.* **30,** 326–327.
3. Kalinske, A. A. (1946). On the logarithmic-probability law. *Transact. Am. Geophys. Un.* **27,** 709–711.
4. Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. Paper No. 1308, *Transact. Am. Soc. Civil Eng.* **77,** 1539–1669.
5. Whipple, G. C. (1916). The element of chance in sanitation. *J. Franklin Inst.* **182,** 37–59, 205–227.
6. Kottler, F. (1950). The distribution of particle sizes. *J. Franklin Inst.* **250,** 339–356, 419–441. (1951). The goodness of fit and the distribution of particle sizes. *ibid,* **251,** 499–514; 617–641.
7. Powell, R. W. (1943). A simple method of estimating flood frequency. *Civ. Eng.* **13,** 105–107. Discussion: W. E. Howland, Estimating flood frequencies, pp. 185; E. J. Gumbel, Exceedance or recurrence intervals in analyzing flood discharge, p. 438.
8. Berkson, J. Codex Book Co. (Norwood, Mass.), paper No. 32,450, logistic ruling; paper no. 32,451, normal ruling; paper no. 34,455, range ruling. Cf. Gumbel, E. J. (1947). The distribution of the range. *Ann. Math. Stat.* **18,** 384–412.
9. Kimball, B. F. (1946). Assignment of frequencies to a completely ordered set of sample data. *Transact. Am. Geophys. Un.* **27,** 843–846. Also discussion by E. J. Gumbel and B. F. Kimball, (1947). *ibid.* **28,** 951–953.
10. Landsberg, H. E. (1946). Note on the frequency distribution of geothermal gradients. *Transact. Am. Geophys. Un.* **27,** 549–551. Discussion: H. Cecil Spicer and H. Landsberg (1947). **28:** 493–494.
11. Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Transact. Roy. Soc. London* **185:** 71–110.

12. Edgeworth, F. Y. (1899). On the representation of statistics by mathematical formulae (Part II). *J. Roy. Stat. Soc.* **62**: 125–140.
13. Pearson, K. (1901). On some applications of the theory of chance to racial differentiation. *Phil. Mag.* 6th ser. **1**, 110–124.
14. Charlier, C. V. L. (1905). Researches into the theory of probability. *Lunds Univ. Årsskrift*, ny följd, afdelningen 2, Vol. 1, No. 5, 51 pp.
15. Court, A. (1949). Separating frequency distributions into two normal components. *Science* **110**, 500–501.
16. Langbein, W. B. (1949). Annual floods and the partial-duration flood series. *Transact. Am. Geophys. Un.* **30**, 879–881. Discussion: Ven Te Chow and W. B. Langbein (1950). **31**, 939–941.
17. Rietz, H. L. (1927). Mathematical Statistics. Carus-Mathematical Monograph No. 3, Mathematical Association of America, Open Court Pub. Co., Chicago, Illinois, 181 pp.
18. von Mises, R. (1941). Probability and statistics. *Ann. Math. Stat.* **12**, 191–205. Doob, J. L. Probability as measure. *ibid.* 206–214. Also disc. by both, 215–217.
19. Uspensky, J. V. (1937). Introduction to Mathematical Probability. McGraw-Hill, New York, 411 pp. (ref. on pp. 103–107).
20. Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413.
21. Eisenhart, C., Hastay, M. W. and Wallis, W. A. (Editors). (1947). (Selected) Techniques of Statistical Analysis. McGraw-Hill, New York, 473 pp. (ref. on pp. 332–333). David, F. N. (1951). Probability Theory for Statistical Methods. Cambridge Univ. Press, London, p. 78. Dixon, F. N., and Massey, F. J. Jr., (1951). Introduction to Statistical Analysis. McGraw-Hill, New York, 322–3.
22. Snedecor, G. W. (1946). Statistical Methods. 3rd ed. The Collegiate Press, Ames, Iowa, 485 pp.
23. Thomas, H. A., Jr. (1948). Frequency of minor floods. *J. Boston Soc. Civ. Eng.* **35**, 425–442. Reprinted as No. 466 of Publications from the Graduate School of Engineering, Harvard Univ., Cambridge, Mass.
24. Linsley, R. K., Jr., Kohler, M. A. and Paulhus, J. H. (1950). Applied Hydrology. McGraw-Hill, New York, 689 pp. (ref. on pp. 548–550).
25. Gumbel, E. J. (1935). Les valeurs extrèmes des distributions statistiques. *Ann. Inst. Henri Poincaré* **5**, 115–158.
26. Kendall, M. G. (1947). The Advanced Theory of Statistics, Vol. I, 3rd ed. Charles Griffin and Co., London, 457 pp. (ref. on pp. 217–224).
27. Gumbel, E. J. (1941). The return period of flood flows. *Ann. Math. Stat.* **12**, 163–190. (1941). Probability-interpretation of the observed return-periods of floods. *Transact. Am. Geophys. Un.* **22**, 836–849, Discussion: Ralph W. Powell, **22**, 849–850. (1942). Statistical control-curves for flood-discharges. *ibid.* **23**, 489–500, Discussion: Bradford F. Kimball, **23**, 501–509. (1943). On the plotting of flood-discharges. *ibid.* **24**, 699–716, Discussion: B. F. Kimball and R. S. Goodridge, **23**, 716–719. (1942). The frequency distribution of extreme values in meteorological data. *Bull. Am. Meteorol. Soc.* **23**, 95–105. (1943). Statistical analysis in hydrology. *Proc. Am. Soc. Civ. Engs.* **69**, 995–1005.
28. Potter, W. D. (1949). Simplification of the Gumbel Method for Computing Probability Curves. U. S. Dept. of Agriculture, Soil Conservation Service (SCS-TP-78), Washington, D. C., 22 pp.
29. Epstein, B. (1948). Statistical aspects of fracture problems. *J. Appl. Phys.* **19**, 140–147.

30. Press, H. (1949). The Application of the Statistical Theory of Extreme Values to Gust-load Problems. National Advisory Committee for Aeronautics, Washington, D. C., Technical Note 1926, 43 pp.
31. Anonymous (1951). Evaluation of Climatic Extremes. Research and Development Branch, Military Planning Division, Office of The Quartermaster General. Washington, D. C., Environmental Protection Section Report No. 175, 36 pp.
32. Gumbel, E. J. (1945). Simplified plotting of statistical observations. *Transact. Am. Geophys. Un.* **26,** 69–82; (1945). Studies on the extremes of statistical variates. Yearbook Am. Phil. Soc. **1944,** 140–141; (1945). Floods estimated by probability method. *Eng. News Record* **134,** 97–101. (1946). *Forecasting Floods. ibid.* **135,** 96. (1948). The Statistical Forecast of Floods. Columbus, Ohio Water Resources Board, 21 pp.
33. Zoch, Richmond, T. (1949). The Climatic Handbook for Washington, D. C. U. S. Dept. of Commerce, Weather Bureau, Technical Paper No. 8, Washington, D. C., 235 pp.
34. Chow, V. T. (1951). A general formula for hydrologic frequency analysis. *Transact. Am. Geophys. Un.* **32,** 231–237.
35. Gumbel, E. J. (1950). The cyclical normal distribution (abstract). *Ann. Math. Stat.* **21,** 143. (1952). The circular normal distribution: applications. *J. Am. Stat. Assn.*, in press.
36. Gumbel, E. J., Greenwood, J. A., and Durand, David. (1952). The circular normal distribution: Theory and tables. *Ann. Math. Stat.*, in press.
37. Leighly, J. (1928). Graphic studies in climatology. II. The polar form of diagrams in the plotting of the annual climate cycle. *Univ. Calif. Publ. Geog.* **2,** 387–407.

Studies of the General Circulation of the Atmosphere

BERT BOLIN

University of Stockholm, Sweden

CONTENTS

		Page
1.	Introduction	87
2.	The Mean State of the Motion of the Atmosphere and Its Seasonal Variations	89
3.	Basic Physical Principles Governing the General Circulation of the Atmosphere	91
4.	The Momentum Balance in the Atmosphere	95
5.	Some Basic Principles for the Energy Balance in the Atmosphere	102
6.	Fluctuations in the Circulation of the Atmosphere. The Index Cycle	103
7.	Principal Aspects of the Approach to a Theory for the General Circulation of the Atmosphere	107
8.	The Barotropic Model	109
9.	The Baroclinic Model	113
10.	Effects of the Non-uniformity of the Surface of the Earth	114
List of Symbols		116
References		116

1. INTRODUCTION

Before we try to present a review of recent research in the field of "the general circulation of the atmosphere," it is necessary to discuss the very meaning of that expression in some detail. A clear formulation of the problems involved will often facilitate the discussion and in some cases even eliminate disagreements that occasionally appear in meteorological literature. Furthermore, in any attempt to give a systematic and consistent picture of a general subject like the present one, it is important to keep the basic questions in mind and proceed from the most fundamental towards more special problems. Such a discussion can be carried through in many different ways, to a certain extent because different opinions exist about what the most fundamental questions are. The presentation below is based on a large number of investigations published recently and may to that extent represent general ideas in the field. In case it disagrees with other contributions (known or unknown to the writer) it is hoped, it will give rise to a fruitful discussion.

We will here consider the problem of the general circulation of the atmosphere as *the description and explanation of the large-scale (time and*

space) behavior of the atmosphere. The subject may be divided as follows:

1) Description of the mean motion of the atmosphere and its seasonal variations as well as characteristic fluctuations in that mean stage.
2) Discussion of basic physical factors that determine the character and scale of atmospheric motion.
3) The momentum balance of the atmosphere.
4) The energy balance of the atmosphere.
5) A rational discussion of the approach towards a dynamic theory for the general circulation of the atmosphere leading up to
6) A model of the atmosphere which is capable of explaining observed conditions as discussed under 1–4.
7) Irregularities in the mean motion of the atmosphere caused by non-uniformity of the surface of the earth (the influence of large mountain barriers and the distribution of land and sea).

A few brief comments should be made here. When discussing mean conditions, it is important to remember that very significant features of the behavior of the atmosphere may be eliminated by an ordinary averaging process. Therefore a close connection must exist between the description and interpretation of the observed conditions. Under certain circumstances it may, for instance, be more instructive to consider extreme conditions than mean values. An excellent illustration to such a case is low and high index (see 6). Of course mean patterns are of interest also in these cases, but they must be computed over periods chosen in relation to the particular phenomenon under study. The consideration of basic physical factors that determine the general character of the motion of the atmosphere therefore should serve as a frame for the discussion of empirical investigations. Having thus obtained a broad picture of the actual behavior of the atmosphere we can proceed to the construction of a simplified model which can be treated theoretically. However, the discussion here must necessarily be incomplete as many of the most fundamental problems have not yet been solved. It is therefore impossible to arrive at a final model of the atmospheric circulation, but we shall analyze the different approaches that have been made.

There are essentially two different models of the atmospheric circulation that have been considered. One is an extension and modification of Hadley's idea of meridional circulation cells [1]. His original idea applies to the trade wind regions: Differential heating between different latitudes gives rise to ascent of air in tropical regions, which leads to equatorward flow at lower levels and flow away from the equator aloft. The combined effect of surface friction and the rotation of the earth

deflects these currents and thus easterly winds are obtained at lower levels and westerlies aloft. From this idea the well-known cellular model of the circulation of the whole atmosphere has been derived [2, 3]. In the other model the daily horizontal circulation patterns are supposed to be the most fundamental features of the motion of the atmosphere for the determination of mean conditions, in other words, the daily disturbances are not merely superimposed perturbations but integral parts of the general circulation. Some aspects of these problems have been considered by Rossby [4], but recent investigations have yielded new information that make a revision of previous ideas necessary. A large portion of this review will be devoted to a discussion of these two models as their relative importance seems to be one of the most basic problems in our attempts to form a consistent picture of the general circulation of the atmosphere.

2. The Mean State of the Motion of the Atmosphere and Its Seasonal Variations

The large-scale features of the motion of the atmosphere are fairly well-known for the northern hemisphere. Lack of data has prevented more detailed investigations of the southern hemisphere.

The mean zonal flow averaged all around the northern hemisphere is shown in Fig. 1 [after 5]. As is well-known from a series of investigations during the last six or seven years, a strong westerly current is found in middle latitudes at about 200 mb. The latitude as well as the intensity of this planetary jet varies in the course of the year. In July there is a maximum mean flow of 20 meters/second at latitude 42°N, while in January the maximum is at latitude 27°N and the speed averages to about 40 meters/second. Westerly winds exist at the surface of the earth from about 30°N to the North Pole (except for a narrow band around 60°N, in winter). Conditions in the vicinity of the North Pole are based on recent Russian investigations [6]. To the south of 20°N easterly winds become predominant. Thus maximum easterlies are found at about 150 mb with a speed averaging 10–15 meters/second, somewhat to the north of the equator in summer, at (or slightly to the south of) the equator in (Northern Hemisphere) winter. At the surface of the earth easterly winds in the mean exist within the whole tropical belt 30°N to 30°S except for possibly a small area just to the north of the equator during August–October. In the Southern Hemisphere the intensity of the zonal flow seems to be consistently higher and furthermore there are indications of a split of the westerlies into two currents [7]. However, it is doubtful if that represents mean conditions around the whole southern hemisphere.

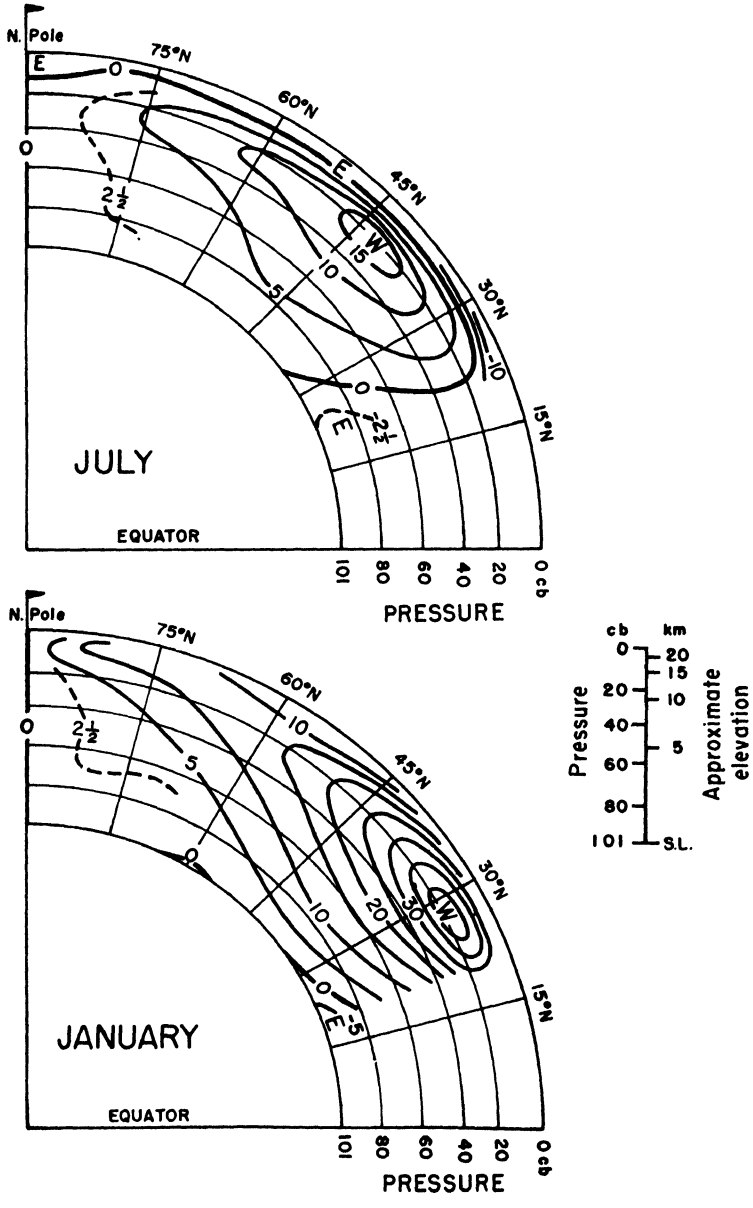


FIG. 1. Mean cross-section through the atmosphere in July and January (after Mintz and Dean, 5). Wind speed in meters/second.

There are considerable variations of the Northern Hemisphere westerly jet-stream from one part of the hemisphere to the other (Fig. 2). The maximum values are found off the Pacific Coast of Asia with a maximum mean velocity of about 60 meters/second in winter and 25 meters/second in summer (the summertime maximum being displaced further out over the Pacific Ocean) and also over the eastern part of the United States, where the maximum values are 50 and 25 meters/second respectively [8]. Furthermore the planetary jet is not a completely zonal current but long waves with fairly small amplitudes are superimposed. In both summer and winter, for example, troughs are located off the Asiatic east coast and over the eastern United States and there is a tendency for a similar trough over central Europe. In summer a fourth trough in the mean is found over the eastern Pacific (cf. Fig. 2). In the southern hemisphere the flow in the mean seems to be more zonal but the existence of a trough to the east of the South American continent has been fairly well established [9].

The patterns described here represent mean conditions and individual years may be very different. Similarly, the transition from summer to winter conditions and vice versa is not a regular displacement and intensification of the described centers, but is sometimes rapid and on other occasions slow [10].

Only a very broad description of the mean flow of the atmosphere has been given here. Details will be given in the course of the following discussion, after some physical concepts have been developed.

3. Basic Physical Principles Governing the General Circulation of the Atmosphere

As long as we disregard the inhomogeneity of the earth, there is no reason for different conditions in the two hemispheres.[1] Therefore all discussions applying to such an idealized system will be carried out for the Northern Hemisphere where data are more reliable. After having established the basic principles of the circulation of the atmosphere, we will consider the effects of the non-homogeneity of the earth and discuss the differences between the two hemispheres.

The following factors seem to be of fundamental importance in determining the general behavior of the atmosphere:

1) The troposphere is a very thin spherical shell, whose vertical dimension is of the order of magnitude of one thousandth of the radius of the earth. This has the effect that in the large-scale motion of the

[1] The difference between the two hemispheres that exist because of the elliptic character of the orbit of the earth around the sun is disregarded here.

atmosphere (horizontal scale comparable with the radius of the earth) one may expect the vertical velocities to be approximately $\frac{1}{1000}$ of the horizontal velocities. They will be of the order of magnitude of a few centimeters per second. However, this does not mean that the vertical velocities are dynamically unimportant. It is merely a characteristic feature of the system that they are small.

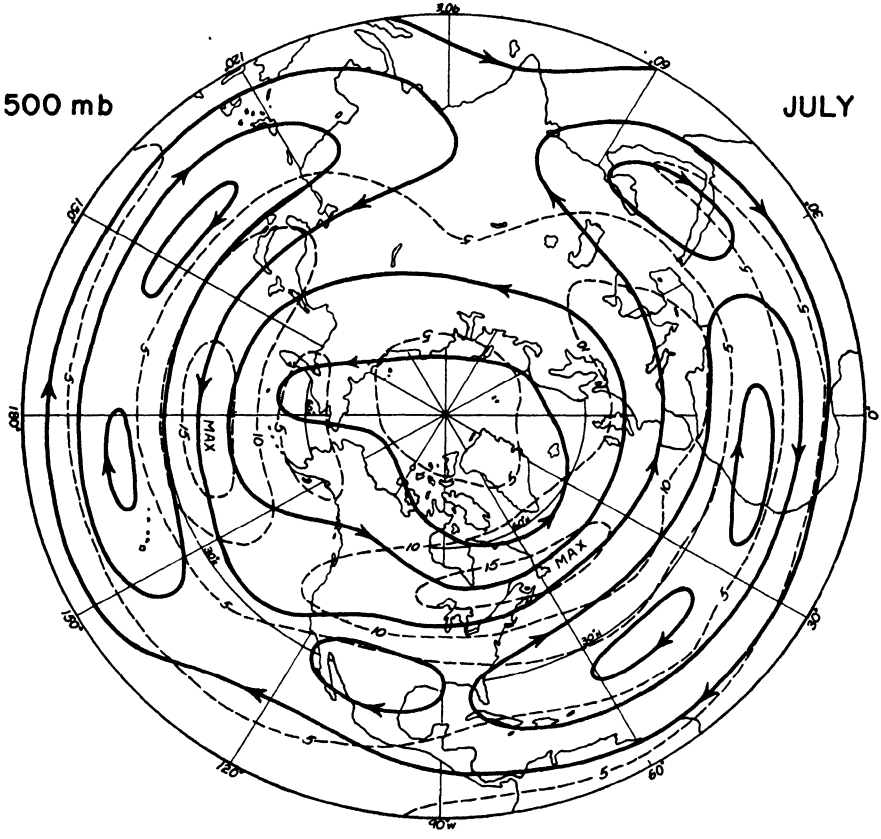


FIG. 2a. Streamlines and isovels of the mean wind at the 500 mb level in July. The wind shown north of lat. 25°N is the mean geostrophic wind; south of lat. 25°N it is the mean pilot balloon wind. Speed in meters per second (after Mintz and Dean [5]).

2) In tropical and sub-tropical regions the atmosphere receives more heat by radiation than it loses while the opposite is true for temperate and polar regions. This gives rise to a *horizontal* temperature contrast between low and high latitudes. In tropical regions there exists a similar distribution of heat and cold sources in the *vertical* in that the lower troposphere gains heat (by contact with the ground and by condensation)

while the upper troposphere loses heat (by radiation). On the average this non-balance between the amount of heat received and given off by different portions of the atmosphere through radiation and small scale convection must be balanced by a transport of heat by the large-scale exchange processes. These large-scale processes must also provide for the maintenance of their own kinetic energy (as well as that of the mean zonal motion) against dissipation by friction. This occurs by the conversion of potential energy to kinetic energy that takes place in connec-

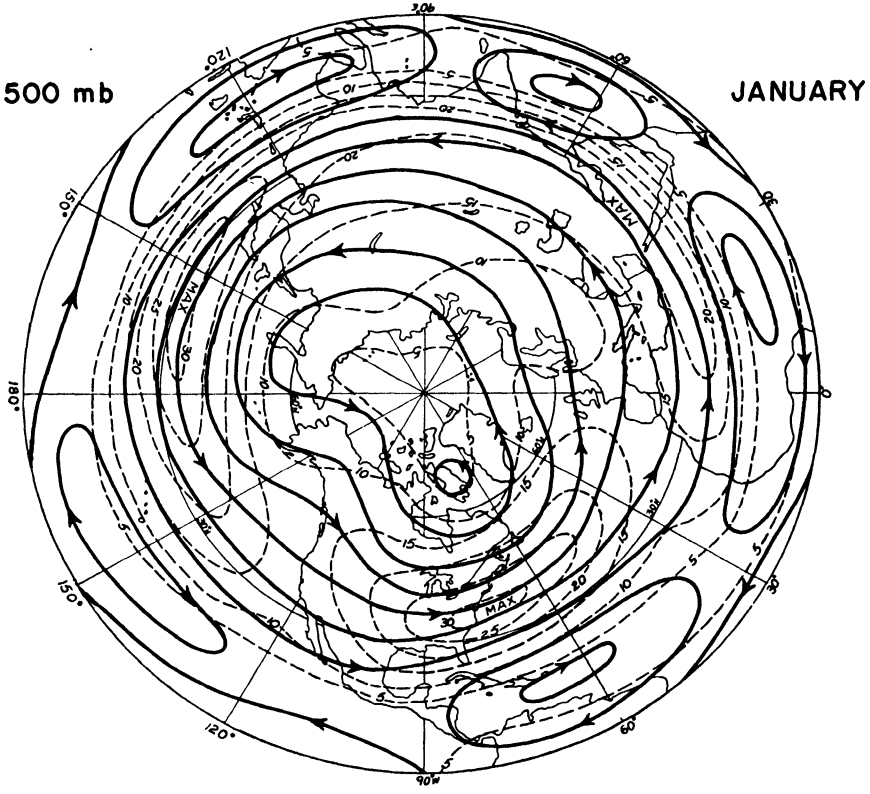


FIG. 2b. Streamlines and isovels of the mean wind at the 500 mb level in January. Cf. Fig. 2a.

tion with these exchange processes. The atmosphere acts as a gigantic heat engine. From this one may not immediately conclude that the characteristic velocity of the flow is completely determined by this differential heating. The flow created must be stable from a dynamic point of view in order to exist for any considerable length of time. Thus the combined effect of thermal processes and dynamic constraints defines the characteristic velocity (U) for the horizontal motion.

3) We have to consider next the rotation of the earth as being one of the most basic dynamic features of the system we are dealing with. It can be demonstrated in various ways that the rotation of a fluid has a pronounced influence upon its motion. Taylor [11] discussed some such problems already thirty years ago and laboratory experiments have recently been taken up by Fultz and Long [12, 13] in order to study these phenomena more closely. Taylor points out that there is a pronounced difference between two-dimensional and three-dimensional flow. In the former case the forces that are introduced by the rotation of the fluid may be completely compensated by pressure forces and in principle the motion may be the same irrespective of whether the fluid is rotating or not. This is not true if the motion is three-dimensional. In such a case a non-balanced Coriolis force is always present. With these facts in mind it is now interesting to observe that the relative motions in a rotating fluid tend to be two-dimensional, provided the vorticity of the relative motion is small compared to the vorticity of the basic rotation. Taylor also presents some considerations that explain this fact. Thus a theoretical discussion of the motion of a rotating fluid may in some cases be simplified considerably from the very beginning by assuming that the flow will be two-dimensional.

It seems reasonable that a similar effect of the rotation of the earth is present in the atmosphere but possibly modified by the tendency for vertical heat exchange, in particular at lower latitudes. The motion is already kinematically restricted to take place in a spherical shell and thus the important parameter for the quasi-horizontal motion in that shell is the *vertical* component of the rotation of the earth or, what in principle is the same, the vertical component of the vorticity of the earth: $f = 2\Omega \sin \varphi$. We notice that this quantity varies with the latitude φ. It has its maximum at the pole and becomes zero at the equator. In analogy to the model experiments referred to above we therefore might expect the motion to be quasi-two-dimensional at higher latitudes, while more pronounced variations of the motion from layer to layer should exist in tropical regions. The relatively successful approach to the forecasting problem in middle latitudes using the two-dimensional model of the atmosphere [14] supports the idea that the dynamics of atmospheric motions in these latitudes to a large extent is controlled by the rotation of the earth. This similarity of the motion from layer to layer in the atmosphere (as well as in the oceans) was already realized by Ekman [15] and he formulated a theorem: "*Der Satz der parallelen Solenoidfelder,*" which also calls the attention to the rotation of the earth as being of fundamental importance for the explanation of this similarity. On the other hand recent investigations of tropical conditions [16] clearly show

that the flow at upper levels seem to be almost independent of conditions at lower levels, which is in qualitative agreement with the discussion above. Further evidence for this distinction between low and high latitudes will be given in the following section.

In middle latitudes, where the motion seems to be controlled by the rotation of the earth we may expect that the Coriolis parameter is a fundamental quantity in determining the scale of the atmospheric flow patterns. From the two parameters U (discussed under 2) and f we now are able to obtain a characteristic length scale, viz., $\lambda_1 = U/f$. This quantity is of the order of magnitude of 200 km. for proper values of U and f in middle latitudes. λ_1, the radius of the inertia circle, plays an important role in the theory for cyclone waves. The size of the cyclones essentially is determined by the rotation of the earth. The energy of these waves, however, depends upon the baroclinic character of the atmosphere.

An important dynamic consequence of the spherical shape of the earth is the fact that the vertical component of the rotation of the earth varies with latitude. One may thus expect the variation of the Coriolis parameter, $\beta = \partial f/\partial y$, to be of importance for the large scale motion of the atmosphere, that is for systems where f really varies significantly between different parts of the same system. As is well-known from the work by Rossby and collaborators [17] this is the case and the length scale that can be constructed from U and β, $\lambda_2 = \sqrt{U/\beta}$, seems to put certain limits on the size of the planetary waves. This agrees with what is observed in reality. Again the energetics of the waves in some way is related to the baroclinicity of the atmosphere.[2]

4. The Momentum Balance in the Atmosphere

The considerations in the previous section were largely qualitative. In the light of direct measurements two particular aspects will be examined more closely in Secs. 4 and 5, viz., the energy and momentum balance. It was pointed out that the non-uniform heating of the earth and the atmosphere is the ultimate cause of the motion of the atmosphere. The temperature distribution is not the one corresponding to radiational equilibrium in that the contrast between equatorial and polar regions is less than what corresponds to such a balance. That means that before this (imaginary) state is ever reached the flow changes its character in such a way that there is a net transport of energy from tropical regions towards the pole. In a similar way a momentum transport must take

[2] It should be mentioned that Eady [18] has put forward the idea that the most unstable wavelength is the predominant one. Also in this case β is an important parameter.

place between different latitudes in order to maintain the mean zonal currents in the atmosphere against frictional losses. Thus the first question that comes up is: How do these exchange processes take place?

It is clear from the previous discussion that there are two basic factors that we have to consider when trying to answer this question:

1) The effect of the differential heating, which would cause one or several meridional circulation cells on a non-rotating earth.

2) The effect of the rotation of the earth which would give rise to two-dimensional horizontal flow if the heating of the atmosphere and friction could be neglected.

As the atmosphere is heated as well as rotating both these processes are present and the question is which one is the dominating one. Obviously the answer depends upon the rate of rotation and the strength of the differential heating. This struggle between two different types of flow can be demonstrated experimentally by studying the behavior of a fluid in a rotating dishpan, which is heated at the periphery. Long [13] thus has found that the flow changes fairly abruptly from the essentially meridional to the quasi-horizontal if the rotation exceeds a certain value while the heating is kept constant. The critical value depends upon the heating of the dishpan. In Fig. 3 a photograph is shown of supercritical flow. The flow has striking similarities with flow in middle latitudes in the atmosphere. Some idea of the reason for this change of circulation may be obtained from the following reasoning: If the circulation takes place in meridional planes we may assume as a first approximation that the angular momentum is conserved (neglecting friction). However, the stronger the rotation of the system the stronger is the shear (both vertical and horizontal) that will develop because of the circulation. From a series of investigations of the stability of laminar flow [18–22] we know that the flow becomes unstable if a certain critical value of the vertical or horizontal shear is exceeded giving rise to the development of quasi-horizontal waves (further comments on this instability will be given in Section 7). In the atmosphere the rotation of the earth around a vertical is zero at the equator and increases towards the north. It therefore seems likely that the main exchange processes in low latitudes should be carried out by meridional circulation cells. In middle latitudes, however, we observe the similarity between atmospheric flow and supercritical flow in the dishpan. Furthermore the critical parameters that are obtained from the stability investigations mentioned above are often exceeded. In fact, Charney and Eady found that the westerlies are the seat of a permanent instability. This indicates that horizontal exchange processes are of primary importance in middle and high latitudes. A look at the direct measurements will clarify this point further.

The work on the momentum balance of the atmosphere was initiated by Jeffreys [24] and has lately been taken up by Priestley [29], at Massachusetts Institute of Technology [25] and at the University of California in Los Angeles [26].

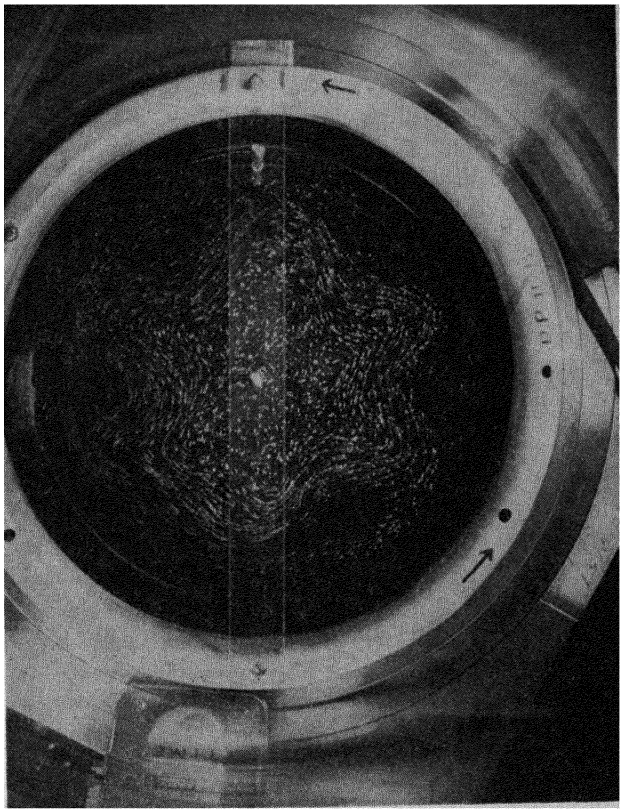


FIG. 3. Streak-flash photo of the dishpan experiment, 5 minutes after the heating at the rim was begun. The temperature difference at the surface between the center and the periphery at the time this picture was taken was about 8°C (after Long [13]).

The angular momentum of a ring of air enclosed between two latitude circles may be changed as a result of three different processes:

1) Transport of angular momentum across the boundaries.

2) Gain or loss of momentum because of frictional interaction between the atmosphere and the earth.

3) Change of momentum because of the pressure exerted by the mountains upon the atmosphere.

Mathematically this may be expressed[3]

$$(1) \qquad \frac{\partial}{\partial t}\int_V \rho M \, dV = \int_S \rho M c_n \, dS + \int_\sigma p \cdot r \cdot d\sigma + \int_S r\tau_x \, dS$$

At the surface of the earth between roughly 30°N and 30°S the prevailing winds are from the east, while we find westerly flow to the north of that area (cf. Section 2). Therefore a positive torque acts on the atmosphere in tropical regions and as the angular momentum of the atmosphere as a whole does not change over long periods of time this gain of momentum at low latitudes must be lost in the middle latitude westerlies. Thus there exists a net transport of angular momentum from equatorial regions towards the north. If we can estimate how large the momentum exchange is between the atmosphere and the earth at different latitudes, we also know the amount of momentum that is carried northward and *vice versa*.

The absolute angular momentum of the atmosphere may be considered as the sum of 1) the momentum due to the motion of the air relative to the earth, and 2) the momentum that depends on the rotation of the earth itself. Accordingly the transport of momentum across a certain latitude (given by the first term on the right hand side of equation (1)) may be considered as the result of two effects, viz., transport of relative momentum given by $R \cos \varphi \int\int \rho uv \, dx \, dz$ and transport of momentum depending on the rotation of the earth: $\Omega R^2 \cos^2 \varphi \int\int \rho v \, dx \, dz$. The latter of these two terms depends only upon the transport of mass across the latitude. This in the long run is zero and it remains merely to consider the transport of relative momentum. This northward flux may be the result of a meridional circulation or horizontal exchange processes. The former is characterized by a net transport of mass towards the equator at lower levels and away from the equator aloft if u increases with height and *vice versa* if it decreases with height. A transport carried out by the horizontal disturbances on the other hand must depend upon a positive correlation between u and v at the same level. Such a correlation gives a characteristic shape of the waves and vortices. Two typical patterns causing a poleward transport of momentum are shown in Fig. 4. By and large these patterns resemble what is observed in the atmosphere.

The most reliable estimate of the surface stress was made by Priestley [29], but his computations are restricted by the fact that only wind data from the oceans were used. Furthermore the relation between wind speed and surface stress is not very well known. The computations of the effect of the mountains also are very uncertain. However, the results are supposed to be accurate enough to permit some general conclusions.

[3] See the List of Symbols at the end of the paper.

Three different methods have been used in estimating the momentum transport that takes place across any one latitude.

1) The meridional circulation cells necessari y are non-geostrophic and by computing the geostrophic flux we can get an estimate of the transport by horizontal exchange. (This is not exact as there may be non-geostrophic components of the flow that are part of the horizontal exchange processes, which are also automatically excluded by this procedure. Presumably their contributions are small.) The evaluation is made from upper air maps and results have recently been published for

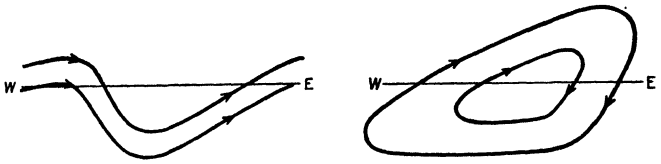


FIG. 4. Two typical horizontal flow patterns that give rise to a large-scale transport of angular momentum towards the north (after Starr [25]).

two different months [27, 28]. Figures 5 and 6 show some of the results. In the computations by Widger an approximate balance against losses by friction is obtained only considering this part of the momentum transport (except at low latitudes), which indicates that the transport by meridional cells is considerably smaller. During the other month the agreement is not so good, but still the major part of the exchange seems to be carried out by horizontal eddies. This method of computing the momentum flux fails to the south of about 20°N, since upper air maps are not available and furthermore the geostrophic assumption becomes questionable.

2) The momentum transport can be evaluated from actual wind data. Thus Priestley [29] found that the meridional cells bring about a considerable transport (approximately 50% of the total transport), but the results are based upon a small number of stations and the representativeness is questionable. Starr and White [30] on the other hand found that the amount of momentum carried northward by meridional cells (at 30°N) is too small to be estimated by this method and probably is less than 10% of the transfer by horizontal eddies. However, the accuracy of wind observations at high levels, where the major flux takes place, is not as good as would be desirable but the results are considerably more conclusive than those of Priestley.

3) An indication of the existence of a mean meridional cell in the tropics can be obtained from surface wind data. Thus Riehl and Yeh [16, 31] have computed the mean meridional surface wind and from that inferred the minimum return flow aloft. They found that even this

meridional cell of minimum intensity is sufficient to account for an approximate momentum balance in these latitudes.

We may conclude: *Measurements carried out so far support the idea that a) horizontal exchange processes take care of the major part of the momentum transport to the north of about 25°N and b) a meridional circulation cell is the predominant mechanism for momentum exchange in the tropics.*[4]

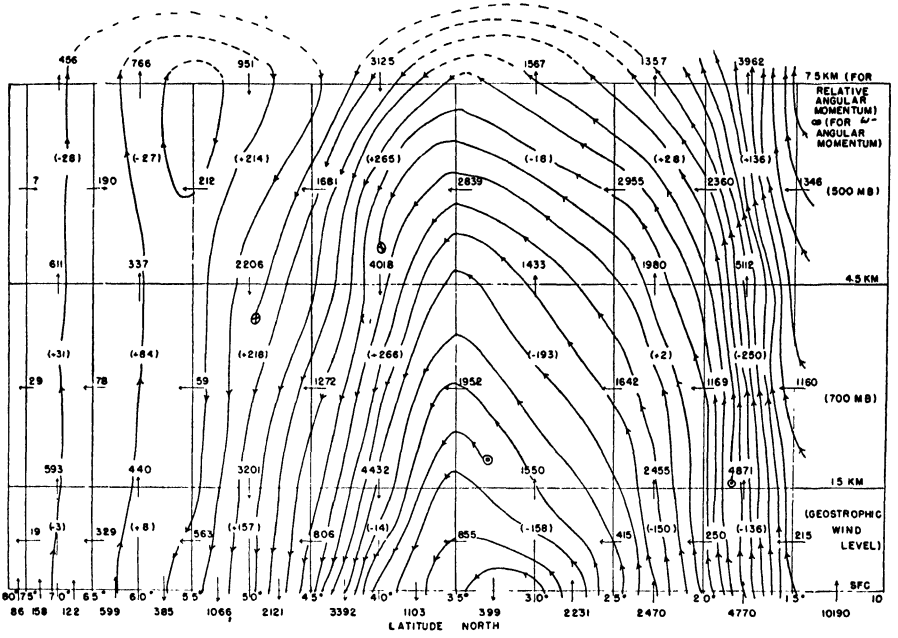


FIG. 5. Net total generation and transport of absolute angular momentum from map time Jan. 1, 1946 to map time Jan. 31, 1946 in CGS-units $\times 10^{-29}$. Small arrows indicate total flow of angular momentum between adjacent vertical or horizontal boundaries. Values in parentheses indicate total change of angular momentum during the month. Streamlines indicate flow of approximately 500×10^{29} CGS-units of angular momentum. Effects of a mean meridional circulation have been disregarded (after Widger [27]).

The transfer of momentum in middle latitudes essentially takes place in the upper troposphere (cf. Fig. 6). In order to complete the picture of this exchange process one must assume an upward transport in the equatorial easterlies and a corresponding downward transport in the middle latitude westerlies. This flux may be the result of small scale turbulence (which certainly is of fundamental importance in the surface

[4] It should be mentioned that Palmén in a recent article [32] stresses the fundamental importance of a meridional circulation in middle latitudes even if the horizontal eddies take care of the major part of the momentum transport.

friction layer) or it may be caused by a certain organization of the vertical motion relative to the horizontal flow patterns. This latter mechanism would mean that in middle latitudes upward motion preferably takes place where the westerlies are weak, while subsidence occurs in connection with strong westerly winds. In the tropics the opposite correlation

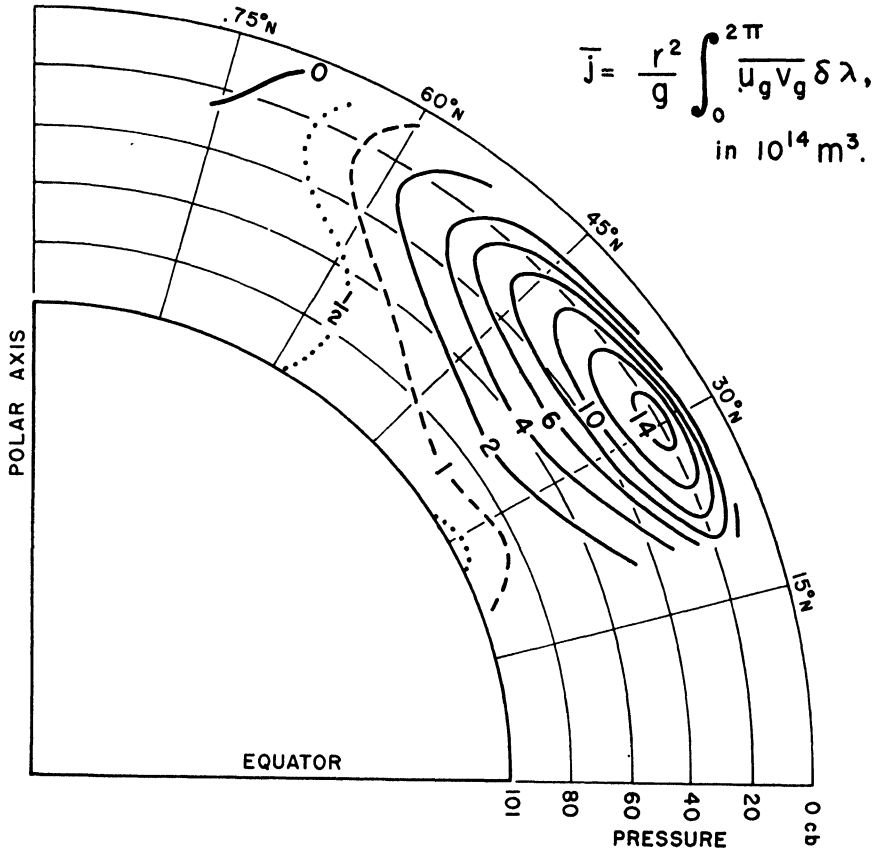


FIG. 6. The mean geostrophic poleward flux of angular momentum, per centibar layer in January 1949 (after Mintz [28]).

between u and w should exist. Some preliminary investigations indicate that such a relation between u and w exists in middle latitudes, but considerably more evidence has to be presented before we can conclude anything definite regarding the dominating process for this vertical momentum transport.

The considerations above are important in that they show how the atmosphere achieves internal dynamic consistency in certain respects,

but they do not *explain* why the existing flow in the atmosphere represents a preferred state. Still, one may get indications of possible approaches to such a dynamic theory for the general circulation. We shall return to these problems in Section 7 and following.

5. Some Basic Principles for the Energy Balance in the Atmosphere

We were able to draw some general conclusions in the last section concerning the momentum balance of the atmosphere. Considerably less is known about the energy balance. The formulation of the balance equations for the total energy as well as for the different energy forms separately (internal energy, kinetic energy and potential energy) has been given by Van Mieghem [33], but very few measurements have been carried out so far. In particular Starr [34] has stressed the importance of studying the different energy forms separately in order to get a better understanding of the internal processes in the atmosphere.

Some measurements of the transport of internal energy have been made by estimating the eddy transport of sensible and latent heat (Priestley [29], White [23]). The total transport of these two energy forms has a maximum in middle latitudes (at $\sim 50°$N) where it approximately balances the loss of energy by radiation. Further to the south, however, the eddy transport of heat is insufficient for the maintenance of the existing temperature distribution. It should also be noted that the eddy transport of sensible heat has a maximum in the middle troposphere, while the transport of latent heat is largest at the surface of the earth, where the moisture content of the air is a maximum.

The balance equation of kinetic energy has been studied in some detail by Starr [34]. The following four processes are responsible for changes of kinetic energy in a given volume:

1) Advection of fluid with new kinetic energy across the boundaries.
2) Work done by the pressure forces on the boundaries of the volume.
3) Production of kinetic energy within the volume.
4) Frictional dissipation.

This may be written

$$\frac{\partial}{\partial t}\int_V E\,dV = \int_S E \cdot c_n\,dS - \int_S p \cdot c_n\,dS + \int_V p \cdot \operatorname{div}_2 \mathbf{V}_2\,dV - \int_V d \cdot dV \tag{2}$$

The first two processes express the interaction between the surroundings and the volume under consideration. Estimates of their sizes show that the first one amounts to only a few per cent of the second, and

consequently may be neglected. The second term is given simply by $R^* \int_S \rho T c_n \, dS$, and is proportional to the transport of internal heat energy. As that transport essentially is from south to north, we may conclude that there is, in the mean, a transport of kinetic energy from equatorial regions towards the poles. The kinetic energy within each latitudinal belt remains constant by and large, and therefore production of kinetic energy takes place at low latitudes and dissipation further toward the poles. The last term in equation (2) always indicates dissipation of energy, and consequently the production is given by the third term in the equation. Examining this term more closely, we notice that at lower levels $\mathrm{div}_2 \mathbf{V}_2 > 0$ where p is large (anticyclones), while $\mathrm{div}_2 \mathbf{V}_2 < 0$ where p is comparatively small (cyclones), giving rise to a net positive contribution. At higher levels in the atmosphere we do not know much about the distribution of divergence and convergence, but the total value of the integral has to remain positive in order to balance the last term. From this analysis one finds that a source region for kinetic energy is found in the subtropical anticyclones, while sinks are found in the belt of cyclones in middle and high latitudes. The difference between this production and destruction of kinetic energy is dissipated by friction. It is interesting to note the similarity between this process and an ordinary heat engine. In both cases the available energy is the difference between a source and a sink, and the energy production cannot take place without both sources and sinks. In a similar way as for a heat engine one may define the efficiency of the process for producing kinetic energy as $(E_+ - E_-)/E_+$, where E_+ is the total intensity of the sources and E_- the total intensity of the sinks. For the atmosphere this ratio is a very small quantity, since the value of p in the regions of divergence (anticyclones) is only a few per cent larger than the value of p in the region of convergence (cyclones).

These considerations give only a few very general ideas about the energy cycle in the atmosphere and these problems need many more detailed investigations before the results can be incorporated in an overall picture of the general circulation of the atmosphere.

6. Fluctuations in the Circulation of the Atmosphere. The Index Cycle

The balanced state that has been discussed in the previous sections refers to mean conditions over long periods of time. For shorter periods such a balance does not exist. In other words, the right side of equations (1) and (2) are not zero and thus an increase and a decrease of momentum and kinetic energy may take place in different latitude belts. In examin-

ing the fluctuations that occur, four time scales of oscillations stand out. One has a period of 1 or 2 days and is associated with the daily cyclones and anticyclones. Another is of the order of magnitude of 5 to 7 days and connected with the long planetary waves in middle latitudes. From the point of view of the general circulation of the atmosphere we are not interested in these individual disturbances themselves, but rather their statistical manifestation in the mean conditions. The mean picture described in the previous sections largely arise as a result of these disturbances. A third oscillation has a period of about 4 weeks and is identical with what has been called the *index cycle*. This phenomenon should in itself be considered as a part of the general circulation and will be discussed in this section. The fourth is identical with the seasonal variations and has been described briefly in Section 2.

The index cycle may be considered as a quasi-periodic redistribution of angular momentum within the hemispheric shell. It has a pronounced influence upon the large-scale flow patterns as well as weather conditions at the surface of the earth. The two extreme conditions, high and low index are described as follows (Rossby and Willett [35]):

a) High index (strong zonal westerlies): The circumpolar vortex is intense, expanding but still located to the north of its normal seasonal latitude. The sea-level westerlies are strong and rapid wave-cyclones move in a west-east direction, but do not give rise to any considerable air mass exchange in north-south direction, in spite of a strong latitudinal temperature gradient.

b) Low index (weak zonal westerlies): Large amplitude troughs and ridges develop in the westerlies and are often cut off, forming cold cyclones in low latitudes and warm anticyclones in high latitudes. The flow at sea level is broken up into a series of semi-stationary cyclones and anticyclones. The maximum temperature gradients are rather in east-west direction from ridge to trough than north-south.

The transition from high to low index is fairly rapid resembling an unstable process with rapidly increasing amplitudes of the waves in the westerlies. The re-establishment is somewhat slower and it is characterized by a dissipation of low latitude cyclones and high latitude anticyclones and a gradual increase of the middle latitude westerlies.

Recent investigations have revealed some further details that are interesting particularly in connection with the consideration of momentum balance given in Section 4. In Fig. 7 the 5-day mean anomaly of angular momentum for each latitude belt (three-quarters of the hemisphere) has been plotted as a function of time (after Riehl *et al.* [10]).

The figures illustrate conditions at 500 mb but essentially the same is observed at 700 and 300 mb. It is found that areas of positive or negative anomalies are systematically displaced from south to north (two-thirds of all cases) or from north to south (one-third of all cases). There seems to be no particular latitude where the anomalies have maximum intensity. The rate of displacement varies between 1 and 5 degrees latitude per day. It should be stressed that the phenomenon described here refers to momentum *deviation* from a seasonal mean and does not necessarily imply a corresponding movement of the mean planetary jet stream itself. With these results in mind the index cycle should rather

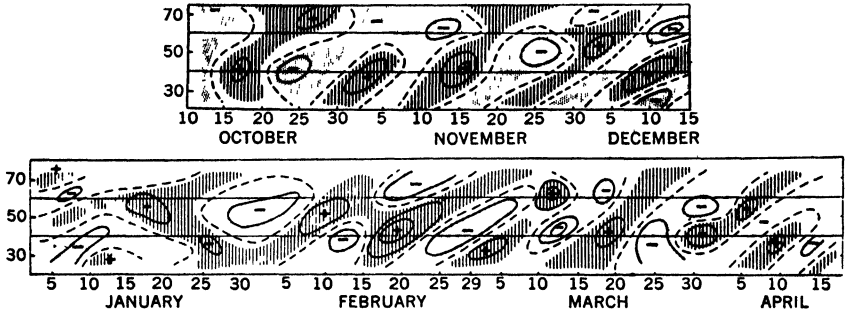


FIG. 7. Time section of the anomaly of angular momentum at 500 mb Oct. 1948–April 1949. Unit is $(4\pi R^2 \times 10^{-4})^{-1}$ CGS. Zero lines are dashed, areas with values of 20 units or more are indicated by vertical hatching, −20 units or less by dotted shading. Solid lines are drawn for intervals of 50 units, positive or negative (after Riehl, Yeh, and La Seur [10]).

be considered as a semiperiodic transfer of momentum from south to north (or north to south) than simply a fluctuation of the middle latitude momentum independent of changes further to the north and south. It should be noted that the total angular momentum for the whole hemisphere remains almost constant (except for seasonal variations). (Cf. also Namias [36].) Thus a low index pattern not only means a decrease of the westerlies in middle latitudes but also an increase of the angular momentum at high and low latitudes. In extreme cases the middle latitude jet-stream is replaced by two currents, one around 60°N and the other between 20 and 35°. This pattern is identical with what is known as blocking action in middle latitudes (cf. Rex [37]) which is the development of one or several warm anticyclones and cold cyclones that block the westerlies and often give rise to easterly currents in middle latitudes throughout the troposphere. It is important to notice that such a pattern very seldom extends around the whole hemisphere but usually is associated with a westerly current in excess of the normal values further

to the west. We here observe a difference between high and low index that may have some bearing on the explanation of the transformation from one stage to the other. High index is characterized by a more or less uniform westerly current all around the hemisphere. Waves exist in this current but do not disturb the principal similarity from one north-south cross section to another. In a low index situation the pattern is

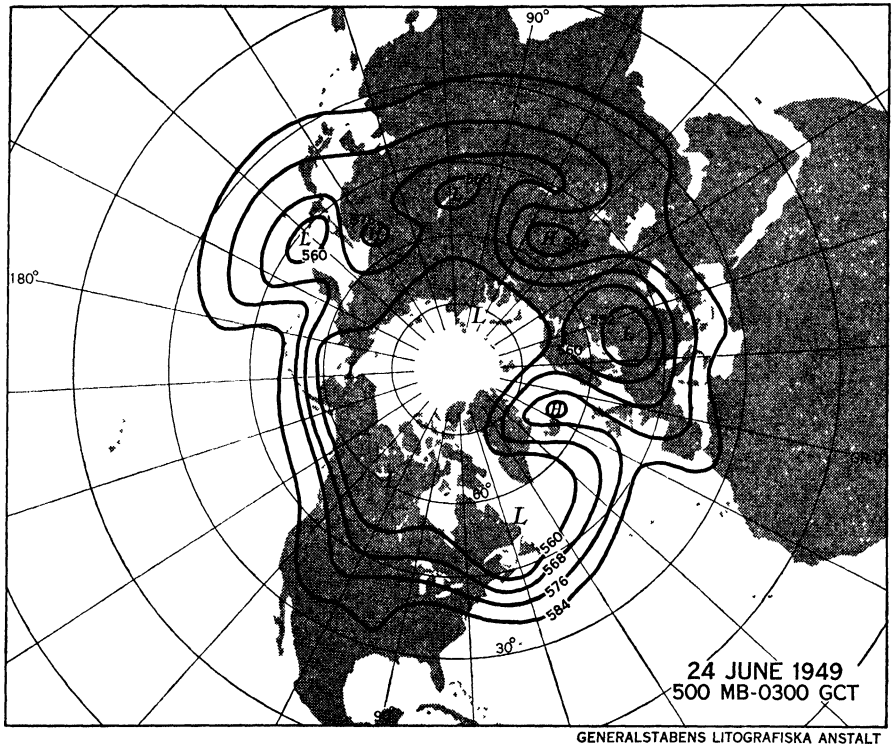


FIG. 8. A clear-cut example of blocking action in middle latitudes, June 24, 1949. Absolute topography of the 500-mb surface is given in decameters (after Rex [37]).

much more irregular in the sense that in some parts of the hemisphere the westerlies are well-developed, in other parts they are replaced by vortex patterns of the type associated with blocking. A clear-cut example of such a low index pattern is shown in Fig. 8. Rex points out another interesting feature of blocking: that there are two preferred regions where blocking is initiated, at 10°W and 150°W. This indicates that the non-uniformity of the earth is of importance for the initiation of blocking and thus also possibly for the generation of low index patterns in general.

7. Principal Aspects of the Approach to a Theory for the General Circulation of the Atmosphere

Because of the complexity of atmospheric motions, there is very little hope for the possibility of deducing a theory for the general circulation of the atmosphere from the complete hydrodynamic and thermodynamic equations. Therefore, the essential problem must be to construct models of the atmosphere that contain the most important characteristics of the behavior of the atmosphere, but still allow a mathematical treatment. This necessarily requires a careful physical discussion of the system we are dealing with. In the preceding sections we have tried therefore to summarize recent investigations of the behavior of the atmosphere in order to obtain a general idea of what phenomena a dynamic theory for the general circulation has to take into account. The problems have been approached from different viewpoints but even the most basic questions have hardly been settled as yet. The following discussions will therefore essentially be qualitative but references will be given to papers that deal with mathematical developments that have been put forward. A very enlightening article on this subject has recently been published by Eady [38].

The previous discussion seems to indicate that we cannot use the same model of the atmosphere for low and high latitudes. In equatorial regions the meridional circulation cell proposed by Hadley seems to be a useful first approximation of actual conditions. In middle and high latitudes, on the other hand, the eddy motion is of predominant importance for the exchange processes. The discussion in this and the following sections will mainly be devoted to a more detailed analysis of this eddy motion. We will also discuss the attempts that have been made to explain the character of the flow in middle latitudes as a result of this eddy motion.

The use of the expressions "eddy transfer" and "eddy motion" when discussing momentum and energy exchange between low and high latitudes already implies some similarity between the over-all behavior of the atmosphere and turbulent motion in the ordinary sense. This concept of large-scale horizontal turbulence in the atmosphere was introduced by Defant [39]. He defined an eddy coefficient and also gave estimates of the magnitude of that coefficient as a function of latitude. However, in the light of recent investigations it now seems necessary to study the mechanism of the turbulence itself in order to be able to answer the most important questions concerning the atmosphere, as for example the explanation for the existence of a strong westerly jet in middle latitudes.

When considering the general circulation of the atmosphere as a

large-scale turbulent process, we mainly ask for long-term statistical behavior and are not interested in the turbulent eddies themselves. Therefore, it may not be necessary to know the detailed structure of these eddies and one may still be able to answer the basic questions of their statistical behavior. In such problems it is often preferable to introduce a considerably simplified model which can be treated mathematically, in this way study the behavior of an idealized medium, in terms of which the real medium then may be discussed. We have examples of such a procedure from the kinetic theory of gases, where an ideal gas serves as a model for real gases, and in statistical mechanics in general.

The driving force of the atmospheric turbulent motion is the thermal contrast between the equator and the poles. As pointed out before this implies that the radiational equilibrium of the atmosphere is dynamically unstable and that the turbulent motion that results from the breakdown of this (imaginary) state of balance gives rise to a transport of energy from equatorial regions towards the poles. One should note, however, that this does not immediately imply a transport of angular momentum in the same direction. The fact that such a transport exists is a result of the dynamic characteristics of the system and the constraints that are given by the boundary conditions (surface friction).

It was briefly mentioned in Section 4 that the probable reason for the break-down of a state of radiational equilibrium is due to the strong wind-shear that must exist in such a case. Charney [19], Eady [18] and Fjørtoft [20] have shown that the middle latitude westerlies become unstable if the vertical shear exceeds a certain critical value (a somewhat different numerical value has been given in the three papers, but qualitatively they are in accord with each other). In other words the break-down is a result of the baroclinicity of the atmosphere. Kuo [21, 22] on the other hand points out that certain critical conditions of the horizontal wind-shear exist under which waves in a westerly current amplify. Thus even a flow of a barotropic fluid with the same velocity distribution as in the atmosphere (in the horizontal) may be unstable and break down into a turbulent motion. It is not clear at present which of these two processes is most important in reality. The settlement of this question is very fundamental, as it answers the question whether or not the atmosphere as a first approximation may be considered as a two dimensional barotropic atmosphere in horizontal motion; in other words, whether the vertical component of the absolute vorticity is approximately conserved.

Several facts seem to indicate that this is a fairly good assumption [14]. The final answer probably will be given when we know more about the behavior of such an idealized atmosphere and can compare it with what is observed in nature. Eady [38] has argued that this approxima-

tion is poor even as a first approach to the basic problem. His conclusion is based on the following reasoning: In middle latitudes a strong gradient of absolute vorticity exists from north to south. Under the assumption that vorticity is conserved in the mixing process there would be a tendency to decrease this gradient, which means a transport of vorticity from north to south. This is, however, in contradiction to existing conditions. As the subtropical anticyclones at the surface of the earth act as source-regions for vorticity and similarly the low pressure belt at 60° latitude is a sink, the transport in reality is from south to north, which also has been verified by actual measurements. Thus the vorticity cannot be conserved in this horizontal mixing process. This reasoning is not valid. Kuo has shown [22] that the direction of the vorticity transport does not necessarily take place along the gradient if the motion of the eddies and the disturbances are controlled by the fact that vorticity is conserved. Under certain conditions the transfer may be just the opposite. This touches upon a very fundamental problem in large-scale turbulence that needs further clarification.

For any discussion of this kind it is not enough to know what properties are conserved during the mixing process, but we also must know what the direct driving mechanism for the turbulence is. If the maintenance of the turbulence is independent of the conservative property there always will be a tendency to destroy the gradient and a first theoretical treatment can be made by introducing an eddy coefficient (constant or varying). If, on the other hand, this break-down into a turbulent motion is controlled by the distribution of this conservative element itself, we cannot draw any conclusions about the direction of the transfer without first analyzing the behavior of the eddies and the disturbances under the influence of a given distribution of the conservative property.

It is true that an explanation of the general circulation based on the principle of conservation of absolute vorticity never can be complete and most problems concerning for example the energy balance of the atmosphere must necessarily depend on the baroclinicity of the atmosphere. But in the light of the previous discussion it seems likely that some questions may be answered by this simpler model. Some results resembling the behavior of the real atmosphere have actually been obtained. We will first discuss these results and thereafter consider the effects of baroclinic processes in the atmosphere to the extent this is possible at the present time.

8. The Barotropic Model

The idea of large-scale turbulence in the atmosphere as introduced by Defant was again taken up by Rossby [40] from the viewpoint of conser-

vation of absolute vorticity. In reality we observe the greatest activity of cyclones and anticyclones in middle and high latitudes, in other words the turbulence is best developed in those parts of the world. Rossby now asked the following question: What limiting stage does the mean flow approach in a hemispherical shell under the influence of mixing during which the absolute vorticity is conserved? If we assume that the mixing is caused and maintained independently of the vorticity distribution

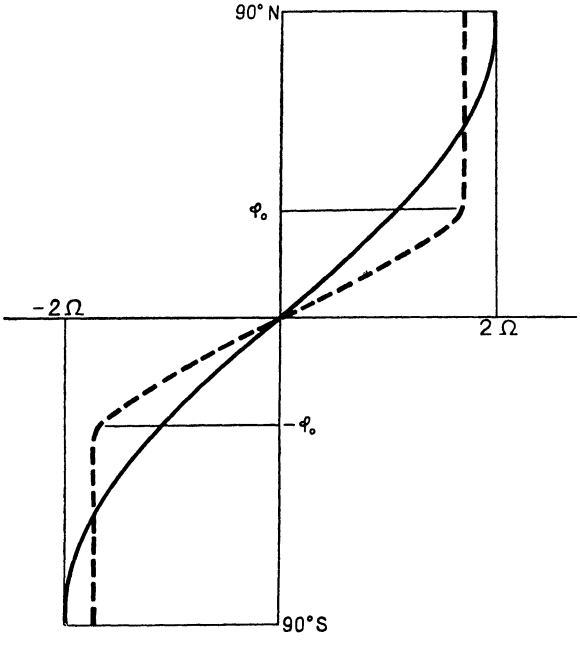


FIG. 9. Vorticity distribution as a function of latitude as a result of lateral mixing (after Rossby [40]). The solid line is the absolute vorticity of a fluid in rest relative to the earth (vertical component), the broken line indicates the vorticity distribution when complete mixing has taken place in the two polar caps.

itself it is obvious that the final stage is that the absolute vorticity is constant in the region of mixing. This process then would try to establish two regions of approximately constant absolute vorticity extending from the poles towards the equator, one with positive absolute vorticity (in the Northern Hemisphere), the other with negative vorticity (in the Southern Hemisphere) (cf. Fig. 9). It is quite obvious from Fig. 9 that such a mixing cannot extend all the way to the equator as a sharp discontinuity would develop in disagreement with the mixing concept itself. Rossby tried to overcome this difficulty by assuming constant transport of vorticity from north to south in equatorial regions. It is interesting

to notice that the velocity profile corresponding to a constant value of the absolute vorticity poleward from approximately latitude 35° is in qualitative agreement with mean conditions in the atmosphere. (It should be pointed out that the theory outlined above cannot explain the velocity distribution in the individual jets that are observed on daily upper air charts. A mixing theory of this type where the cyclones and anticyclones are the turbulent elements can only be applied to mean conditions over a time period that is large compared with the time-scale of the eddy motion itself.) However, some objections can be made against the results obtained in this way. By and large the conditions as indicated in Fig. 9 correspond to a transport of vorticity from the pole towards the boundary of the regions of mixing, because surface friction constantly tries to increase the vorticity of the air in the vicinity of the pole as it is less than the vorticity of the earth, while the opposite is true farther south. In reality the transport is in the opposite direction at least in middle latitudes. Secondly the southern boundary of the region of mixing coincides with the maximum westerlies aloft. From inspection of upper air charts one gets the impression that the belt of maximum westerly flow is the place for the most pronounced large-scale turbulence rather than the southern boundary of such an area. In fact it cannot be considered as settled what role the variation of the Coriolis parameter plays for the generation of a jet stream in middle latitudes. Even in the dishpan experiments a jet is formed which shows many similarities with the jet stream in the atmosphere and in those experiments the basic rotation obviously does not vary from one place in the fluid to the other. Finally it seems likely that the motion and development of the turbulent eddies should at least to some extent be controlled by the principle of conservation of absolute vorticity. As pointed out previously we cannot conclude that the absolute vorticity is constant within an area of mixing. Recent investigations by Kuo [22] and Fjørtoft [20] shed some light on these problems.

Kuo adopts the following model of the atmosphere: The atmosphere may be divided into two layers. In the lower layer (some 3–5 km. deep) the heat transfer plays the most important role and the thermal structure may give rise to the instability we know exists as the deepening of polar front cyclones. In the upper layer the motion of the atmosphere is essentially controlled by the vorticity equation (quasi two-dimensional motion). There is mutual interaction between the two layers essentially by divergence and convergence. Kuo proposes that these impulses may be assumed to take place intermittently and that a study of the upper layer between such impulses may give us some clue to the behavior of the atmosphere. In reality this interaction is not such a discontinuous

process, but if the behavior of the upper layer and the influence from below may be considered as two relatively independent processes, a separation of the mathematical treatment in this way should be a first approximation with some resemblance to the real atmosphere. Kuo essentially has been studying the upper layer of this atmospheric model assuming that it is controlled by the vorticity equation and that the horizontal divergence is zero.

In a theory based upon results obtained by Lin [41], Kuo finds that waves in an atmospheric jet amplify if the absolute vorticity profile of the basic current has one or several extreme values and the wavelength exceeds a certain critical value (of the order of magnitude of 5000 km). If these conditions are not fulfilled disturbances in the current are neutral or damped. The amplification of waves means a transfer of kinetic energy from the basic current into the disturbances, momentum is carried away from the center of the current and a transfer of vorticity takes place across the jet from north to south. Damping on the other hand implies just the opposite: the kinetic energy of the basic current increases, momentum is concentrated into the center of the current and vorticity is transferred from south to north. In the former case the vorticity gradient becomes weaker; in the latter case it is steepened. As was mentioned before the vorticity transport in this way may take place against the gradient, when the behavior of the disturbances is controlled by the basic vorticity distribution.

These theoretical results for a barotropic non-divergent fluid can be applied in a qualitative way to the model of the atmosphere, that was outlined above. Under normal conditions the absolute vorticity profile of the westerlies in the upper troposphere does not have any extreme points and thus all waves are damped or neutral. Energy is fed into the zonal flow and the intensity of the westerlies increases. Obviously the waves cannot be generated in this layer by any mechanism discussed so far. Kuo points to the influence of the lower layer for the formation of these disturbances. In a qualitative way we are thus able to increase the energy of the system and energy is organized into a westerly flow by the action of the waves. Gradually extreme points may develop in the vorticity profile as there is a constant transport of vorticity from south to north. Then, if an impulse is given to the current from below with a wavelength greater than the critical wave length the wave will amplify and possibly cause a complete breakdown of the westerlies. Actual integration of such a case using the nonlinear equations (carried out at The Institute for Advanced Study, Princeton) shows that the mean jet may even split into two branches separated 20 to 30° latitude from each other. Ultimately this instability causes its own destruction in that the

vorticity transport from north to south removes the extreme points in the vorticity profile. The waves in the westerlies again are stable and the intensification of the basic current starts over again. The time scale of such a cycle agrees fairly well with the observed index cycle of about four weeks. The theory also may explain some of the details of this cycle according to the results obtained by Riehl *et al.* [10]. Furthermore preliminary investigations by White (personal communication) indicate that the long waves in the westerlies (at 45°N) seem to give rise to a momentum transport in the opposite direction to the short waves, but that the effect of the latter is dominating and responsible for the direction of the mean momentum transfer (from south to north). This change of the direction of the momentum transfer, when a certain critical value of the wavelength is exceeded, is in qualitative agreement with Kuo's theoretical results. However, some doubt may be raised regarding the effectiveness of the damped waves in building up the jet. Furthermore, for a complete discussion of this model of the flow in the upper troposphere it is necessary to consider the vertical momentum transport caused by these eddies or by other means. Not until then will one be able to discuss the momentum balance of the upper troposphere theoretically. Such a discussion is necessary in order to explain, for example, the fact that the maximum northward momentum transport takes place at approximately the latitude of the maximum westerlies [28]. In spite of the fact that it is impossible to reach even definite qualitative conclusions at present, there are enough similarities between the behavior of these barotropic waves and the flow patterns at upper levels in the atmosphere to encourage further investigations along these lines.

9. The Baroclinic Model

Some of the attempts to incorporate the effects of the baroclinicity of the atmosphere for these large-scale mixing processes are presented in the following discussion. It is true that the strongest baroclinicity exists in the lower parts of the troposphere, but it is by no means zero aloft. It has already been pointed out that a zonal current is unstable if a critical value of the vertical shear is exceeded and that this is almost always the case in nature. This more or less permanent instability may be taken as the cause of the large-scale turbulent motion of the atmosphere. It is, however, somewhat surprising that the actual vertical wind shear even in the mean is several times larger than the critical value at which instability occurs. We know for example from the theory of convection that the limiting lapse rate, when fully developed vertical turbulence exists, is approximately equal to the critical lapse rate and that a superadiabatic lapse rate very seldom is observed over a large area and for any consider-

able length of time. One therefore would expect that the mean vertical shear in the atmosphere should be less than or equal to the critical shear. Eady [38] has pointed out that the regeneration of the basic conditions causing turbulence takes place continuously. The mean state of the atmosphere therefore expresses a balance between these regenerating factors and the destruction of the basic field by the turbulence. Such an equilibrium may not be reached until the critical conditions are exceeded to a considerable extent. This raises the question when such a balance is reached. This cannot be answered without a discussion of the formation of the basic current itself. We here notice a principal difference between the interpretation of Kuo's results for a barotropic atmosphere and the results that have been obtained for a barocline atmosphere. In the former case the turbulence (or better the damped disturbances) builds up the basic flow, occasionally the critical conditions may be exceeded causing a breakdown of the quasi-zonal motion into a large-scale eddy motion, but stable conditions are predominant. In the baroclinic case we derive almost permanent instability but in spite of that the basic flow is maintained considerably stronger than what corresponds to initiation of the turbulence. The basic current therefore must be maintained by some other process.

The very fact that there are pronounced changes in the circulation of the atmosphere from time to time suggests that it is operating in the vicinity of the critical values of one or several parameters. Thus the index cycle usually means a rapid breakdown from high to low index and then a more or less gradual building up of the westerlies again. In the light of the discussion above it seems desirable to combine some of the features of the barotropic and baroclinic models treated so far. It would be helpful to know the stability criteria for a jet with both horizontal and vertical shear as well as the characteristics for waves on such a current. This is a very difficult problem that as yet has not been solved successfully.

10. Effects of the Non-uniformity of the Surface of the Earth

In the previous discussion we have assumed that the surface of the earth is uniform. The considerations therefore may explain certain general features of the atmospheric circulation, but they do not account for differences between the Southern and Northern Hemisphere, nor irregularities in the zonal flow around the hemisphere. Here factors as the thermal contrast between land and sea, the large mountain barriers in middle latitudes, variations in the surface stress etc. presumably play an important role.

For a long time the thermal influence from the surface of the earth

has been considered to be of basic importance for the explanation of the irregularities mentioned above (cf. Figs. 1 and 2). The reasoning goes as follows: A solenoidal field is established along the coasts. In an equilibrium state this solenoidal field must be balanced by a wind change with height, but in order to compensate for frictional loss of kinetic energy a direct circulation around the solenoids must also take place. Finally the deviations from geostrophic flow because of friction should be consistent with this solenoidal circulation. We thus arrive at the following picture: Anticyclonic circulation at lower levels and cyclonic circulation aloft around regions that are colder than the environments and *vice versa* if the surroundings are colder than the area considered. The mean zonal flow is superimposed upon this pattern giving rise to a wavelike pattern at upper levels. The ridges are located over the areas, where heat is supplied from below, while troughs develop, where the atmosphere is cooled.

Undoubtedly such a monsoonal circulation exists. The large monsoon wind systems at the surface of the earth are conclusive evidence for that. There are, however, several features of the flow aloft that seem to indicate that the orographic influence also plays an important role for the development of the final pattern [42].

First of all, there are many features of the upper flow that do not change from summer to winter even if their intensity may vary. Thus the stationary mean wave pattern extending from mid-Pacific across the American continent and the Atlantic Ocean (Figs. 1 and 2) has approximately the same phase in the two extreme seasons. This similarity is difficult to explain from a purely thermal point of view as the thermal field by and large is reversed from summer to winter. The existence of a trough over western Europe is hardly accounted for by thermal contrasts even if its structure varies to a certain extent from season to season. A comparison between the two hemispheres gives some further evidence for the importance of the mountains for the structure of the mean upper flow. It is well-known that an upper trough in the mean is located to the east of the Andes Mountains [9]. It is found in approximately the same relative position to the Andes as the trough over the eastern United States is to the Rocky Mountains, in spite of the fact that the thermal conditions are quite different.

It is at present impossible to draw conclusions about the relative importance of the two processes. A quantitative theory for the thermal synoptic idea is still missing. Furthermore it is very likely that there is a certain interaction between them. The mountains may for instance initiate a disturbance (barotropically) but the intensity of it may still be determined by the baroclinic character of the atmosphere.

List of Symbols

c_n inward component of velocity at the boundary
d rate of decrease of kinetic energy by (small scale) turbulence and viscosity per unit volume
E kinetic energy per unit volume
f $2\Omega \sin \varphi$, Coriolis parameter
M absolute angular momentum per unit mass
p pressure
r distance to the axis of the earth
R radius of the earth
R^* gas constant for dry air
T temperature
dS surface element
dV volume element
U velocity of mean zonal flow
u velocity towards the east
v velocity towards the north
w velocity upwards
β $\partial f/\partial y$
φ latitude
λ characteristic length-scale
Ω angular velocity of the earth
ρ density
$d\sigma$ projection of surface element on a meridional plane
τ_x eastward frictional stress acting on the air at the surface of the earth

References

1. Hadley, G. (1735). Concerning the cause of the general trade winds. *Reprinted in* The Mechanics of the Earth's Atmosphere, *Smithsonion Inst. Misc. Coll.* **51,** No. 4, 1910.
2. Bergeron, T. (1928). Über die dreidimensional verknüpfende Wetteranalyse, I. *Geofys. Publik.* **5,** No. 6, 111 pp.
3. Rossby, C.-G. (1941). The scientific basis of modern meteorology. Yearbook of Agriculture, United States Department of Agriculture, pp. 599–655.
4. Rossby, C.-G. (1949). On the nature of the general circulation of the lower atmosphere. *In* The Atmospheres of the Earth and Planets. Univ. of Chicago Press, Chicago, Illinois, pp. 16–48.
5. Mintz, Y., and Dean, G. (1951). The observed mean field of motion of the atmosphere. *In* Investigation of the General Circulation of the Atmosphere, Part II, Report 7. Univ. of California, Los Angeles, 55 pp.
6. Dzerdzevski, B. L. (1946). On the distribution of atmospheric pressure over the central regions of the Arctic. (In Russian.) *Meteorol. i Gidrol.* No. 1.
7. Hutchings, J. W. (1950). A meridional atmospheric cross section for an oceanic region. *J. Meteorol.* **7,** No. 2, 94–100.
8. Namias, J., and Clapp, P. F. (1949). Confluence theory of the high tropospheric jet-stream. *J. Meteorol.* **6,** No. 5, 330–336.
9. Boffi, J. A. (1949). Effect of the Andes Mountains on the general circulation over the southern part of South America. *Bull. Am. Meteorol. Soc.* **30,** No. 7, 242–247.

10. Riehl, H., Yeh, T. C., and La Seur, N. E. (1950). A study of variations of the general circulation. *J. Meteorol.* **7,** No. 3, 181–194.
11. Taylor, G. I. (1921). Experiments with rotating fluids. *Proc. Roy. Soc. London* **A100,** 114–121.
12. Fultz, D. (1949). A preliminary report on experiments with thermally produced lateral mixing in a rotating hemispheric shell of liquid. *J. Meteorol.* **6,** No. 1, 17–33.
13. Long, R. (1951). Research on experimental hydrodynamics in relation to large-scale meteorological phenomena. Rept. No. 5, Hydrodynamic Laboratory, Dept. of Meteorol, Univ. of Chicago, Chicago, Illinois, 55 pp.
14. Charney, J., Fjørtoft, R., and von Neumann, J. (1950). Numerical integration of the barotropic vorticity equation. *Tellus* **2,** No. 4, 237–254.
15. Ekman, V. W. (1932). Studien zur Dynamik der Meeresströmungen. *Gerl. Beitr. Geophys.* **36,** 385–438.
16. Riehl, H. (1950). On the role of the tropics in the general circulation of the atmosphere. *Tellus* **2,** No. 1, 1–17.
17. Rossby, C.-G. and collaborators. (1939). Relations between variations in the intensity of the zonal circulation of the atmosphere and the displacement of the semi-permanent centers of action. *J. Mar. Res.* **2,** No. 1, 38–55.
18. Eady, E. T. (1949). Long waves and cyclone waves. *Tellus* **1,** No. 3, 33–52.
19. Charney, J. (1947). The dynamics of long waves in a barocline westerly current. *J. Meteorol.* **4,** No. 5, 135–163.
20. Fjørtoft, R. (1950). Application of integral theorems in deriving criteria of stability for laminar flows and for the baroclinic circular vortex. *Geofys. Publik.* **16,** No. 5, 52 pp.
21. Kuo, H. L. (1949). Dynamic instability of two-dimensional, non-divergent flow in a barotropic atmosphere. *J. Meteorol.* **6,** No. 2, 105–122.
22. Kuo, H. L. (1950). Dynamic aspects of the general circulation and the stability of zonal flow. Report No. 4, General Circulation Project, Massachusetts Institute of Technology, Cambridge, Mass.
23. White, R. M. (1951). The meridional eddy flux of energy. *Quart. J. Roy. Meteorol. Soc.* **77,** No. 332, 188–199.
24. Jeffreys, H. (1926). On the dynamics of geostrophic winds. *Quart. J. Roy. Meteorol. Soc.* **52,** 85–104.
25. Starr, V. (1948). An essay on the general circulation of the earth's atmosphere. *J. Meteorol.* **5,** No. 2, 39–43.
26. Bjerknes, J. (1948). Practical applications of H. Jeffreys' theory of the general circulation. *Procès-Verbaux des Séances l'Association de Météorologie.* UGGI, Oslo, pp. 53–55.
27. Widger, W., Jr. (1949). A Study of the flow of angular momentum in the atmosphere. *J. Meteorol.* **6,** No. 5, 291–299.
28. Mintz, Y. (1951). The geostrophic poleward flux of angular momentum in the month of January 1949. Report No. 7. Investigation of the General Circulation of the Atmosphere. University of California, Los Angeles.
29. Priestley, C. H. B. (1951). Physical interaction between tropical and temperate latitudes. *Quart. J. Roy. Meteorol. Soc.* **77,** No. 332, 200–214.
30. Starr, V. P., and White, R. M. (1951). A hemispheric study of the atmospheric angular momentum balance. *Quart. J. Roy. Meteorol. Soc.* **77,** No. 332, 215–225.
31. Riehl, H., and Yeh, T. C. (1950). The intensity of the net meridional circulation. *Quart. J. Roy. Meteorol. Soc.* **76,** No. 328, 182–188.

32. Palmén, E. (1951). The role of atmospheric disturbances in the general circulation. *Quart. J. Roy. Meteorol. Soc.* **77,** No. 333, 337–354.
33. Van Mieghem, J. (1949). Les equations génćrales de la mécanique et de l'énergétique des milieux turbulents en vue des applications à la météorologie. *Inst. Roy. Météorol. Belg., Mémoires* **34,** 60 pp.
34. Starr, V. P. (1948). On the production of kinetic energy in the atmosphere. *J. Meteorol.* **5,** No. 5, 193–196.
35. Rossby, C.-G., and Willett, H. C. (1948). The circulation of the upper troposphere and lower stratosphere. *Science* **108,** 643–652.
36. Namias, J. (1950). The index cycle and its role in the general circulation. *J. Meteorol.* **7,** No. 2, 130–139.
37. Rex, D. F. (1950, 1951). Blocking action in the middle troposphere and its effects upon regional climate. *Tellus:* Part I, **2,** No. 3, 196–211; Part II, **2,** No. 4, 275–301; Part III, **3,** No. 2, 100–112.
38. Eady, E. T. (1950). The cause of the general circulation of the atmosphere. *Centenary Proc. Roy. Meteorol. Soc. London,* pp. 156–172.
39. Defant, A. (1921). Die Zirkulation der Atmosphäre in den gemässigten Breiten der Erde. *Geograf. Ann.* **3,** 209 ff.
40. Rossby, C.-G. (1947). On the distribution of angular momentum in gaseous envelopes under the influence of large-scale mixing processes. *Bull. Am. Meterol. Soc.* **28,** No. 2, 253–268.
41. Lin, C. C. (1945). On the stability of two-dimensional parallel flow. I, II, III. *Quart. Appl. Math.* **3,** 117–142, 218–234, 277–301.
42. Bolin, B. (1950). On the influence of the earth's orography on the general character of the westerlies. *Tellus* **2,** No. 3, 184–196.

Exploration of the Upper Atmosphere by Meteoritic Techniques*

FRED L. WHIPPLE

Harvard College Observatory

CONTENTS

	Page
1. Introduction	119
1.1. Delimitation of the Problem	119
1.2. Definitions and Astronomical Background	120
2. Techniques of Observation and Astronomical Results	122
2.1. Visual Methods	122
2.2. Photographic Methods	123
2.3. Meteor Spectra	124
2.4. Radio Methods	127
2.5. Micro-Meteorites	131
3. Theory of the Meteoric Process	133
3.1. Basic Principles	133
3.2. Basic Relationships	134
3.3. The Mass of the Meteoroid	136
3.4. Determination of Atmospheric Densities	137
3.5. Comments on the Nature of the Meteoroid	138
4. Results Concerning the Upper Atmosphere	139
4.1. Densities in the Upper Atmosphere	139
4.2. Temperatures and Pressures in the Upper Atmosphere	142
4.3. Variations in the Upper Atmosphere	145
5. Circulation in the Upper Atmosphere	147
5.1. Winds from Meteor Trains	147
5.2. The Radio-Meteor Group of Stanford University	149
List of Symbols	151
References	151

1. INTRODUCTION

1.1. Delimitation of the Problem

Particles of solid material from space are known to encounter the earth's atmosphere. They can be detected directly by their interaction with the atmosphere, which results in radiation of visual or photographic

* Prepared partly as an investigation under Office of Naval Research Contract No. N5ori-07647 and Air Force Contract No. AF 19(122)-482.

light, the production of electron clouds which reflect radio waves, the production of clouds which reflect sunlight or radiate for a short time after the body has passed, or by the production of transient magnetic fields. Observations of these phenomena are a source of information concerning the upper atmosphere; the conclusions that can be drawn from these observations represent the subject of this article.

Because of the earth's gravitational attraction, the minimum velocity of fall from space to the atmosphere is about 11.1 km/sec. The maximum velocity of fall depends upon the origin of the particle and the circumstances of encounter with the earth. The earth is moving in a nearly circular orbit about the sun at a nearly constant velocity of 29.7 km/sec; the velocity of escape from the sun at the earth's distance is about 42.1 km/sec, corresponding to a rest velocity at an infinite distance from the sun. Hence, for particles permanently confined to the gravitational attraction of the sun, the maximum velocity of encounter with the earth's atmosphere is nearly 73 km/sec. The gravitational potential of the earth contributes very little to the velocity of high-speed particles but it predominates in determining the velocity of low-speed particles.

Particles from interstellar space, if such exist, might well reach the earth with velocities considerably greater than the above upper limit.

As yet we do not know whether the air glow of the night sky is affected directly or indirectly through the action of meteoric particles, but it is generally believed that the night-sky radiations of the Zodiacal Light and Gegenschein arise from the scattering of sunlight by such particles in space. Probably also a certain amount of the scattered sunlight in the daylight sky arises from very small particles that are falling slowly through the earth's atmosphere.

1.2. Definitions and Astronomical Background

The term *meteor* is customarily used to designate the visual, photographic, or other electromagnetic *phenomena* associated with the passage of a small particle through the atmosphere. The particle sizes involved in the phenomenon depend greatly upon the velocity of the particles; roughly the range in diameter is from less than a millimeter to several centimeters. It is often convenient to define the active particle as a *meteoroid* so that the phenomenon and particle can be easily distinguished.

Meteors sufficiently bright to cast shadows are generally called *fireballs* and detonating fireballs are frequently called *bolides*. When meteoroids are sufficiently large to withstand the ablation arising from interaction with the earth's atmosphere and to fall to the surface in sizeable pieces, they are known as *meteorites*. The phenomenon is known

as a meteorite fall. Microscopic particles accompanying such falls are generally known as *meteoritic dust* or as *micro-meteorites*. The latter term, however, is conveniently reserved to designate the very small meteoroids that are able, because of their large ratio of surface to mass, to radiate away the energy of interaction with the atmosphere so rapidly that they are stopped in the atmosphere without appreciable ablation. Evidence concerning the nature and numbers of micro-meteorites is increasing very rapidly at the present time.

The term *meteor trail* is generally used to refer to a meteor's visible or photographed path across the sky. The term *meteor train* is used to designate the persistent luminosity left along the trail after the moving meteoroid has passed. For most meteors the train lasts for a very short interval of time but in some cases it may be observable for seconds, minutes, or even as much as an hour. Daylight trains are visible because of reflected sunlight while night trains are self-luminous. The luminosity of night trains has not yet been explained by any comprehensive theory.

An appreciable fraction of observed meteors is comprised in *meteor showers* which may recur, from year to year, with variable intensity at the same positions of the earth in its orbit. These meteor showers arise from the earth's encounter with *meteoritic streams* of material moving about the sun in elliptical orbits of considerable cross-sectional area. Because the meteoroids in a stream strike the earth in nearly parallel paths, the resultant meteors appear to radiate from a *radiant* in the sky, the position of which depends upon the velocity vectors of the encounter. The various meteoric streams or showers are normally named for the constellations in which the radiants are located. A number of these streams have been definitely identified with the orbits of known comets. For example, the most conspicuous and reliable of present-day showers, the Perseid shower, observable during the first half of the month of August, was first shown by Schiaparelli to be associated with Comet 1862 III. Most astronomers believe, though they cannot absolutely prove, that all meteor showers or streams have originated from the disintegration of comets. Streams as yet unidentified with known comets are much more numerous than those for which identification has been possible.

Meteors not associated with recognized showers are called *sporadic* meteors. Most of the very bright fireballs, bolides, and apparently all of the meteoritic falls belong to the sporadic class.

For general background information on meteors and meteorites the reader is referred to treatises by Watson [1], Olivier [2], and Hoffmeister [3].

2. Techniques of Observation and Astronomical Results

2.1. Visual Methods

Since visual techniques for the observation of meteors are largely being superseded for definitive studies of the upper atmosphere, the present account of the visual methods will be relatively brief. Only in the case of persistent meteor trains are the visual techniques still extremely valuable. They have provided the best determinations yet made of wind velocities in the atmosphere at heights from 30 to 110 km.

In the more advanced techniques of visual observation, two or more observers a number of miles apart make simultaneous accurate records of the apparent paths of meteors. From these records it is possible, then, to determine the height, radiant point and atmospheric trajectory of a given meteor observed by both with a degree of accuracy depending upon the ability of the individual observers, the brightness of the meteor, and the geometrical circumstances. A precision of one mile in altitude is excellent. Visual estimates of angular velocity, however, are quite poor and lead to spurious results concerning the velocities of meteors in the atmosphere. The most refined visual technique is that utilized by Öpik [4, 5, 6] in the Harvard-Arizona Meteor Expedition. The observers view the sky through a mirror which is rocked by a synchronous motor so that the perpendicular bisector of the plane mirror describes a right circular cone of small amplitude. An observer looking at the reflected skylight in the mirror sees a meteor describe a type of cycloid motion which may show open loops or cusps. He can use both the criteria of the shape of the apparent trail and the angular separation of loops or cusps as a measure of the angular velocity of the meteor.

From his analysis of these rocking mirror observations, Öpik concluded [7] that a large fraction of the fainter meteors describe hyperbolic paths about the sun. As we shall see from the more recent photographic and radio observations of meteors, this conclusion has not been substantiated.

The most extensive series of visual observations of meteor trails is that compiled by Denning [8] and his coworkers in England. Lindemann and Dobson [9] were the first to develop a meteoric theory which, combined with the Denning observations, led to certain conclusions concerning the density and temperature of the upper atmosphere. We shall discuss their results briefly in Section 4. Various attempts to use the visual observations of meteoric heights as a measure of seasonal variations in the density of the upper atmosphere have led to rather controversial results and will not be described here because better conclusions

can be drawn from photographic work (see, for example, McIntosh [10], Porter [11], and Link [12]).

2.2. Photographic Methods

The first systematic application of the photographic method for observing meteors by the use at two stations of cameras equipped with rotating shutters was carried out by Elkin [13] at the Yale Observatory, 1893–1909. Simultaneous photography of the meteor trail from two stations leads obviously to a determination of the precise trajectory above the earth's surface. The use of rotating shutters to cover the lenses at known short intervals of time provides direct measures of the

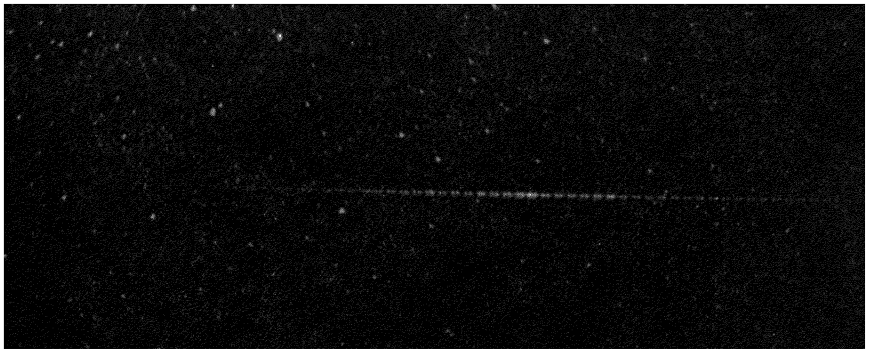


FIG. 1. Meteor trail over New Mexico, Harvard observatory photograph.

angular velocity of the meteor at various points along its trail. Incidentally, meteor trails are extremely straight, the earth's gravity rarely producing a measurable deflection. The resistance of the atmosphere is sufficiently small that only with the most precise techniques can the deceleration of the meteoroid be measured. The first systematic measurements of deceleration were made in the Harvard Observatory program [14]. A photographic trail showing shutter breaks is reproduced in Fig. 1.

Fundamentally, the two-camera photographic technique of observing meteors leads to the following measured quantities: (a) the trajectory of the meteor referred to the surface of the earth, (b) the velocity of the meteor with respect to the surface of the earth, (c) the rate of change of the velocity of the meteor, and (d) the brightness as a function of time along the trail.

From the astronomical point of view, the observed velocity vector of the meteor can be corrected for atmospheric resistance, rotation of the earth, attraction of the earth, and orbital motion of the earth, in order to

reconstruct the particle's original motion with respect to the sun. No hyperbolic meteors have been observed with certainty in about seventy doubly-photographed meteors observed at Harvard and elsewhere. This result applies only to the very brightest meteors since faint ones could not be photographed until recently. Table I shows the velocity distribution of forty-two Harvard meteors both with respect to the atmosphere, V_{obs}, and with respect to the sun, V_H.

TABLE I. Velocity distribution for meteors.

Velocity range	No. V_{obs}	No. V_H
11–20 km/sec	4	0
20–30 km/sec	10	0
30–40 km/sec	9	21[a]
40–50 km/sec	1	17[b]
50–60 km/sec	6	0
60–70 km/sec	6	0
70–72 km/sec	2	0
Total No.	38	38

[a] Min. V_H = 33.6 km/sec.
[b] Max. V_H = 42.18 km/sec.

In 1951 the first Super-Schmidt meteor camera, of aperture 12¼ inches, focal length 8 inches (F/0.65) and field 52°, was put into action by the Harvard Observatory for the U. S. Naval Bureau of Ordnance. This camera, designed by Dr. James G. Baker, and manufactured by the Perkin-Elmer Corporation, has proved to be an extremely powerful instrument for the photography of meteors. It is shown at its site near Las Cruces, New Mexico, in Fig. 2.

2.3. Meteor Spectra

The spectra of the brightest meteors can be photographed by the use of an *objective prism* or *objective grating* placed in front of the lens or mirror of an astronomical telescope. The trail of the meteor effectively provides a slit. Extremely small dispersions of the order of several hundred angstroms per millimeter are the greatest that can be used in this work because of the rarity of extremely bright meteors. Millman [15] has recently reported on 104 meteor spectra obtained largely at various Canadian observatories and the Harvard College Observatory. These 104 spectra comprise by far the major portion of those in existence.

The most striking result from these spectra is the fact that no atmospheric lines or bands are present. The continuous spectrum, if present

at all for any meteor, is relatively faint compared to the bright-line spectrum; the latter consists mostly of the lines of neutral atoms of the elements more abundant in meteorites. The lines of FeI are the most numerous but among fast meteors the H and K lines of CaII often dominate the spectrum while among slow meteors the D lines of NaI are

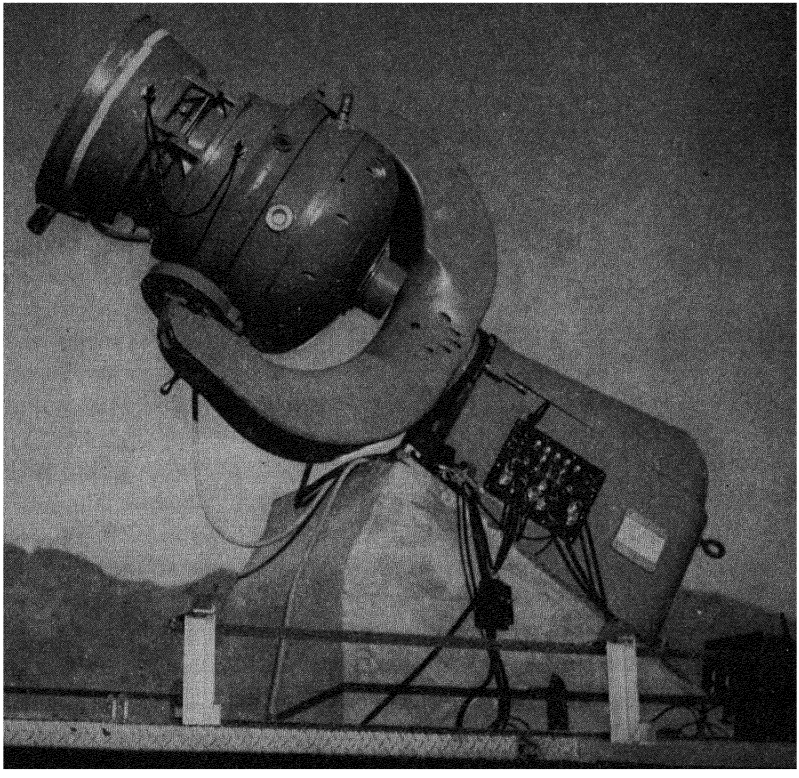


FIG. 2. The first super-Schmidt Meteor camera at Harvard station in New Mexico.

extremely strong. Lines from spectra of the following elements have been identified by Millman: FeI, CaI, MgI, MnI, CrI, NiI, AlI, CaII, MgII and SiII. Perhaps bands of FeO are present in a spectrum obtained at Mount Wilson. In addition, low excitation lines of FeII have been identified by Vyssotsky [16] in a relatively high dispersion spectrum of a fast meteor observed at the Leander-McCormick Observatory in Virginia.

Millman finds that the major factor which determines the nature of

the meteor spectrum is the velocity. Low-velocity meteors show only low excitation lines of neutral atoms, while meteors of intermediate velocities may show the CaII lines faintly. Meteors of the highest velocities show the CaII lines strongly and some other lines of the ionized metals. Among the sporadic meteors approximately half show the CaII lines as prominent features.

Where color or excitation changes can be observed over the course of a meteor, Millman has found that in nine out of ten examples the progressive change with time involves either a relative increase in the ratio of blue to red light or an increase in the general state of excitation.

No photographic spectra of long persistent trains have been obtained but Millman [17] has obtained for one meteor the spectrum of the persistent luminosity between rotating shutter breaks lasting 0.05 sec. This particular meteor was a Perseid (high velocity) in which the normal spectrum showed strong CaII lines. The spectrum during the shutter breaks showed no CaII and only extremely low excitation lines of FeI, CaI, MgI, and NaI. No nitrogen afterglow could be observed.

These results indicate generally that bright meteors contain most of the more abundant elements known to be present in meteorites. No analysis of relative composition has yet been attempted. The absence of lines or bands from atmospheric gases is probably a result of the inability of these gases at low stages of excitation or ionization to produce strong lines in the visual or photographic regions. The one element very abundant in meteorites that does not show spectroscopically in meteors is oxygen, unless the uncertain FeO band really occurs in meteor spectra. The low states of excitation and ionization in meteor spectra would be surprising in terms of the higher stagnation temperatures one might expect to find under adiabatic conditions at the surface of the meteoroid. On the other hand, there are other reasons to believe that these very high temperatures do not occur and, consequently, that the temperature at the surface is not greatly in excess of the boiling point of meteoritic material.

The fact that the same elements occur in meteor spectra and in meteorites might lead one to expect similarity in chemical structure and possibly even in origin. One must note, however, that these elements are usually good radiators in relatively low stages of excitation and that there is no assurance from the identity in actual elements that the minerals or abundances are at all comparable among the two groups of objects. The writer, in fact, concludes from the orbital elements that most of the meteors are of cometary origin, which is probably quite different from that of the meteorites. From the structure of meteorites many investigators suspect that they are fragments of a broken planet.

2.4. Radio Methods

Observations of meteors by radio techniques is a field of research which has grown at such an enormous pace since World War II that no historical account will be attempted here. The reader who wishes to study the field more thoroughly may refer to summary articles by Lovell [18], Herlofson [19], and Hey [20], or to the many individual contributions in the field.

The basic process that makes radio detection possible is the production of a column of electrons by the meteoroid as it reacts with the atmosphere. For very bright meteors an electron column in the neighborhood of the meteoroid moves with the body and is dense enough to reflect a radio beam. For all meteors the electron column as it grows with the moving meteoroid (or in an appreciable time thereafter) produces radio reflections in the proper frequency range.

Radio "whistles" were first detected by Chamanlal and Venkataraman [21] using a continuous wave transmitter separated some distance from the receiver. The beat frequency between the ground wave and the growing electron column produces a modulation with a frequency in the audible range, which can be detected by ordinary audio receiving systems.

McKinley [22] has shown, however, that in most cases of radio whistles or amplitude-time variations of a returned radio signal, the predominating effect is not the beat frequency with the ground wave or transmitter wave; it is a variation in signal amplitude resulting from the increasing length of the ion column. We note, as Pierce [23] suggested, that most radar echoes from meteors are observed near the point where the axis of the antenna beam is nearly perpendicular to the space trajectory of the particle. The radio reflection, as first suggested by Blackett and Lovell [24] results largely from coherent scattering of the electrons near the trajectory. This is particularly true for observations made at frequencies in the range from about 10 mc/sec to 100 mc/sec. Hence, as the narrow ion column grows during the progress of a meteor, the vector sum of the returning radio amplitudes received by the antenna changes in much the same fashion as the diffraction of light past a straight edge. Hence, the Fresnel pattern, changing with time, leads to an amplitude variation in the received signal accounting for the "whistle" effects.

With regard to the determination of velocities from radio observations, McKinley [22] divides the methods into two classes: (a) range-time variations (pulses only) and (b) amplitude-time variations (pulses or continuous wave). The range-time method was first used by Hey,

Parson, and Stewart [25] during the Giacobinid shower of 1946. The method utilizes the range-time oscilloscope presentation of the ordinary radar where the transmission is by pulses. An example is shown in Fig. 3. The range-time method is less used than the amplitude-time method for the reason that only the very brightest meteors show sufficient ion density near the head to produce the necessary radar effect. On the other hand, for very bright meteors the method is extremely valuable and has been used to a considerable extent by McKinley.

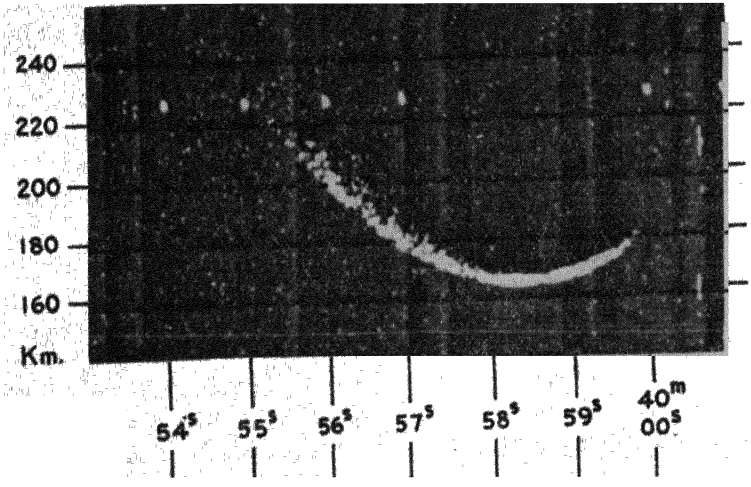


FIG. 3. Time (abscissa)-range (ordinate) diagram of radar meteor by D. W. R. McKinley, courtesy National Research Council of Canada.

The amplitude-time method depends upon the variations in the amplitude of the reflected radio beam according to the growth of the Fresnel pattern as described above. For measuring the velocities of meteors this method has been employed extensively by Davies and Ellyett [26], who utilize a pulse transmitter, and by Manning, Villard, and Peterson [27] and also very extensively by McKinley, who use both pulse and continuous-wave transmitters. An example is shown in Fig. 4.

The chief results obtained from the radio-meteor research as yet concern more the nature of meteoritic orbits and to a lesser extent characteristics of the upper atmosphere. It is probable that this tendency will be reversed as the physical interpretation of the radio observations reaches a better state of development.

Of outstanding astronomical value have been the results concerning meteor orbits. The group at Manchester, England, under the direction of Lovell [26, 28] and the Canadian group, under the direction of

McKinley [22] have demonstrated that less than 1 %, if any, of the visual meteors (and well below the visual limit) were derived from closed orbits about the sun. There is no definite evidence for the existence of *any* hyperbolic meteors from interstellar space. This conclusion extends the photographic results to many more examples and to much fainter magnitudes.

The studies of daylight meteors by radio techniques, begun in 1945 by Hey and Stewart [29], have been continued with great success by the Manchester group. The results are of great interest astronomically. The fact noted by Lovell [30] that there is no observable change in the frequency of meteors from a given shower during the twilight transition

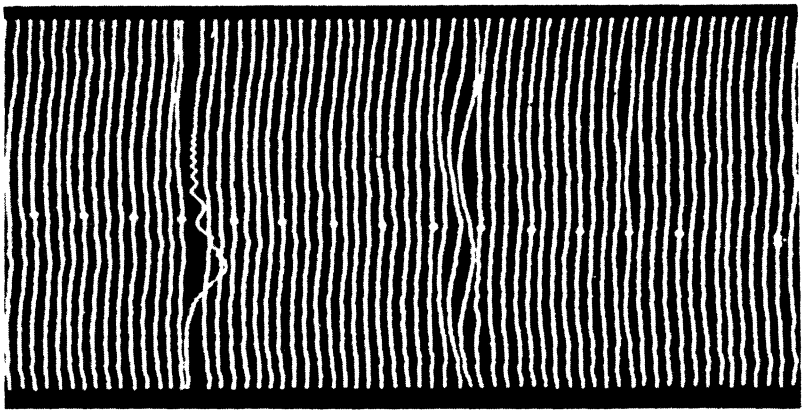


FIG. 4. Amplitude (abscissa)—time (ordinate) diagram of radar meteor by D. W. R. McKinley, courtesy National Research Council of Canada.

period between darkness and sunlight indicates that the production of ionization by the meteoroid is not dependent upon the state of the ionization in the atmosphere. According to Millman [31] meteoric ionization occurs in the region from 80 to 120 km with mean heights in the region from 97 to 103 km. This region is not far removed in general level from the maximum of the E-layer. As yet there is no clear-cut evidence that the degree of ionization in the E-layer affects in any way the production of ion trails or the persistence of ion trails produced by meteors, although it does produce absorption of echoes in the daylight at lower frequencies.

The theoretical determination of the number of ions produced by a meteor moving through the atmosphere has not yet been attacked seriously. Herlofson [19], following the older discussions, particularly Öpik's [32], concludes that of the original meteoric energy approximately 10^{-2} is used in radiation and approximately 10^{-4} in ionization. McKinley [33] has recently improved this result observationally by

obtaining the decelerations of three meteors by radar techniques alone. Applying the theory, proven to be successful for photographic meteors, he has been able to calculate the efficiency of ion production. For a meteor moving at 60 km/sec through the atmosphere, he estimates that 10^{-5} to 10^{-6} of the energy is used in ionization, while at 20 km/sec only the fraction 10^{-8}, is so utilized. In estimating the number of ions in the trail, he adopts the generally accepted theory by Lovell for the electron density per unit length along the trail.

Liller [34] has shown from the Harvard meteor-radar observations at 3.5 mc that the ionization concurrent with meteors of the same visual apparent magnitude varies as a relatively high power (4 to 6) of the meteoric velocity. McKinley's result generally confirms Liller's conclusion; both are consistent with the negligible degree of ionization observed in the spectra of slow meteors and the appreciable ionization in the spectra of the fastest meteors. Furthermore, no meteor spectrum indicates a high degree of ionization.

A valuable contribution concerning ionization in the meteor process has been made by Millman [31]. In particular, he compares the duration of persistent meteor trains to the duration of the ion trails in meteors, both durations as functions of the apparent visual brightness of the meteor. He finds that with logarithmic scales for both duration and brightness the relationship is linear whether the duration applies to ionization or train. With increasing velocity the meteors show a greater duration both for the train and for the electron column. Furthermore, at a given velocity the slope of the log-duration versus log-brightness curve is the same for the persistence of trains as for the persistence of electrons. The observed ionization at 30 mc/sec lasts some 20 times as long as the visual train. Lovell [35], in a summary paper on meteoric ionization, discusses a number of the problems and demonstrates an inverse linear relationship between velocity and echo duration to one-half amplitude at 72 mc/sec.

The general consensus is that meteors do not contribute appreciably to the ionization of the night sky except on rare occasions such as that of the Giacobinid comet shower in 1946. For a few hours a commercially usable E-layer was maintained by the action of the meteor shower. On the other hand, it is certain that ionization in small columns is being produced almost continuously near the E-layer; ionization can be detected there at all times. The extent to which the variations in the E-layer affect the persistence of meteor trains or of meteor ionization has not yet been determined. As noted above, such variations appear not to affect the frequencies of radar-observable meteors.

A number of major problems in the area of radio-meteor studies

remain to be solved theoretically. Of particular interest is the short duration of the electron cloud that shares the motion of the meteoroid in contrast to the long duration of the electron column that persists after the body has passed. A clear theoretical understanding of the electronic processes in meteors will undoubtedly involve a much clearer understanding of the various atmospheric dissociation, recombination, and associated processes in the region below 120 km. Until such understanding has been attained, our knowledge of conditions in this region of the atmosphere can be considered as only fragmentary. It is clear that the observations of meteor ionization and decay by radio techniques will become an increasingly important factor in solving these difficult problems.

By the use of a new type of equipment, Kalashnikov [36] has reported the observation of transient magnetic fields apparently associated with a few of the brightest meteors in the major showers of 1948–1950. A further investigation of such effects is clearly required.

The study of upper atmospheric winds by means of radio techniques, particularly by the Stanford University group, will be discussed in Section 5. The method depends upon the Doppler velocities measured for the persistent electron columns associated with meteors.

2.5. Micro-Meteorites

Micro-meteorites were defined earlier as very small meteoroids that are able, because of their large ratio of surface to mass, to radiate away the energy of interaction of the atmosphere sufficiently rapidly that they are stopped without appreciable ablation. The possibility of such a process, bringing very small particles to the earth's surface without damage, was first suggested by Öpik [37]. The writer [38] has investigated the subject further from the theoretical point of view and has shown that the calculated dimensions of such particles are in good agreement with those of the few magnetic particles observed by Landsberg [39] and suspected by him to have fallen from the great Giacobinid comet shower of October 1946. The expected dimensions at the atmospheric velocity of 23 km/sec are a few microns in radius for spheres or in cylindrical diameter for long thin particles.

Buddhue [40] has discussed the observations of meteoritic dust at some length. On the basis of an extrapolation of meteoroid masses from the photographic and visual meteors (see for example Watson [1]), one would not expect an appreciable mass of micro-meteorites to exist in the earth's atmosphere. This older viewpoint was changed markedly, however, by the independent researches by van de Hulst [41] and Allen [42], who showed that the light of the solar corona giving a Fraunhofer

spectrum must arise from diffraction by small particles lying nearly on the line of sight between the earth and the sun. By extending their theory for this diffraction and scattering to include light from similar small particles spread over the plane of the ecliptic close to the earth's orbit, they were able to include also the *Zodiacal Light* as a part of the same phenomenon. The results demand, however, that the particle-size distribution include a high concentration of very small particles generally less than 10^{-1} or 10^{-2} cm in radius. One estimate by van de Hulst places the space density of such material at 5×10^{-21} gm/cm^3 in the neighborhood of the earth and would correspond to an accretion of perhaps 2000 tons of material per day on the entire surface of the earth. Allen's estimate, based upon the assumption of uniform particle size of 10^{-3} cm radius requires only 5×10^{-23} gm/cm^3 density and reduces the total accretion of matter by the earth to the order of 20 tons per day. Watson [1], on the other hand, from an integration of all of the meteor and meteorite accretion, found that only 1 ton per day of such matter is accumulated by the earth. It must be noted specifically that this observable meteoritic material will all be very large compared to the dimensions of the particles postulated by van de Hulst and Allen.

More recently another source of information suggests that the higher estimate of the daily accretion of meteoritic material by the earth may be more nearly correct. Pettersson and Rotschi [43] have found good evidence from the peculiar nickel content of deep sea deposits both in the Atlantic and the Pacific Oceans that several thousand tons per day of meteoritic material are accumulated by the earth.

If, then, we accept, for order-of-magnitude calculations, that around 2000 tons per day of material are accreted by the earth and, for the sake of a simple calculation, divide this up into particles of diameter 10 microns and density 4 gm/cm^3, the fall would correspond to 1 particle per square centimeter of the earth's surface in approximately 10 days. Since the rate of fall through the atmosphere would require the same order of magnitude of time, one particle of 10 micron radius should be present in roughly a square centimeter vertical column of the atmosphere; the space density there should be of the order of 1 particle per 10^7 cm^3.

The general order-of-magnitude confirmation of this concentration of micro-meteorites in the atmosphere is suggested by three rather unusual lines of evidence. Burnight [44] of the Naval Research Laboratory has found evidence which suggests that small polished surfaces exposed at high altitudes from V-2 rockets become pitted, presumably by collisions with small particles. These small craters have diameters of the order of 10 to 100 microns and would require a density of about 1 particle per 10^6 cm^3. Bohn and Nadig [45] have also studied the

occurrence of high-frequency (approximately 60 kc) sounds on the nose of a rocket at high altitude. Their results suggest that there is a particle large enough to activate the very sensitive recording equipment in a volume of approximately 10^8 cm^3.

Collections of atmospheric dust by aircraft at high altitudes have been made by Crozier and Seely [46] of the New Mexico School of Mines. They have no clear-cut evidence that any extraterrestrial particles are present but they can set an upper limit to the space density by the collections made from unusually clean northern air. They find under these most ideal conditions that less than 10 particles of diameters greater than 10 microns are present in 10^6 cm^3 of such air. This limit is determined entirely by unavoidable contamination. In normal air the occurrence of large particles is much more frequent than this figure. Hence, we see that an upper limit is set at approximately 1 particle in 10^5 cm^3.

Although these latter three sources of information are not very consistent as to the order of magnitude of particle density, nevertheless, they tend to confirm directly the more quantitative but less direct evidence. There appears to exist a surprising quantity of interplanetary material in the form of small particles with dimensions of only a few microns. It is very likely that these particles may contribute to non-Rayleigh scattering observed at high altitudes in the atmosphere. Further investigation of the general subject of micro-meteorites and scattering of sunlight in the earth's upper atmosphere from the theoretical point of view is planned at Harvard.

It is difficult, as yet, to assess the effects that micro-meteorites may have in the earth's upper atmosphere, but it is possible that they contribute to various optical phenomena. Their energy contribution, however, is negligible as compared to the normal black-body radiation of the night sky.

3. Theory of the Meteoric Process

3.1. Basic Principles

I shall not attempt a presentation of the historical development of the methods now used in determining upper atmospheric densities from photographic observations of meteors but will present the theory in its present state of development. Pioneer contributions were made by Lindemann and Dobson [9], Öpik [32, 37], Maris [47], Sparrow [48] and by Hoppe [49]. The form of the basic equations as presented below is essentially that derived by Hoppe with fundamental concepts contributed by the other investigators mentioned, particularly Öpik.

The first basic principle is that the loss of mass of the meteoroid is

proportional to the energy available from the impinging air molecules. One equates the rate of loss of mass to a constant times the kinetic energy of the air column encountered by the cross sectional area of the meteoroid, and divides by a quantity of heat per unit mass necessary to melt, vaporize, or disintegrate the surface. The details of the aerodynamic processes of heat transfer are neglected in this generalized approach.

The second basic assumption is that the radiant energy observed is proportional to the rate of mass loss multiplied by the kinetic energy of this mass with respect to still air. This concept is one in which the luminosity of the meteor does not arise in large part from the surface of the meteoroid but from the interaction of the escaping material of the meteoroid with the surrounding atmosphere. The luminosity may actually originate a considerable distance behind the meteoroid as the escaping material is stopped upon striking the air molecules. The observed luminosity is known to be produced by atomic recombinations and de-excitations, while the ionization or excitation must probably have been produced by atomic or molecular collisions.

A third basic assumption concerns the drag of the meteor by the resistance of the atmosphere and is formulated according to the classical Newtonian "putty-ball" model.

Some general remarks may be of interest. Because of the low atmospheric densities involved, the Reynolds' number is generally of the order of 10^4 or below, indicating a non-turbulent situation. It is generally recognized, however, that in such circumstances of extremely high Mach numbers the Reynolds' number usually does not play a critical role. In the present formulation of the theory the concept of a shock wave has not as yet proved useful, although there is, presumably, a shock wave attached to the gaseous envelope near the surface of the meteoroid. Even for the highest and fastest meteors the aerodynamic situation does not become that of free molecular flow. That is to say, a sufficiently thick air (and meteoritic) cap is formed in front of the meteoroid that the mean free paths of incoming molecules are less than the thickness of the cap. It is probable, however, that near the beginning of the faintest and fastest meteor trails the situation does not deviate greatly from free molecular flow.

3.2. Basic Relationships

The notations used in this chapter will be summarized at the end. The *drag* equation for a meteoroid of mass, m, moving with velocity, v, through an atmosphere of density, ρ, is given by

$$m\frac{dv}{dt} = -\Gamma\mathfrak{A}\rho v^2 \tag{1}$$

where $\mathfrak{A}$ is the cross-sectional area of the meteoroid and Γ the drag coefficient. This equation is based upon the principle that the mass of air encountered in time, dt, is given by $\mathfrak{A}\rho v$, which when multiplied by the velocity gives the momentum possibly transferable to the surface. The dimensionless *drag coefficient*, Γ, measures the efficiency of this process. We shall neglect here certain possible refinements of the theory based upon the added momentum carried by outgoing particles, both air and meteoritic, from the surface.

The second basic equation is that for the loss of mass. Here the element of air encountered in time, dt, multiplied by $\frac{1}{2}v^2$, measures the available amount of energy. Of this energy a fraction, Λ, the *heat transfer coefficient*, succeeds in vaporizing an element, dm, of mass, which mass element is given by the above product divided by a quantity of heat per unit mass, ζ. The resulting equation is

$$\frac{dm}{dt} = -\frac{\Lambda}{2\zeta}\mathfrak{A}\rho v^3 \tag{2}$$

The third basic equation is that for the luminous intensity. Here a fraction, τ, of the energy of the escaping mass with respect to air, $v^2dm/2$, is efficient in producing the intensity, I. The resultant equation is

$$I = -\frac{\tau}{2}\left(\frac{dm}{dt}\right)v^2 \tag{3}$$

A secondary relationship involves the *shape factor*, A, of the meteoroid. The quantity A is defined by

$$\mathfrak{A} = Am^{\frac{2}{3}} \tag{4}$$

For a sphere of density ρ', $A^3 = 9\pi/(16\rho'^2)$. If $\rho = 4$ gm/cm^3, the value of A is about 0.5 cm^2 gm$^{-\frac{2}{3}}$ for a sphere and about 0.7 for a brick of dimensions $2 \times 3 \times 6$.

Were it necessary here to apply a numerical value to the drag coefficient Γ, I would choose the value 0.50. No discussion of this numerical value will be given; the reader who is interested may refer to a fuller discussion in a paper by Thomas and Whipple [50].

A fundamental equation is that for the luminous efficiency factor τ, given by

$$\tau = \tau_0 v \tag{5}$$

where τ_0 is a constant. Equation (5) was derived by the writer numerically from certain calculations by Öpik [32]. More theoretical work on the nature of this relationship is highly desirable. To date there is no

further justification for it except the general degree to which the meteoric theory gives correct results in determining atmospheric densities.

3.3. The Mass of the Meteoroid

It is clear from inspection of the first three basic equations that some evaluation of the mass of the meteoroid from instant to instant is necessary in determining atmospheric densities. Such a relationship can be established from equations (3) and (5). An integration for the original mass, m_∞, of the meteor leads to the equations

$$m = \frac{2}{\tau_0} \int_t^\infty \frac{I}{v^3}\, dt$$

and

$$m_\infty = \frac{2}{\tau_0} \int_{-\infty}^{+\infty} \frac{I}{v^3}\, dt \tag{6}$$

in which the meteoroid has mass, m, at time, t. Since the intensity and velocity of the meteor are continuously observed, equation (6) can be integrated numerically from the observational data to provide numerical values of the mass in terms of the luminous efficiency factor, τ_0.

A second valuable relationship relating mass and velocity is due to Hoppe and may be derived by dividing equation (2) by equation (1), thus eliminating the unknown area of the body and the unobservable atmospheric density. The resultant equation is

$$\frac{1}{m}\frac{dm}{dt} = \frac{\Lambda v}{2\Gamma\zeta}\frac{dv}{dt} \tag{7}$$

which, upon integration, assumes the form

$$m = m_\infty \exp\left[\frac{\sigma}{2}(v^2 - v_\infty{}^2)\right] \tag{8}$$

where $\sigma = \Lambda/(2\Gamma\zeta)$.

As we shall see below, σ is an extremely important quantity in the meteoric theory, measuring the rate at which meteoric mass disintegrates, in terms of the density and deceleration. It is of great interest that the quantity σ can be evaluated directly from the observational data for each individual meteoroid. From equations (1), (2), (4), (5) and (6) one can derive the relationship

$$\sigma = \frac{\Lambda}{2\Gamma\zeta} = \left(\frac{I}{v^3}\right)\left(\int_t^\infty \frac{I}{v^3}\, dt\right)^{-1}\left(v\frac{dv}{dt}\right)^{-1} \tag{9}$$

All of the quantities involved in the right member of equation (9) are obtained directly from the observations.

3.4. Determination of Atmospheric Densities

The most precise determinations of atmospheric densities from photographic meteor observations are obtained from equation (1) when the deceleration, dv/dt, is known. When one substitutes the expression for the cross-sectional area from equation (4) and the mass from equation (6) into equation (1), one derives the following relationship for the atmospheric density

$$\rho = -\frac{2^{\frac{1}{3}}}{\Gamma A \tau_0^{\frac{1}{3}}} \left(\int_t^{\infty} \frac{I}{v^3}\, dt \right)^{\frac{1}{3}} \frac{1}{v^2} \frac{dv}{dt} \tag{10}$$

All of the quantities in the righthand member of equation (10) are observable except for the factor $2^{\frac{1}{3}}/(\Gamma A \tau_0^{\frac{1}{3}})$, which Jacchia designates by the quantity K. Although the theoretical values of this constant are rather close to those that can be obtained by known atmospheric densities at relatively low heights, we may better consider the atmospheric densities determined in this fashion as subject to a constant correction in the logarithm. Thus we arbitrarily fit the observed density curves in the altitude range from 40 to 60 km. It is of considerable interest in equation (10) that the least accurately determined quantity, the intensity, enters only to the $\frac{1}{3}$-power while the most accurately determined quantity, the velocity, enters effectively to the 3rd power. It is for this reason that the densities determined by the meteoric method have a considerable degree of reliability in spite of the many uncertainties in theory and in the fundamental character of the bodies themselves.

A second useful method for determining atmospheric densities can be applied at a point near the beginning of the meteor trail. From equations (2) and (3), evaluating the mass by equation (6), we find the following expression for atmospheric density

$$\rho = \frac{4\zeta}{\Lambda A \tau_0} \frac{I}{v^6} \frac{1}{m^{\frac{1}{3}}} \tag{11}$$

In practice, the determinations of density by equation (11) are less certain than those from equation (10), chiefly because the intensity of the meteor enters to the first power instead of the one-third power and because irregularities in the meteoroid are apt to introduce more serious discrepancies. On the other hand, the use of equation (11) is valuable because it carries the determinations to higher altitudes than equation (10).

Two other methods for determining densities have been used but will not be described in detail here. It is possible to determine the atmospheric pressure from the observational data at the apparent end of the

meteor trail and the density at the position of maximum light. The four methods give results that are in very close agreement [51] and establish considerable confidence in the theory generally, even though it is very simple in form. One fact is of considerable interest—that the various methods of determining the density involve in all only the two general constants, K and σ, listed above. No astroballistic quantity other than σ can be obtained directly from the photographic observations of meteors nor does the theory require the use of any other combinations of the basic constants than are given in the coefficients K and σ.

3.5. Comments on the Nature of the Meteoroid

No method exists for measuring the physical structure of a meteoroid directly. Since such knowledge is very important in the general development of the theory as well as in the evaluation of the constants, we shall mention briefly some of the observational data that bear on this problem.

Jacchia's study [52] of individual photographic meteors has shown that the value of σ varies over nearly a factor of 10 from one meteor to another with a mean value of 1.8×10^{-11} (in cgs units). A long discussion of the general problem of heat transfer and the significance of σ has been made by Thomas and Whipple [50] and the repetition of the discussion would be out of place in this article. Nevertheless, it is worth pointing out that for an individual meteor for which σ can be determined at several altitudes, the range of numerical values is far smaller than from one meteor to the next. This fact appears to be best explained as a consequence of fragmentation of the meteoroid as well as of fusion and vaporization.

Other evidence for fragmentation arises from the frequent occurrence of bright flares in the meteor trails. Jacchia [52] has demonstrated rather conclusively that these flares arise from breakage or large-scale fragmentation of the meteoroid. His argument rests on the fact that meteors without flares follow the theoretical light curves as indicated by the theory given here, with a surprising precision. On the other hand, meteors that show increasingly greater degrees of irregularity in brightness tend to follow the theoretical curves in the early stages of the phenomenon but fade away progressively sooner than theory would predict. This result indicates that the substance of the meteoroid is wantonly expended in producing the flares. It is reasonable to ascribe this loss of substance and increase in brightness to an actual breakage of the body in flight. Dust and debris carried to relatively low altitudes and high atmospheric densities by the parent mass would then be ablated very rapidly and produce a large increase in brightness. Other considerations concerning the nature of cometary masses would also suggest

that fragmentation should be a common phenomenon among the meteoroids of cometary origin.

If, on the other hand, one accepts the hypothesis that meteoroids arise from the same material as meteorites, then in most cases one would expect to be dealing with small stones, which are generally relatively weak as compared to iron meteorites. As a working hypothesis, then, it is probably best to assume that the meteoroids are relatively fragile and that they break by forces of resistance in the atmosphere, by the strain set up due to external heating, or by the penetration of heat into the interstices and irregularities of a none-too-solid body.

It is possible, too, that the production of droplets instead of vapor at the surface plays an important role in the variations of σ from one meteor to the next. Öpik [37] has discussed this question in great detail for irons and stones, but in the determination of atmospheric densities the distinction between irons and stones cannot yet be made.

Readers who are further interested in the meteoritic processes and particularly in the problem of heat transfer at high velocities may refer to a paper by Cook, Eyring, and Thomas [53] as well as the above-mentioned paper [50]. The term *astroballistics* has been coined to cover the problems of heat transfer and general ballistic problems when the moving body suffers appreciable ablation as a result of its motion through the air.

4. Results Concerning the Upper Atmosphere

4.1. Densities in the Upper Atmosphere

The first results concerning densities in the upper atmosphere obtained by meteor methods were presented by Lindemann and Dobson [9]. These investigators showed clearly that the older concepts of a negative temperature gradient in the high atmosphere or a constant stratospheric temperature to great altitudes were inadequate to account for the meteor observations. Since the frequency distribution of the end heights of meteors showed a minimum in the region from 50 to 60 km, they concluded that the atmospheric temperature must rise abruptly at a level near 60 km. Although their detailed argument must be revised in terms of later theory, their general conclusion stands. Their estimates of atmospheric densities and temperatures above this region were generally much too high.

The accumulation of appreciable data by the Harvard double-station photographic technique, in which the photographic cameras are equipped with rotating shutters, has led to a number of results concerning the densities in the upper atmosphere. The results to 1943 were published

by the writer [51] and showed clearly the existence of a temperature maximum in the neighborhood of 60 km and a temperature minimum near 80 km. The general adequacy of the theory as presented in Section 3 was demonstrated by the consistent results obtained from four different methods of determining the atmospheric density or pressure. These various methods applied at the points on a meteor trail where deceleration could be measured, at an early point in the trail, a point near the end and the point of maximum light. This internal consistency of the theory and the general consistency of the results with those obtained by other methods indicated that the density determinations by meteor decelerations could be relied upon statistically to give a rather good determination of upper-atmospheric densities as a function of height in the range from roughly 60 to 100 km. The density curve could be extended to about 120 km with somewhat less precision by the method of beginning points. The Tentative Atmosphere adopted by the National Advisory Committee on Aeronautics in 1947 [54] was greatly influenced by the meteor results above an altitude of roughly 60 km.

Since 1946 the Harvard photographic meteor program has been greatly assisted and enlarged by a contract with the U. S. Naval Bureau of Ordnance, which also financed the reductions of the photographic meteors by the Center of Analysis at the Massachusetts Institute of Technology. These reductions have been under the general direction of Z. Kopal and actively carried out under the direction of L. G. Jacchia. Summaries of the upper-atmospheric results from meteors photographed by Harvard at its stations in Massachusetts have been prepared by Jacchia [52, 55, 56]. The mean determinations from the decelerations of some 40 meteors are shown by circles in Fig. 5 while those from the beginning points are indicated by x's. It will be seen that these density determinations are in fairly close agreement with the NACA Tentative Standard Atmosphere [54], represented by a dashed curve.

Beginning in 1947, the United States Army Ordnance rocket program, involving instrumentation by various government agencies and research institutions, has led to remarkable progress in our knowledge concerning the pressures, densities, and temperatures in the upper atmosphere. Some results of this program, primarily those by Havens, Koll, and Lagow [57] of the Naval Research Laboratory, are shown by a short-dash curve in Fig. 5. Measures by Dow and Spencer [58] of the University of Michigan, under a U. S. Air Force contract, also contributed to this curve.

It will be seen from Fig. 5 that the meteoric results and those obtained by rockets begin to diverge appreciably above approximately 65 km. For some years the reason for this divergence was uncertain. The

uncertainty has, however, been removed partially by the reduction of meteor observations made at Harvard stations supported by the U. S. Naval Bureau of Ordnance near Las Cruces, New Mexico, only a few miles from the White Sands Proving Ground where the rocket firings were conducted. Tentative unpublished reductions of these data by Jacchia show that densities measured by meteors over New Mexico are

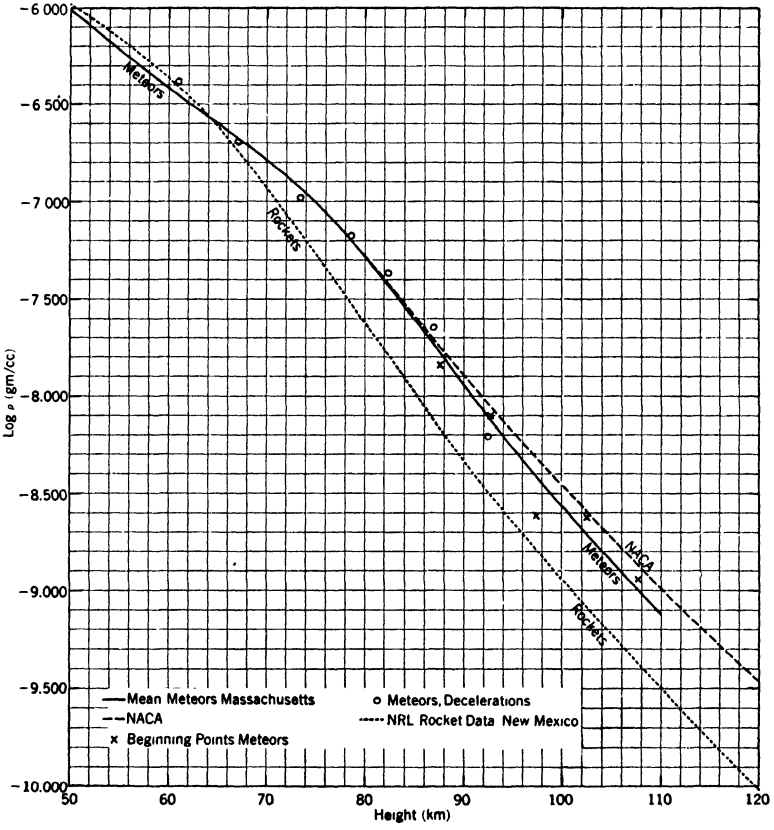


FIG. 5. Height-density results for upper atmosphere.

in good agreement with those obtained in the rocket program (the meteor values of log ρ are systematically higher by 0.08).

There is every reason to believe that the discrepancy between the rocket and meteoric results as indicated in Fig. 5 arises largely from an intrinsic difference in the upper atmosphere over New Mexico and over Massachusetts. The latter observations were made at a latitude of approximately 42° while the former were made near latitude 32°. A geographical effect is clearly indicated; at present there is not sufficient

evidence to attribute it purely to latitude rather than to some more general effect conceivably associated with a mid-continental region as compared to a seaboard region.

It appears, then, that at an altitude of approximately 70 km the densities in the upper atmosphere over New Mexico decrease somewhat more rapidly with height than those over Massachusetts. No simple or obvious explanation for this effect is apparent.

The use of the Super-Schmidt meteor cameras, capable of providing precision observations of many more meteors than could be observed with the smaller classical photographic cameras, should make possible the detailed simultaneous comparison of atmospheric densities by rocket and meteor techniques under identical atmospheric conditions. Jacchia [52] has found that the probable error of a single determination of $\log \rho$ is approximately ± 0.08 from a single deceleration measure.

4.2. Temperatures and Pressures in the Upper Atmosphere

From the perfect gas laws it follows that measures of density, pressure and the velocity of sound in the upper atmosphere may be interchangeably interpreted so long as the measures extend over a finite range in altitude. Uncertainties or variations in the mean molecular weight do not affect the interrelationships among these three quantities except as the ratio of the specific heats for a gas may be involved. Since this ratio, γ, does not change markedly for the various gases which may be expected in the upper atmosphere, little uncertainty is introduced by transferring measures of pressure, density, or the velocity of sound to either of the other two parameters.

To date no method has been developed for the *direct* measurement of temperature in the upper atmosphere by rockets or meteors. Hence, the temperature must always be deduced from measures of pressure, density or the velocity of sound (including Mach number determinations). In all three cases temperature determinations involve the mean molecular weight, in the sense that the deduced temperature is linearly proportional to an assumed mean molecular weight.

Figure 6 represents various determinations of temperature as a function of altitude, on the assumption that the mean molecular weight is independent of altitude. The "NRL" curve is calculated from the rocket pressure data over New Mexico by Havens, Koll, and Lagow [57] while the "NACA" curve represents the Tentative Standard Atmosphere [54]. The "Meteors" points were derived by Jacchia [52, 56] from Harvard photographic meteors over Massachusetts. The "S.C.E.L. Grenade" points are from unpublished data obtained by the Evans Signal Laboratory at Belmar, New Jersey. The "SCEL" data were obtained

by measuring the vertical velocity of sound from grenades ejected from an Aerobee rocket over New Mexico.

More recent temperature data (to October 1951) combined from all the rocket research at the White Sands Proving Ground indicate good consistency among the various methods. The mean Temperature-Height curve averaged over some 16 firings reaches a maximum value of some 270°K near a height of 50 km and a minimum of about 200°K near 80 km. For a mean molecular weight of 29 the temperature is about 240°K at 100 km and rises to perhaps 500°K at 150 km.

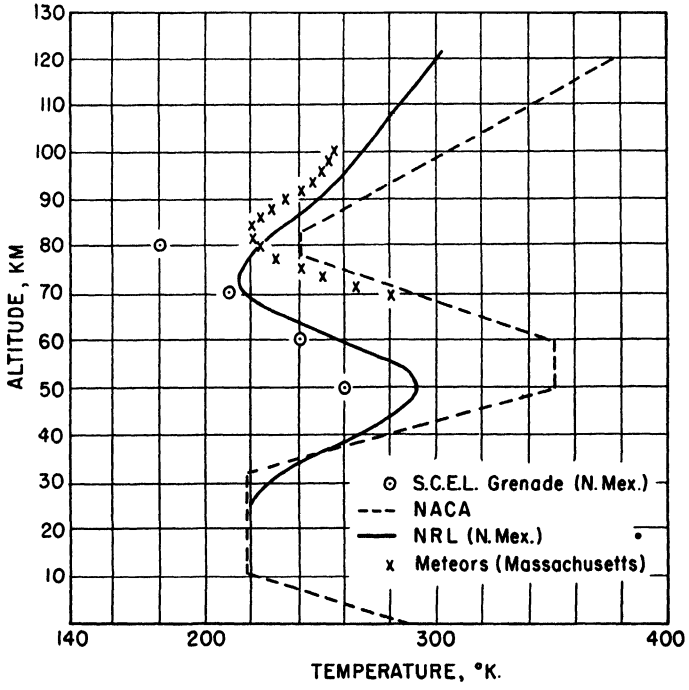


FIG. 6. Temperature-altitude diagram for upper atmosphere.

Near the 100-km level, of course, there is reason to believe that an appreciable dissociation of O_2 should occur, reducing the mean molecular weight and, consequently, the *derived* temperatures (but not the actual temperatures). Also, there is now some evidence for diffusive separation in the atmosphere beginning between 64 and 72 km. Three atmospheric samples have been obtained recently in this height range by L. M. Jones and E. A. Wenzel of the University of Michigan under a U. S. Army Signal Corps contract. Analysis by Chackett, Paneth, Reasbeck, and Wiborg [59] indicates that the He/N_2 ratio increases by as much as a

factor of 2 near 70 km. Furthermore, the Ne/N_2 ratio increases slightly and the A/N_2 ratio decreases slightly, in the directions to be expected from diffusive separation.

The decrease in the mean molecular weight of the atmosphere may become quite appreciable in the 150-kilometer level and above, so that the temperature will remain moderate. Very high temperatures (1000–1500°K) have been deduced for the exosphere (see, e.g., Spitzer [60]). From present indications the exosphere, where molecular collisions become unimportant, may begin as low as 300–400 km.

Returning to Fig. 6, we note that the temperature extremes (near 50 and 80 km) represent averaged values. It is highly probable that at any given time the actual extremes are much greater than indicated here and that the heights of the extremes vary appreciably. The actual maximum may well be 20°K higher or the minimum 20°K lower. Detailed refinement in this regard is observationally difficult.

The existence of a temperature maximum at the 50-kilometer level is generally explained as the result of the beginning absorption of ozone. Even though the concentration of ozone is very low in this region, it is a very effective absorber of solar ultraviolet light. The O_2 absorption presumably raises the temperature near the 100-km level, but at 80 km the O_2 absorption is complete while the O_3 absorption has not begun; hence, the minimum temperature occurs near 80 km. Various strong atomic and molecular absorptions occur above 100 km, raising the temperature further.

There is little doubt that the atmospheric temperature in the general level of 50–60 km is higher on the average over Massachusetts (lat. 42°.5 N) than over New Mexico (lat. 31–32°N). The meteor data indicate this effect clearly, although the amount of the effect is undoubtedly smaller than indicated in Figs. 5 and 6.

Some pressure data in the upper atmosphere are presented in Table II.

TABLE II. Atmospheric pressures over New Mexico.

Height	Pressure	Height	Pressure
km	mm Hg	km	mm Hg
0	7.6×10^{2}	80	1.1×10^{-2}
10	2.1×10^{2}	90	2×10^{-3}
20	4.2×10	100	6×10^{-4}
30	9.5	110	2×10^{-4}
40	2.4	120	6×10^{-5}
50	7.6×10^{-1}	130	2×10^{-5}
60	2.2×10^{-1}	140	7×10^{-6}
70	5.5×10^{-2}	160	2×10^{-6}

These values were obtained from the data presented by Havens, Koll, and Lagow [57] of the U. S. Naval Research Laboratory. The last entry in Table II was kindly made available by these investigators in advance of publication.

4.3. Variations in the Upper Atmosphere

We have already discussed the differences in the densities and temperatures of the upper atmosphere over Massachusetts as compared to New Mexico. Newell [61], in charge of the rocket research at the U. S. Naval Research Laboratory, finds that atmospheric pressure measures made by rocket near the equator from the U. S. naval vessel Norton Sound are in good agreement with those made over New Mexico. It might appear, then, that the geographical effect observed between New Mexico and Massachusetts is not purely attributable to differences in latitude.

Furthermore, we have mentioned that the temperature extremes (near 50 and 80 km) are undoubtedly underestimated in a mean-temperature curve. A perusal of the rocket data [e.g., 57] indicates that appreciable changes in pressure and density have occurred from time to time. It is still too soon, however, to make quantitative measures of these variations. The rocket data are insufficient in number and the internal comparisons are too few to make such determinations significant.

The Harvard meteor data over Massachusetts, however, show clear-cut evidence for a seasonal effect. Whipple, Jacchia and Kopal [62] have shown that the density at a mean height of 78 km follows the normal mean ground temperature variation with season. The total amplitude is about 0.4 in log density with a correlation coefficient of 0.91 in the direct sense; i.e., the density is greatest in summer and least in winter. The correlation coefficient is reduced to 0.85 with actual ground temperature and to 0.66 with insolation.

The seasonal variation is illustrated in Fig. 7. The abscissa is the normal mean ground temperature in Boston and the ordinate is the residual in density, $\Delta \log \rho$, from the mean. Subdivisions of the data according to velocity groups show that the seasonal variation does not, in appreciable measure, arise from chance seasonal correlations with meteor velocities.

Jacchia [56], in a later study, finds that the total amplitude of the seasonal variation decreases with height, becoming very uncertain above about 85 km. His values of the total amplitude in log density are 0.43 at 64 km, 0.29 at 77 km, 0.14 at 87 km and 0.12 at 95 km. Wexler [63], from a theoretical study, finds evidence to support the meteor results concerning seasonal variations in the upper atmosphere.

Basing his conclusions on visual meteor studies made near Flagstaff, Arizona, Öpik [64] found a seasonal variation in the mean heights of meteors. Numerically the result is in good agreement with those found photographically; the total amplitude in height is 3.7 kilometers. From the unpublished tentative photographic meteor data in New Mexico, Jacchia finds evidence for a similar seasonal effect there.

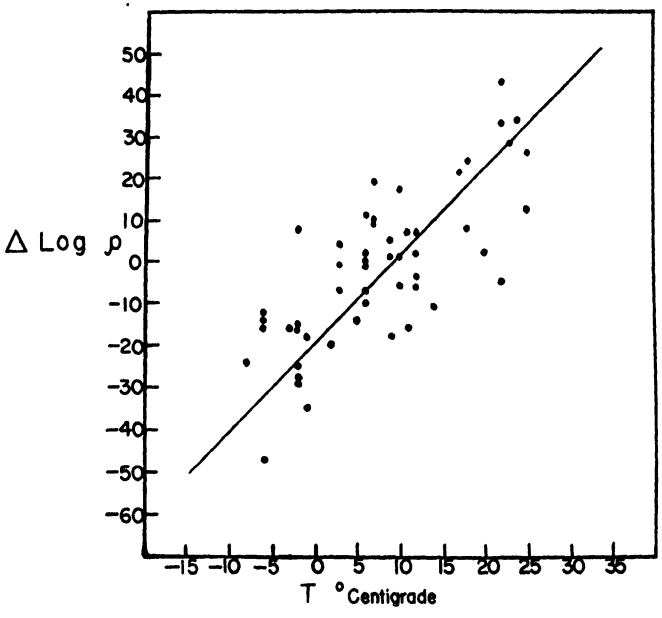


FIG. 7. Seasonal effect in upper-atmospheric densities.

Hence, the meteor data consistently demonstrate a seasonal effect in which ground temperature and upper-atmospheric densities are directly correlated. The pressures and densities determined by Havens, Koll and Lagow [57] do not indicate a seasonal effect, but only a few rocket firings were involved. More recent rocket results reported to the Upper Atmosphere Rocket Research Panel appear to be more in agreement with the meteor results as regards the seasonal effect.

Possible correlations with a number of other variables have been sought in the meteor log-density residuals. Among the variables are synoptic weather fronts, radiosonde data to 23 kilometers, sunspot activity, lunar hour-angle, solar hour-angle, hours after sunset, and others. Correlations with stratospheric temperatures are poor. There is some slight indication of a six-hour period during the night, with minimum density near local midnight. Otherwise no significant or

encouraging correlations have been found. The mean virtual height of the E-layer at a constant altitude of the sun shows a seasonal amplitude of the same order as the amplitude obtained from meteor data, but with a large phase shift in season. The rocket data show no evidence of a night-day effect in upper-atmospheric pressures or densities.

5. Circulation in the Upper Atmosphere

5.1. Winds from Meteor Trains

It has long been known that persistent trains of meteors become greatly distorted in a very few minutes. The most comprehensive study of this subject has been made by Olivier [65, 66] who has compiled observations of 1492 persistent meteor trains lasting more than one minute. Winds and heights could be estimated for only a fraction of the total observations but components of wind direction are given quite frequently. A brief summary of Olivier's compiled data is given in Table III. The trains are divided into three groups, those observed

Table III. Winds from meteor trains (After Olivier).

	Day	Twilight	Night
Height beg. (km)	45	77	104
Height mean (km)	36	61	92
Height end (km)	27	45	80
Number	12	55	27
Mean speeds (km/sec)	173		203
Number	16		50

during daylight hours, those in the twilight period, presumably illuminated by the sun at high altitude, and those observed during the night hours. The table gives mean heights of the trains at the beginning, end and middle points, mean speeds in kilometers per hour and the number of trains observed in each category. The mean speeds are all based upon observed heights obtained by triangulation from two or more stations. Olivier reduced twenty-five additional observations of speed from angular velocities by adopting a mean height; from this he derived a mean speed of 182 km/hr. for twenty-five night trains.

It will be seen from the table that the height of the persistent trains increases from day to night; Olivier believes that the corresponding increase in velocity is also real. He states that 166 trains were described with drawings or photographs and that all of them show a zigzag shape.

In some cases he finds good evidence for vertical components of motion in the upper atmosphere. There seems no doubt that in regions of the atmosphere from 30 to over 100 kilometers the winds are generally high and variable with height. Furthermore, it is very likely that a considerable degree of turbulence exists in addition to horizontal winds.

Olivier has made no effort to analyze the observed data of wind directions according to season but has made divisions according to geographical position and time of day. He includes a number of wind vector diagrams based upon directions only, not including velocities. A rough tabulation of his results is given in Table IV showing the most common component of wind vector and the second strongest component (or its sector), as well as the direction of minimum activity. These directions are given as *vectors of motion* indicating the direction *towards* which the motion occurs rather than the direction *from* which the wind appears to come.

Similar data have been collected by Fedynsky [67] on a systematic program of observation in the U.S.S.R. He finds the strongest single prevalent vectors in the northerly direction with a comparable number falling in the southeast quadrant. There are very few velocity vectors with a westward component. Generally speaking, Fedynsky includes very much higher winds than Olivier (as much as 1200 km/hour). Olivier is of the opinion that many of these observations depend upon too short a time interval and that the unusually high speeds must be accepted with reserve.

The data in Table IV are not in very good agreement with Hulburt's [68] analysis of meteor-train observations collected by S. Kahlke [69],

TABLE IV. Frequencies of wind vectors (directions of motion).

	Time of Day								
	Day and Twilight			Night Average			Night after Midnight		
Area	Max	2nd	Min	Max	2nd	Min	Max	2nd	Min
America	N & E	W	...	N	E	SW	N	E	SW
W. Europe	..	...	...	W	SE	NE	W	E	...
Europe and W. Asia	E	S	SW	...	...	..	...	...	.
E. Europe and W. Asia	...	...	...	ESE	SW	N	ESE	N	SW

both of whom find a generally westerly drift for daylight trains. Hulburt developed a seasonal as well as hourly analysis and presented a simple circulation theory based on the principle of daylight heating in

the upper atmosphere. Probably the data are too few to give a satisfactory statistical result.

More recently Hoffmeister [70] has summarized his previous studies of upper-atmospheric circulation as a part of a summary paper. His observations have been of *Leuchtstreifen*, or irregularities in brightness of the non-polar aurora. He concludes that at Sonneberg (lat. 50°N) the air motion at an altitude of some 120 km is in the NE quadrant during the summer (May to September) with a mean speed of 50 m/sec; in winter (October to March) the motion is almost entirely with an East component, ENE at 65 m/sec and S at 89 m/sec. In Southwest Africa (lat. 23°S) the winds are more uniform with a stronger S component. He interprets these upper-atmospheric motions (40–150 km) as representing circulation towards the poles from the equator, becoming stronger and more easterly to a maximum velocity entirely eastward at a high latitude. Correspondingly, there is a circulation from the pole southward to this discontinuity. He believes that the discontinuity is farther north in summer than in winter. He does not discuss the night-day variations. A more complete analysis of all data now available may clarify the general circulation problems.

5.2. The Radio-Meteor Group of Stanford University

The Radio-Meteor group of Stanford University, headed by Manning and Villard [27], have concentrated on the problem of measuring upper atmospheric winds by the motion of the ion columns remaining after meteors. Their radio system operates generally at 23 mc/sec and they use a wide-beam antenna with a direction-finding system to locate the azimuth from which meteor reflections occur. In earlier work they depended upon continuous wave transmission, measuring the Doppler effect by means of the beat frequency with an oscillator at two identical receivers, the beat frequency being retarded 90° in phase in one as compared to the other. They found that this method of measuring the so-called "Body-Doppler" is difficult to apply during daylight hours because of the interference of reflections from the E-layer. As a consequence they have more recently added a pulsed system which enables them to measure the Body-Doppler of the persistent meteor ionization more effectively during daylight hours.

They find a systematic and positive Body Doppler radial motion for meteor ion columns averaged over the sky. Their normal procedure for determining the mean wind vector in the upper atmosphere is to observe the Body-Doppler for a large number of meteors at all azimuths over the sky. There is a sufficient number in one hour to make a significant determination. They average the wind vectors with respect to

azimuth and subtract the systematic positive component to obtain the mean vector during the interval. They can also determine the average wind speed from the same source material. As we have seen, the heights of meteor ion columns do not vary greatly, so the results apply in the general region of 100 km altitude.

From such measures made during the summer of 1950, Manning and Villard [71] obtained an average wind velocity of 125 km/hour and an average wind speed of 275 km/hour. From 18 mean wind velocities obtained in the interval from May 29 to September 11, 1949, they found the velocities in the range from 47 to 210 km/hour with an average of 105 km/hour. The majority of the motions during the summer seem to concentrate along the north-south line, perhaps closer to the magnetic axis. For the entire season they now find that the average winter drifts tend to be smaller in the early morning than in the late evening while during the summer the mean drifts are larger in the early morning.

They suggest as a possible explanation of the systematic recession effect that variation in height (and corresponding wind velocity) with meteor velocity may be responsible. Even though the most common wind directions are magnetic north and south, they can find no dependence of the growth of the trail upon the direction of magnetic field nor any indication of the *dip* of the compass. They believe that neither the motion nor the rate of growth of the electron clouds are affected materially by the earth's magnetic field.

Their general view concerning the nature of the circulation of the upper atmosphere changed appreciably as experimentation proceeded. They have become of the opinion that horizontal winds in the upper atmosphere are highly stratified in various regions of thickness only 5–10 km in altitude. In these regions they believe that the winds are of the order of 300 km/hour. In the intermediate thin layers of comparable thickness they believe that turbulent motions again are of the order of 300 km/hour.

The Stanford group has in the course of the analysis of the general radio technique of wind determination made a great many studies bearing upon the structure and growth of the electron column produced by a meteoroid. Any reader particularly interested in diffusion, fading, decay, and other phenomena of such ion columns will find a wealth of information collected in the research by the Stanford group and in papers by Greenhow [72] and Ellyett [73]. A comprehensive theory of the detailed processes has not yet been developed but we may anticipate great theoretical progress in the near future. A short report of a conference on "Winds and Turbulence in the Upper Atmosphere" is contained in *The Observatory* [74] and includes discussion by several investigators.

List of Symbols

ρ	Atmospheric Density
ρ'	Density of the Meteoroid
m	Mass of the Meteoroid at Time t
m_∞	Original Mass of Meteoroid
$\mathfrak{A}$	Cross Sectional Area of the Meteoroid
A	Shape-Density Factor of the Meteoroid
v	Velocity of the Meteoroid at Time t
v_∞	No-Atmosphere Velocity of Meteoroid
Γ	Drag Coefficient
Λ	Heat Transfer Coefficient
ζ	Heat of Vaporization and/or Fusion
I	Light Intensity
τ_0	Luminous Efficiency Factor
τ	$\tau_0 v$
σ	$\Lambda/(2\Gamma\zeta)$
k	$2^{\frac{1}{2}}/(\Gamma A \tau_0^{\frac{1}{2}})$
NACA	National Advisory Committee for Aeronautics
NRL	Naval Research Laboratory
SCEL	Signal Corps Engineering Laboratories.

References

1. Watson, F. G. (1941). Between the Planets. The Blakiston Company, Philadelphia, 214 pp.
2. Olivier, C. P. (1925). Meteors. Williams and Wilkins Co., Baltimore, 272 pp.
3. Hoffmeister, C. (1937). Die Meteore. Akademische Verlagsgesellschaft m.b.H., Leipzig, 129 pp.
4. Shapley, H., Öpik, E., and Boothroyd, S. (1932). The Arizona expedition for the study of meteors. *Proc. Natl. Acad. Sci. U. S.*, **18**, 16–23.
5. Öpik, E. (1934). Results of the Arizona expedition for the study of meteors. II. Statistical analysis of group radiants. *Circ. Harvard Coll. Obs.*, No. 388, 38 pp.
6. Öpik, E. (1936). Results of the Arizona expedition for the study of meteors. VI. Analysis of meteor heights. *Ann. Harvard Coll. Obs.*, **105**, 549–600.
7. Öpik, E. (1934). Results of the Arizona expedition for the study of meteors. III. Velocities of meteors observed visually. *Circ. Harvard Coll. Obs.*, No. 389, 1–9.
8. Denning, W. F. (1899). General catalogue of the radiant points of meteoric showers and of fireballs and shooting stars observed at more than one station. *Mem. Roy. Ast. Soc.*, **53**, 203–292.
9. Lindemann, F. A., and Dobson, G. M. B. (1922). A theory of meteors, and the density and temperature of the outer atmosphere to which it leads. *Proc. Roy. Soc. (London)*, **102**, 411–437; (1923). A note on the temperature of the air at great heights. *ibid.*, **103**, 339–342.
10. McIntosh, R. A. (1940). Seasonal variation in the height of meteors. *Monthly Not. Roy. Ast. Soc.*, **100**, 510–528.
11. Porter, J. G. (1944). An analysis of British meteor data: Part 2. Analysis. *Monthly Not. Roy. Ast. Soc.*, **104**, 257–272.
12. Link, F. (1936). Exploration météorique de la haute atmosphère, *Časopis p. pestovdni mat. fiys., roč.*, **71**, 79–90.

13. Elkin, W. L. (1900). The velocity of meteors as deduced from photographs at the Yale observatory. *Astrophys. J.*, **12**, 4–7.
14. Whipple, F. L. (1938). Photographic meteor studies. I. *Am. Phil. Soc.*, **79**, 499–548.
15. Millman, P. M. (1949). One hundred meteor spectra (Abstract). *Ast. J.*, **54**, 177–178.
16. Vyssotsky, A. N. (1940). A Meteor Spectrum of High Excitation, *Astrophys. J. U. S.*, **91**, 264–266.
17. Millman, P. M. .(1950). Spectrum of a meteor train, *Nature*, **165**, 1013–1014.
18. Lovell, A. C. B. (1948). Meteoric ionization and ionospheric abnormalities. *Rept. Prog. Phys.*, **11**, 415–442.
19. Herlofson, N. (1948). The theory of meteor ionization. *Rept. Prog. Phys.*, **11**, 444–453.
20. Hey, J. S. (1949). Reports on the progress of radio astronomy. *Monthly Not. Roy. Ast. Soc.*, **109**, 179–214.
21. Chamanlal, C., and Venkataraman, K. (1941). Whistling meteors—Doppler effect produced by meteors entering the ionosphere. *Electrotechnics*, **14**, 28–40.
22. McKinley, D. W. R. (1951). Meteor velocities determined by radio observations. *Astrophys. J.*, **113**, 225–267.
23. Pierce, J. A. (1938). Abnormal ionization in the E-region of the ionosphere. *Proc. I.R.E.*, **26**, 892–902.
24. Blackett, P. M. S., and Lovell, A. C. B. (1941). Radio echoes and cosmic ray showers. *Proc. Roy. Soc. (London)*, **177**, 183–186.
25. Hey, J. S., Parson, S. J., and Stewart, G. S. (1947). Radar observations of the Giacobinid meteor shower, 1946. *Monthly Not. Roy. Ast. Soc.*, **107**, 176–183.
26. Davies, J. G., and Ellyett, C. D. (1949). The diffraction of radio waves from meteor trails and the measurement of meteor velocities. *Phil. Mag.*, Ser. 7, **40**, 614–626.
27. Manning, L. A., Villard, O. G., Jr., and Peterson, A. M. (1949). Radio doppler investigation of meteoric heights and velocities. *J. Appl. Phys.*, **20**, 475–479.
28. Almond, M., Davies, J. G., and Lovell, A. C. B. (1950). On interstellar meteors. *The Observatory*, **70**, 112–113.
29. Hey, J. S., and Stewart, G. S. (1947). Radar observations of meteors. *Proc. Phys. Soc. (London)*, **59**, 858–883.
30. Lovell, A. C. B. (1950). Notes on Mr. Rigollet's Paper. *Doc. d. Obs., Inst. d'Astrophys. Paris, Bull. No. 6–7*, 45.
31. Millman, P. M. (1950). Meteoric ionization. *J. Roy. Ast. Soc. Can.*, **44**, 209–220.
32. Öpik, E. (1933). Atomic collisions and radiation of meteors. *Harvard Coll. Obs. Reprint*, No. 100, 1–39.
33. McKinley, D. W. R. (1951). Deceleration and ionizing efficiency of meteors. *J. Appl. Phys.*, **22**, 202–213.
34. Liller, W. (1949). Radio detection of meteors at 3.5 megacycles, Cruft Lab., Harvard Univ. *Tech. Report* No. 65, 1–9.
35. Lovell, A. C. B. (1950). Meteor ionization in the upper atmosphere. *Science Progress*, **38**, 22–42.
36. Kalashnikov, A. (1949). On the induction methods for observations of magnetic fields in meteors. *Compt. rend. acad. sci. URSS*, **66**, 373–376.
37. Öpik, E. (1937). Research on the physical theory of meteor phenomena. III. Basis of the physical theory of meteor phenomena. *Tartu. Obs. Pub.*, **29**, 1–67.

38. Whipple, F. L. (1950). The theory of micro-meteorites, Part I. In an isothermal atmosphere. *Proc. Natl. Acad. Sci., U. S.*, **36**, 687–695; (1951). Part II. In heterothermal atmospheres, *ibid.*, **37**, 19–30.
39. Landsberg, H. E. (1947). A Report on dust collections made at Mount Weather and Arlington, Virginia, 1 October to 20 November, 1946. *Pop. Ast.*, **55**, 322–325.
40. Buddhue, J. D. (1950). Meteoritic Dust. The University of New Mexico Press, Albuquerque, New Mexico, 96 pp.
41. van de Hulst, H. (1947). Zodiacal light in the solar corona. *Astrophys. J.*, **105**, 471–488.
42. Allen, W. C. (1946). The spectrum of the corona at the eclipse of 1940 October 1. *Monthly Not. Roy. Ast. Soc.*, **106**, 137–150.
43. Pettersson, H., and Rotschi, H. (1950). Nickel content of deep-sea deposits. *Nature*, **166**, 308.
44. Burnight, T. R. (1950). Private communication.
45. Bohn, J. L., and Nadig, F. H. (1950). Research in the Physical Properties of the Upper Atmosphere With Special Emphasis on Acoustical Studies with V-2 Rockets. Research Inst. of Temple Univ., Report No. 8, 1–26.
46. Crozier, W. D., and Seely, B. K. (1950). Some techniques for sampling and identifying particulate matter in the air. *Proc. First Natl. Air Pollution Symposium*, Pasadena, Calif., 45–49; and private communication.
47. Maris, H. B. (1929). A theory of meteors. *Terr. Magn.*, **34**, 309–316.
48. Sparrow, C. M. (1926). Physical Theory of Meteors. *Astrophys. J.*, **63**, 90–110.
49. Hoppe, J. (1937). Die physikalischen Vorgänge beim Eindringen meteoritischer Körper in die Erdatmosphäre. *Ast. Nachr.*, **262**, 169–198.
50. Thomas, R. N., and Whipple, F. L. (1951). The Physical Theory of Meteors. II. Astroballistic heat transfer. *Astrophys. J.*, **114**, 448–456.
51. Whipple, F. L. (1943). Meteors and the earth's upper atmosphere. *Rev. Mod. Phys.*, **15**, 246–264.
52. Jacchia, L. (1949). Photographic meteor phenomena and theory, Tech Report No. 3. *Harvard Coll. Obs. Reprint Series*, **II-31**, 1–36.
53. Cook, M. A., Eyring, H., and Thomas, R. N. (1951). The physical theory of meteors. I. A reaction-rate approach to the rate of mass loss in meteors. *Astrophys. J.*, **113**, 475–481.
54. Warfield, C. N. (1947). Tentative tables for the properties of the upper atmosphere. *Natl. Adv. Com. Aeronaut., Notes*, No. 1200, 1–50.
55. Jacchia, L. (1948). Ballistics of the upper atmosphere, Tech. Report No. 2. *Harvard Coll. Obs. Reprint Series*, **II-26**, 1–30.
56. Jacchia, L. (1949). Atmospheric density profile and gradients from early parts of photographic meteor trails, Tech. Report No. 4, *Harvard Coll. Obs. Reprint Series*, **II-32**, 1–12.
57. Havens, R., Koll, R., and Lagow, H. (1950). Pressures and temperatures in the earth's upper atmosphere. *Reprint Naval Research Lab.*, 1–13.
58. Dow, W. G., and Spencer, N. W. (1950). Final Report Pressure and Temperature Measurements in the Upper Atmosphere, Engineering Research Inst., Univ. Mich.
59. Chackett, K. F., Paneth, F. A., Reasbeck, P., and Wiborg, B. S. (1951). Variations in the chemical composition of stratosphere air. *Nature*, **168**, 358–360.
60. Spitzer, L. (1947). The terrestrial atmosphere above 300 km. In The Atmospheres of the Earth and Planets. Ed. by G. P. Kuiper, The University of Chicago Press, Chicago, Chapter VII, 213–249.
61. Newell, H. (1951). Private communication.

62. Whipple, F. L., Jacchia, L., and Kopal, Z. (1947). Seasonal variations in the density of the upper atmosphere. In the Atmospheres of the Earth and Planets. Ed. by G. P. Kuiper, The University of Chicago Press, Chicago, Chapter V, 149–158.
63. Wexler, H. (1950). Annual and diurnal temperature variations in the upper atmosphere. *Tellus*, **2**, 163–273.
64. Öpik, E. (1936). Meteor heights from the Arizona expedition. *Proc. Natl. Acad. Sci., U. S.*, **22**, 525–530.
65. Olivier, C. P. (1942). Long enduring meteor trains. *Proc. Am. Phil. Soc.*, **85**, 93–135.
66. Olivier, C. P. (1947). Long enduring meteor trains: Second paper. *Proc. Am. Phil. Soc.*, **91**, 315–327.
67. Fedynsky, V. V. (1944). Results of observations of meteor trains in Jadjikistan (1934–1938). *Ast. J. Sov. Un.*, **21**, 291–306.
68. Hulburt, E. O. (1932). On winds in the upper atmosphere. *Pub. Ast. Soc. Pacific*, **44**, 178–182.
69. Kahlke, S. (1921). Meteorschweife und hochatmosphärische Windströmungen. *Ann. Hydrog.*, **49**, 294–299.
70. Hoffmeister, C. (1951). Spezifische Leuchtvorgänge im Bereich der mittleren Ionosphäre, *Erg. exakt. Naturw.*, **24**, 1–53.
71. Manning, L. A., and Villard, O. G., Jr. (1947–1951). Quarterly Status Reports Nos. 1 to 14 to the Office of Naval Research, Electronics Research Laboratory, Stanford Univ., Calif.
72. Greenhow, J. S. (1950). The fluctuation and fading of radio echoes from meteor trails. *Phil. Mag., Ser. 7*, **41**, 682–693.
73. Ellyett, C. D. (1950). The influence of high altitude winds on meteor trail ionization. *Phil. Mag., Ser. 7*, **41**, 694–700.
74. Various authors. (1951). Winds and turbulence in the upper atmosphere. *The Observatory*, **71**, 104–109.

Unsolved Problems in Physics of the High Atmosphere

N. C. GERSON

Geophysics Research Division, Air Force Cambridge Research Center, Cambridge, Massachusetts

CONTENTS

	Page
1. Introduction	156
2. The Terrestrial Atmosphere	158
2.1. General	158
2.2. Temperature-Altitude Relationship	164
2.3. The Ionosphere	167
2.4. Atmospheric Emissions	169
2.5. High Altitude Winds	176
3. Static Properties and Processes of the High Atmosphere	179
3.1. Temperature	182
3.1.1. Theoretical Determination of Temperature	183
3.1.2. Gas and Rotational Temperatures	185
3.1.3. Interferometric Determinations	187
3.2. Composition	188
3.2.1. Compendium: Spectra of Atmospheric Gases	190
3.2.2. Emission Spectra of the Atmospheric Gases	190
3.2.3. Absorption Spectra of the Atmospheric Gases	192
3.2.4. Solar Infrared Absorption Spectrum	194
3.2.5. Emission Altitude of the Airglow	196
3.2.6. Photochemical Equilibrium	197
3.3. Collisional Phenomena	199
3.3.1. Collision Frequencies in the Ionosphere	202
3.3.2. Cross Section for Elastic Collision	204
3.3.3. Cross Section for Energy Absorption	206
3.3.4. Cross Section for Collisional Excitation and Ionization	208
3.3.5. Cross Section for Recombination and Association	211
4. Dynamics of the Ionosphere and Mesosphere	212
4.1. Wind Observations	215
4.1.1. Movement of Ionospheric Irregularities	217
4.1.2. Movement of Clouds	219
4.2. Tides	221
4.2.1. Theory of Tidal Oscillations	222
4.2.2. Statistical Analysis of Atmospheric Tides	224
4.3. Diffusion	226
4.3.1. Diffusion in Magnetic and Electric Fields	228
4.3.2. Diffusive Equilibrium	229

Page
5. Conclusions 230
Acknowledgments 234
General References 234
References 235

1. Introduction

Unsolved problems in geophysics and particularly in the physics of the high atmosphere are numerous. Their solution will require a more extensive observational program, intensified laboratory research, and broadened theoretical treatments. The purpose of this paper is to discuss briefly several of the many problems which obscure our understanding of the high atmosphere. The solution of problems lying in the field of atmospheric statics and dynamics is a prerequisite to the development not only of certain phases of geophysics, but of physics as well. Although a group of entirely different problems could have been considered, those included below lie mainly in the field of physics. In some instances, as in the case of cross sections for collision at low energies, basic research must be begun in physics before any advances in clarifying collisional events in the high atmosphere will be possible. The problems discussed in this paper were chosen because their treatment may to some extent be foreseen and outlined. In many cases the problem is sufficiently broad to become a focus for additional research.

Our understanding of the events occurring in the high atmosphere of the earth has great scientific and practical importance. This region, although very small in relative mass, contains a wealth of uninterpreted phenomena. The research necessary borders on physics and astrophysics. The high atmosphere has been studied through analyses of data obtained from radio propagation and radio probings; geomagnetic and auroral activity; airglow and meteor observations; earth currents and audio-frequency radio waves (induced in the earth); solar phenomena discerned in the microwave, infrared, and optical regions of the electromagnetic spectrum; and direct probings by means of rockets. Past studies have emphasized the prognostication of radio propagation conditions, while neglecting the basic properties of, and the fundamental processes in, the high atmosphere. Answers to many presently unsolved problems will facilitate our future use of the atmosphere. What are the properties of the medium through which rockets will ultimately pass? Preliminary evidence indicates that relatively small amounts of energy are stored daily in the high atmosphere by the excitation, ionization, and dissociation of the atmospheric particles. However, the energy and power brought into the higher atmospheric regions by solar radiation is very much larger. Can these energies be tapped for practical purposes?

This energy, available daily in the flux of solar radiation, far exceeds the energy absorbed at the earth's surface, and may be available for almost limitless centuries in the future.

The terrestrial atmosphere may be utilized as a physical or astrophysical laboratory. This region is like a huge absorption cell lying between the earth and the sun. Spectroscopic observations on the solar radiation passing through this "cell" increase our knowledge of the atmosphere. Analyses of these data also provide information of great importance in physics and chemistry. Thus, observations on the solar absorption spectrum have yielded information on the rotational constants and the equilibrium moment of inertia of the carbon dioxide molecule, and supplied the first evidence of the oxygen isotopes of mass 17 and 18. Observations on the aurora and airglow have yielded new band systems for some molecules (nitrogen, oxygen, and hydroxyl) which have not yet been duplicated in the laboratory. In addition, as our knowledge of high atmospheric reactions and mechanisms increases, any change in the nature of this region may ultimately be employed to infer changes taking place on the sun.

The many photochemical reactions occurring under the action of sunlight create virtually an ocean of highly excited particles in the atmosphere; these complex processes and their rates must be untangled and the entire field carefully examined.

Although a great many investigations have been accomplished, research on the high atmosphere is only in its initial stages. There is still insufficient information to support a unified theory combining geomagnetic, ionospheric, and auroral observations with thermodynamic and radiative equilibrium, atmospheric movements, photochemical and collision processes, and such transient phenomena as ionospheric storms.

In this paper, problems of communications and routine ionospheric probings have generally been ignored; these are being actively investigated by several proficient groups and need no comment here. A few desirable types of geophysical observations have been suggested, not only to relate them to the laboratory and theoretical investigations, but also to point out deficiencies in the present observational program.

A brief survey of the extent of current knowledge in this field is given in Section 2. During the course of the investigations yet to be conducted, numerous presently accepted "results" will undoubtedly be revised. Many of the problems require fundamental research in quantum mechanics, gas kinetics, hydrodynamics, and spectroscopy, all capable of solution at the university level. For the experimental approach a moderately well-equipped spectroscopic, low-pressure discharge or physical chemistry laboratory is necessary. It is hoped that the various

topics given in the text may serve as a guide for thesis problems to be used by candidates for advanced degrees. A study of the dynamics of the rarefied regions of the atmosphere and of the aurora is an intriguing and challenging field, promising very fruitful yields.

2. The Terrestrial Atmosphere

2.1. General

Several texts which contain comprehensive information on the conditions and phenomena of the high atmosphere are given in the general references [A–G].

Although man is most familiar and normally most concerned with its lowest two meters, the terrestrial atmosphere extends outwards from the earth for a distance of probably 10,000 km. In spite of the extent of this gaseous envelope, one-half its mass of 5×10^{15} metric tons is located below an altitude of 6 km. The number density of the atmosphere decreases approximately exponentially with height above the earth's surface, finally attaining at the outermost extremities the density of interplanetary space. Although the uppermost limits of the atmosphere may hardly be regarded as a surface in the usual sense, an isopycnic envelope in these regions would probably have a tear-drop appearance, having been formed by the interaction between interplanetary debris and the terrestrial atmosphere. This tear-drop shape would be directed generally tangentially to the earth's orbit, depending upon the interaction of radiational, gravitational and frictional forces.

The atmosphere may be considered to consist of several strata, each identified by some characteristic feature. Though several systems of nomenclature have been proposed to define these various strata, there is general agreement regarding the use of the suffix "sphere" on terms indicating atmospheric subdivisions or shells, and the suffix "pause" on terms defining the separating surfaces between shells. The terminology adopted in this report considers the atmosphere to be composed of six shells, being, from the lithosphere outwards, the troposphere, stratosphere, chemosphere, ionosphere, mesosphere, and exosphere.* The

* A proposal now before but not adopted by a subcommittee of the U. S. National Advisory Committee for Aeronautics suggests that the regions defined above as stratosphere and chemosphere be consolidated into "stratosphere," and that the mesophere defined above be renamed the suprasphere. There may be some argument for considering the stratosphere in the broader, suggested sense. However, the term mesosphere as defined above has already been employed in the scientific literature of Argentina, Canada, England and the United States, and has been the subject of an International Colloquium on this topic. Contrarily, the term suprasphere does not seem to have been previously employed in connection with the atmosphere. Logically, the mesophere as defined above seems well named, inasmuch as the region to which it refers may be considered as the middle atmosphere.

approximate location of these shells in middle latitudes is given in Table I.

The lowest atmospheric shell, the troposphere, is characterized by large-scale convective air movements and marked frontal activity involving the movement of fairly well identified air masses. In the troposphere the average vertical temperature gradient is negative, large quantities of water vapor condense to form clouds and precipitation, and the latitudinal differences in the meteorological characteristics of the air are sharply defined.

TABLE I. Atmospheric subdivisions.

Atmospheric region	Altitude* (km) (middle latitudes)	Dividing surface (between the shell concerned and the next higher shell)
Troposphere	0–11	Tropopause
Stratosphere	11–32	Stratopause
Chemosphere	32–80	Chemopause
Ionosphere	80–400	Ionopause
Mesosphere	400–1000	Mesopause
Exosphere	Above 1000	

* The altitudes given are approximate, particularly those of the upper shells.

The term stratosphere signifies the somewhat isothermal region of the atmosphere lying above the troposphere. It contains the ozone region, is generally free of water vapor clouds, and thus exhibits little or no manifestation of "weather." The chemosphere is a shell wherein chemical activity becomes increasingly important. Within the chemosphere the dissociation of water vapor becomes complete, some atmospheric emissions occur and the temperature attains a maximum near 60 km and a minimum near 80 km. The chemosphere also contains the D ionic layer. The ionosphere is characterized by the presence of appreciable numbers of ions and electrons, the latter being of a sufficient density to reflect low and medium frequency radio waves. Within the ionosphere, the temperature gradient with altitude is positive. In the mesophere, the number density of electrons and ions, although relatively high, is less than that found in the ionosphere. Auroral luminescences generally emanate from both the ionospheric and mesospheric regions. The temperature gradient in the mesosphere is negative. In the exosphere the atmospheric particles experience few collisions. The mean free path depends upon direction. Particles moving away from the earth may reach extreme altitudes before the earth's gravitational field reduces their speed and returns them to the lower regions of the exosphere and the mesosphere.

Each of the boundary surfaces, i.e., tropopause, chemopause, stratopause, ionopause, and mesopause, may be considered as a zone of transition between the two shells concerned.

An additional term of great usefulness in discussing the terrestrial atmosphere is the airglow. The airglow includes all emissions arising in the atmosphere except those produced through auroral activity or

TABLE II. Composition of the (dry) atmosphere in the troposphere and stratosphere.

Constituent	Percentage by volume	Amount (atm-cm at NTP)	Reference
N_2	78.09	625,000	[1]
O_2	20.95	168,000	[1]
A	0.93	7,440	[1]
CO_2	0.03	240	[1]
Ne	0.00182	14.6	[2]
He	0.000524	4 2	[2]
CH_4	0.00014	1.1	[3]
Kr	0.0001	0.8	[1]
CO	0.0001	0.8	[4]
H_2	0.00005	0.4	[1, 5]
N_2O	0.00005	0.4	[6, 7]
Xe	0.000008	0.06	[1]
O_3	>0.000001	>0.01	[1]
I_2	Variable (maximum about 3×10^{-8})		[8]
Dust			
Bacteria			
Upper limits to concentration of some possible atmospheric constituents:			
C_2N_2	0.0002		[9]
NO	0.00005		[9]
NH_3	0.00001		[9]
C_2H_4	0.000001		[10]

lightning. As atmospheric emissions may occur throughout a twenty-four hour period, day-, twilight-, and night-airglows may be distinguished.

Samples of the air have been obtained to the top of the stratosphere by means of balloon-borne containers, and above this level by sampling devices aboard rockets. Within the troposphere and particularly within the first few kilometers of the ground, the percentage composition of dry air is remarkably constant. Analyses have been made at many geographic locations over the surface of the globe, employing both chemical examination of air samples and spectrographic observations of the solar absorption spectrum. The results, summarized in Table II [1–10], indicate practically no change in composition near the surface. However, many of the percentage values given for the atmospheric constituents

must be redetermined with greater accuracy. Appreciable but varying quantities of water vapor are also found in the troposphere, the proportion being as high as 3–4% in the tropics. Over the oceans, the percentage of carbon dioxide varies between 0.015 and 0.036.

Isotopes of the atmospheric particles may be detected through an examination of samples and by spectroscopic studies. Most of the

TABLE III. Normal percentage of isotopes in the atmosphere.

Element	Atomic number	Mass number	Percentage
H in H_2O	1	1	99.98
		2	0.02
He	2	3	1.1×10^{-4}
		4	100
C in CO_2	6	12	98.9
		13	1.1
N	7	14	99.62
		15	0.38
O	8	16	99.757
		17	0.039
		18	0.204
Ne	10	20	90.0
		21	0.27
		22	9.73
A	18	36	0.307
		38	0.061
		40	99.632

information already available has been obtained from the latter technique. The very long absorption-path lengths involved when observing the solar spectrum permit the determination of isotopic bands which are usually very weak in comparison to the main absorption bands of the molecule concerned. All isotopes of the atmospheric particles may be expected. Those of $O^{16}O^{18}$, $O^{16}O^{17}$, $N^{14}N^{15}$, CH_3D, $C^{13}O^{16}$ and HDO have been positively identified [11–15]. The normal relative proportion of isotopic constituents in the atmosphere is given in Table III [16].

From acoustical and radio-wave probings, it would appear that the atmosphere is well mixed by wind systems and turbulence, probably to the altitude of the ionosphere. This evidence suggests that, neglecting sinks and sources for any of the atmospheric constituents, the percentage

composition of the atmosphere should remain constant to altitudes of about 300–400 km. However, recent analyses of air samples collected from the chemosphere with improved rocket sampling techniques provide a basis for believing that some gravitational separation of the atmospheric gases may begin in the altitude interval 64–72 km. It was found that the content of helium and neon with respect to nitrogen increased,

TABLE IV. Ionization and dissociation potentials of some molecules found or possibly existing in the atmosphere.

Molecule	First ionization potential (eV)*	Dissociation energy (eV)*	Dissociation products
CH_4	14.5; 15.2	4.4	CH_3, H
CO_2	14.4	5.5	CO, O
CO	14.01	11.11; 9.61	.....
		9.14; 6.90	
$(CO)^+$	.....	9.90; 8.2	
		6.8; 6.40	
D_2	.....	4.55	
H_2	15.42	4.48	
$(H_2)^+$	.....	2.65	
H_2O	12.56; 13.0	5.1; 5.0	H, OH; H_2, O
He_2	4.25	0	....
$(He_2)^+$	.....	3.1	.....
N_2	15.58	9.76; 7.37	
$(N_2)^+$	....	8.72; 6.34	.....
N_2O	12.9	1.7	N_2, O
NO	9.5	6.49; 5.30	
$(NO)^+$	.....	10.6; 9.4	
NO_2	11.0	3.1	NO, O
Na_2	.....	0.77; 0.73	
O_2	12.2	5.08	.....
$(O_2)^+$	.....	6.48	
OH	$\leq$13.6	4.40; 4.35	.....
$(OH)^+$	.....	$\geq$4.4	.....

* Where no agreement has been reached, the several proposed values are listed.

whereas that of argon decreased [17]. Because of the presence of a heavy constituent, e.g., molecular nitrogen, at an altitude of 1000 km, it seems possible that the atmosphere may alternate between regions of gravitational separation followed by remixing. This condition implies the possible existence of alternating calm and windy strata in the chemosphere, ionosphere, and mesosphere.

The capture of meteoric debris and cosmic flotsam introduces many metals as trace contaminants of the chemosphere and ionosphere. These contaminants exist in a total concentration of possibly $10^5/cm^3$ or about

10^{-8} of the total atmospheric density. At higher altitudes a greater proportion of the metals may be found. The origin of sodium, which exists in similar concentrations, is uncertain; it may be introduced through extraterrestrial sources or through the upward diffusion and subsequent dissociation of oceanic sodium chloride. The very small concentration of sodium is noteworthy because of the remarkable intensity of its emission lines in the airglow. Other minor constituents of the chemosphere and ionosphere, such as the molecule OH, also have very strong emissions.

A study of composition at the higher altitudes is further complicated by the formation or destruction of some atmospheric constituents because of photochemical and collisional reactions. Water vapor dissociates

TABLE V. Probable constituents in the chemosphere and ionosphere.

A. Of atmospheric origin			
N_2	A	H_2	NH
N	Ne	H	NH_2
O_2	He	N_2O	CO
O_3	Na	NO	Kr
O	NaO	OH	Xe
B. Of meteoric origin			
Fe	Ca	Mn	Cu
Si	Al	K	H
Mg	Co	P	O
Ni	Na	Ti	C
S	Cr	Cl	N

above about 70 km. A variety of hydrogen-oxygen compounds, such as OH, HO_2, etc., are possible. (Dissociation potentials of some atmospheric compounds are given in Table IV.) Atmospheric ozone is mainly found (and is presumably mainly formed) above an altitude of 20 km, but its existence from that altitude to the earth's surface is well known. It has been suggested that the nitrogen soil cycle [18] is responsible for atmospheric nitrous oxide, but its formation in higher regions is not precluded. In the chemosphere and ionosphere, many reactions involving nitrogen-oxygen compounds may exist in modified equilibrium with incoming solar radiation. With regard to the major constituents, dissociation of molecular oxygen begins somewhat above 90 km and rapidly becomes complete [19–21]. The possible constituents of some of the higher atmospheric regions are shown in Table V.

The greatest number of particles introduced into the atmosphere by cosmic rays are hydrogen and helium; the next greatest number are in the carbon-nitrogen-oxygen group (B, C, N, O, F, Ne) followed by the

magnesium-silicon group (Na, Mg, Al, Si, P). Evidence of the heavier elements has also been found [22].

The mean atmospheric molecular weight is 29.98 below the altitude of oxygen dissociation and 23.95 above. The altitude at which molecular nitrogen dissociates and the rate at which it dissociates are controversial. The presence of molecular nitrogen at an altitude of 1000 km suggests that atomic nitrogen may not be an important constituent of these regions.

Observations of sunlit aurorae provide a basis for some interesting speculations regarding the extent and composition of the atmosphere. Apparently the uppermost portion of sunlit aurorae begins somewhat *sharply;* such observations, if verified, possibly imply that the atmosphere at this altitude may be in diffusive equilibrium, and that molecular nitrogen exists from this height (about 1000 km) downwards.

2.2. Temperature-Altitude Relationship

The temperature of the terrestrial atmosphere has been studied by a variety of means below an altitude of about 100 km. The values obtained by each of the methods are generally consistent so that, except for geographical variability, the broad temperature-altitude relationship to 100 km may be considered known. The technique employed may be listed as (a) direct observations using meteorographs on balloons and aircraft or special devices on rockets, (b) determinations from the anomalous propagation of (explosive) sound waves, (c) analysis of visual meteor trains, and (d) examination of tidal oscillations in the atmosphere. Above this altitude temperatures have been evaluated from spectroscopic observations of the aurorae and airglow, theories of radiative and thermal equilibrium in the atmospheric shells, and deductions from ionospheric parameters.

Without a rather high positive temperature gradient above 100 km, the atmospheric number density would fall so rapidly with altitude that there would be difficulty in explaining the formation of the F_2 ionospheric layer and the occurrence of high altitude aurorae. Electron densities at altitudes of 300–400 km are of the order of $10^6/cm^3$, thus setting a lower limit to the concentration on the justifiable assumption of singly ionized particles. The occurrence of visible aurorae at 1000 km requires a minimum number density of about $10^2/cm^3$ ionized nitrogen molecules and probably a higher concentration of neutral particles. Such a number density could not be maintained at this altitude unless temperatures in the intervening region were high. For example, if the atmosphere were isothermal at 300°K from 100 km upwards, the number density would be about $100/cm^3$ at 420 km and $10^{-15}/cm^3$ at 1000 km.

The temperature probably attains a maximum near the ionopause after which it decreases to meet the temperature of the exosphere. The maximum in temperature undoubtedly arises from the increased number of energy absorption processes occurring within the ionosphere. At different strata within this shell, the rate of heat change (loss or gain) through advection or conduction probably remains essentially constant. Although the rate of heat loss through radiation increases with altitude, the rate of energy absorption undoubtedly increases much more rapidly,

TABLE VI. Temperature, pressure, number density, mean free path and kinematic viscosity at various altitudes (middle latitide conditions).

Altitude km	Temperature °K	Pressure mb	Number density cm^{-3}	Mean free path* cm	Kinematic viscosity $cm^2\ sec^{-1}$
0	288	1013.25	2.5×10^{19}	8.6×10^{-6}	1.7×10^{-1}
11	218	2.3×10^{2}	7.8×10^{18}	2.8×10^{-5}	4.8×10^{-1}
32	218	8.6	2.9×10^{17}	7.7×10^{-4}	1.3×10^{1}
62	330	2.0×10^{-1}	4.5×10^{15}	4.9×10^{-2}	1.0×10^{3}
84	200	1.2×10^{-2}	4.4×10^{14}	5.0×10^{-1}	8.3×10^{3}
94	262	2.8×10^{-3}	7.8×10^{13}	2.8	5.3×10^{4}
100	300	1.5×10^{-3}	3.6×10^{13}	6.1	1.2×10^{5}
200 (Aug.)	1500	3.9×10^{-5}	1.9×10^{11}	1.2×10^{3}	5.2×10^{7}
(Jan.)	1150	2.1×10^{-6}	1.3×10^{11}	1.7×10^{3}	6.6×10^{7}
300 (Aug.)	2700	1.1×10^{-5}	3.1×10^{10}	7.2×10^{3}	4.3×10^{8}
(Jan.)	2000	4.0×10^{-6}	1.4×10^{10}	1.5×10^{4}	7.9×10^{8}
400 (Aug.)	3900	5.7×10^{-6}	1.1×10^{10}	2.1×10^{4}	1.5×10^{9}
(Jan.)	2850	1.5×10^{-6}	3.9×10^{9}	5.6×10^{4}	3.5×10^{9}
500 (Aug.)	3500	3.2×10^{-6}	6.6×10^{9}	3.3×10^{4}	2.3×10^{9}
(Jan.)	2625	7.1×10^{-7}	2.0×10^{9}	1.1×10^{5}	6.6×10^{9}
600 (Aug.)	3100	1.7×10^{-6}	4.0×10^{9}	5.5×10^{4}	3.5×10^{9}
(Jan.)	2400	3.1×10^{-7}	9.5×10^{8}	2.3×10^{5}	1.3×10^{10}
700 (Aug.)	2700	8.6×10^{-7}	2.3×10^{9}	9.5×10^{4}	5.7×10^{9}
(Jan.)	2175	1.3×10^{-7}	4.3×10^{8}	5.1×10^{5}	2.7×10^{10}
800 (Aug.)	2300	3.9×10^{-7}	1.2×10^{9}	1.8×10^{5}	9.8×10^{9}
(Jan.)	1950	5.1×10^{-8}	1.9×10^{8}	1.2×10^{6}	6.0×10^{10}

* Assuming a constant value of collisional cross section.

thereby producing the temperature maximum found near 400 km. The energy absorption processes may not be detectable by means of radio probing techniques.

The probable temperature-altitude function to 800 km is shown in Table VI and Fig. 1. Different summer and winter temperatures are considered above 100 km, with the greatest seasonal difference occurring at the region of maximum temperature. The table provides the temperature, pressure, number density, mean free path and kinematic viscosity at various levels.

Seasonal changes in temperature also occur below the ionosphere. Recent studies show that the winter-summer variation of temperature in the ozone layer is about 100°C [23] at 68°N. This variation seems to be latitude-dependent, being greater at the higher latitudes [24, 25]. The

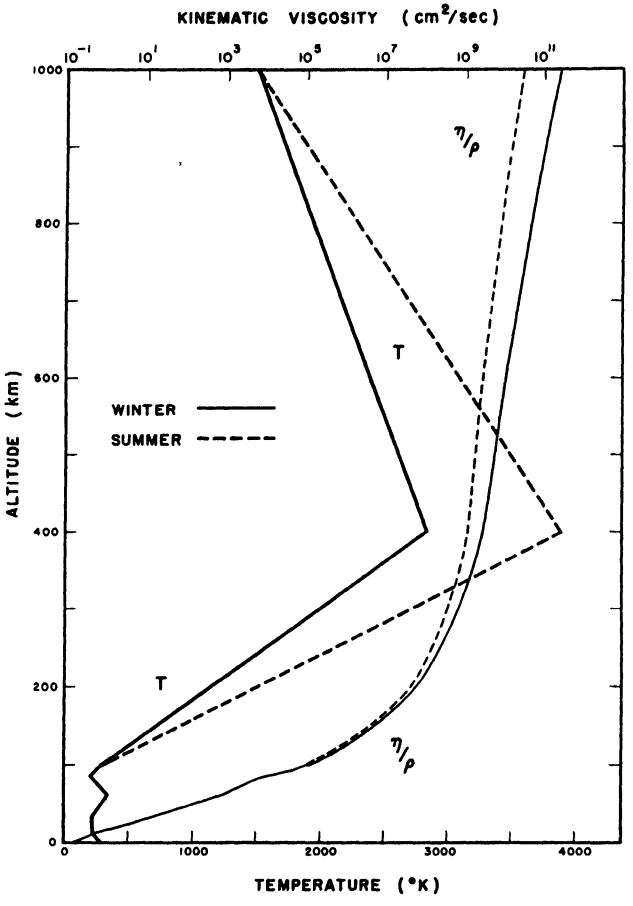


FIG. 1. Probable relationship between altitude and temperature and altitude and kinematic viscosity from the surface of the earth to the mesopause.

temperature increase from winter to summer in middle latitudes seems to be small below 25 km, relatively large in the altitude range 35–45 km and small in the stratum 50–55 km [26].

A latitudinal temperature gradient exists in the several atmospheric shells. In the troposphere, temperature increases towards the equator; in the stratosphere and chemosphere, towards the poles; in the lower ionosphere (to 120 km), towards the equator; and from 120–300 km, towards the poles [27].

2.3. The Ionosphere

Appreciable concentrations of electrons exist in the atmosphere at altitudes above about 70 km and extending through the mesosphere. Several ionic layers or regions, the D, E, F1, F2 and G, are known. The F1 and F2 layers merge during darkness to form a single layer.

The E, F1 and F2 regions have been extensively studied and their properties seem fairly well known. Much less information is available regarding the remaining layers. All layers except sporadic E (E_s), which occurs near 110 km, are formed entirely through the action of solar ultraviolet and shorter wavelength radiation.

TABLE VII. Ionization potentials of some atmospheric elements.

Atom	First ionization potential (eV)	Second ionization potential (eV)
H	13.595	
D	13.598	
He	24.58	54.40
C	11.26	24.38
N	14.54	29.60
O	13.61	35.15
Ne	21.56	41.07
Na	5.14	47.29
Mg	7.64	15.03
Al	5.98	18.82
Si	8.15	16.34
S	10.36	23.4
A	15.76	27.62
K	4.34	31.81
Ca	6.11	11.87
Fe	7.90	16.24

The ionospheric layers are formed through photochemical reactions under the influence of sunlight at the appropriate altitudes. The maximum electron density is greatest at about noon and reaches a minimum near sunrise for the layer concerned. The E region most closely follows Chapman's theory describing the formation of the layers, but the F2 layer deviates widely from this theory. It is still uncertain as to which particular atmospheric constituents are ionized to form each of the various layers [28]. Ionization potentials of the atmospheric constituents are given in Tables IV and VII.

The ionic concentration and altitude of maximum electron density of the layers vary with time of day, with season and with sunspot activity, and are different from day to day and at different geographical areas. It also seems that the ionospheric characteristics depend to some extent upon the geomagnetic latitude.

The rate of change of electron density is generally considered to be governed by the quadratic law of recombination. However, other factors also may be operative in reducing the electron concentration: electron attachment to and detachment from neutral particles, diffusion, winds, etc.

The D region is found in the chemosphere and may be regarded as a "ledge of ionization" below the E ionospheric region. With regard to radio wave propagation, the D layer may be considered as an energy absorption layer in contrast to the remaining regions which have mainly reflection properties. The D region is greatly intensified during solar flares and is generally considered responsible for the disruption of long distance radio communications during flare periods. At the time of a solar chromospheric eruption, a radio fadeout occurs lasting from minutes to hours over the sunlit portion of the earth.

TABLE VIII. Average value of collisional frequency between electrons and other particles in the ionosphere (experimental determination).

Altitude	Collisional frequency
km	sec^{-1}
100	4.0×10^5
150	3.5×10^4
200	1.1×10^4
250	5.0×10^3
300	2.7×10^3
350	1.6×10^3
400	1.1×10^3

Sporadic E may be regarded as clouds of very high electronic concentration imbedded in the regular E layer. It has also been proposed that E_s arises from the scattering of radio waves by the fine structure occurring within local "turbulent regions" of the E layer. Sporadic E is very seasonal in its occurrence, displaying a sharp major maximum during summer and a minor maximum during winter. It may suddenly appear during darkness or daylight. These cloudlike regions, observed through radio probings, indicate motions and movements, some of which appear to be cyclonic or anticyclonic in nature.

Although the G layer has been postulated on several occasions, further extensive study is needed to establish its properties, location, periods of occurrence and diurnal characteristics.

The frequency of collision between electrons and other particles in the ionosphere has been obtained for the E, F1 and F2 layers by means of certain types of radio-wave probing measurements. Results of these studies indicate that the collision frequency is about 4×10^5/sec at

100 km, and decreases to about 1×10^3/sec at 400 km as shown in Table VIII. It would be desirable to redetermine the values of collisional frequency in the ionospheric regions to obtain both their diurnal and seasonal variations.

It is interesting to note that through observations on the ionosphere a means exists for investigating the sun. For such a study, the ionosphere may be viewed as an extensive absorption cell whose variations give an index of solar activity.

2.4. Atmospheric Emissions

The atmospheric emissions which will be considered in this report are those of the airglow and the aurora. The latter have been observed for millenia in a variety of forms (coronas, arcs, draperies, rays, bands, and pulsating surfaces), and in many colors (red, yellow, green, blue, violet, and white). Rapid time variations may occur in the color and intensity. On occasion the intensity of illumination can equal that of the full moon, but generally, it is weak if not feeble. The aurora is mainly found in an approximately circular zone of "maximum frequency of occurrence," having a mean radius of about 23° and centered at the geomagnetic poles. Its occurrence decreases to the north and south of the zone. The number of days per year during which aurorae are observed at any given location is directly related to sunspot activity. A very high correlation exists between the occurrence of magnetic storms and auroras, both of which frequently recur in phase with the solar rotation period of 27 days.

It is generally believed that the auroral luminescences arise from the bombardment of the terrestrial atmosphere by charged particles which are ejected from the sun during active periods. The auroral problem encompasses four phases which lie in the fields of solar physics and stellar dynamics, electromagnetics and magnetohydrodynamics, spectroscopy and quantum mechanics, and low pressure discharges and plasma oscillations. These phases, the physics of which is not well understood, are, respectively (a) the method of ejection of material from the sun; (b) the interplanetary travel of this neutral stream of ionized particles towards the earth, including the interaction of the stream with the geomagnetic field and the subsequent deflection of the particles into the polar regions; (c) the penetration of the particles into the earth's atmosphere with the resulting excitation of the atmospheric constituents; and (d) the interaction of the ionized plasma penetrating the atmosphere with the magnetic field.

On bombarding the atmosphere, the penetrating particles cause the auroral luminescences; these have been found, by triangulation against a star background, in the altitude range 60–1100 km in the northern

hemisphere. The most common altitude, however, is 100 km. Spectral observations of the aurora provide invaluable information on the constituents of the chemosphere, ionosphere and mesosphere as well as on the excitation processes occurring therein during auroral activity. Spectra have been sought from 2950 A to about 11,000 A.

Molecular spectra possibly present in the aurora are those of nitrogen and oxygen, viz., of nitrogen: first negative system, first positive system, second positive system, Vegard-Kaplan bands, Goldstein-Kaplan bands, and the Meinel bands; of oxygen: the atmospheric bands and the Meinel bands. Of those possible molecular spectra, the following are firmly established: of nitrogen, the first negative system, first positive system, second positive system and the Meinel bands; of oxygen, the atmospheric and Meinel bands. The presence in the aurora of the Vegard-Kaplan bands is questionable, and that of the Goldstein-Kaplan bands doubtful [29].

TABLE IX. Atomic oxygen lines observed in the aurora.

Wavelength, A	Transition	
3693	$3s\ {}^3S^0—5p\ {}^3P$	
3948	$3s\ {}^5S^0—4p\ {}^5P$	
4368	$3s\ {}^3S^0—4p\ {}^3P$	
5577	$2p^4\ {}^1D_2—2p^4\ {}^1S$	Forbidden
6157	$3p\ {}^5P—4d\ {}^5D^0$	
6300	$2p^4\ {}^3P_2—2p^4\ {}^1D_2$	Forbidden
6364	$2p^4\ {}^3P_1—2p^4\ {}^1D_2$	Forbidden
6455	$3p\ {}^5P—5s\ {}^5S^0$	
7255	$3p\ {}^3P—5s\ {}^3S^0$	
7774	$3s\ {}^5S^0—3p\ {}^5P$	
7987	$3p\ {}^3P—3s^1\ {}^3D^0$	
8446	$3s\ {}^3S^0—3p\ {}^3P$	

The atomic lines possibly present in the auroral spectrum are those of hydrogen, helium, sodium, oxygen, and nitrogen. Hydrogen lines are strong in homogeneous arcs and flaming aurorae. The existence of HeI and HeII lines in the aurora is not conclusive. The sodium spectrum does not arise in the aurora but in the airglow. The lines of OIII and NIII are not present. The presence of NI, NII and OII is doubtful and requires further study. The lines of OI are firmly established in the auroral spectrum. Atomic oxygen lines observed in the aurora are given in Table IX.

Probably the most important emissions of the aurora are the green and red lines of atomic oxygen (5577 and 6300 A), the first and second positive systems of nitrogen, the molecular oxygen emissions, and the hydrogen Balmer lines. While the first negative bands are strong, the

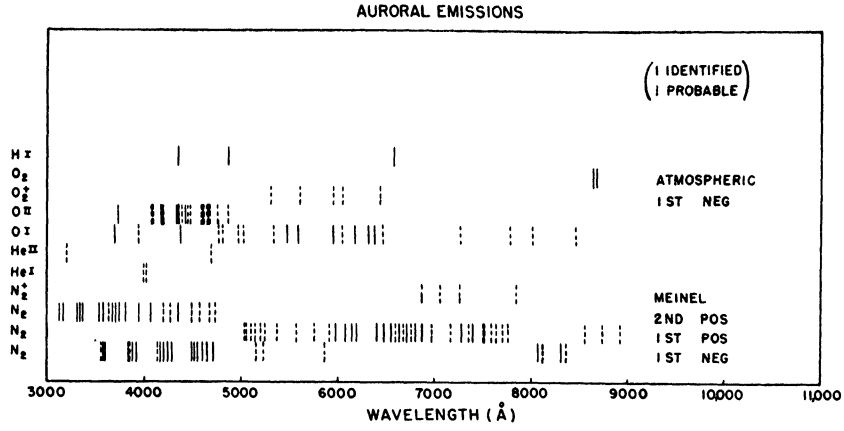


FIG. 2. Spectra found in the auroral emissions arranged according to the atom or molecule excited and the wavelengths emitted. (Identified lines are shown solid; probable lines, dashed.)

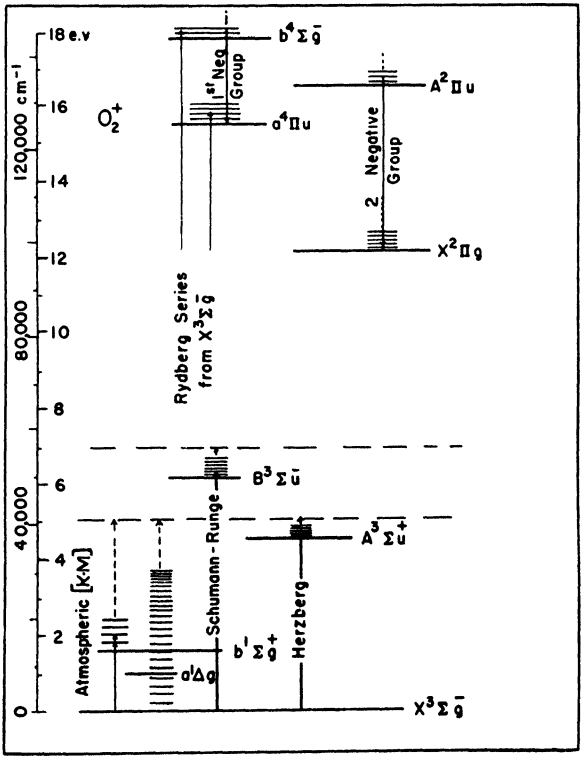


FIG. 3. Energy level diagram for molecular oxygen.

first positive system of molecular nitrogen is one of the most pronounced features of the aurora in the visible region of the spectrum. In the violet, the most prominent emissions arise from the second positive system of N_2 although the first negative bands are also strong [30].

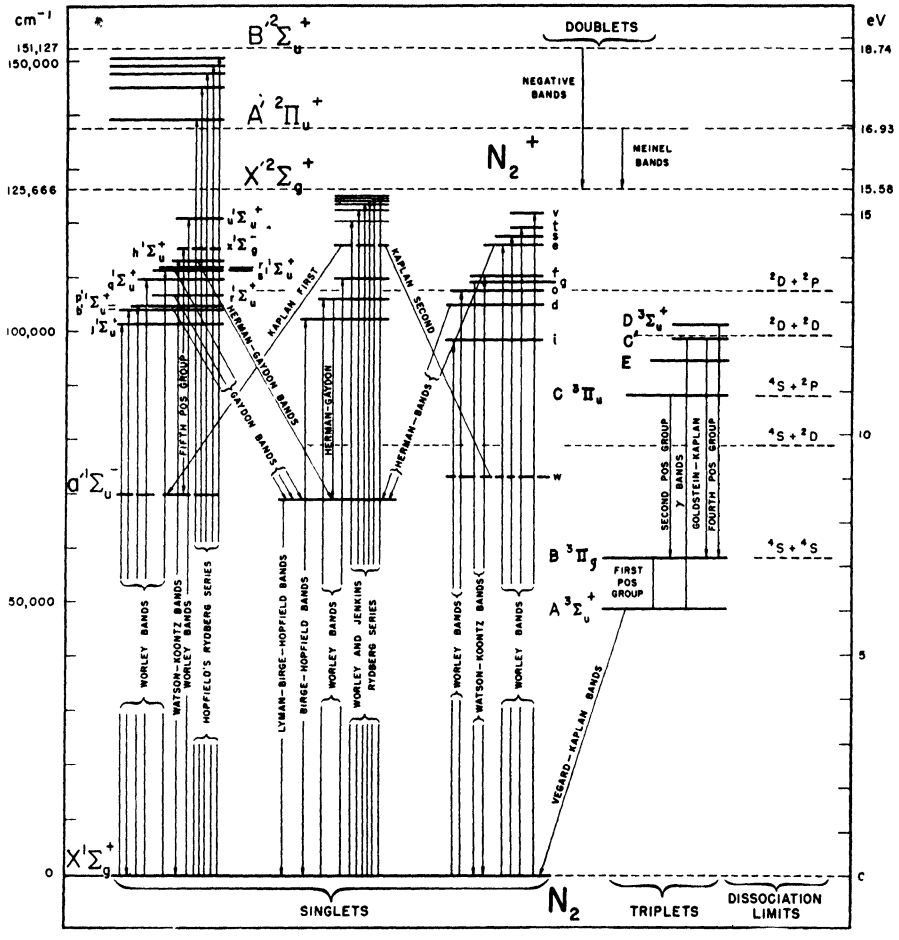


FIG. 4. Energy level diagram for molecular nitrogen.

Verified spectra from the aurora are shown in Fig. 2. The energy level diagrams for molecular oxygen and nitrogen and atomic oxygen are depicted in Figs. 3–5. It may be noted in passing that, in general, the intensity distribution among the various lines and bands in the auroral spectrum has not been reproduced in the laboratory.

Recent reports have indicated that electromagnetic radiation at radio frequencies is emitted by the aurora [31]. However, these radio wave

emissions have not been confirmed. The results, therefore, should be considered as tentative and viewed with caution until verified.

The airglow is constantly emitted in the terrestrial atmosphere. The day airglow, however, is obscured by the intense Rayleigh scattering occurring in the troposphere, and therefore cannot be observed at the earth's surface (except perhaps during favorable solar eclipses). The night airglow arises from the release of energy stored during daylight;

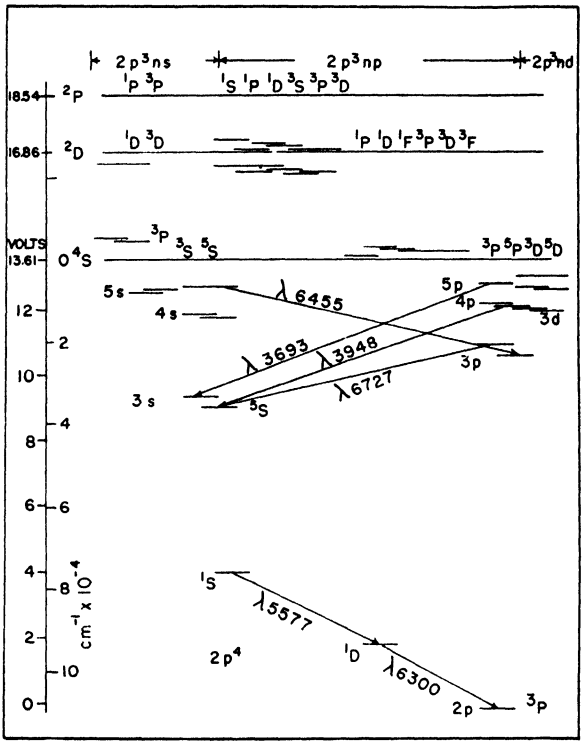


FIG. 5. Energy level diagram for atomic oxygen.

e.g., from recombination of ionized particles, three-body collisions of atoms to form molecules, and two-body reactions between atoms and molecules. The latter two mechanisms are probably the most important in causing the night airglow. The same processes are also operative in the twilight airglow, but in this case, the most important emissions may be attributed to resonance and fluorescent scattering. It is obvious that the spectra of the aurora and the airglow are quite distinct, the former arising from high and the latter from low excitation energy conditions. Sources of radiation contributing to the brightness of the night sky emissions in the spectral region 3600–4500 A are given in Table X.

Many attempts have been made to determine the altitude of emission of given spectral features of the night airglow, but the results have been rather discordant. While the observations may not be difficult, the interpretation of the results is ambiguous. One theory employed to determine emission altitudes considers uniform emission in a thin atmospheric stratum. However, the luminosity is not emitted uniformly and probably does not arise from a thin or even a single layer. The emissions

TABLE X. Radiation sources contributing to the emissions of the night sky.

Cause	Percentage of brightness
Resolved bands	20
Terrestrial continuum	53
Zodiacal light	15
Unresolved stars	12
Cerenkov radiation	10^{-2}

TABLE XI. Airglow emission altitudes.

Constituent	Emitted line or bands	Calculated height
	A	km
O	5577	75–100; Second layer at 1000 100; 103; 110; 200; 215; 250; 260; 291; 500 (Approx.)
O	6300	100; Second layer at 750 180; 500 (Approx.)
O_2	Herzberg	100; 350
O_2	Kaplan-Meinel	70–80
N_2	6580	500 (Approx.)
N_2	Vegard-Kaplan	300 (Doubtful); 900
OH	Meinel	70
OH	6550	100; Second layer at 750
Na	5890–5896	80; 275 300 (Doubtful)
...	3500–4500 continuum	500–1600
...	Infrared	125

appear to be in the form of bright patches or clouds, whose movements can sometimes be easily followed. The altitudes determined for the night airglow emissions are listed in Table XI.

The spectrum of the night airglow shows a prominent continuum in the spectral region 3900–4900 A. An unambiguous interpretation of this continuum has not been presented although it has been suggested that it results from wings of the Vegard-Kaplan bands of molecular nitrogen, emissions from the negative oxygen atom, and radiative association of

neutral atoms. The probable molecular emissions are those from the hydroxyl molecule, the Vegard-Kaplan bands, and the Herzberg and Meinel bands of oxygen. The atomic lines present are those of oxygen and sodium. The most prominent features of the night airglow are (a) the OH lines, particularly the intense emission at 10,400 A; (b) the sodium doublet at 5890–96 A; (c) the green and red atomic oxygen lines at 5577 and 6300 A; and (d) the numerous bands which are very distinct at wavelengths less than 4000 A but which seem somewhat lost in a continuum in the blue-violet region. Ultraviolet bands of OH are probably masked by the Herzberg bands of O_2. If recent work on band blending is accepted, the Vegard-Kaplan bands appear rather weakly in the night airglow. Spectra present in the night airglow are depicted in Fig. 6.

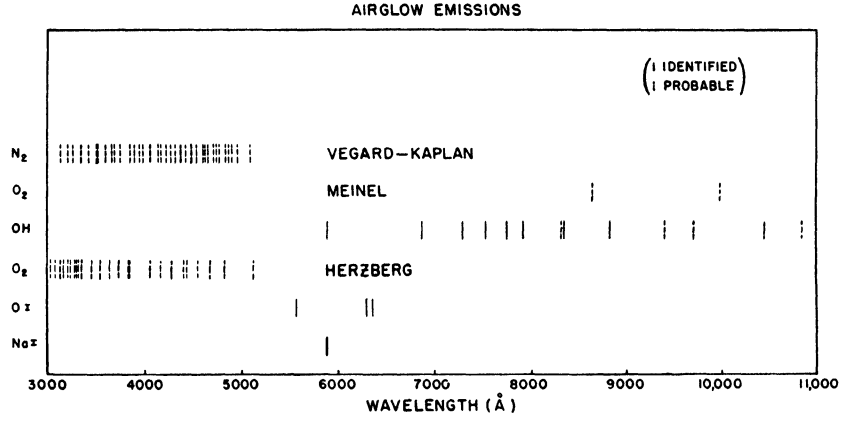


FIG. 6. Spectra found in the night airglow arranged according to the atom or molecule excited and the wavelength emitted. (Identified lines are shown solid; probable lines, dashed.)

Observations on the rotational structure of some of the bands of oxygen and nitrogen indicate temperatures of 200–300°K, but the altitudes of emission are in doubt. Temperatures deduced from the Vegard-Kaplan bands in the blue region of the spectrum are probably inaccurate because of the presence of the background continuum.

The twilight airglow results from the sharp enhancement of the sodium D doublet, the N_2^+ flash, and the red lines of atomic oxygen at 6300 and 6364 A. The D lines are a resonance phenomenon emitted when solar radiation illuminates the atmospheric region containing sodium. The emission altitudes obtained from this study indicate the presence of sodium at 70 km and probably also at 200–600 km. The negative bands of N_2^+, absent in the night airglow, are prominent in the twilight spectrum. They are emitted at 85–130 km. The spectrum of

the 6300 A line is enhanced for several hours after the end of evening twilight although the intensity decreases exponentially with time. A corresponding but less intense condition prevails at sunrise. The infrared emissions of the night airglow do not seem to be enhanced during twilight.

It should be noted that the non-observance of given spectra in the airglow or the aurora do not *a priori* imply the absence of particular constituents from the chemosphere, ionosphere, or mesophere. Spectra are not emitted unless efficient excitation mechanisms and favorable transition probabilities exist.

2.5. High Altitude Winds

Although practically no theoretical investigations on the general circulation of the high atmosphere have been made, some experimental results are available on winds to an altitude of the F layer. Below 70 km,

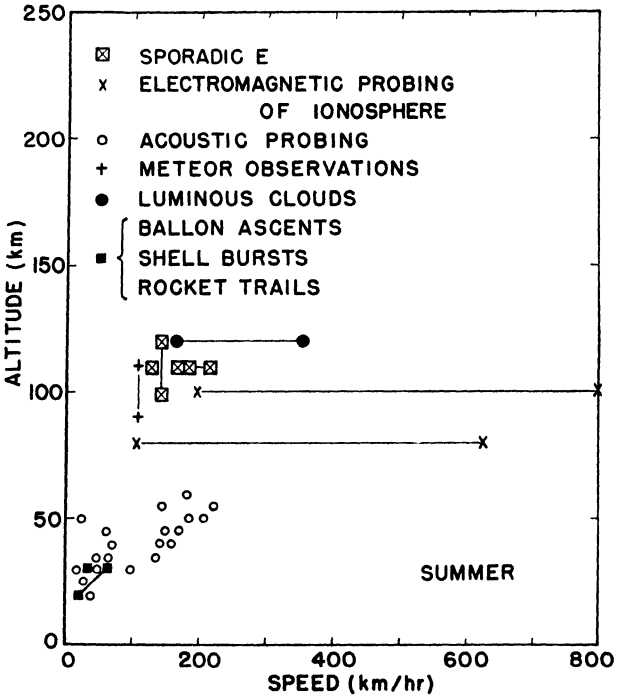


FIG. 7. Summary of wind speeds observed at various altitudes in the atmosphere (summer data).

air movements have been studied by means of observations on ascending balloons, smoke shells and rocket ejecta [32–35], and through acoustical probings of the atmosphere [36-38]. Above 70 km, the techniques employed include radio-wave probings of the motions of ionospheric

irregularities [39–46] and sporadic E [47–52], observations on luminous clouds [53–55], and analyses of magnetic variations. A study of meteor trains provides information on movements in the altitude range 40–150 km [41, 56–59]. Statistical treatments of geophysical data permit deductions on the magnitude of tidal oscillations from the troposphere to the ionosphere. From a knowledge of tides in the atmosphere, it may be possible to infer the velocity of tidally-produced winds.

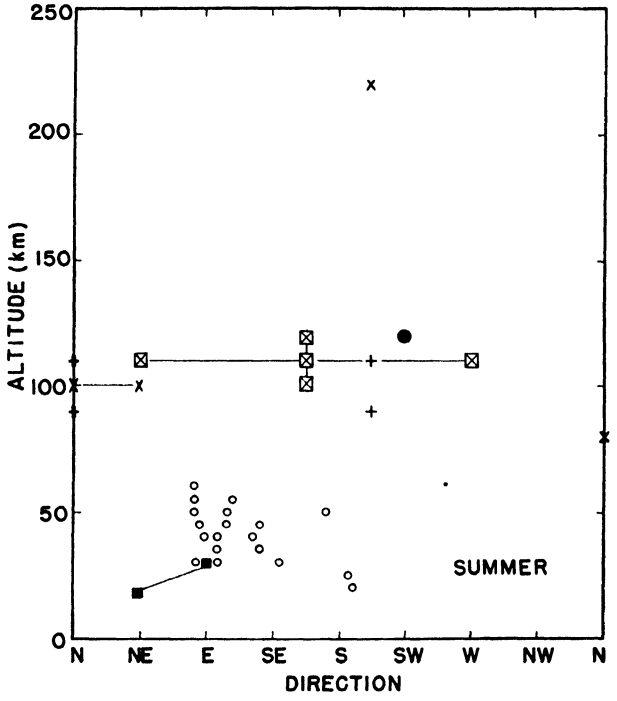


FIG. 8. Summary of wind directions observed at various altitudes in the atmosphere during summer. Directions are given in the meteorological sense; i.e., *from* the direction indicated.

In general, in middle latitudes, the wind is about W 110 km/hr (directions are given in the meteorological usages; i.e., *from* the west) at the tropopause. A "monsoon" effect with reversal of direction of the wind occurs from the tropopause to about 45 km. In this layer, wind directions change from westerly to easterly depending upon season, becoming entirely easterly in the altitude range 45–80 km [27]. Westerly winds are found between 80–190 km, becoming easterly above that altitude.

These general results are supplemented by the summary of wind velocities shown in Figs. 7–8 for summer and Figs. 9–10 for winter.

These summaries include only the available seasonal information, and indicate that wind speeds increase in the chemosphere to about 200 km/hr during summer and 250 km/hr during winter. Studies on the movements of ionospheric irregularities in the E, F1 and F2 layers indicate average speeds of about 250 km/hr. Winds near 100 km height seem to be generally about 175 km/hr during summer and 350 km/hr during

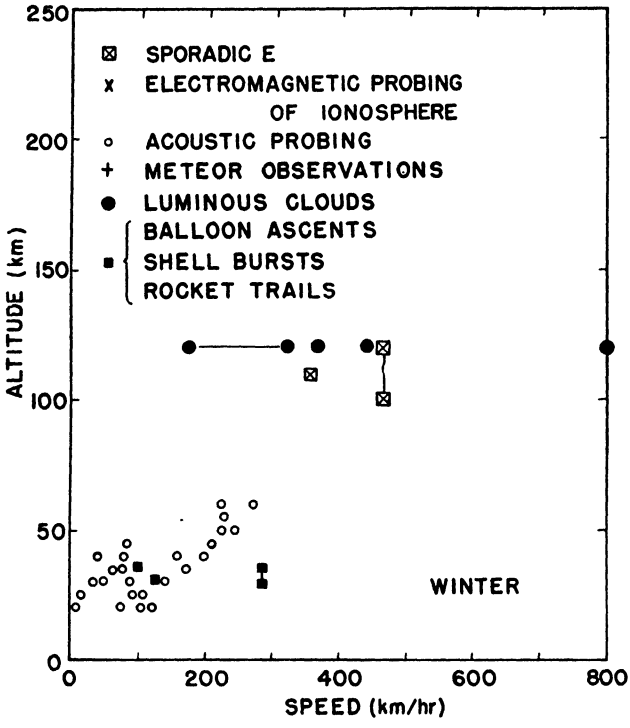


FIG. 9. Summary of wind speeds observed at various altitudes in the atmosphere (winter data).

winter, but the number of observations is small. Some results on ion cloud movements give speeds as high as 3600 km/hr; such high values are doubtful and, if true, indicate that the wind is in a non-steady state condition. The maximum value which the horizontal geostrophic wind can attain in the terrestrial atmosphere is about 740 km/hr [60]. A summary of wind velocities at different altitudes, irrespective of season, is shown in Fig. 11.

Meteor observations indicate the existence of appreciable wind shear and turbulence in the chemosphere and lower ionosphere.

Sporadic E observations reveal the existence of striking large-scale

movements, sometimes cyclonic or anticyclonic in nature, over rather extensive geographic areas. An example of one such movement is depicted in Fig. 12. This E_s area was followed for a distance of about 2300 km; its average apparent speed was 321 km/hr.

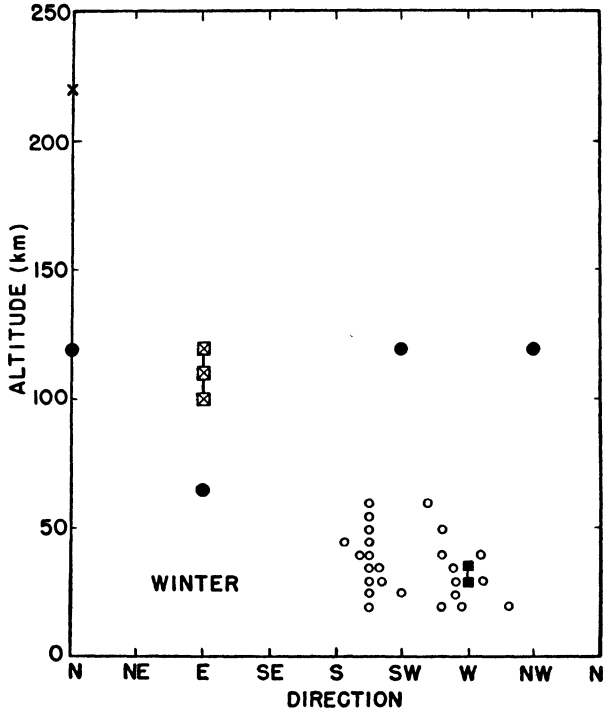


FIG. 10. Summary of wind directions observed at various altitudes in the atmosphere during winter. Directions are given in the meteorological sense; i.e., *from* the direction indicated.

In general, considerably more investigations of fluid movements and tidal oscillations in the ionospheric and mesospheric regions are needed. The entire field of auroral dynamics remains untouched. Vertical movements, subsidence, and general turbulence have important bearings on the establishment of diffusive equilibrium in the atmospheric regions; but these effects have scarcely been investigated. Diffusion of ions across the magnetic and electric fields found in the atmospheric regions also warrants further study.

3. STATIC PROPERTIES AND PROCESSES OF THE HIGH ATMOSPHERE

While a variety of methods may be employed to examine the properties of the lower atmosphere, an investigation of the ionosphere and

mesosphere is more restricted by the fewer available techniques. Radio wave probings allow some inferences to be gained regarding the collision frequency, magnetic field, and number density in the ionic layers. Spectrographic observations of the airglow and aurora may be utilized to infer

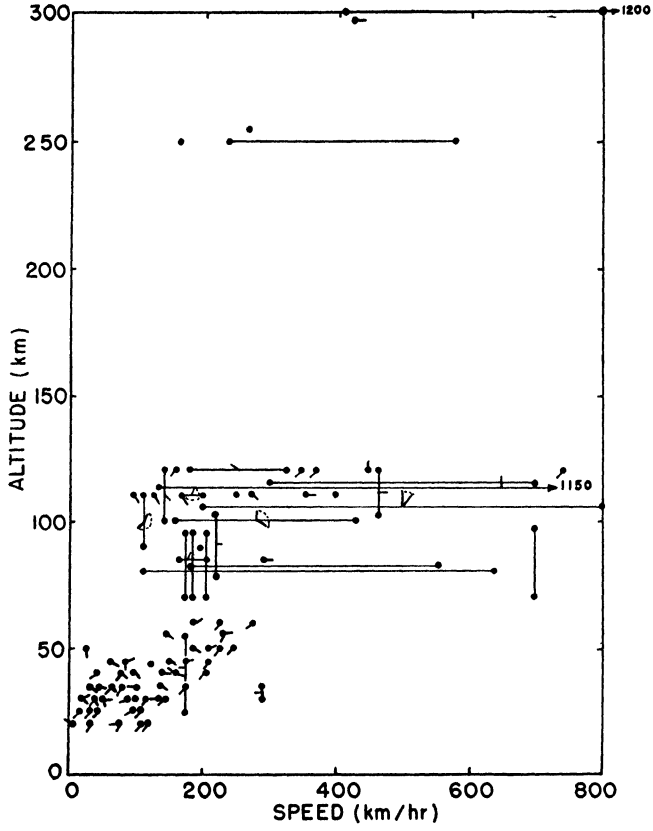


FIG. 11. Mass plot of wind speeds and directions observed irrespective of season at different altitudes. The small tail indicates the direction from which the wind came. Extreme wind speeds of 3600 km/hr were not plotted. (A solid bar connecting two points gives the range of wind or altitude values reported; in this case, the tail indicating direction is placed at the center of the bar. When a range of direction was given, the quadrant is shown. Directions are as follows: top = North, left = West, bottom = South, right = East.)

the temperature, altitude, and composition of the emitting strata, and to study the excitation conditions giving rise to the observed spectra. Similarly, rocket observations allow a direct determination of the pressure, temperature, composition, magnetic field, etc. Results of these several observational techniques provided the information on the general state of the atmosphere given in Section 2.

However, while past observations have been successful in providing a macroscopic impression of the high atmosphere, a complete, coherent picture of this region is noticeably lacking. The many inconsistencies still found in the present state of knowledge prevent a grasp of the detailed mechanisms and processes occurring within the higher regions. There still remain many difficulties in interpreting the limited observational data. To remove these deficiencies in knowledge and of interpretation, additional theoretical and laboratory studies will be necessary; only after these have

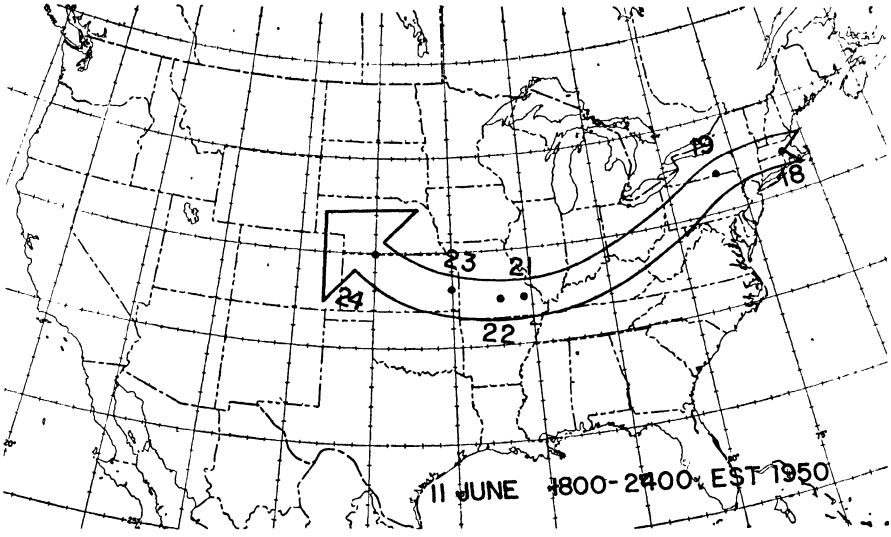


FIG. 12. Type of movement determined from sporadic E observations.

been accomplished will it be possible better to interpret the observational data, and so obtain a more detailed, unified concept of the high atmosphere.

Several difficulties in the present interpretation of geophysical observations may be mentioned. While collisions between electrons and other particles in the ionosphere may be measured, analyses of the data require knowledge of the cross section for collision of electrons with the other atmospheric constituents; this information is not available. From results of radio probing and spectroscopic observations, innumerable attempts have been made to determine the temperature of the high atmosphere. Analyses of the former data are in some cases so strongly circumscribed by the assumptions involved that the results are very questionable. Further information is necessary on the relationship between recombination coefficient and temperature. In deducing temperatures from rotational band spectra a serious discrepancy exists between gas and rotational temperatures. The identification of lines

and bands in the airglow, aurora, and solar absorption spectrum is still controversial in some respects. Additional experiments are required to obtain in the laboratory all the emissions observed in the high atmosphere, or the absorptions found in the solar spectrum. In this connection, further theoretical and laboratory research is even necessary on such effects as the pressure broadening of absorption lines, the structure and interatomic distances for some atmospheric molecules (e.g., ozone), and transition probabilities for various emissions. Although these topics are fundamentally problems in physics, they directly affect the interpretation of geophysical observations.

In the discussion which follows, emphasis has generally been placed upon experimental and theoretical studies in physics which must be undertaken if a better understanding of the high atmosphere is desired. Some of these investigations are so fundamental to geophysics that, until they are accomplished, further specialized investigations of the high atmosphere are almost completely blocked. Solution of the problems mentioned in this section will allow accurate determinations of the temperature, number density, specific heat, viscosity, pressure, mean molecular weight, etc., of the atmosphere to great heights. Further, diurnal and seasonal variations of these quantities, also as a function of altitude, could be obtained. The amount of energy absorbed and the availability of this energy could be determined. The many processes and mechanisms occurring under the action of sunlight would become better known, and knowledge of the reverse chemical reactions which take place during darkness would become available. In short, much of the desired detailed knowledge of the high atmosphere and its mechanisms will be provided only if many of the background researches are completed.

3.1. Temperature

A knowledge of the temperature in the ionosphere and mesosphere is of primary importance since from it may be determined other static and dynamic properties of these regions.

Although a large number of attempts has been made to study temperatures above 100 km, the results deviate widely among researchers, depending upon the method of investigation and the specific assumptions involved. At these altitudes, theoretical deductions of temperatures have been made from observed values of ionospheric parameters, radiative equilibrium, deceleration of meteors, and from considerations of thermal conduction in the atmosphere. Direct spectroscopic observations of the rotational structure of molecular bands in the aurora and the airglow, and interferometric determinations of some airglow emission lines, also have been employed.

Unfortunately, the results are not in agreement. The differences involved may not be attributed to a lack of thermodynamic equilibrium in the high atmosphere. Spectral observations on some auroral emissions indicate high temperatures, but the values of about 7000°K (obtained from the molecular vibrational levels) probably show the kinetic temperature of the particles bombarding the atmosphere rather than deviations of the atmospheric particles from a Maxwellian velocity distribution. In general, the number density in the mesosphere is probably high enough so that even at altitudes of 600–700 km thermodynamic equilibrium exists and deviations from a Maxwellian distribution probably are rare.

The term, thermodynamic equilibrium, as used here, considers only the condition of the free atmospheric particles—atoms, molecules, ions, and electrons—and does not consider the energy distribution of the bound electrons within a given atom or molecule.

Some gross upper and lower limits to the temperature may be readily obtained. Since a planetary and not a stellar atmosphere is involved, the maximum temperature must be less than 5000°K. Radio wave probing measurements indicate an electron density of about $10^6/cm^3$ at 400 km. If all atmospheric particles were ionized, the temperature lapse rate in the stratum 100–400 km necessary to support such a particle density at the higher altitude would be about 1.3°K/km. The temperature would be approximately 680°K at 400 km. Various methods have been proposed for computing the temperatures in the ionosphere from ionospheric measurements. However, although all of the techniques have been utilized, some of the widely quoted results are worthless. A critical survey of the several procedures employed to deduce temperatures from ionospheric parameters is overdue.

Spectroscopic observations seem to offer the best hope of obtaining reliable temperatures at the altitude of the emission layer. However, in this regard the large disparity between gas and rotation temperatures under certain types of excitation conditions must be examined more fully. Measurements of the Doppler broadening of certain atomic lines, although somewhat difficult to undertake, will provide accurate temperatures of the emitting constituents if interferometers of sufficient resolution are employed.

It is also possible to determine temperatures in the altitude range 40–170 km from the number density, which in turn may be deduced from meteoric observations.

3.1.1. Theoretical Determination of Temperature. It is tempting to think that a new and relatively simple procedure may be found for determining temperatures from ionospheric data. If accurate theoretical or empirical relationships could be obtained, data from the ever increasing

global network of ionospheric stations could be employed on a synoptic basis to initiate a study of ionospheric temperature and pressure conditions and, eventually, to study weather and climate.

A variety of methods already has been proposed to deduce temperature from an ionospheric parameter. The methods relate temperature to (a) collision frequency, (b) recombination coefficient, (c) ionospheric scale height, (d) diurnal and nocturnal variations in electron concentration and (e) seasonal or latitudinal variations in the solar zenith angle.

Unfortunately, however, some of the proposed relationships rest upon rather tenuous foundations and others require extensive refinement before accurate results may be expected through their use. The law relating temperature to recombination coefficient, for example, has not yet been given in a form suitable for employment in ionospheric studies. Further, even if this law were firmly established, the several methods now utilized for obtaining recombination coefficients in the ionospheric layers (particularly during daylight) are open to serious criticism. The method of obtaining temperature from the scale height or the electron concentration rests upon the assumption that a Chapmanian layer (or its parabolic approximation) is involved. The application of the Chapman theory to the higher layers would be open to question, even if the theory did not rest upon the initial assumption of an isothermal atmosphere. Variations of electron concentration (not attributable by simple theory to changes in the solar zenith angle) may be considered as indicating an influx or efflux of electrons brought about by a temperature change in the layer concerned. A relationship of this type for daylight conditions is sensitive to small changes in the recombination coefficient, which in turn is difficult to estimate accurately. Use of the theory for night conditions appears somewhat simpler and more reliable. The determination of seasonal or latitudinal temperature changes, through utilization of the Chapman theory, is open to the objection that the ionospheric layers are not simple, and that the relationship between temperature and recombination coefficient is not known. Attempts may be made to determine temperatures from measured values of collisional frequency in the ionospheric layers, providing (a) the cross section for elastic collision is known and (b) the collisional frequencies are accurately determined (see Section 3.3).

It is obvious that new and greatly improved relationships between temperature and an ionospheric parameter must be formulated before reliable information may be regularly obtained from ionospheric data.

Other theoretical determinations of temperature may be undertaken in addition to those explicitly related to an ionospheric parameter. These studies are concerned with the energy balance of the atmosphere at

various altitudes. In the ionosphere, energy is absorbed in producing ionization, dissociation and metastable particles. Radio-wave probing techniques, however, may not be employed to determine the radiant energy input into the ionosphere; these probings do not reveal the non-ionizing energy absorbed, but only those newly produced electrons which have appreciable lifetimes. Ions which disappear rapidly after their production will not be detected. Radio-wave probing studies are limited by the screening effect of the F2 layer to altitudes below about 400 km.

Although the observations do not provide full information on the energy absorbed by ionization and other processes, heat balance studies will allow a determination of the energy input and therefore the temperature. From the heat input into various levels of the ionosphere, a better knowledge of the processes occurring therein may be obtained. Such investigations must include the energy loss by radiation from all the atmospheric constituents, considering the possibility of different concentrations at different altitudes. The role of conduction in affecting the heat budget of the higher shells warrants much more study. Any comprehensive theory must, of course, consider the effect of heat transport through advection to other regions of the atmosphere, or through convection into higher strata. A determination of temperature from an examination of solely radiative equilibrium seems too restrictive to be applied to the atmospheric shells unless it can be shown that the effects of conduction and advection are negligible. All possible emissions of the atmospheric constituents must be considered in determining heat loss by radiation.

Obviously, although a variety of techniques may be employed to determine the temperature of the high atmosphere theoretically, the results must be intercompared when applied to the same model. Additional information on the theoretical determination of temperature in the atmospheric shells is given in the references [61–64].

3.1.2. Gas and Rotational Temperatures. Observations on some of the rotational bands of nitrogen in the aurora and in the airglow, and of oxygen in the night airglow, indicate temperatures of the order of 200–300°K for the former and 150–200°K for the latter. These temperature values appear reasonable if the emission altitude of these spectra is about 100 km and about 80 km, respectively. However, temperatures of about 220°K have also been obtained from resolved rotational line intensities of the aurora at very much higher altitudes. From these observations it has been concluded that the ionosphere is essentially isothermal to rather high altitudes. A temperature of this value would imply an extremely rapid decrease of the number density in the higher regions of the ionosphere. Electron densities in the F2 layer, then, could only be explained

if every atmospheric particle were at least singly ionized. Also, the existence of aurorae at 1000 km would be difficult to explain, possibly requiring a very rapid temperature increase in the range 300–1000 km. This possibility seems remote; it is more probable that the temperature of the upper ionosphere and the mesosphere is considerably higher than 300°K. From this reasoning, a large discrepancy then exists between the computed gas or kinetic temperature and the rotational temperature.

Normally, it would be expected that the gas temperature controls the distribution of the molecules among their rotational energy levels in accordance with Boltzmann's law for thermodynamic equilibrium. However, the validity of rotational temperatures as indicative of the kinetic temperature is dependent upon several factors, one of which is the lifetime of the excited state. The lifetime is important because chemical or other selective mechanisms may cause large discrepancies to occur between the rotational and the gas temperature. However, up to the middle of the mesosphere, number densities are still sufficiently high, and it may be expected that the rotational temperatures of forbidden transitions are fairly close to the gas temperature. A second factor of probably greater importance is the fact that thermodynamic equilibrium in the excited rotational state (of a low pressure discharge) is justified only under certain restrictive conditions; i.e., that the nuclear separation in the molecules concerned is the same in the excited and the ground states. Generally, this condition has been satisfied for most molecules studied but for some the nuclear separation is appreciably greater in the excited than in the ground state. It is interesting to note that nitrogen, whose rotational bands have been examined to determine atmospheric temperatures in the aurora, is one of those molecules for which there is a much larger nuclear separation in the excited state.

It is well known that temperatures derived from band rotational spectra are not necessarily equal to the kinetic temperature of the gas. Many experiments have been performed where higher rotational than gas temperatures were observed. Spectra showing *lower* rotational than kinetic temperatures have been found in the laboratory on only a few occasions because of the special conditions required for their excitation.

Much more theoretical and experimental research is necessary on anomalous molecular rotation leading to conditions where rotational temperatures are lower than kinetic temperatures. In addition to investigations of this effect when nitrogen is excited by a discharge, excitation of the gas at low pressure by bombardment of electrons, protons, helium ions and calcium ions should also be considered. Such studies will allow a resolution of the discrepancy between these temperatures as

deduced from auroral spectra. Undoubtedly the study will reveal new methods of attack upon the auroral problem; e.g., from simultaneous measurements of the rotational and gas temperatures (see Section 3.1.3), excitation conditions in the aurora may be inferred.

In addition to laboratory measurements designed to investigate more fully the deviations between rotational and gas temperatures, field investigations on the aurora must also be undertaken. Analyses of these observations should attempt to redetermine rotational temperatures of the molecular nitrogen bands and simultaneously should measure the altitude of emission. Admittedly the observations are difficult, inasmuch as stray light from different altitudes (particularly from the more intense lower portion) of the aurora than that being studied must be eliminated.

Additional background information on theoretical and laboratory methods [65–70] and field observations [71] may be found in the references.

3.1.3. Interferometric Determinations. Probably the most accurate method for determining temperatures in the high atmosphere lies in an examination of the forbidden atomic lines. Such a determination is relatively simple in principle inasmuch as the spectral width is simply related to the molecular weight and temperature of the emitting particles. Although observations on the Doppler widths present some difficulties, the results are unambiguous providing adequate instruments are employed. The transition probability of the emission determines the natural line width, and in order to detect any thermal broadening, those atomic lines must be examined which have small widths.

Early interferometric measurements of this type indicated temperatures of about 900°K when observations were made on the 5577 A line of oxygen. A Fabry-Perot etalon of comparatively small order was used. Additional observations of a similar nature must be undertaken, taking full advantage of the tremendous improvements in resolution and technique which have occurred since that time. It would be highly desirable, for example, to use an instrument having a resolution of at least 500,000 in order to obtain the necessary reliable measurements. As the atomic lines extend to high altitudes in both high altitude auroral rays and low latitude aurorae, a means exists for the optical probing of the atmosphere to determine temperatures in the ionosphere and mesosphere.

Interferometric observations on the atmospheric emissions should be undertaken at various latitudes from the equator to the poles on atomic spectra of the twilight airglow, the night airglow and, when possible, of the aurora.

The first interferometric determination of temperature by Babcock [72] is widely quoted, but later results are also available [73].

3.2. Composition

In determining the atmospheric constituents and their relative concentrations, three general procedures have been used: (a) direct analysis of air samples obtained by means of some type of upper-air vehicle; (b) interpretations of the emission spectra of the airglow and the aurora; and (c) interpretations of the solar absorption spectrum.

Although, in principle, air samples may be obtained at any altitude (providing a suitable vehicle is available), the practical limitations become excessive in the higher atmospheric regions where the gas pressure is exceedingly small. The problems of selective absorption on the walls of the container, absorption during the sealing process, or contamination by leakage or by residual air are very troublesome, to say nothing of possible chemical reactions occurring within the container after sealing. For these reasons, successful sampling techniques using conventional methods have been used mainly with balloon-borne containers. However, recent results with rocket-borne containers are very gratifying and it is hoped that fresh thought and new experimental techniques will allow samples to be obtained from much higher altitudes.

Examination of the atmospheric emission spectra has yielded important information on the composition of the atmosphere at the altitude of emission. However, the results are not conclusive. While countless spectrograms have been collected, the identification of a fair portion of the lines and bands is still in doubt. Some of the confusion in the identifications may be traced to the observational techniques which are necessarily difficult because of the weak light intensities involved. On the other hand, many of the spectra of the atmospheric gases are also lacking; these vitally needed spectra may be obtained in the laboratory. Indeed, the compilation of a compendium of the emission (and absorption) spectra of the atmospheric constituents is urgently required. Such a compendium would aid immensely in the positive identification of some of the lines and bands observed.

The problem of excitation mechanisms for the airglow and auroral spectra is one of great scope, and much more extended investigation in the laboratory is necessary. Many of the forbidden lines of the atmospheric particles are found in the observations, and some of them may be reproduced in the laboratory under carefully chosen conditions. Although the atmospheric particles may be raised to the necessary excited state rather easily, most of them in discharge tubes are usually deactivated at the walls of the tube. In the high atmosphere, however, walls are lacking and the excited particles must either radiate or lose their excess energy in collisions. Thus the relative intensity of given spectral lines in the

laboratory may be quite different from that observed in the airglow or aurora.

It is also desirable to investigate the relative intensities of the lines and bands in nitrogen-oxygen mixtures excited by several different methods. Intensities should be compared in spectra obtained through bombardment of the gas by electrons or positive ions having different energies, as well as through different discharge conditions. The effect of traces of a foreign gas upon the intensity and type of these spectra must be critically studied, considering as contaminants any of the minor atmospheric constituents. Pronounced effects on discharge tube spectra may be produced by the introduction of contaminants (e.g., the quenching of resonance radiation) and it is conceivable that the metallic and other contaminants in the high atmosphere may have similar important influences. The movement of these contaminants through the high atmosphere by winds or diffusion may then give rise to remarkable effects and changes in the emitted spectra.

Deduced emission heights of the airglow are markedly inconsistent. Additional improvements in the observational techniques and refinements in the theoretical approach must be undertaken to allow a better interpretation of the observations.

In connection with an observational program, attempts must be made to broaden the latitudinal coverage. At high latitudes, intensive investigations of the airglow are required, particularly over the polar caps. Observations made in the spectral range 2000–40,000 A are very desirable. The observations should include a spectroscopic program, a study of emission altitudes, and an examination of the movement of luminous clouds (see Section 4). It is possible that the atmospheric cutoff at 2950 A, caused by ozone absorption, may be shifted during winter near the geographic poles. Measurements at high dispersion and resolution are required not only in polar areas, but also at lower latitudes, near the equator.

Studies of the solar absorption spectrum have been made on various occasions with emphasis on the infrared spectral region. These observations have been very fruitful, indicating the presence of isotopes and minor atmospheric constitutents. However, as in the case of the atmospheric emission spectrum, many more bands are obtained than have been identified. A compendium of the absorption spectra of the atmospheric gases would alleviate a basic deficiency. As many of the constituents have not been fully examined, comprehensive laboratory studies of the absorption spectra of the atmospheric constituents must be undertaken, preferably simulating conditions found in the atmosphere (i.e., long paths

and pressures below 760 mm of Hg). These laboratory examinations as well as additional solar observations should be made with spectrographs of high resolution, a condition hitherto mainly lacking. A program of continuous observations on the solar absorption spectrum is also needed in order to determine diurnal and seasonal variations of water vapor and other constituents, and their possible variation or correlation with weather. Observations of the solar absorption spectrum must be extended to the arctic regions where very long absorption paths may be obtained during winter. Further improvements in instrumentation and the observational techniques are required to permit deductions on the conditions of the terrestrial atmosphere *above* 50 km.

3.2.1. Compendium: Spectra of Atmospheric Gases. The phenomena occurring in the chemosphere, ionosphere and mesosphere might be better understood if full information were available on (a) the physical characteristics (such as the emission and absorption spectra, molecular structure, etc.) of the atmospheric constituents, and (b) the effects of the sun and solar bombarding particles on these constituents. The compilation of a compendium of the emission (and absorption) spectra of the atmospheric gases would be a step in this direction. Although a considerable amount of information already exists in the literature, the data must be critically examined and summarized in a form useful to the atmospherici-ist. The ideal compendium would be an extremely valuable aid for comparing and identifying the observed spectra. It should contain the emission spectra of all the atmospheric particles (listed in Section 3.2.2) in their normal, singly ionized and doubly ionized states as well as the spectra of the negative ions of atomic oxygen, molecular oxygen, nitric oxide, hydrogen, hydroxyl molecule, etc.

If possible, the compendium should include the following data: (a) the wavelength for each line or band emitted; (b) the relative intensity of emission of each line in a given series; (c) microdensitometer traces for each line and band and (d) an analysis of the spectrum in terms of the energy level, series relationships, and electron configurations for the particle concerned. Similar wavelength tables should be compiled for the absorption spectra.

A literature search requires the critical investigation of innumerable papers already published. It is not possible here to give as references the many hundreds of reports dealing in some fashion with the spectra of the atmospheric constituents. The texts by Moore [74], Pearse and Gaydon [75], Herzberg [76], and Price [77], however, provide an initiation into the desired bibliography.

3.2.2. Emission Spectra of the Atmospheric Gases. To assist in the identification of many of the lines present in the aurora and the airglow,

and to aid in the interpretation of the mechanisms and reactions which lead to the emission of these lines, the emission spectra of all the atmospheric constituents must be examined in detail. The analysis should follow the objectives listed for the compilation of a compendium of emission spectra of the atmospheric particles. Emphasis should be placed not only upon the comparatively simple determination of wavelengths but also upon an examination of the relative intensities of the emitted lines and bands. In the interpretation of observed data on the aurora and airglow, knowledge of the relative intensities of given bands may indicate which of several possible mechanisms has occurred (to bring the particle concerned into its observed emission state). The importance of obtaining accurate relative intensities, both experimentally and theoretically, of the bands of the ionospheric and mesospheric molecules under different excitation conditions cannot be overestimated.

Care must be exercised in both the laboratory experimentation and the theoretical computations. Experimentally it is possible to obtain in emission given band systems (e.g., the Vegard-Kaplan bands of molecular nitrogen) under a variety of excitation conditions; however, the relative intensities of the bands will differ with each situation. It is necessary to undertake the experiments under *very carefully controlled* laboratory conditions, considering, for example, microwave, arc, spark, and discharge sources, both direct and afterglow, and making intercomparisons of the spectra obtained in each instance. Unless the excitation conditions in the laboratory are accurately known, the results cannot be applied without ambiguity to the observed airglow and auroral spectra.

By employing various excitation techniques for the laboratory emissions, the excitation processes in each of the different sources may be clarified. Thought should be given to the use of flames as emission sources. In investigations on flames, the conditions in the flame already have been correlated with the relative intensities observed spectroscopically; no corresponding examinations of the excitation kinetics in glow or electrodeless discharges have yet been accomplished. The study of afterglows also merits further attention.

It might be noted that in laboratory experiments, long period flashing of the discharge tubes (of the duration of months or longer) changes the emission spectrum. Although the conditions within the tube are not accurately known after long flashings, this method, in some instances, allows the production of spectra which were absent in the fresh tube but which are identifiable with airglow emissions. With respect to technique, new methods may be required to improve the intensity of emission of some of the lines and bands. It would be desirable to obtain the spectra at temperatures of about 300 and 3000°K respectively. In the atmos-

phere, the spectra are produced at pressures of 10^{-3} to 10^{-7} mm of mercury.

In the theoretical determinations, the numerical evaluation of some of the wave equations is very difficult, and recourse to machine computation will probably be necessary. Quantal calculations are required to determine absolute intensities. While these evaluations are necessarily approximate their usefulness should be extended. Improved methods of computation should be found. Fortunately, the results are very useful (even though based upon rough calculations) because the transition probabilities vary by several orders of magnitude.

The atmospheric particles are defined as those bombarding as well as those present in the atmosphere. Emission spectra should be examined for the normal, singly ionized and double ionized states of the particles listed below.

(a) Molecules:

O_2	NO	CO	O_3
N_2	N_2O	OH	
H_2	NaO	NH_2	
Na_2	H_2O	NH	

(b) Atoms:

O	A	Ca
N	He	Ne
H	Na	

In addition to the general references given in Section 3.2.1, the texts by Bacher and Goudsmit [78] and Weizel [79] may also be mentioned. It is impossible in a paper of limited scope to include all references on the subjects of emission or absorption spectroscopy.

3.2.3. Absorption Spectra of the Atmospheric Gases. In penetrating the terrestrial atmosphere, solar radiation suffers considerable absorption, giving rise to the telluric absorption bands. These bands, entirely molecular in character, are sufficiently intense in some cases to absorb almost all radiation in their particular spectral interval. At the earth's surface, the absorption may be observed from the ultraviolet cutoff to the long infrared portion of the spectrum. Ozone contains a few diffuse absorption bands below 3000 A followed by a very strong continuum completely absorbing radiation below 2950 A. The absorption by atmospheric ozone extends to 2200 A. Regardless of the presence of ozone, however, molecular oxygen absorbs strongly in the region 2400–2200 A. As seen from the troposphere, the solar spectrum from 1950–1300 A is completely opaque because of absorption by the Schumann-Runge bands of molecular oxygen.

Other strong absorptions occur in the visible and infrared regions of the solar spectrum. Some faint telluric bands of ozone are found in the

visible region. Absorption by the rotational spectrum of water vapor is very large above 24,000 A and partially complete above 16,000 A. Ozone, oxygen, carbon dioxide, and water vapor give rise to distinct absorption bands between 7000–16,000 A. Observations on the solar absorption spectrum with proper instruments constitutes a powerful method of detecting the minor atmospheric constituents.

For the interpretation of the observed absorption spectra, it is vital to produce corresponding absorption spectra in the laboratory under as nearly similar conditions as possible. Some past experimentation in this regard has attempted to obtain absorption intensities comparable with those obtained in the atmosphere by employing high pressure absorption cells together with fairly long path lengths. However, there are disadvantages to using high pressures in the laboratory; the pressure broadening effect, for example, obliterates and smears the fine structure of the absorption bands, making a comparison with the intensity distribution in the fine structure of the solar absorption spectrum difficult and ambiguous. The laboratory measurements must be made with various pressures and temperatures. Pressures should not exceed 760 mm of mercury, and the temperatures should range from about 300 to 3000°K. With low pressures, long path lengths are required.

The experimental aspects involve a determination at high resolution of the absorption lines of the atmospheric constituents. The line widths, relative intensities, and line profiles should also be obtained. The necessary path lengths at the lower pressures require the use of multiple traversal absorption cells having means for temperature and pressure regulation.

An accurate knowledge of the conditions under which the data are obtained and an accurate determination of the absorption spectra are needed not only to aid in identifying solar absorption measurements but also for other purposes. Careful studies on ozone, for example, will yield the interatomic distance, clarify the assignment of fundamental frequencies, and accurately define the apical angle. Absorption profiles may serve as a basis for more careful investigations of radiative equilibrium in the stratosphere, chemosphere, and possible higher shells. In the experimental determinations of absorption spectra, intensities, and the pressures under which they are obtained must be accurately known. Careful measurements on the absolute absorption coefficients in the various bands of the atmospheric constituents are required. These measurements should be made (for a given gas at various pressures and temperatures) with varying amounts of the remaining atmospheric gases as contaminants. Additional experiments are needed on such factors as (a) the pressure broadening introduced in absorption bands both, through an

increase in the amount of the same gas and through the introduction of a second gas; (b) the thermal increase in overall absorption of a given band; etc. The widths of the absorption lines also should be obtained.

The theoretical determination of the absorption spectra and the relative intensities involved may be calculated by means of quantum mechanical methods.

The gases for which complete absorption spectra are desired are listed below. It should be noted that the compounds in this list are not identical with those given in Section 3.2.2. The difference between the two lists arises because one deals with emission and the other with absorption spectra. Absorption spectra in the atmosphere are most affected by the lower shells where the gas density is greater, whereas almost all emission arises in the chemosphere or higher regions. Thus, emphasis is placed in this section upon constituents of the lower atmosphere (see Section 2.2). Complete absorption lines and bands in the spectral range 2000–40,000 Å are required for the following molecules:

H_2	CH_4	O_3	CH	NaO	SO_2
O_2	CO	H_2O	NH	NO	NH_3
N_2O	CO_2	H_2S	NH_2	N_2O_5	

In addition to determining absorption spectra for the normal molecules, the spectra should also be obtained where possible for singly and doubly ionized molecules and isotopes.

In studying the heat balance of the atmosphere, the relative importance of the various absorption bands (of all possible atmospheric constituents) in the absorption and emission of radiation must be known. For this purpose accurate measurements on the absolute absorption coefficients in the principal bands of the atmospheric constituents are necessary at different temperatures and pressures and in the presence of varying proportions of the other atmospheric gases. Further information regarding the line breadths is required, however, before extensive studies of radiative equilibrium may be reliably undertaken.

Although many references may be given to the type of research desired, only a few need to be given here [80–83] (see also Section 3.2.2).

3.2.4. Solar Infrared Absorption Spectrum. The fraction of a second during which radiation traverses the solar and terrestrial atmospheres is sufficient to leave an indelible imprint upon the solar spectrum. This fact, long recognized, has led to various examinations of the solar absorption spectrum, particularly in the infrared, in an attempt to gather a wealth of information regarding the condition of both atmospheres. With respect to the earth's atmosphere, such factors as composition, temperature, radiative equilibrium, etc., and their diurnal and seasonal

variations may be studied. As diatomic molecules having no permanent dipole moment do not absorb infrared radiation, molecular oxygen, nitrogen, and hydrogen are mainly transparent in this region. Most of the radiation is absorbed by the triatomic and polyatomic molecules, such as water vapor, carbon dioxide, ozone, nitrogen-oxygen compounds, methane, and deuterium hydroxide.

The earliest investigations of the infrared solar spectrum were initiated by Langley in 1881. Since that time the studies were gradually extended so that by 1900 a fairly complete but rough absorption map was available to about 5.3 microns. The spectrum could be described in terms of broad absorption bands separated by gaps or windows. This type of endeavor continued so that at the present time the infrared solar spectrum is known with fine detail from about 0.8–2.5 microns and from 7–13 microns. This mapping has been performed with both prismatic and grating instruments. Maps of the spectrum also extend to about 24 microns with considerably lower resolution. Preliminary attempts are now under way to examine the spectrum to 100–400 microns.

The solar spectrum in the range 0.8–14 microns, when examined with a low dispersion spectrograph, could be almost entirely attributed to absorption by water vapor and carbon dioxide. With greater resolution a large number of individual absorption lines and bands, generally of the rotation-vibration type, were found; many of these bands are unresolved and others unidentified. In the transparent regions centered at about 1.65, 2.20, and 3.3 microns, the absorption lines are divided between those originating in the terrestrial and the solar atmospheres. However, simple observational techniques allow the identification of the solar spectrum. There is a progressive reduction with increasing wavelength in the observance of solar lines.

Analyses of the infrared solar spectrum not only allow a determination of those constituents absorbing infrared radiation, but also permit estimates to be made of (a) their relative concentration in the atmosphere; (b) a determination of rotational temperatures; and (c) a rough derivation of the height of the absorption layer. Because of the very long atmospheric path, the spectrum also contains lines and bands which could not be reproduced in any laboratory absorption cell. From high precision solar spectra, bands not yet conclusively known may be identified, and independent redeterminations may be made of the equilibrium moment of inertia and the rotational constants of some constituents.

From a geophysical aspect, additional long-term studies of the solar infrared spectrum are required in high and low latitudes, preferably at high mountain stations. Other studies may indicate the presence of industrial contaminants in manufacturing areas, or of unusual isotopic

distributions, provided, of course, the concentration of the particles is sufficiently great. Higher resolution instruments may allow an examination of absorption conditions in the chemosphere, where the air mass traversed is relatively small. In order to determine the nocturnal variation in the concentration, temperature, and possibly height of some constituents, additional observations are necessary on the infrared spectrum of lunar-reflected sunlight. Where feasible, the absorption produced in the airglow and auroral spectra should also be examined. A serious lack in the observational program is experimental data on the percentage absorption of the separate rotation lines in the solar absorption spectrum. Accurate measurements of the atmospheric transmission in the infrared are extremely important in determining temperatures of the other planets as well as in undertaking a comprehensive study of the earth's heat balance.

Many fundamental problems in solar physics may also be studied through the infrared solar spectrum. Investigations are urgently needed on the far infrared spectroscopy of solar features; i.e., sunspots, limb darkening, prominences, faculous areas and the lower chromosphere.

Several general texts and papers which will be useful in extending the studies mentioned above are given in the references [3, 84–86].

3.2.5. Emission Altitude of the Airglow. Although many investigations have been made of the altitude at which the airglow emissions occur, the results vary widely. Further, the temperatures deduced at most of the altitudes do not agree with the suspected distribution. The markedly discordant and irreconcilable altitude values hitherto obtained necessitate considerable improvement in instrumentation, technique, and theory.

In general, three methods may be employed to determine the altitude of the emission strata: (a) Van Rhijn's method (the most widely used technique); (b) twilight observations on the emission above the earth's shadow; and (c) triangulation measurements made upon given features of the emitting layer. Method (c) is confined to auroral investigations where the intensity of emission is comparatively large.

With Van Rhijn's method it is assumed that the emitting layer is thin, of uniform density, and at a constant height above the earth. The theory then predicts the emission altitude from the intensity ratio of a given spectral line (a) at a given zenith angle and (b) at the zenith. Unfortunately, however, for a small change in intensity of emission near the horizon a large difference in emission altitude is indicated. Van Rhijn's method fails because of the idealizations invoked. The luminescences of the night airglow are not emitted uniformly over the sky, but are concentrated in patches or clouds, some of which show distinct movements. Several emission layers, each of non-homogeneous intensity, may

be present. Neither the scattering of light by lower, stratified layers of the atmosphere, nor the absorption within those layers, has been considered. The disturbing effects caused by scattering and absorption, although difficult to evaluate, must be included in a more generalized theory if reasonable accuracies are to be obtained. It should also be noted that the present theory does not consider the background luminosity arising from zodiacal light and unresolved stars, both of which are equivalent to a thin emission layer located at an infinite altitude.

One recent improvement in technique is the development of automatically recording photoelectric photometers which can survey the entire sky in a series of eight horizon-to-horizon sweeps. On the assumption that a given spectral feature is always emitted at the same altitude, this instrument may be employed with interference filters to obtain the average intensity at the zenith and fixed zenith angles. On this basis a modified Van Rhijn method may be employed for the analysis. However, the question arises as to whether a given spectral line is always emitted at a constant altitude above the earth; the emitting stratum may be non-concentric with respect to the geoid and it may contain undulations; emissions may occur in several distinct strata, etc. In any event, no satisfactory corrections for the presence of the background continuum, scattering or absorption have yet been made.

Much more accurate altitude determinations are possible from examinations of the twilight airglow, which originates from resonance and fluorescent scattering. The relative intensity variation of a given spectral line with time may be observed and correlated with the altitude of the earth's shadow. Thus, from examinations of the twilight enhancement of certain radiations, the vertical distribution of the emitting particles may be calculated.

In connection with determining the emission altitude of the airglow, some classification of the radiations according to the population of the excited molecular and atomic levels is needed. A study of the altitude of emission of the various lines and bands gives a positive indication of the existence of particular constituents at definite heights. If the temperature is simultaneously determined by interferometric means, accurate determination of the emission altitudes will provide information on the temperature-altitude relationship.

References to investigations of the type desired or to extensions of present methods may be found in the papers of Barbier [87], Chandrasekhar [88, 89], and Roach and Barbier [90].

3.2.6. Photochemical Equilibrium. Photochemical equilibrium in an atmosphere defines the equilibrium established between the atmospheric constituents, incoming radiation, the physical state of the constituents,

and outgoing radiation. The problem is similar for both planetary and stellar atmospheres although probably more difficult for the former. Specifically, the photoequilibria of the following are necessary as a function of altitude: (a) an atmospheric molecule and its atoms; (b) a neutral particle and its ionization products; and (c) a negative ion and its corresponding neutral particle and electron. For one constituent in thermodynamic equilibrium with a radiation field at the same temperature, the equilibrium condition among the incoming radiation, parent particles, and daughter particles is given by the well-known reaction isochore. Employing the reaction isochore and knowing the necessary atomic and molecular constants, the density of the molecule dissociated, for example, may be determined as functions of the pressure and temperature. If a homogeneous, static atmosphere consisting of a single molecule is considered, it becomes possible to determine dissociative equilibrium as a function of height.

In the terrestrial atmosphere, the problem is much more complicated than in the simple case mentioned above. The atmosphere does not consist of a single constituent but a mixture of gases. A temperature gradient with height exists. Strata lying above the reaction zone probably absorb some of the incoming radiation, which then decreases in intensity and energy as it penetrates further into the atmosphere. Horizontal and vertical wind systems undoubtedly are present. Even in a single-molecule atmosphere these winds mix and redistribute the parent and daughter particles, making a determination of their distribution with altitude difficult. Further, the atmosphere is not in thermodynamic equilibrium with the incoming radiation; the radiation field of the sun is at a much higher temperature than that of the ionosphere and mesosphere. Attempts to solve the general static problem face certain characteristic difficulties inherent in treatments of systems not existing in a true equilibrium but only in a stationary state. A complete thermodynamic theory of stationary states is much to be desired.

Because of its complexities, the treatment of this problem has been simplified considerably when applied to the terrestrial atmosphere. Usually one or more of the following assumptions are implicitly adopted: (a) that the rate of the forward and reverse reactions are equal (e.g., the rate of photoionization equals the rate of recombination, so that the parent and daughter particles remain in a dynamic equilibrium); (b) that the energy absorbed equals the energy radiated, so that an energy balance exists; (c) that the particles are distributed in hydrostatic equilibrium with the temperature gradient; and (d) that no wind or turbulence is present. It should be noted that assumptions (a) and (c) must be consistent. Although the assumptions (a), (b) and (c) idealize the

problem, they provide some inkling of the magnitude of photochemical equilibrium in the atmosphere. Refinements in the theory ultimately should include simple hydrodynamic effects disregarded in (d).

Even with a relatively simple treatment based upon the previous assumptions, care must be taken not to invoke procedures and factors of doubtful accuracy. The reaction leading to the formation of the parent particle may not be the reverse of that producing the daughter particles. It is obvious that before calculations can be undertaken the actual processes involved must be known. The reactions which occur in the high atmosphere are not known with certainty. Thus, dissociation may occur through the direct absorption of energy greater than the dissociation potential. However, the final particles will exist in different energy states depending upon the particular energy-absorption processes involved. Association may be of the two-body or three-body type. In each instance, the cross section for the process is different. Not only is there doubt regarding the actual processes which take place, but the cross section for the processes producing the daughter as well as the parent particles is generally not known. In some cases, the amount of energy involved in a particular process (e.g., the dissociation of molecular nitrogen) is not known. Satisfactory calculations must include as a minimum the most prominent processes leading to the end products.

For the terrestrial atmosphere, the dissociation of molecular oxygen has been considered on several occasions. Calculations on attachment equilibrium and ionization equilibrium have also been undertaken. The results for oxygen, although performed with several different assumptions, generally indicate dissociation in a zone beginning somewhat close to 100 km. The treatment of dissociative equilibrium for nitrogen should also be attempted and must be attempted for other possible atmospheric gases. The difficult problem of a mixture where several gases are simultaneously undergoing dissociation at different rates eventually should be considered. Similarly, attachment equilibrium for the negative ions of molecular and atomic oxygen (and other negative ions) must be examined. With regard to ionization equilibrium, much remains to be redone, and each of the several atmospheric constituents must be treated. In the case of ionization equilibrium the electrostatic forces involved are rather small.

Past research on photochemical equilibrium is given in the references [20, 21, 88, 89, 91–96].

3.3. Collisional Phenomena

It is unfortunately true that an adequate understanding of the atomic and molecular processes occurring within the high atmosphere is almost

completely lacking. For example, the particular constituents ionized to form each of the ionospheric regions is not known. The general problem involves a determination of the probability for each of the many possible reactions. At any impact between two particles, a certain probability exists for the occurrence of a given process (e.g., elastic collisions, excitation, dissociation, ionization, attachment, association, recombination, and detachment). This probability may be conveniently expressed in terms of the effective collision cross section for the process involved. At some energies and under some conditions the collisions may be mainly elastic; but on other occasions only inelastic collisions occur. For a given bombarding and target particle both the total cross section for collision as well as the effective cross section for the reaction involved (e.g., excitation to various levels, collisions of the second kind, etc.) must be determined as a function of energy within the energy ranges conceivably present in the atmosphere.

Values of cross sections for most collisional processes, including collision with photons, are not available. Information on the probability of many energy-transfer processes (e.g., energy transfer from one of the forms of translational, vibrational, rotational, electronic, etc., to another of these forms) is lacking. Quantitative results are available for some very high-energy inelastic collisions involving heavy particles; however in the atmosphere the energies involved are small.

Collisional processes in the high atmosphere may be considered to occur in several distinct energy ranges. With regard to photochemical reactions induced under the influence of solar radiation, the reaction may be considered as the collision of an atmospheric particle with photons having energy probably not exceeding 600 eV. The high-energy particles, which on penetration into the atmosphere give rise to the aurorae, have either constant speeds of about 10^8 cm/sec or constant energies of about 500,000 eV. Present theories on the aurora advocate one or the other of these hypotheses. Although auroral theory is unsettled in this regard, it may be clarified if the required cross sections for both energy ranges were available. Another important process in the high atmosphere is that involving mutual collisions, both elastic and inelastic, of the atmospheric particles in the energy range 0–0.5 eV.

Some of the most important collisional processes may occur under the action of solar radiation which raises many of the atmospheric particles in the high sunlit atmosphere to metastable states. The energy available in radiation is undoubtedly several times greater than that absorbed at the surface of the earth. By collisions the energy absorbed in the high atmosphere may be interchanged and transferred among the atmospheric particles. In this fashion quantities of energy which are eventually lost

by reradiation, advection, conduction, etc. may be stored in the daylight hemisphere; however, little is known regarding the probabilities and cross sections for most of the reactions which may take place. This aspect of the problem is particularly intriguing, especially because of the possibility of injecting a foreign body in the form of a vehicle into the terrestrial atmosphere. It is conceivable that the vehicle could extract, utilize or even transmit some of the energy found in the higher shells.

The high speed (auroral) bombarding particles cause excitation and ionization of the atmospheric particles. The mutual collisions of the atmospheric atoms and molecules mainly favor the occurrence of association, recombination, some excitation, elastic collision, etc. Collision between the constituent particles allows the high atmosphere to attain a steady state condition. Knowledge of the cross sections for collision, though fundamentally needed for an explanation of many of the processes in the chemosphere, ionosphere, and mesosphere, is not available.

In general, collisional processes in molecular gases are very complicated. Those in the upper reaches of planetary atmospheres, where temperatures are high, contaminants are many, mean free paths extend from centimeters to kilometers, and metastable particle states persist for days and weeks, are even more involved. The logical method of examining the conditions in the terrestrial atmosphere consists initially of a study of the rate and cross section constants for the possible processes involved. If enough of these constants were available, it would be possible to deduce the atmospheric mechanisms. The method of qualitatively assigning the probable mechanisms from gross observations of the airglow, aurora, disappearance of electrons, etc., may hide many pitfalls that quantitative knowledge of cross sections (for the possible atomic and molecular processes) would immediately reveal.

It is not possible at present to separate all the different types of collisions for study, and to assign to each a cross section. The cross sections discussed below represent wide areas of deficient physical knowledge of extreme importance to research in physics, astrophysics, geophysics, and chemical kinetics. In this connection, a minimum listing includes a study of elastic collisions, excitation and ionization, recombination of ions to form neutral particles, and association of atoms to produce molecules. The important problem of attachment to neutral particles to form negative ions has been implied briefly in the section on recombination but has not been treated in greater detail. Similarly the necessity for considering charge transfer reactions is indicated in the sections on excitation and recombination. Obviously, however, the cross sections for attachment and charge transfer also warrant a full investigation.

The problems discussed so lightly here seem complicated in practice, perhaps because so little attention has been given them. Satisfactory theories on dissociative collisions or on very low pressure recombination are completely lacking. Both experimental and theoretical approaches to the problem are required, but both, because of their complexities, have scarcely been initiated. The experimental difficulties associated with very low pressure discharges or cross section studies are well known and need not be mentioned here; nevertheless, laboratory studies are vitally necessary to confirm the theoretical investigations and to give an insight into possible complicating or competing reactions not immediately obvious. The design of experiments which will allow the study of a single reaction or a determination of a given cross section requires careful thought. By no means is it desirable merely to repeat a technique which provides a result without giving a clue as to the reaction involved. The nature of active nitrogen (associated with the Lewis-Rayleigh afterglow in nitrogen discharges) is controversial and must be further investigated. The de-excitation of particular levels of a given particle by contaminating particles requires more attention.

The theoretical treatments are beset perhaps with even more obstructions. A solution of the problem requires (a) extensions of the basic theory to obtain more accurate approximations for the wave functions; (b) improvements in mathematical techniques to permit treatment of the resulting equations; and (c) machine calculations involving, if necessary, the construction of special computers designed specifically to handle the final equations. Many collisional problems involving molecules can hardly be treated with existing approximations to the wave functions. It cannot be emphasized too strongly that new assumptions for the wave functions of the reacting particles and new advances in the mathematical manipulation of the resulting equations must be sought. Extensive efforts are necessary to enlarge, generalize, and advance existing concepts. Only in this fashion will the desired information on cross sections become available.

3.3.1. Collision Frequencies in the Ionosphere. The ionospheric observations which yield information on collisional frequencies attempt to measure the amount of energy absorbed from a probing radio wave. The type of measurement employed is based upon the condition that no energy is absorbed from a radio wave when it penetrates an ionized medium wherein collisions are absent. If, however, collisions are sufficiently frequent, the electrons and other particles will absorb and dissipate energy from the penetrating radio wave. The amount of absorption, which may be appreciable, depends upon the collisional frequency of the electrons with all other atmospheric particles. As the density or tem-

perature of a gas may be obtained from the collision frequency, a means exists for determining the temperature of the ionic layers if the collision cross sections are known and if accurate ionospheric collisional frequencies may be determined.

The experimental values of collision frequency in the ionospheric regions have been determined by means of two techniques. In one type of experiment the reflection coefficient of a given ionic layer is measured. In a second type of experiment, the influence of the ionosphere (disturbed by the absorption of energy from a first probing radio wave) upon a second probing wave, is measured. In the latter method, two transmitters and one receiver are employed. One transmitter and its receiver form the termini of the propagation path, and the second transmitter is located somewhat close to the midpoint of the great circle path passing through the terminal points. It is found that low frequency radiations from the second transmitter induce a modulation on the transmissions from the first transmitter. This phenomenon of cross modulation, known as the Luxembourg effect, is produced because of a local "forced heating effect" which occurs when energy from the probing radio wave of the second transmitter is absorbed. The absorbed energy increases the local collision frequency in the ionic layer concerned. Measurements of the degree of modulation of the second signal upon the first give an indication of the collision frequency of electrons in that portion of the ionosphere where the interaction occurs. It is important that the altitude of reflection of the first wave be closely the same as the altitude of maximum energy absorption from the second wave. The effect of the geomagnetic field should be considered. Either pulse or continuous wave transmissions may be employed, and all three equipments may be located at the same site.

The first method yields an estimate of the collision frequency from observations on the reflection coefficient. The frequency of the probing radio wave used in these measurements should be close to the critical (or penetration) frequency for the ionospheric layer involved. From simultaneous observations of the equivalent path and the relative intensity of the reflected wave, the absorption may be estimated. Under suitable assumptions the collision frequency may then be obtained.

Past experiments have yielded some information on the collision frequencies at discrete altitudes from 85–400 km. However, much more extensive experimentation is required to determine the collision frequencies in the ionic layers as a function of altitude, geographic location and time, both diurnally and seasonally. In this fashion the diurnal and seasonal changes in density and temperature at various altitudes and latitudes may be obtained. The information could be employed to

determine the temperature and density patterns at altitudes above 85 km. More accurate determinations of the ionospheric collisional frequency are needed.

The experimental techniques are well described in the literature [97–104].

3.3.2. Cross Section for Elastic Collision. In order to interpret clearly the results of radio probing measurements which provide elastic collision frequencies, it is necessary to determine the magnitude of the collision cross section of electrons with all atmospheric particles and their ions. Without accurate cross sections, grossly misleading densities or temperatures may be computed from the experimentally determined collisional frequencies. It should be noted that the temperature varies with the fourth power of the collision cross section.

Some early experimental and theoretical work indicated that for both atomic and molecular nitrogen, and for molecular oxygen, classical values of the cross section for collision may be employed. However, in other instances (e.g., atomic oxygen) the determination of an accurate value for the cross section is somewhat difficult.

The cross section for elastic collision must be determined not only between electrons and neutral particles, but also between electrons and the positive ions of these particles. All constituents existing in the atmosphere should be considered because of the possibility that some particles or ions may have extremely large cross sections. Because of the probability that large negative ion concentrations exist in the lower extent of the ionosphere, the elastic collisional cross section between these particles and electrons should also be obtained.

As temperatures in the ionosphere and mesosphere probably do not exceed 5000°K, elastic cross sections are needed at most for particles within the energy range 0–0.5 eV. Because the total collisional frequency is measured by the radio probing techniques, theoretical values of the cross sections must be evaluated for several groups of particles:

a. Electrons with the following atoms and positive ions:

O	O^+	O^{++}	A	A^+	Ca	Ca^+
N	N^+	N^{++}	H	H^+		
Na	Na^+		He	He^+		

b. Electrons with the following molecules and positive ions:

N_2	N_2^+	N_2^{++}	NO	NO^+	NO^{++}
O_2	O_2^+	O_2^{++}	N_2O	N_2O^+	N_2O^{++}
Na_2	Na_2^+	Na_2^{++}	NaO	NaO^+	NaO^{++}
A_2	A_2^+	A_2^{++}	CO	CO^+	CO^{++}
H_2	H_2^+	H_2^{++}	OH	OH^+	OH^{++}
			NH_2	NH_2^+	NH_2^{++}

c. Electrons with the following negative ions:

O^-	NO^-
O_2^-	H^-
OH^-	CO^-

d. Negative ions included in (c)with all particles listed in (a) and (b).

e. Electrons with the atmospheric atoms and molecules (listed in (a) and (b) above) raised to metastable states.

The problem of collisions of electrons with metastable particles has much greater importance in the high atmosphere (where the mean free paths are very long and where no walls exist) than in the laboratory. In the mesosphere particles may persist in metastable states for appreciable periods of time.

Many quantal calculations are required for the theoretical determination of collision cross sections. The general scattering problem involves the solution of an infinite set of simultaneous equations; however, by judicious approximations which attempt to reduce the number of terms and thus the number of equations involved, it is possible to obtain practical solutions of varying degrees of accuracy. Consideration has already been given in solving these problems to the use of Born's and Jeffrey's approximations, Hartree's unmodified field, the Hartree-Fock field with and without exchanges and with polarization corrections, Hulthen's method, variational techniques, numerical solution of the differential equations, etc., to list but a few.

For the lighter atoms, it may be necessary to employ better approximations and to consider the possibility of electron exchange between the particles involved and the colliding electron.

The variational method may, perhaps, be the most accurate and provide the best procedure in certain cases. However, in attempting to solve some of the equations for the evaluation of specific cross sections, it undoubtedly will prove necessary to develop new approximations to the wave function. The calculated values of some atomic properties are very sensitive to the wave functions employed. The best available wave functions in some instances are far too inaccurate; radical improvements to these functions must be sought before the desired research may take place.

While the theoretical cross sections for collision must be evaluated, the experimental values should also be obtained in the laboratory to verify the calculations. Some experimental techniques for determining the cross sections of slow electrons and negative ions have utilized accelerated electron beams, and recent experiments have been performed using microwave breakdown techniques. However, considerably more

thought and ingenuity is still necessary in devising suitable experimental methods for determining the collisional cross section, for example, of electrons and negative ions to each other or to the particles given above.

Some of the references given in this section also apply to succeeding sections dealing with collisional problems [105–113].

3.3.3. Cross Section for Energy Absorption. The terrestrial atmosphere may be likened to a gas periodically exposed to a radiation field under whose influence many photochemical reactions occur. Most of the absorbed radiation raises some particles to excited states, ionizes and dissociates others, causes photodetachment, etc. The entire cycle is repeated daily, storing fair quantities of energy. Obviously, the atmosphere is a radiant heat engine driven by solar energy absorbed through photochemical reactions; it is therefore of extraordinary importance to determine quantitatively the details by which the sun's radiation is initially absorbed by the atmospheric constituents.

Unfortunately, little is known regarding the total energy or the relative distribution of high- and low-energy quanta emitted by the sun. It is certain that the sun does not radiate as a black body at 6000°K. On the basis of theoretical studies and rocket experiments it seems wise to adopt arbitrarily ca. 600 eV as the maximum photon energy responsible for the ordinary photochemical reactions taking place in the atmosphere. In these mechanisms, the reacting particles acquire electronic energy as a result of the absorption of radiation. The particles are raised to higher energy levels accompanied by dissociation, ionization, etc., which may occur directly or through subsequent radiations. (The atoms, ions, and molecules so produced may then enter into other reactions determined by purely thermal considerations; see Section 3.3.4.) In effect, an almost unlimited number of possible reactions, both photochemical and thermal, become possible among the various atmospheric constituents, all caused primarily through the absorption of solar radiation.

Although many photochemical processes have been suggested as occurring in the chemosphere, ionosphere, mesosphere and exosphere, no more than a small fraction of the total has been proposed or quantitatively considered. The photoionization of a variety of substances (e.g., excited atomic oxygen, molecular oxygen, nitric oxide, sodium, and others) has been put forth as causing the D ionic region. Photoionization of these constituents (except sodium) as well as atomic or molecular nitrogen has been suggested at various times for the remaining ionospheric layers. Numerous secondary compounds (probably oxides, hydrides, nitrides, etc.) form and react in this chemical kitchen of the terrestrial atmosphere, some of them with singular importance. (The emissions of the hydroxyl molecule and atomic sodium, for example, are among the most intense features of the night airglow.)

To examine the absorption of solar radiation by the atmosphere, the cross section for absorption of radiant energy of each constituent particle must be determined both theoretically and experimentally. The cross section values may be directly applied to theoretical determinations on the formation of the ionospheric layers, regions of dissociation, radiative equilibrium processes, etc.

In examining the absorption of radiation, consideration must be given to such factors as pre-dissociation and pre-ionization of the atmospheric particles. In the laboratory, pre-dissociation of diatomic molecules is comparatively rare, even though this condition may exist for all discrete states that lie above the dissociation limit of a molecule. Generally, the probability of a radiationless transition into the dissociative state is small; the molecule emits radiation and drops into a stable state before dissociation might have occurred. In the ionosphere and mesosphere, however, the probability of pre-dissociation may be very much greater, and this type of reaction may be of considerable importance. Further study is needed. A very important pre-dissociation mechanism for the diatomic molecules of the earth's atmosphere is that wherein the dissociation continuum belonging to a second electronic state (vibrational or rotational level) overlaps an electronic state. In pre-ionization the particle is first raised to an excited discrete level located above the ionization potential; spontaneous ionization may ensue if the level is unstable.

The cross sections for the absorption of radiation by the atmospheric particles are known in only a few instances. Some progress has been made in determining this factor for the atoms of oxygen and nitrogen. Calculations of the absorption coefficient have been made in some instances through the use of Kramers' formula. Objections may be raised against this method, because it extends a formula originally derived for X-ray absorption far beyond its range of applicability. Experimental determinations of the cross sections are few, mainly because of the difficulties involved in making measurements in the desired energy range. Numerical evaluations may be attempted for some atomic particles, but for many others and especially the molecules, considerable extensions of the theory are first necessary.

In examining the cross section for absorption, the processes leading to dissociation, pre-dissociation, ionization, pre-ionization, photodetachment, and excitation into various states must all be treated. Without a set of accurate wave functions, almost all of which are lacking for the atmospheric particles, little progress can be made in determining the cross sections theoretically. Approximations to the wave functions for molecules are exceptionally troublesome. Even with some atoms severe restrictions of the method are encountered. Thus, in the case of atomic

potassium the cross section is extremely sensitive to the form of the wave functions employed in the basic formula. Because of this fact laborious calculations do not yield a value correct even in order of magnitude. Fortunately, an estimate of the reliability of the results may be made rather early in the calculations. In other instances, as the results are but slightly sensitive to the form of the wave functions adopted, fairly reliable estimates of the cross section may be obtained.

Obviously, considerable improvement and more refined assumptions to the wave functions must be made in the theoretical studies before the final equations may be used with any degree of assurance. The effect of electron exchange should be included if possible. Extensive research is required to formulate methods for the computation of absorption cross sections for molecules. Normalizations of the continuous wave functions are inadequately tabulated at present, making individual computations extremely tedious. After the methods have been extended and generalized, machine calculations will be necessary.

Experimental investigations of the absorption cross section are also needed, especially since for some constituents existing deficiencies hardly permit an accurate theoretical examination of the cross section. Photoionization, photodissociation, photodetachment, and other processes must be critically examined in the laboratory under conditions approximating those found at 100 km or higher. Again, as with elastic collisions, considerable thought is necessary to devise new and less ambiguous experiments for obtaining complete and accurate data. With regard to molecular oxygen, past laboratory evidence suggests that ordinary photoionization is very inefficient. If so, pre-ionization may be important and should be studied. Absorption cross sections for all atmospheric constituents should be re-examined experimentally.

Cross sections for the absorption of radiation by the normal, singly- and doubly-ionized states of the following particles as a function of energy in the energy range 0-600 eV are ultimately necessary if an understanding of the atmospheric mechanisms is to be obtained:

Elements:	O	H	Ne	Na	N	He	A				
Molecules:	O_2	H_2	NaO	CO	NH_2	N_2O	N_2	NO	OH	NH	H_2O

Some past works which may be useful in pursuing this topic further are given in the references [76, 114–121].

3.3.4. Cross Section for Collisional Excitation and Ionization. Although many spectral observations have been made of the airglow and the aurora, little emphasis has been placed upon their interpretation in terms of the atomic and molecular processes involved. For this purpose an exhaustive investigation of the collisional mechanisms involving the atmospheric

particles must be initiated. As excitation almost invariably accompanies ionization when produced by particle collisions, both processes will be considered in this section.

The auroral emissions arise because of the bombardment of the high atmosphere by solar ejecta (probably electrons, protons, helium ions, calcium ions and undoubtedly others) having either speeds in the range 10^7–10^9 cm/sec or constant energies of about 500,000 eV. These primary bombarding particles traverse the mesophere and ionosphere, causing ionization and excitation of some of the atmospheric constituents. As the primary positive ions penetrate the atmosphere they may be neutralized by the acquisition of electrons which subsequently may be torn loose with increasing penetration. In this fashion the particles may alternate between ionic and atomic forms as they stream through the atmosphere. It is also possible that the primary particles produce large quantities of secondary electrons through multiple ionization of the atmospheric gases. These electrons in their downward movement may cause an appreciable excitation of the atmospheric particles, comparable to conditions found in a discharge plasma. Thus, for the auroral problem, an investigation of the cross section for excitation and ionization of the atmospheric gases as a function of energy by the particles listed above (and their atoms) must be made.

A second, but quite different type of inelastic impact in the high atmosphere, applies mainly to the airglow emissions. The airglow emissions arise from the mutual collision of the atmospheric particles which exist in normal or metastable states and which move with energies well within the range 0–0.5 eV. For the latter problem, the mutual and self cross sections of all the atmospheric particles must be determined, also as a function of energy.

As a few examples of some of the observed effects, the bands of the first negative system of O_2^+ (in the auroral spectrum) probably are produced by collisions which both excite and ionize. Nitrogen molecules in the $X^1\Sigma_g^+$ ground state may be excited to a singlet state followed by de-excitation collisions which bring the molecule to the $C^3\Pi_u$ state. Subsequent transition to the $B^3\Pi_g$ levels causes emission of the bands of the second positive group. Obviously, in this particular case a knowledge of the collisional cross section and the mean lifetime of such levels as $a^1\Pi_g$ are necessary for a quantitative treatment of the problem. Many transitions of the atmospheric gases are involved; in all cases the cross section for the reaction in question must be obtained.

The effective cross section for excitation by electrons, protons, and other positive ions varies widely according to the molecule involved and the level to which it is excited. The optimum energy (relative to the

excitation potential) of the bombarding particle, and the sharpness of this maximum, depend upon the transition concerned. The processes which occur at different energies of the bombarding particles are not clear. In this connection, the cross section for the transfer of excitation, in which electronic or other excitation is transferred from one of the colliding systems to the other, must also be determined more carefully.

Another collision process of importance in studies of the high atmosphere is that wherein an electron is transferred during a collision between a neutral and a charged particle. During the collision the original ion becomes a neutral particle and vice-versa. The mechanism involves the transfer of an electron as well as energy. As the number of charged particles in the ionosphere and mesosphere is appreciable, collisions involving charge transfer may be of great importance, particularly in conjunction with subsequent recombination mechanisms. One reaction which should be more thoroughly treated, for example, is that involving the cross section for electron transfer between an atomic oxygen ion and all other possible constituents of the atmosphere. The entire problem should be fully treated and the cross section eventually determined for charge transfer between any ion and any neutral particle existing in these atmospheric regions.

It would generally be expected that for the transfer of excitation or charge in slow collisions, the probability of transfer attains a maximum when the energy difference between the two involved states is zero. Also, the largest cross section (which may be materially greater than the gas kinetic cross section) and the sharpest resonance would be expected to accompany excitation involving optically-allowed transitions in both systems.

Practically no quantitative data are available on the cross section for excitation of the atmospheric gases as a function of electron energy. Because proton spin does not change during a transition, protons may induce transitions only between terms of the same multiplicity. Thus, many energy levels of the atmospheric particles may not be directly excited from the ground level during proton impact. However, this restriction is removed if the proton acquires an electron to form a hydrogen atom. The possibility of resonance effects should be fully explored.

In the course of their bombardment of the atmosphere, the solar particles undoubtedly raise some atmospheric particles to metastable states. These metastable atoms and molecules may appreciably affect subsequent reactions and contribute a significant proportion of the total ionization and excitation. The effectiveness of metastable particles partially lies in their long lifetime, during which they may store energy with a high probability of transferring this energy at the next encounter with a second particle.

The laboratory research is exacting. Experiments must consider the effect of contaminants. Striking effects in the apparent cross sections have been produced (in low pressure discharges) by impurities in the gases studied. The theoretical determination by quantum mechanical methods of inelastic collisions for slow electrons is much more complicated than for fast particles. In general, however, a satisfactory quantum mechanical method of treating slow, inelastic electrons does not exist. Approximations of varying degrees of accuracy may be utilized, but in this connection much more refinement in the theory is in order. Theoretical determination of the cross sections involving heavy particles are also in a most unsatisfactory state. The theoretical calculations involved are intricate and are beset with difficulties similar to those outlined in Sections 3.3.2 and 3.3.3.

Further information on this subject is given in the references [110, 112, 122–131].

3.3.5. Cross Section for Recombination and Association. The reverse actions of photoionization, photodissociation and photodetachment are recombination, association, and attachment. A theoretical determination of the former cross sections should also yield the latter, inasmuch as the same matrix element is employed in both the forward and reverse actions. Similarly, the reverse action of impact ionization and dissociation is three-body recombination and association, respectively.

The experimental study of recombination is complex. Recombination may occur under a wide variety of circumstances, each having its own process and its own coefficient. Early investigators did not recognize these factors, and in their results the different types of recombination were inextricably confused. Notwithstanding marked advances in experimental techniques, the interpretation of recent experiments is clouded by many complications, such as (a) the nature of the ion carriers involved; (b) the chemical and other processes occurring in the gas under study; (c) the atomic or molecular levels into which an electron is captured; (d) the quantitative effect of impurities; and (e) the effect of the walls upon the reactions under study. With regard to the latter, for example, radiationless recombination may occur at the walls of the discharge tube. It should be noted also that results obtained at "low pressures" in the laboratory may not necessarily apply to the ionosphere and mesosphere where pressures may be lower by several orders of magnitude.

New types of experiments dealing with recombination and dissociation are required, preferably in the pressure range 10^{-3} to 10^{-7} mm of mercury and at energies of 0–0.5 eV.

Few theoretical examinations of recombination have been made. Electrons may recombine, for example, into any one of many levels in the atom; for each level a value of the effective cross section as a function of

the (relative) energy of the electron may be found. The sum of these cross sections for recombination gives the total cross section from which the recombination coefficient could be determined as a function of the electron velocity distribution. Not only will the theoretical investigations supplement and clarify the laboratory experiments, but they will also give an insight into possible recombinations occurring in the high atmosphere. Both two- and three-body recombination mechanisms must be considered, the latter being generally restricted to the lower ionic regions where number densities are much higher than in the F layer. Because of the long free paths found in these regions, metastable particles may exist for long periods of time; their role in recombination should be carefully examined.

The types of processes to be examined both experimentally and theoretically are given below:

a. Electronic recombination with the following:

O^+	N^+	Na^+	H^+	H_2^+	Ne^+
O_2^+	N_2^+	A^+	He^+	NO^+	OH^+

b. Neutralization recombination of negative ions with the positive ions given above. While the negative ions of atomic and molecular oxygen may be the most important in the upper atmosphere, other possible negative ions include:

$$H^-,\ CO^-,\ OH^-,\ NH^-,\ \text{and}\ NO^-$$

c. Three-body recombination involving a neutral particle or ion and (a) electrons and positive ions and (b) positive and negative ions.

In the two-body process radiation may carry away the excess energy; with three-body recombination the neutral particle serves the same purpose. However, at very low pressures the probability of the three-body collisions becomes very small.

Dissociative recombination may be very important in the lowest ionospheric layer. While emphasis in this section has been placed upon recombination, similar remarks may be made for association. Further information on the research desired may be found in the references [132–138]. (See also Sections 3.3.3 and 3.3.4.)

4. Dynamics of the Ionosphere and Mesosphere

A major portion of the studies regarding the characteristics and properties of the high atmosphere has been based upon the assumption of a tranquil, quiescent region. That the ionosphere, particularly, and the mesosphere are far from calm, but are in constant and violent motion, has been intimated from several types of observations. The rapid

distortion of long enduring meteor trains, observed visually and photographically, provides striking evidence of the existence of strong winds or ebullient eddies between 45–100 km. The periodic fluctuations in barometric pressure indicate the existence of tides of appreciable amplitude in the high atmospheric regions. The definite wavy arrangement of noctilucent clouds (near 80 km) and their rapid and continuous change in shape and form suggest the presence of vertical currents. Observed deviations between the expected static and dynamic pressures on some rocket flights may be explained if strong updrafts exist. Inferences from observations on the diurnal variation in the terrestrial magnetic field lead to the conclusion that the magnetic changes could be explained relatively simply if ion currents were present within the chemosphere or ionosphere. More recent observations on the movements of ionospheric irregularities (80–350 km), meteor ionization trains (about 90 km), luminous clouds and electron clouds (about 110 km) also verify the existence of some type of motion in these regions.

The above deductions provide an entrance into the difficult and almost totally neglected field of high atmospheric dynamics. This evidence requires that the concept of a uniform, regular, and static atmosphere be replaced by one where turbulence, movements and, perhaps, rapid changes are common. One cause of the movements may be the unequal heating in the high atmosphere; i.e., greater heating in the sunlit than in the dark hemisphere and greater heating during summer than during winter. The unequal heating sets up thermal and pressure gradients which then initiate both local and global circulation systems. It is possible that these insolation-produced winds are of the same order of magnitude as those created through gravitational attraction of the sun and moon. It is customary to consider both the gravitational forces (which are global in character) and the global thermal forces as the tide-producing forces. The wind systems which the sun and moon originate are termed the solar and lunar tidal winds, respectively. Thus, the term "tides" indicates those large-scale atmospheric circulations attributable to both solar and lunar gravitational attraction as well as to solar heating. The tidal movements may dominate all other motions in the ionosphere and mesosphere. Contrary to expectation, the atmospheric solar tides, observed at the surface of the earth, are about fifteen times greater than the lunar tides. This apparent anomaly arises because the free period of oscillation of the atmosphere is very closely a harmonic of the solar day.

The wind movements at the higher altitudes may be markedly affected by the presence of charged particles. Ions are imbedded in the atmospheric fluid from about 70 km upwards. If no geomagnetic field

were present, the air flow (including the ion flow) would be determined mainly by the thermal and gravitational forces described above. However, the magnetic field affects the movements of the ions as they are carried with the air. The tidal forces set up predominantly horizontal (concentric with the geoid) winds which intersect the geomagnetic lines of force because of the inclination of the latter to the horizontal. Under these circumstances, the charged particles develop a vertical velocity component which may be appreciable at some latitudes. It is therefore possible for an essentially horizontal wind to impart a rather large vertical motion to the ions and electrons in the higher shells. The effect of the magnetic field may be considered to be similar to that of a frictional drag upon the entire wind motion. In this sense the magnetic field increases the air viscosity of the high atmosphere.

An investigation of winds, tides, and other movements within the ionosphere and mesosphere will provide a better understanding of conditions to be expected within these regions, and undoubtedly will clarify some presently known peculiarities. Thus, the anomalous behavior of the F2 ionospheric layer may be partially attributed to winds. Tides and winds affect the height and maximum ion production in each of the several ionic layers and may influence the altitude of maximum auroral luminosity. In general, it is difficult to isolate the solar tidal movements from ionospheric data statistically because of coincident periods in ion production arising under the influence of solar radiation.

The process of diffusion increases in importance with increasing altitude, and its effects must be considered carefully in theories on the high atmosphere. In the ionosphere, where mean free paths may attain hundreds of meters, it would be suspected that molecular diffusion is much more significant than eddy diffusion, but additional research is necessary to demonstrate the effect and magnitude of each. In general, turbulence is probably appreciable and the principal factor involved in mixing the constituent gases is molecular diffusion.

A careful examination of frictional effects arising from the molecular viscosity of the air is fundamental in a study of the large-scale circulations of the ionosphere and mesosphere. The kinematic viscosity increases with altitude to values much greater than that found near the surface. If the pressure gradient is a function of altitude, the velocity of the air at different altitudes must also vary. Under this condition frictional drag (caused by the shearing stresses exerted between upper and lower layers) prevents the actual winds from attaining the gradient wind velocity at the altitude considered. The steady state circulation system will, of course, be different from the initial wind patterns. At first, air motions will be directed along the pressure gradient, possibly at very high speeds,

from a high- to a low-pressure area. The final steady state wind system which results will have rather lower speeds when the pressure gradient, frictional forces, and the Coriolis force are balanced.

The equation of continuity describing the fluid movements in the higher regions becomes more involved inasmuch as it must include sources and sinks of the atmospheric constituents. Molecules dissociate to their constituent atoms and vice-versa, ions are produced (through photo-ionization and photodetachment) and removed (through attachment of electrons to neutral particles, recombination of ions to form neutral particles, diffusion, etc.). The effect of divergence and the inflow or efflux of particles (brought about through thermal contractions or expansions) further complicate the theory.

The entire problem of fluid flows in the higher atmospheric regions is very complicated and stands as a virgin research field offering an exciting challenge to the hydrodynamicist. The problems arising in this connection are many and complex. Only several of the large number of problems concerning the dynamics of the high atmosphere will be discussed here. The hitherto neglected field of auroral dynamics, and the important topic of ion currents flowing near the ionospheric regions are not discussed for lack of space. These currents produce the diurnal variations in geomagnetic field strength. Each subject warrants much further study.

4.1. Wind Observations

In any particular atmospheric stratum where solar energy is strongly absorbed, the heat gain may be balanced by strong convective activity (providing, of course, that other heat sinks, such as conduction, advection, and reradiation, are relatively small or absent). Above this absorbing layer, conditions are somewhat similar to those found in the troposphere, except that in the higher regions no boundary in the usual sense is present. Undoubtedly, Bénard cells may develop whose size depends upon the thickness of the convective layer and, possibly, upon the degree of instability. Even with some changes in the amount of heat transport, a regular convective cell pattern probably would be maintained, providing the wind shear was not too great. Above a certain critical value of wind shear, the cellular characteristics would tend to disappear.

Because of the many energy processes and the corresponding large energy absorption occurring within the ionosphere, it is quite possible that Bénard cells form in the upper ionosphere or in the lower mesosphere. This heat absorption zone may become somewhat modified by mixing which arises from the local winds produced through convective

activity. However, under the action of solar radiation, the energy absorption layer would tend to persist and maintain itself, although possibly shifted somewhat in altitude. While this proposed model is qualitatively simple, its quantitative examination is difficult and requires knowledge of many additional factors including the coefficients of eddy viscosity and eddy diffusivity. (Although all the necessary data are not immediately available, and many undetermined factors are still present in the model, preliminary calculations on the convection region would be of great interest at this time.)

The horizontal wind in the upper atmosphere, as in the troposphere, initially tends to flow from a high pressure area toward a low pressure area. However, through the action of the Coriolis force the flow is deflected in the usual manner (i.e., towards the right in the northern hemisphere and towards the left in the southern). As in the lower regions the circulation pattern is not global, but is broken down into several smaller "regional" circulation patterns, each possibly indicating a diurnal and seasonal effect. Indeed, observations on the motion of ionospheric irregularities confirm the existence of flow patterns which roughly reverse themselves from day to night and from summer to winter. It would be anticipated that regional circulations would shift their position with respect to the earth with season; these effects may be difficult to determine by a single ionospheric wind observing station. Seasonal wind shifts in the mesosphere and ionosphere may be appreciable at high latitudes where the energy absorption within the ionosphere must change markedly from winter to summer. At the poles, for example, no sunlight falls upon the atmosphere to an altitude of 100–200 kilometers for several consecutive months. Thus, within this region the seasonal change in absorbed radiation is very large. Changes in wind velocity originating at the higher latitudes may be correspondingly great, and may be expected to manifest themselves at much lower latitudes.

It is also interesting to consider that purely horizontal winds at the higher latitudes must eventually result in diverging or converging flows. The density change within the mesosphere is not large, and the horizontal winds near the polar regions may induce large upward or downward directed currents in this shell.

The above speculation regarding wind systems in the mesosphere has neglected the effect of gravitational tides, and has confined itself solely to speculation on the thermally produced flows. Both types of winds must be examined more critically, particularly to determine the relative magnitude of each. Obviously, observational methods of determining winds in these regions provide information only on the total winds (i.e., those produced both through gravitational action of the sun and moon,

and through the absorption of solar radiation). Theoretical determinations are necessary in order to separate the relative importance of the thermal and gravitational components.

The presence of winds above about 90 km has been confirmed by observations made on luminous and ion clouds, ionization and visible trails of meteors, and radio wave probings on ionospheric irregularities.

Many additional observations are required before knowledge on the global wind regimes in the ionosphere will be possible. Continuous observations are necessary at low, middle, and high latitudes at numerous stations on the earth. Radio probing techniques should be employed and coordinated recordings taken during the same time period at many stations located over extensive geographic areas.

Although emphasis has been placed in the following sections upon wind observations in the ionospheric regions, a theoretical investigation of winds in this and higher atmospheric shells also merits extensive study. The general circulatory system is influenced by the viscous stress, pressure gradient, temperature lapse rate, horizontal temperature distribution and the time necessary to attain a steady state wind condition. From available information on the horizontal and vertical temperature patterns, it is possible to undertake preliminary evaluations of the expected wind components normal to and along the isobars. Simple meridional circulations as a function of altitude may then be proposed for the ionosphere and mesosphere.

A study of traveling disturbances in the ionosphere has led to the concept that they involve vertical cellular waves, rotational in type, which travel between two horizontal bounding surfaces. A treatment of these high altitude disturbances may be very similar to those proposed for waves in the lee of mountains or for the cellular wave theory of clouds.

4.1.1. Movement of Ionospheric Irregularities. The radio wave probing of ionospheric irregularities allows investigation of ionospheric movements during both daylight and darkness from a single station. The station is arranged so that its equipment, consisting of one or more transmitters, three receivers and a recorder, is distributed at the corners of a right triangle. The sides of the triangle are several hundred meters in length, the actual distance being somewhat dependent upon the radio frequency employed for the observations. A given pulse emission from one of the transmitters is received at each corner of the triangle after reflection from the ionosphere. The received signals are brought to the same locations and recorded continuously but independently upon the same chart. From these simultaneous observations of a given transmitted pulse as received at three sites, it becomes possible to determine motions in the reflecting medium.

The principles involved in the theory and analysis are straightforward. Consider an irregular medium containing innumerable scattering centers. A single reflected wave returned from such a medium is composed of a series of wavelets scattered by the centers. The medium is not homogeneous so that the diffraction pattern (as measured in a surface at a given distance from the medium) varies from point to point but is constant at any given point. If the irregularly spaced scattering centers are themselves in a random or directed motion, or if the medium itself is moving, the diffraction pattern, measured at a point in the given surface, will be time dependent.

The analogy with radio probings of ionospheric winds is obvious: the scattering medium is considered to be the ionosphere, and the surface is identified with the ground. When the ionosphere moves bodily in a direction along one side of the observing triangle, the displacement in the fading patterns received at the two receivers would be interpreted as the existence of a steady wind at the reflecting region. When, however, the ionospheric scattering centers are in random motion, the fading observed at the two receivers would also be random, and there would be no indication of a steady movement. Both the random motions and the regular movements are frequently found in the radio wave probing data. On some occasions, the random patterns are so marked and distinct that it is impossible to determine whether or not a wind is present. In this fashion a regular ionospheric motion may be totally masked when heavy turbulence causes violent movements of the scattering centers.

Several factors complicate the analyses of the observations. One arises because the ionosphere is a doubly refractive medium. Thus a reflected wave is not necessarily a single integrant, but may be composed of two overlapping magnetoionic components with opposite senses of polarization. If the phase between the two components varies because of a change in the equivalent paths of the two components, an apparent change in signal strength would be noted at the receiver. An improper analysis of the resulting fading pattern might attribute the change to a hypothetical wind. A second annoying factor in the analysis arises from the combination of fading patterns recorded; some indicate the regular winds and others merely a random pattern. Great care is necessary to decide which observations are suitable for accurate wind determinations. In this respect, use is made of the cross-correlation coefficient to facilitate reduction of the data.

A third factor which disturbs the analysis occurs when similar fading patterns are found at the three receiving sites. This similarity may arise not only because of an ionospheric drift in the reflection layer, but

because of a uniform motion existing in a *lower* ionic layer. Another difficulty in interpretation occurs when probing a uniform or homogeneous ionospheric region. In this case, all diffraction patterns received at the observing site are similar, whether the region is in motion or is stationary.

Observations to determine ionospheric winds should be extended by the implementation of additional ionospheric wind stations. These should be placed at various geographic regions, so located that information concerning regional circulatory systems at ionospheric altitudes may readily become available. Stations in both the northern and southern hemispheres, and in high and low latitudes are necessary in order to obtain information on the global wind system in the altitude range 70–400 km. The recordings at the different localities should preferably be made during given time intervals and on a coordinated basis. The results could then be compared with computed values of winds based upon somewhat simple models.

The experimental details are given in several papers dealing with this topic [39, 43, 139–143]. A basis for the theoretical study of probable movements to be expected in the ionosphere and mesosphere may be obtained from an extension of cellular wave theories [144–146].

4.1.2. Movement of Clouds. In this section, the term "cloud" will be employed to denote such phenomena as (a) noctilucent clouds; (b) luminous (auroral) clouds; (c) clouds in the airglow; and (d) sporadic E ionizations. The noctilucent clouds are well known. They usually are found during summer at the latitude range 45–60°N when the sun is 10–20° below the horizon. Their mean altitude is about 80 km. The movement of noctilucent clouds has been followed by means of theodolite observations from which the cloud velocity may be determined.

The brighter areas in the airglow also have been shown to exhibit large scale movements during the night. The motions have been determined from spectrophotometric observations of the airglow. Filters may be placed in front of the spectrophotometer to determine the motion of luminous clouds emitted from different atmospheric particles and presumably from different altitudes. The *Leuchtstreifen* observed in Europe and Africa may also be included in this category. The luminous auroral clouds have been seen at altitudes of about 100 km; altitudes of the night airglow clouds are not so well known and seem to range from 100–300 km. The movement of sporadic E areas has been determined from radio probing techniques utilizing a large number of probing stations. Sporadic E "clouds" are areas of high refractive or scattering ability at an altitude of about 110 km.

Observations on the various types of clouds are obviously limited to

those periods when they are present. The luminous clouds may be seen only during darkness while sporadic E may be tracked in daylight as well as darkness.

Sporadic E observational techniques are quite different from those for tracking luminous clouds. The extent, form, and intensity of luminous clouds may be readily ascertained because of the spectral range of the eye, filters, or photographic plates. The photometers which are sometimes employed are compact and rather small instruments. A similar method may be used for sporadic E observations by employing a rotating antenna and operating at a frequency of perhaps 50 mc/s. Sporadic E movements may be followed by such techniques as (a) the establishment of a network of radio stations over given geographic areas operating at frequencies of 50 mc/s, 100 mc/s etc., and (b) the use of back scatter measurements. Stations within the network are so located that if sporadic E appears anywhere over the geographic region under study, it may be discerned immediately and followed until it leaves the observational area. With back scatter observations, the presence of sporadic E over extensive areas may be determined from a single station.

Although this ideal E_s reporting system has never been established, it has been approximated over North America by the voluntary cooperation of large numbers of amateur radio operators. The typical operation of the cooperative system was as follows: Two amateur radio stations operating at a frequency of 50 mc/s were so located that with normal ionospheric conditions they could not effect radio contact. When sporadic E appeared approximately over the midpoint of the path between the two stations, however, these amateurs were able to establish radio contact. This contact could be maintained as long as some portion of the sporadic E area were near the midpoint of the path. When these conditions prevailed, the amateurs submitted reports from which the location of the E_s area could be determined in each case. When a large number of observers supplied sufficient reports, the location and extent of sporadic E could be determined and its motion and growth followed from hour to hour. Although the cooperation of the amateurs has been excellent and the data extremely useful, the method suffers from several deficiencies, such as the irregular geographic distribution of the amateurs, their operating habits, etc. These deficiencies introduced a bias into the data the removal of which would require an ideal network having optimally spaced stations operating at several frequencies on a *continuous* basis. It should be noted, however, that several back scatter stations would accomplish the same purpose.

The analysis of the sporadic E data is relatively simple. It seems highly probable that the data represent a true movement of "clouds"

carried along by the wind and not the propagation of some type of ionospheric disturbance. Also, the movement of the E_s clouds probably represents the air movement at the altitude concerned and not the motion of ions through the air. However, more study of this aspect is in order. It would be anticipated that the geomagnetic field develops polarizations within the sporadic E cloud which in turn might hinder the ion movement. Further investigations also are needed to determine the nature of sporadic E.

An examination of sporadic E movements at various geographic regions would be very desirable. Similar studies near the auroral zones are complicated by the existence of the aurora and the phenomenon of auroral interaction. Further, in high latitudes several types of sporadic E not found in lower latitudes are present. In general, the phenomenon loosely termed sporadic E in high latitudes is much more complicated than that found in middle latitudes.

Background literature on luminous clouds and sporadic E is given in the references [51–53, 147–151].

4.2. Tides

Under the influence of the gravitational forces of the sun and moon, periodic oscillations are imparted to the terrestrial fluids. Oceanic tides are well known and are usually influenced more by the moon, which is closer, than by the sun. Differences between the ocean and atmosphere, however, cause the reverse condition to apply to the latter where the solar semi-diurnal tide is about fifteen times greater than the lunear tide and about 100 times greater than expected. The cause for this remarkable occurrence lies in resonance; the free period of oscillation of the earth's atmosphere is but a few minutes less than twelve hours. Atmospheric tides manifest themselves as regular variations in the barometric pressure as observed at the earth's surface.

Additional factors which need not be considered with respect to the oceans apply to the atmosphere: (a) the atmospheric fluid is compressible and (b) thermal forces, arising from the heating effect of the sun, are very important. These effects make it possible for tidal motions in the atmosphere to have, at different altitudes, distinct speeds and in some instances, opposite directions. The main cause of tides in the atmosphere is the simultaneous occurrence of solar heating and solar gravitational attraction, both of which are applied simultaneously at a harmonic of the free resonant period of the earth's gaseous envelope. Both effects may be about of equal magnitude in the ionospheric and mesospheric regions. Lunar effects are solely gravitational in character.

Because the lunar and solar periods are different they may be isolated

and studied independently. One method of studying their effects has been through statistical analyses of ionospheric data obtained at many stations for long periods of years. However, analysis of these data is subject to so many pitfalls that, prior to such an investigation, a thorough understanding must be had of the procedure to be followed (particularly with regard to missing data) and the difficulties to be avoided. Using statistical analyses, attempts have been made to isolate lunar tides in auroral height data, ionospheric observations and signal intensity measurements. However, with respect to the isolation of solar tides in ionospheric data, it is extremely difficult to disentangle tidal components from ion production components, as both are caused by the sun at the same time. Present knowledge regarding the processes involved is almost too deficient to allow purely solar oscillations to be separated from the measurements (see Section 3.3).

A study of tides in the terrestrial atmosphere may be undertaken from three distinct aspects: (a) a theoretical treatment of tides to be expected on the basis of specific temperature-altitude relationships assumed for the atmosphere; (b) a statistical analysis of ionospheric, barometric or perhaps other geophysical data to determine the magnitude of possible tidal components; and (c) a study of the influence of tidal components found in (b) on the variations and fluctuations of the geomagnetic field. Some investigations have already been accomplished in these three categories but all still require additional effort.

The interpretation of tidal oscillations in the ionosphere in terms of geomagnetic variations is a large and comprehensive subject which, although of extreme importance, will not be treated here. There is some question regarding the levels in the ionosphere at which lunar magnetic variations originate, and to a lesser degree, the solar magnetic variations. An understanding of the physics of this subject becomes clearer after studying solar and lunar tides. Undoubtedly the worker engaged in statistically determining the magnitude of ionospheric tides, or in theoretically determining the modes of oscillation for given temperature distributions, will of his own accord seek to relate his results with the observed geomagnetic variations.

4.2.1. Theory of Tidal Oscillations. Tidal theory for incompressible fluids has been adequately treated in the literature. From this theory, the free period of oscillation of the ocean may be readily determined. Complicating influences arising because of the comparatively slow velocity of tidal waves and because of the angular rotation of the earth may also be included to generalize the simple theory. For compressible fluids the theory of tidal oscillations is generally much more involved than the simple theory for incompressible fluids. This condition is true

for the terrestrial atmosphere, although in two instances results similar to the simple compressible theory are found. These cases include the assumption of (a) an isothermal atmosphere with isothermal pressure variations, and (b) an adiabatic atmosphere having adiabatic pressure changes. With both conditions the propagation of long waves in the atmosphere is similar to that predicted by the simple theory for the ocean. No vertical standing waves exist; the wave motions are purely horizontal and no vertical energy flow takes place.

For every other atmospheric model, however, the vector curl components do not vanish; vortices appear and the particle velocity varies with altitude. In the general case, vertical circulation systems or cells develop. This condition is markedly different from that of the simple theory, where the motion is the same throughout the vertical extent of the fluid. With long atmospheric waves the motion of the air at one altitude may have a direction opposite to that at a second altitude; phase changes in the oscillations occur with a change in altitude, and at some heights nodes may exist. The problem of atmospheric oscillations when considering any probable temperature gradient which may be found in the atmosphere is complicated, and obviously three dimensional.

The problem is similar to that presented by a cavity resonator enclosed between two concentric spheres whose radii are about equal and much larger than their difference. In this resonator, the refractive index of the enclosed medium is anisotropic, being different in the radial and tangential directions. Although the boundary surfaces are rigid for the resonator, the corresponding surfaces (one of which may be the ground) in the atmosphere are defined by the temperature and the temperature gradient. The free mode of oscillation in the atmosphere depends upon several factors such as the distance between the surfaces, the length of the atmospheric wave involved, etc.

In the adiabatic and isothermal cases noted above, the horizontal energy flow decreases exponentially with altitude. With any other case, however, the horizontal energy flow (according to present limited theory) under certain conditions becomes somewhat uniform with altitude. The amplitude of the tidal oscillations would then increase with altitude, so that tidal air velocities and relative pressure changes in the ionosphere and mesosphere may be very much larger than those found at the ground.

The present theory is limited to pressure changes which are small compared to the total pressure at the altitude considered and to wind velocities small compared with the velocity of sound. The theory cannot be expected to apply when large pressure changes occur. However, to obtain greater accuracy the non-linear and the second order terms must be considered in the theory. From the large value of the kinematic

viscosity in the ionosphere and mesosphere and the effect of magnetic viscosity* in these regions, it would be expected that tidal oscillations would become damped with increasing altitude.

Additional information on theories of tidal oscillations with various temperature altitude lapse rates is given in the references [152–158].

4.2.2. Statistical Analysis of Atmospheric Tides. Evidence for the existence of tidal oscillations in the atmosphere was first obtained from analyses of the semi-diurnal variations in atmospheric pressure at the surface of the earth. Since that time many statistical analyses of geophysical phenomena have been made in order to determine the magnitude of the tidal oscillation at various sites over the surface of the globe and at various altitudes in the atmosphere. Thus, tidal oscillations have been sought in the height of the lower auroral surfaces, the altitude of appearance of meteors, virtual and actual heights of the various ionospheric layers, radio wave signal strength measurements, barometric pressures, and magnetic variations. A large number of analyses also have been made of the variations caused by the moon in such geophysical parameters as surface temperature, atmospheric pressure and terrestrial magnetism. However, additional determinations of tidal effects from geophysical data, obtained at all altitudes of the atmosphere, and particularly at both low and high altitudes, are nevertheless still very worthwhile.

The presence of tidal motions in the atmosphere has induced workers to search for similar effects in ionospheric data. Analyses of the hourly records (of the virtual reflection height, the altitude of maximum ionization for the layer concerned and the critical radio frequency) have shown that these values oscillate with the solar and lunar tides. The motions of the charged particles produced by tidal motions of the air are rather involved because of the complicating influence of the geomagnetic field. Because of this reason, it is not always simple to relate tidal variations (in data on ionospheric heights and critical frequencies) directly in terms of tidal air motions. Attempts also have been made to relate a change in the value of the recombination coefficient with tidal effects.

Before the investigations may be undertaken, ionospheric data for long periods of time are required. The tidal variations produced by the sun are usually much greater than those caused by the moon. However, these solar tides are also more difficult to isolate because the ionospheric variations have a strong twelve hour harmonic arising from the solar-produced ionizations. The lunar tidal effects in the ionosphere, although having a much smaller amplitude, are determined more readily because of the absence of these complicating effects.

* Magnetic viscosity is considered to arise when charged particles are carried by a moving fluid in the presence of a magnetic field.

The procedure employed for deducing the magnitude of the lunar tide in ionospheric data is the same as that used in isolating this effect in surface pressures and in magnetic variations. It is first necessary to remove the effect of the solar diurnal variation from the data, after which they are rearranged by months in terms of lunar time. The lunar monthly data are then subjected to harmonic analysis after which the mean value and its probable error are determined.

While the required statistical analyses may be undertaken without great difficulty, the snares which may invalidate the study must be carefully understood and avoided. Several statistical procedures have been developed which are particularly suitable for the analysis of tidal variations in geophysical data; these are fully described in the literature [159]. The methods consider such factors as the theory of errors, conservation, and quasi-persistence, which, if ignored, may lead to spurious and unreliable results. Many early conclusions are misleading because of the use of unsatisfactory statistical methods.

Lunar tidal variations have been found for the E, F1 and F2 ionic layers at selected locations. However, practically no analyses have been made for the D region. Additional studies are required for the E layer, especially with respect to critical frequency, before a full interpretation of the global results may be made.

Lunar variations have been examined fairly thoroughly for the F2 layer at a large number of stations. The effects seem well marked and, in addition, a type of "luni-solar" variation seems present. The latter effect is one wherein the phases and amplitudes of the lunar variations depend upon the solar time in a complex fashion. Although lunar tides have also been sought for the F1 region, no statistically dependent results are yet available in the analyses of critical frequency; probably the small amount of data utilized for the studies is responsible for the lack of dependability. It is interesting to note that the phase of the virtual height in the F1 layer does not, on the basis of present determinations, show any regular variation with latitude.

Although solar tidal variations have not been obtained satisfactorily, preliminary attempts for the E and F1 layers reveal certain semi-diurnal features which may be attributed to solar tidal effects. Preliminary studies of solar tidal variations in F2 layer ionospheric data seem somewhat successful in clarifying major anomalies of this layer. The theory indicates that the solar tide acting with the geomagnetic field could cause a vertical distortion of an ionized medium.

Additional investigations are needed to isolate lunar tides in the ionospheric layers at those stations where analyses have not yet been undertaken. More concentrated attention must also be given to the deter-

mination of solar tides in the ionosphere, considering the effects of divergence and the equation of continuity. The distortion of a regular ionic region by vertical ion motions must be considered. It is of course essential to undertake these studies with data obtained from stations having a wide latitude range. The determination of solar and lunar tides from magnetic data gathered at additional stations is also highly desirable.

Information on the isolation of tidal effects from ionospheric and geophysical data are given in the references [154, 160–163].

4.3. Diffusion

Diffusion may be defined as a process by which a fluid permeates its environment. More explicitly, diffusion is the average motion or drift of one species of atoms or molecules relative to a second species. The mathematical examination of diffusion problems is usually difficult. In the simplest case found, i.e., a binary mixture of gases in a closed system, the resulting equations may be complicated and cumbersome. The classical model of diffusion is solely concerned with the movements of two constituents (atomic or molecular particles) having known kinetic properties. The movements occur in a closed system where the physical state is explicitly defined and where the total molecular concentration of each constituent gas is constant.

Obviously, diffusion in the atmosphere is far more complex than the elementary example given above. In addition to a large variety of atoms and molecules, the high atmosphere contains cosmic dust and ionized matter, each of which further entangles the treatment. The atmosphere does not constitute a closed system at any altitude. Its physical condition is uncertain in many respects, being under the influence of time-dependent, large- and small-scale circulation patterns which cannot be explicitly defined. For certain species of atomic particles, sources (arising from radioactivity, dissociation, ionization, or detachment) or sinks (caused by association, recombination, or attachment) may be present, the intensity of each varying with time, altitude, latitude, and longitude. Thus the total number density may be constant for some constituents but may vary appreciably in time or space for others. The sources and sinks may be internal or external to given atmospheric shells; e.g., atomic oxygen is produced within the ionosphere; but helium is produced by the lithosphere and lost to outer space, both processes occurring external to the atmosphere as a whole.

In general, therefore, atmospheric diffusion is a three dimensional problem having diurnal, seasonal, and spatial variations. Because the atmosphere is turbulent to great heights, a coefficient of eddy diffusion

may be defined. The coefficient of molecular diffusion is a function of density and temperature, and is a monotonically increasing function of altitude. Contrarily, the coefficient of eddy diffusion, which depends upon the horizontal and vertical gradients of pressure and temperature, may vary from point to point.

The general theory of diffusion in binary mixtures considers four types of diffusion: thermal, pressure, ordinary, and forced. Of these four types one or more may be peculiarly effective for a given atmospheric constituent. Thus, with regard to ions, forced diffusion may be the major diffusive agent in the high atmosphere. For the remaining molecular or atomic particles, however, forced diffusion is negligible. Each type of diffusion must be considered in the atmosphere. Although the thermal diffusion component is inconsequential in the troposphere and stratosphere, it may be important in the ionosphere and mesosphere where the temperature may change with altitude by a factor of about five to ten.

An additional diffusive agency, not defined by kinetic theory but of importance in atmospheric problems, is that of dust particle diffusion. In this case, the coefficient of dust particle diffusion is required when the particle size is very much greater than that of the molecules. In a calm atmosphere, where winds and turbulence are absent and where gravity is the only force present, the dust will fall. When movements are present, as in the actual atmosphere, the downward diffusion of dust may be markedly slowed at some levels giving rise to dust strata. As winds and turbulence are related to the vertical temperature distribution, the dust strata may be expected to be found at temperature discontinuities. The rate of descent of dust in planetary atmospheres has not been fully studied and warrants much more investigation. The usual formulae which might be employed, such as Stoke's law, fail because the assumptions involved in their derivation do not apply to dust. Dust particles are not spherical. Further, at certain heights the size of the particle equals the mean free path. New treatments on the diffusion of dust obviously are necessary.

With regard to diffusion in the high atmosphere, few observations have been made, and these entirely of visual or ionized meteor trains. An interpretation of the results is not clear, however, inasmuch as it is not known whether the observed distortions arise mainly from winds, wind shear, or diffusion. These observations provide information in the altitude range about 50–150 km; no data are available for higher altitudes. Because of the paucity of observational data and the difficulties in their interpretation, greater emphasis must be devoted to theoretical analyses. However, the difficulties inherent in a mathematical treatment of time varying diffusion in three dimensions are so great that

simplifying assumptions frequently must be made; these assumptions in the past have so oversimplified the model that the results could scarcely be applied to the atmosphere.

Basic treatments of diffusion, of a more thorough and critical fashion than hitherto utilized, are necessary before appreciable progress may be made. Future investigations must not only consider more realistic atmospheric models, but must extend present theory. Past studies mainly have considered only vertical diffusion, implicitly assuming either that lateral diffusion did not exist or that it could be absorbed into the eddy diffusion term of local disturbances. However, the effects of lateral diffusion should be considered, at least from the surface to the ionospheric layers. From the chemosphere upwards, the study of diffusion is complicated by photochemical, tidal, and thermal processes. The relative importance of tides (with respect to diffusion) in dispersing ionized or neutral constituents must be further examined. The entire problem of forced diffusion and diffusive equilibrium should also be investigated. Although the latter condition may be considered as a static property of the atmosphere, the treatment of diffusive equilibrium is so closely allied with the analysis of diffusion that this topic is best placed in this section.

4.3.1. Diffusion in Magnetic and Electric Fields. The greatest component of forced diffusion in the high atmosphere undoubtedly arises from the effect of the geomagnetic field upon the diffusion of charged particles. Although electric fields may be present under certain conditions, their intensity is believed to be small. The magnetic field has its origin in the solid earth. However, electron currents at different altitudes and in the possible ring current (at a distance of about five earth radii from the earth) produce time-dependent variations in the magnetic field at different altitudes and different locations. At the extremely low number densities found in the high atmosphere, the geomagnetic field may exert a profound effect upon the motions of ions and electrons. At the densities found at these altitudes, the ions and electrons gyrate around the magnetic lines of force between the times of collisions with molecules. Therefore, the transverse diffusion across the magnetic field is greatly impeded, while diffusion along the direction of the lines of force is not affected.

While many of the difficulties presented in the section on Diffusion also apply to forced diffusion, several specific problems for study may be mentioned. One such problem concerns the diffusion of an electron or ion in the upper ionosphere or the mesosphere immediately after ionization. In this case, it may be assumed that the electron is ejected with sufficient energy so that the electrostatic field between it and the newly

formed ion is negligible. Complicating factors such as space charge distributions may be considered as refinements to the problem.

Another effect which should be carefully examined is the diffusion of ionized meteor trains. In this case the ionization trail formed by the incoming meteor at essentially a point source begins to diffuse immediately. Because of the meteor's high velocity, the ionization trail is essentially a cylinder of ionization. The diffusion of electrons from this trail normal to the geomagnetic field must proceed more slowly than along the field. Thus the distortion of the trail caused by diffusion is partially controlled by the angle between the ionized trail and the magnetic field. The length of the meteor trail in the atmosphere extends through a region where the mean free path changes by two or three orders of magnitude; diffusion at the higher altitude therefore would better indicate the effect of forced diffusion.

It might be mentioned that rapid distortions occur in the meteor train under the combined action of diffusion, turbulence, and wind shear. Probably the most effective dispersing mechanism causing the rapid dilution of the ionized trail is wind shear. However, the relative importance of diffusion in diluting the ionized trail should be determined.

Diffusion of various molecular and atomic ions, and electrons from the higher ionized layers and from the mesosphere, must also be investigated.

Additional information on forced diffusion with applications to the ionosphere may be found in the references [164–173].

4.3.2. Diffusive Equilibrium. In the absence of turbulence, winds and other disturbing effects, a uniformly mixed planetary atmosphere would eventually stratify and attain a state of diffusive equilibrium. This condition, also termed a "Dalton atmosphere" or "gravitational equilibrium," results when the component gases of the atmosphere settle out according to their respective molecular weights. The heaviest gases then would be found in the lowest strata and the lightest gases in the outermost strata. However, in the terrestrial atmosphere, winds and mixing occur through a transition zone. Within this zone, winds and eddy diffusion gradually become weaker until at the top of the zone no motion takes place. Below this zone the atmosphere is found as a turbulent mixture whereas above this zone diffusive equilibrium occurs. In the transition zone, the atmosphere gradually passes from the former state to the latter. The settling out of the atmospheric gases according to their relative molecular weights would begin above the transition zone. For constituents not having sources or sinks, the separation would depend upon the temperature, pressure and turbulence, and their variation with altitude.

Several computations have already been made on the altitude of the transition zone in the terrestrial atmosphere. However, these investiga-

tions have neglected internal sources and sinks arising from photochemical or collisional processes; have adopted arbitrary pressure or temperature distributions; and have overlooked the effect of eddy diffusion. These omissions so oversimplify the problem that the results are greatly impaired. It might nevertheless be mentioned that when the above factors are considered, the problem of diffusive equilibrium becomes very complicated. The rate of production or destruction of the appropriate constituents must be known as a function of altitude. These factors depend in an involved fashion upon solar radiation, the state of the atmosphere, the relative effects of molecular and eddy diffusion, the magnitude of the general circulation, etc. The problem becomes even more involved when the probable diurnal and seasonal differences of these quantities at different altitudes and latitudes are considered. In fully treating diffusive equilibrium in the atmosphere, many factors have been mentioned, but the most important may easily be winds and eddy diffusion.

Further information on this topic is given in the references [174–177].

5. Conclusions

It should be stressed that the topics discussed in the previous pages by no means exhaust the list of unsolved problems related to the high atmosphere. Innumerable other problems exist. All stand as a challenge to the researcher and as an inspiration to the student of physics and geophysics. The problems are not only directly concerned with the atmosphere of the earth but with planetary atmospheres in general. Many of the results may undoubtedly be applied to the gaseous envelopes of Mars and Venus. Of the many problems not previously discussed, a few of the more conspicuous may be briefly mentioned.

Probably the greatest obstacle preventing a better understanding of the processes and mechanisms occurring within the atmosphere is the lack of sufficient information on the solar emission spectrum. Planetary atmospheres are essentially heat engines driven by radiation from the parent stellar body; if the energy input into the working fluid is unknown, how can the mechanisms occurring within it be determined or the heat budget be studied? Although the sun is usually considered to radiate as a black body at 6000°K, this temperature is taken in desperation and only because a more accurate value has not been provided. Energy emissions from the sun in the ultraviolet are believed to be much more intense than those emanating from a black body at 6000°K. In the solar corona, atoms of many metals are stripped of their outer electrons and ionized to a degree not possible of attainment in present terrestrial laboratories. Extraordinarily high velocities have been found in the

eruptive prominences and polar-zone spicules. These conditions of extreme turbulence and strong radiation occurring in the solar surface and chromosphere profoundly affect the terrestrial atmosphere. However, the radiations emitted during such periods of solar activity are not completely known. It is much to be hoped that the major objective of the rocket experiments will be that of obtaining complete data on the emission spectrum of the quiet sun and the active sun. Even after this information becomes available, much remains to be done on clarifying effects produced in the terrestrial atmosphere. How does the active sun induce electric currents in the high atmosphere, for example, and how do these currents act to produce a fluctuating magnetic field that disturbs the quiet component? The precise solar source of the radiations causing any of the ionic layers in the terrestrial atmosphere is unknown.

The zodiacal light presents another interesting aspect for study. Although several theories have been presented to explain its occurrence, none are completely acceptable. Presumably it arises from dust clouds far from the earth's surface either within the atmosphere or in interplanetary space. Although the location of the dust cloud is controversial, a dust layer at the mesopause may account for the observed scattering. Additional observations are required to determine the spectrum, the degree of polarization and the scattering of this light. More accurate measurements of seasonal and other intensity variations of zodiacal light also would be helpful in determining the possible influence of extraterrestrial material on abnormal ionization in the ionosphere or mesosphere.

Very little is known regarding that curious phenomenon known as sporadic E. The mechanisms regarding its genesis, movement and apparent dissipation are completely unknown. The marked irregularity in the occurrence of sporadic E presumably arises from the variability of the causative agent. Sporadic E exhibits a pronounced seasonal maximum of occurrence in summer and a minor maximum in winter. Vertical incidence measurements made at sites as close together as 100 km are often uncorrelated, indicating that sporadic E may be found in the form of clouds similar to the fair weather cumulus of the troposphere. Oblique incidence reporting networks or back scatter stations delineate the extent of these clouds, and allow particular cloud systems to be tracked over great geographical distances. The phenomenon is different from the remaining ionic layers, and shows no consistent relationship with solar zenith angle or sunspot activity. Except in isolated instances, it cannot be attributable to meteors. A study of E_s is complicated by the fact that the term "sporadic E" includes a variety of abnormal ionizations near 110 km. The middle and low latitude variety, which has been discussed

above, may be termed true sporadic E. At high latitudes and particularly near the auroral zone, a variety of abnormal E ionizations are present; these have also been termed sporadic E. A preliminary classification has already indicated that auroral sporadic E may be differentiated into at least five distinct types. In addition the phenomenon of "auroral interaction" (i.e., radio wave reflections from the aurora itself at frequencies of 50–150 mc/s) is also loosely included within the term "sporadic E." A rather good correlation exists between magnetic activity and auroral sporadic E. In any event, the fundamental problems regarding the origin of sporadic E, the agencies responsible for its movement, production and dissipation, its marked seasonal appearance, etc., are not completely solved.

The method whereby the regular ionospheric layers are formed is by no means clear. A satisfactory explanation of the stratification of the several ionic layers has not been presented. The specific atmospheric constituents ionized to form each of the layers are not known with assurance. Altitudes wherein the ionization of such constituents as atomic nitrogen, atomic sodium, nitric oxide, etc. should be considered as a possible factor in forming the layers are purely speculative. Several studies have attempted to show that the absorption of solar radiation in specific spectral ranges by certain atmospheric constituents may account for the observed stratifications of the ionic layers. The results, however, are not conclusive. An examination of the complications introduced by the process of attachment, which is very important for the D and E layers, must be more carefully considered. In general, the ionic layers have been attributed to a variety of possible atmospheric constituents, to coronal radiations of high energy, to altitude-dependent recombination and attachment coefficients, etc. With regard to observations, additional eclipse measurements may determine whether the F1 and F2 layers arise from the same cause but become differentiated through the action of different recombination mechanisms.

Anomalies found in the F2 region have puzzled ionosphericists for many years. This region exhibits a number of pecularities not found in the lower, more regular ionic layers. The altitude of maximum ionization is higher in summer than winter, but the electron density is greater in winter. Interpretations of solar eclipse measurements made on the F2 layer are far more difficult than for the lower layers. Further, the solar cycle variations in the electron concentration during winter exceed the corresponding variations found in summer.

A marked asymmetry is found in the electron concentration of the F2 layer for conditions of symmetrical solar illumination. Ion densities at equinoctial noon and at the same latitudes but different longitudes

may be very different. However, when the results are reconsidered with respect to geomagnetic coordinates, the anomalies disappear, thereby suggesting the existence of a geomagnetic control over the F2 layer. During equinoctial noon, a belt of low ionization density is apparently found over the magnetic equator; this condition is associated with a marked bifurcation of the F2 layer. The cause for the geomagnetic control of the F2 region is not known but may possibly be linked with the greater mean free path found at altitudes concerned.

Sudden ionospheric disturbances (*SID*) directly interfere with daily activities. During these periods radio blackouts occur over the sunlit hemisphere of the earth. The ultraviolet radiation from the flare apparently increases the electron concentration in the D ionic layer. As this layer absorbs energy from penetrating radio waves, this increase in ionization is often severe enough to completely absorb incident radio waves.

It has been known for many years that solar activity was responsible for a *SID* and for magnetic storms; however, a precise correlation between these two phenomena still has not been established. Recent studies show that wherever a flare is about three times as intense as the background, a sudden ionospheric disturbance will occur regardless of the total area of the disturbance or of the location of the flare on the solar disc. While a correlation may thus be found for the very bright flares, the fainter ones present problems. For example, an instance is known where two flares occurred a few hours apart. The second flare was more intense than the first, covered an area about three times greater and lasted for a longer period of time. Nevertheless, the first flare was accompanied by a *SID*, but the second flare was not. Undoubtedly, flares are merely indications of a more deep-seated solar phenomenon that produces a *SID*, but it may also be possible that physical differences occur between particular types of flares. To resolve the problem more detailed information such as the accurate measurement of the variation of flare intensity with time together with observations of the flare spectra is needed. Some preliminary studies indicate that the time variation in the light intensity for individual flares may be quite different.

Much more additional research remains to be undertaken before the relationships between the magnitudes of the flare, the amount and duration of the ionospheric absorption and the associated geomagnetic effect may be found. Some flares produce intense ionospheric absorption with only a slight magnetic effect but others produce marked geomagnetic effects with only moderate ionospheric absorption. The cause for the difference is yet to be determined. Considerable work is also required to determine the nature of those solar flares which produce changes and

heating in the ozone layer. Movements produced in this layer by the heating also warrant additional study.

The entire problem of magnetohydrodynamics has many applications in geophysics and astrophysics. The problem deals with a moving compressible ionized medium, either dense or diffuse, subject to a magnetic field. A diffuse gas is defined as one where the time between collisions is much larger than the precession time. With a diffuse medium, a fundamental question arises: how are the charged particles accelerated? Answers to this question have a direct bearing on such events as the aurora, cosmic rays, cosmic noise and auroral noise. A definition of the conductivity of a diffuse gas is not yet clear. It has been found that, with a beam of ionized particles, diffusion transverse to a strong magnetic field is much greater than expected, probably because turbulence and plasma fluctuations provide additional scattering centers. Similar considerations may apply to the particles which bombard the terrestrial atmosphere and produce the aurora. In considering the effect of magnetic fields on conductivity in the dynamo theory (which has been proposed to account for the diurnal magnetic variations), recent studies have shown that the transverse conductivity is inadequate to account for the current; additional studies are necessary. In connection with the geomagnetic field, it might be mentioned that there are strong differences of opinion regarding its origin. Important differences also exist in the theories regarding the formation of the aurora, and no theory yet proposed is completely acceptable.

ACKNOWLEDGMENTS

The author is pleased to acknowledge the many helpful comments and criticisms of Drs. R. Craig, R. J. Donaldson, H. E. Landsberg, H. Lettau, and R. Penndorf of the Geophysics Research Division, and J. Kaplan of the University of California. The author also wishes to express his thanks to Mrs. N. C. Gerson for her invaluable assistance during the preparation of this paper.

GENERAL REFERENCES

A. Chapman, S., and Bartels, J. (1940). Geomagnetism. The Clarendon Press, Oxford.

B. Fleming, J. A. (1939). Terrestrial Magnetism and Electricity. McGraw-Hill, New York.

C. Gassiot Committee. (1948). The Emission Spectra of the Night Sky and Aurorae. The Physical Society, London.

D. Gerson, N. C. (1951). Proceedings of the colloquium on mesospheric physics. *Geophys. Res. Pap. USAF No.* **8.**

E. Kuiper, G. P. (1952). The Atmospheres of the Earth and Planets 2nd ed., Univ. Chicago Press, Chicago.

F. Mitra, S. K. (1948). The Upper Atmosphere. The Royal Asiatic Society of Bengal, Calcutta.

G. Underhill, B. B., and Donaldson, R. J., Jr. (1950). Proceedings of the conference on ionospheric research (June 1949). *Geophys. Res. Pap. USAF No.* **7.**

References

1. Paneth, F. A. (1937). The chemical composition of the atmosphere. *Quart. J. Roy. Meteorol. Soc.* **63,** 433.
2. Glückauf, E. (1946). A micro-analysis of the helium and neon contents of air. *Proc. Roy. Soc. (London)* **A185,** 98.
3. Goldberg, L. (1950). Recent advances in infrared solar spectroscopy. *Rept. Prog. Phys.* **13,** 24.
4. Shaw, J. H., and Bates, D. R. (1951). Private communication.
5. Harteck, P., and Suess, H. E. (1949). Der Deuteriumgehalt des freien Wasserstoffs in der Erdatmosphäre. *Naturwissenschaften* **36,** 218.
6. McMath, R. R., Pierce, A. K., Mohler, O. C., Goldberg, L., and Donovan, R. A. (1950). Nitrous oxide bands in the solar spectrum. *Phys. Rev.* **78,** 65.
7. Slobud, R. L., and Krogh, M. E. (1950). Nitrous oxide as a constituent of the atmosphere. *J. Am. Chem. Soc.* **72,** 1175.
8. Cauer, H. (1948). Meteorologie und Physik der freien Atmosphäre, ed. R. Mügge. *FIAT Review of German Science* **19,** 277.
9. Shaw, J. H. (1951). Identification of bands of gases in the infrared solar spectrum. Symposium on Molecular Spectroscopy. Ohio State Univ., Columbus, Ohio.
10. Shaw, J. H., and Claassen, H. H. (1951). The absence of atmospheric ethylene. *Phys. Rev.* **81,** 462.
11. Babcock, H. D., and Herzberg, L. (1948). Fine structure of the red system of atmospheric oxygen bands. *Astrophys. J.* **108,** 167.
12. Chapman, R. M., and Shaw, J. H. (1950). Fine structure of HDO near 3.7μ in the solar spectrum. *Phys. Rev.* **78,** 71.
13. Gebbie, H. A., Harding, W. R., Hilsum, C., and Roberts, V. (1949). Atmospheric HDO. *Phys. Rev.* **76,** 1534.
14. Lagemann, R. T., Nielsen, A. H., and Dickey, F. P. (1947). The infra-red spectrum and molecular constants of $C^{12}O^{16}$ and $C^{13}O^{16}$. *Phys. Rev.* **72,** 284.
15. McQueen, J. H. (1950). Isotopic separation due to settling in the atmosphere. *Phys. Rev.* **80,** 100.
16. Glückauf, E. (1950). Composition of Atmospheric Air. Atomic Energy Research Establishment, Harwell, England.
17. Chackett, K. F., Paneth, F. A., Reasbeck, P., and Wiborg, B. S. (1951). Variations in the chemical composition of stratospheric air. *Nature* **168,** 358.
18. Adel, A. (1951). Atmospheric nitrous oxide and the nitrogen cycle. *Science* **113,** 624.
19. Rakshit, H. (1947). Distribution of molecular and atomic oxygen in the upper atmosphere. *Ind. J. Phys.* **21,** 57.
20. Penndorf, R. (1949). The vertical distribution of atomic oxygen in the upper atmosphere. *J. Geophys. Res.* **54,** 7.
21. Moses, H. E., and Wu, T. Y. (1950). The distribution of atomic and molecular oxygen in the upper atmosphere. *Phys. Rev.* **78,** 333.
22. Bradt, H. L., and Peters, B. (1950). The heavy nuclei of primary cosmic radiation. *Phys. Rev.* **77,** 54.
23. Vassy, A. and E. (1950). Recherches sur l'ozone atmosphérique et la température de la stratosphère en Laponie Suédoise. *Tellus* **2,** 69.
24. Wexler, H. (1950). Annual and diurnal temperature variations in the upper atmosphere. *Tellus* **2,** 262.
25. Vassy, A. and E. (1939). Temperature of the stratosphere in high latitudes. *Nature* **144,** 284.

26. Gowan, E. H. (1947). Ozonosphere temperature under radiative equilibrium. *Proc. Roy. Soc. (London)* **A190,** 219.
27. Sheppard, P. A. (1949). The exploration of the upper atmosphere. *Science Progress* **37,** 488.
28. Gerson, N. C. (1950). The colloquium on mesospheric physics. *J. Franklin Inst.* **250,** 472.
29. Oliver, N. J. Proceedings of the colloquium on auroral physics. *Geophys. Res. Pap. USAF* (in press).
30. Meinel, A. B. (1951). The spectrum of the airglow and the aurora. *Rept. Prog. Phys.* **14,** 124.
31. Forsyth, P. A., Petrie, W., and Currie, B. W. (1950). On the origin of ten centimeter radiation from the polar aurora. *Canadian J. Phys.* **A28,** 324.
32. Brasefield, C. J. (1950). Winds and temperatures in the lower stratosphere. *J. Meteorol.* **7,** 66.
33. Johnson, N. K. (1946). Wind measurements at 30 km. *Nature* **157,** 24.
34. Ockenden, C. V. (1939). High altitude pilot balloon ascents at Habbaniya, Iraq. *Quart. J. Roy. Meteorol. Soc.* **65,** 551.
35. Scrase, F. J. (1949). Wind and temperature measurements up to 30 km. *Nature* **164,** 572.
36. Crary, A. P. (1950). Investigation of stratosphere winds and temperatures from acoustical propagation studies. *Geophys. Res. Pap. USAF No.* **5**; and *J. Meteorol.* **7,** 233.
37. Crary, A. P. (1951). Unpublished data (private communication).
38. Murgatroyd, R. J. (1951). Anomalous sound reception experiments: April 1944–April 1945, London. *Meteorol. Res. Comm. Pap. No.* **611.**
39. Beynon, W. J. G. (1948). Evidence of horizontal motion in region F_2 ionization. *Nature* **162,** 886.
40. Eyfrig, R. (1940). Über Echomessungen bei Fernübertragung und ihre Beziehung zu Zenitreflexionen. *Hochfrequenztech. u Elektroakust.* **56,** 161.
41. Gerson, N. C. (1951). Proceedings of the colloquium on mesospheric physics. *Geophys. Res. Pap. USAF No.* **8,** 78.
42. Krautkrämer, K. (1943). Über Wanderungserscheinungen von Feldstärkeschwankungen bei Ionosphärenechos. *Deut. Luftfahrtforsch.* Forsch Nr. 1761.
43. Mitra, S. N. (1949). A radio method of measuring winds in the ionosphere. *Proc. Inst. Elec. Engrs. Pt. III,* **96,** 441.
44. Munro, G. H. (1949). Short-period variations in the ionosphere. *Nature* **163,** 812.
45. Waynick, A. H. (1951). Basic ionospheric research. *Penn. State Coll. Quart. Rept. No.* **10,** 13.
46. Briggs, B. H., and Phillips, G. J. (1952). Proceedings of the conference on ionospheric physics. Ed. by N. C. Gerson and R. J. Donaldson, Jr. *Geophys. Res. Pap. USAF No.* **11,** 199.
47. Fennell, P. (1944). Winds in the ionosphere indicated by radio reflecting "clouds" of high ionic density. *Bull. Am. Meteorol. Soc.* **25,** 371.
48. Ferrell, O. P. (1947). Note on the sporadic-E layer. *Proc. Inst. Radio. Engrs.* **35,** 493.
49. Ferrell, O. P. (1948). Upper-atmosphere circulation as indicated by drifting and dissipation of intense sporadic-E clouds. *Proc. Inst. Radio Engrs.* **36,** 879.
50. Gerson, N. C. (1950). Large-scale sporadic movements of the E-layer of the ionosphere. *Nature* **166,** 316.
51. Gerson, N. C. (1951). Abnormal E region ionization. *Canadian J. Phys.* **29,** 251.

52. Meek, J. H. (1949). Sporadic ionization at high latitudes. *J. Geophys. Res.* **54**, 284.
53. Katz, L., and Gerson, N. C. (1951). Proceedings of the conference on ionosphere physics. *Geophys. Res. Pap. USAF No.* **12**.
54. Hoffmeister, C. (1946). Atmospheric currents at a height of 120 km. *Z. Meteorol.* **1**, 33.
55. Schilling, G. F. (1950). Survey of Data and Theoretical Analysis of the Upper Atmosphere. Part I. Ed. J. Kaplan, Univ. California, Los Angeles, p. 15.
56. Manning, L. A., Villard, O. G., Jr., and Peterson, A. M. (1950). Meteoric echo study of upper atmospheric winds. *Proc. Inst. Radio Engrs.* **38**, 877.
57. Olivier, C. P. (1942). Long enduring meteor trains. *Proc. Am. Phil. Soc.* **85**, 93.
58. Olivier, C. P. (1947). Long enduring meteor trains. *Proc. Am. Phil. Soc.* **91**, 315.
59. Whipple, F. L. (1943). Meteors and the earth's upper atmosphere. *Revs. Modern Phys.* **15**, 246.
60. Lettau, H. (1948). Maximalwerte der Windgeschwindigkeit als Function der Erdrotation. *Meteorol. Runds.* **1**, 451.
61. Bates, D. R. (1951). The temperature of the upper atmosphere. *Proc. Phys. Soc. (London)* **B64**, 805.
62. Elsasser, W. M. (1942). Heat transfer by infrared radiation in the atmosphere. *Harvard Meteorol Stud. No.* **6**.
63. Gerson, N. C. (1951). A critical survey of ionospheric temperatures. *Rept. Prog. Phys.* **14**, 316.
64. Godfrey, G. H., and Price, W. L. (1937). Thermal radiation and absorption in the upper atmosphere. *Proc. Roy. Soc. (London)* **A163**, 228.
65. Branscomb, L. M. (1950). Anomalous molecular rotation and the temperature of the upper atmosphere. *Phys. Rev.* **79**, 619.
66. Oldenberg, O. (1934). On abnormal rotation of molecules. *Phys. Rev.* **46**, 210.
67. Schüler, H., and Gollnow, H. (1937). Über die Verteilung der Rotationszustände bei einem Elementarprozess der Molekülbildung (keine Boltzmann-Verteilung) und die Änderung der relativen Übergangswahrscheinlichkeit. *Z. Phys.* **108**, 714.
68. Schüler, H., and Gollnow, H. (1938). Berichtigung zu der Arbeit: Über die Verteilung der Rotationszustände bei einem Elementarprozess der Molekülbildung (keine Boltzmann-Verteilung) und die Änderung der relativen Übergangswahrschenilichkeit. *Z. Phys.* **109**, 432.
69. Schüler, H., and Gollnow, H. (1939). Über Molekülbildungprozesse mit und ohne Boltzmann-Verteilung und über Umwandlung von Translations-in Rotationsenergie. *Z. Phys.* **111**, 484.
70. Wakeshima, H. (1942). Über die abnormale Rotation des OH Moleküls (II). *Proc. Phys.-Math. Soc. Japan* **24**, 367.
71. Vegard, L., and Tönsberg, E. (1944). Results of auroral spectrograms obtained at Tromso observatory during the winters 1941–42 and 1942–43. *Geofys. Publik.* **16**, No. 2.
72. Babcock, H. D. (1923). A study of the green auroral line by the interference method. *Astrophys. J.* **57**, 209.
73. Dufay, J., Cabannes, J., and Gauzit, J. (1942). L'analyse interferentielle des raies brillantes due ciel nocturne. *Astronomie* **56**, 159.
74. Moore, C. E. (1949). Atomic energy levels. *Natl. Bur. Standards U. S. Circ.* **467**.

75. Pearse, R. W. B., and Gaydon, A. G. (1941). The Identification of Molecular Spectra. Chapman and Hall, London.
76. Herzberg, G. (1950). Spectra of Diatomic Molecules. 2nd ed., D. Van Nostrand, New York.
77. Price, W. C. (1943). Absorption spectra and absorption coefficients of atmospheric gases. *Rept. Prog. Phys.* **9**, 10.
78. Bacher, R. F., and Goudsmit, S. (1932). Atomic Energy States as Derived from the Analyses of Optical Spectra, McGraw-Hill, New York.
79. Weizel, W. (1931). Bandenspektren. Akademische Verlagsgesellschaft, Leipzig.
80. Cowling, T. G. (1943). The absorption of water vapor in the far infrared. *Rept. Prog. Phys.* **9**, 29.
81. Dennison, D. M. (1940). The infrared spectra of polyatomic molecules. Part II. *Revs. Modern Phys.* **12**, 175.
82. Gaydon, A. G. (1947). Dissociation Energies and Spectra of Diatomic Molecules. Chapman and Hall, London.
83. Moore, C. E. (1945). A Multiplet Table of Astrophysical Interest. Princeton Univ. Obs., Princeton, New Jersey.
84. Babcock, H. D., and Moore, C. E. (1947). The Solar Spectrum, λ6600 to λ13495. *Carnegie Inst. Wash. Publ.* **579.**
85. Minnaert, M., Mulders, G. F. W., and Houtgast, J. (1940). A Photometric Atlas of the Solar Spectrum. Schnabel, Kampert and Helm, Amsterdam.
86. Shaw, J. H., Oxholm, M. L., and Claassen, H. H. (1951). The solar spectrum from 7–13 microns. *Ohio State Univ. Res. Found. Rept.* **IA-4.**
87. Barbier, D. (1949). Comparaison des formules donnant les corrections d'extinction dans les observations due ciel nocturne. *Ann. Geophys.* **5**, 83.
88. Chandrasekhar, S. (1944). On the radiative equilibrium of a stellar atmosphere. *Astrophys. J.* **100**, 76.
89. Chandrasekhar, S. (1950). Radiative Transfer. The Clarendon Press, Oxford, England.
90. Roach, F. E., and Barbier, D. (1950). The height of emission in the upper atmosphere. *Transact. Am. Geophys. Un.* **31**, 7.
91. Bhar, J. N. (1938). Stratification of ionosphere and origin of E1 layer. *Ind. J. Phys.* **12**, 363.
92. Dirac, P. A. M. (1924). Dissociation under a temperature gradient. *Proc. Cambridge Phil. Soc.* **22**, 132.
93. Pannekoek, A. (1926). Ionization equilibrium in stellar atmospheres and in the earth's atmosphere. *Proc. Koninkl. Akad. Wetenschap. Amsterdam* **29**, 1165.
94. Rosseland, S. (1936). Theoretical Astrophysics. The Clarendon Press, Oxford, England.
95. Wildt, R. (1936). Equilibrium of stellar atmospheres under a temperature gradient. *Astrophys. J.* **83**, 202.
96. Woolley, R. v. d. R. (1947). Radiative equilibrium in the ionosphere. *Proc. Roy. Soc.* (*London*) **A189**, 218.
97. Bailey, V. A., and Martyn, D. F. (1934). The influence of electric waves on the ionosphere. *Phil. Mag.* **18**, 369.
98. Cutolo, M. (1947). Gyro-interaction of radio waves obtained by the pulse method. *Nature* **160**, 834.
99. Farmer, F. T., and Ratcliffe, J. A. (1935). Measurements of the absorption of wireless waves in the ionosphere. *Proc. Roy. Soc.* (*London*) **A151**, 370.

100. Ginsburg, V. L. (1944). On the absorption of radio waves and the number of collisions in the ionosphere. *J. Phys. (USSR)* **8,** 253.
101. Huxley, L. G. H. (1950). Ionospheric cross-modulation at oblique incidence. *Proc. Roy. Soc. (London)* **A2C0,** 486.
102. Huxley, L. G. H., and Ratcliffe, J. A. (1949). A survey of ionospheric cross-modulation. *Proc. Inst. Elec. Engrs. Pt III* **96,** 433.
103. Perkeris, C. L. (1940). The vertical distribution of ionization in the upper atmosphere. *Terr. Magn.* **45,** 205.
104. Ratcliffe, J. A., and Shaw, I. J. (1948). A study of the interaction of radio waves. *Proc. Roy. Soc. (London)* **A193,** 311.
105. Bohr, N. (1948). The penetration of atomic particles through matter. *Kgl. Danske Vidensk. Selsk. Mat.-fys. Medd.* **18,** No. 8.
106. Fisk, J. B. (1936). Theory of the scattering of slow electrons by diatomic molecules. *Phys. Rev.* **49,** 167.
107. Hartree, D. R., Hartree, W., and Swirles, B. (1939). Properties of neutral and ionized atomic oxygen and their influence in the upper atmosphere. *Trans. Roy. Soc. (London)* **A238,** 229.
108. Howard, R. R., and Smith, W. V. (1950). Microwave collision diameters. I. Experimental. *Phys. Rev.* **79,** 128.
109. Hulthén, L. (1944). Variational problem for the continuous spectrum of a Schrödinger equation. *Kgl. Fysiograf. Sallskap. Lund Förh.* **14,** No. 21.
110. Massey, H. S. W. (1949). Collisions between atoms and molecules at ordinary temperatures. *Rept. Prog. Phys.* **12,** 248.
111. Massey, H. S. W., and Bates, D. R. (1943). Self consistent fields including exchange and superposition of configurations including some results for oxygen. *Rept. Prog. Phys.* **9,** 62.
112. Massey, H. S. W., and Burhop, E. H. S. (in press). Electronic and Ionic Impact Phenomena. The Clarendon Press, Oxford, England.
113. Yamanouchi, T. (1947). Elastic collision cross section of oxygen atom for slow electrons. *Progr. Theoret. Phys.* **2,** 33.
114. Bates, D. R. (1939). The quantal theory of continuous absorption of radiation by various atoms in their ground state. I. The atoms from boron to neon. *Monthly No. Roy. Ast. Soc.* **100,** 25.
115. Bates, D. R. (1946). An approximate formula for the continuous radiative absorption cross section of the lighter neutral atoms and positive and negative ions. *Monthly No. Roy. Ast. Soc.* **106,** 423.
116. Bates, D. R. (1946). Calculation of the cross section of neutral atoms and positive and negative ions toward the absorption of radiation in the continuum. *Monthly No. Roy. Ast. Soc.* **106,** 432.
117. Chandrasekhar, S. (1945). On the continuous absorption coefficient of the negative hydrogen ion. *Astrophys. J.* **102,** 223, 395.
118. Coster, D., Van Dijk, E. W., and Lameris, A. J. (1935). Predissociation in the upper level of the second positive group of nitrogen. *Physica* **2,** 267.
119. Ditchburn, R. W., and Jutsum, P. J. (1950). Continuous absorption of light in sodium vapor. *Nature* **165,** 723.
120. Flory, P. J. (1936). Predissociation of the oxygen molecule. *J. Chem. Phys* **4,** 23.
121. Flory, P. J., and Johnston, H. L. (1946). Predissociation in nitric oxide. *J. Chem. Phys.* **14,** 212.
122. Bates, D. R., Fundaminsky, A., Leech, J. W., and Massey, H. S. W. (1950).

Excitation and ionization of atoms by electron impact—The Born and Oppenheimer approximations. *Trans. Roy. Soc.* (*London*) **A243,** 93.

123. Curran, S. C., Cockroft, A. L., and Insch, G. M. (1950). Investigation of soft radiations—VII. Energy expenditure per ion-pair for slow electrons in various gases. *Phil. Mag.* **41,** 517.
124. Hagstrum, H. D., and Tate, J. T. (1941). Ionization and dissociation of diatomic molecules by electron impact. *Phys. Rev.* **59,** 354.
125. Mott, N. F., and Massey, H. S. W. (1949). The Theory of Atomic Collisions. 2nd ed., The Clarendon Press, Oxford, England.
126. Nicholls, R. W. (1950). Intensity distribution of the second positive band system of molecular nitrogen. *Phys. Rev.* **77,** 421.
127. Tate, J. T., and Smith, P. T. (1934). Ionization potentials and probabilities of the formation of multiply charged ions in the alkali vapors and in krypton and xenon. *Phys. Rev.* **46,** 773.
128. Watanabe, M., and Milda, J. (1950). Ionization of the negative ion by electron impact—III. *J. Phys. Soc. Japan* **5,** 149.
129. Williams, E. J. (1945). Application of ordinary space-time concepts in collision problems and relation of classical theory to Born's approximation. *Revs. Modern Phys.* **17,** 217.
130. Yamanouchi, T. (1950). Probabilities of excitation and de-excitation of metastable states of oxygen atom by collision of slow electron. *J. Phys. Soc. Japan* **5,** 154.
131. Yamanouchi, T., and Kotani, M. (1940). Excitation of atoms by electron collision. *Proc. Phys.-Math. Soc. Japan* **22,** 14.
132. Bates, D. R., Buckingham, R. A., Massey, H. S. W., and Unwin, J. J. (1939). Dissociation, recombination and attachment processes in the upper atmosphere. II. The rate of recombination. *Proc. Roy. Soc.* (*London*) **A170,** 322.
133. Biondi, M. A., and Brown, S. C. (1949). Measurement of electron-ion recombination, *Phys. Rev.* **76,** 1679.
134. Craggs, J. D., and Hopwood, W. (1947). Electron-ion recombination in hydrogen spark discharges. *Proc. Phys. Soc.* (*London*) **59,** 771.
135. Holt, R. B., Richardson, J. M., Howland, B., and McClure, B. T. (1950). Recombination spectrum and electron density measurements in neon afterglows. *Phys. Rev.* **77,** 239.
136. Loeb, L. B. (1939). Fundamental Processes of Electrical Discharge in Gases. J. Wiley and Sons, New York.
137. Stuckelberg, E. C. G., and Morse, P. M. (1930). Computation of the effective cross section for the recombination of electrons with hydrogen ions. *Phys. Rev.* **36,** 16.
138. Yamanouchi, T., and Kotani, M. (1940). Photo-ionization and recombination of oxygen atom. *Proc. Phys.-Math. Soc. Japan* **22,** 60.
139. Berg, H. (1951). Bemerkung zur Meteorologie der Ionosphäre. *Geofis. Pura Appl.* **19,** 33.
140. Booker, H. G., Ratcliffe, J. A., and Shinn, D. H. (1950). Diffraction from an irregular screen with applications to ionospheric problems. *Trans. Roy. Soc.* (*London*) **A242,** 579.
141. Briggs, B. H., Phillips, G. J., and Shinn, D. H. (1950). The analysis of observations on spaced receivers of the fading of radio signals. *Proc. Phys. Soc.* (*London*) **B63,** 106.
142. Krautkrämer, J. (1950). Über Wanderungsercheinungen rascher Feldstärke-Schwankungen von Ionosphären-Echos. *Arch. Elekt. Übertragung.* **4,** 133.

143. Mitra, S. N. (1949). Statistical analysis of fading of a single downcoming wave from the ionosphere. *Proc. Inst. Elec. Engrs. Pt. III.* **96,** 505.
144. Scorer, R. S. (1949). Theory of waves in the lee of mountains. *Quart. J. Roy. Meteorol. Soc.* **75,** 41.
145. Sekera, Z. (1948). Helmholtz waves in a linear temperature field with vertical wind shear. *Meteorol. J.* **5,** 93.
146. Wasiutyński, J. (1946). Studies of hydrodynamics and structure of stars and planets. *Astrophys. Norvegica* **4,** 1.
147. Archenhold, F. S. (1928). Die leuchtenden Nachtwolken und bisher unveröffentliche Messungen ihrer Geschwindigkeit. *Weltall* **27,** 137.
148. Gerson, N. C. (1950). Sporadic E movements. *Geofis. Pura. Appl.* **18,** 162.
149. Hoffmeister, C. (1934). Leuchtstreifen, Ionization der oberen Luftschichten, und Ausbreitung der elektromagnetischen Wellen. *Forsch. u. Fortschr.* **10,** 322.
150. Roach, F. E., and Pettit, H. B. (1951). On the diurnal variation of (OI) 5577 in the nightglow. *J. Geophys. Res.* **56,** 325.
151. Störmer, C. (1935). Measurements of luminous night clouds in Norway 1933 and 1934. *Astrophys. Norvegica* **1,** 87.
152. Hough, S. S. (1898). On the application of harmonic analysis to the dynamical theory of tides. II. On the general integration of LaPlace's dynamical equations, *Trans. Roy. Soc. (London)* **A191,** 139.
153. Lamb, H. (1945). Hydrodynamics, 6th ed. Cambridge Univ. Press, Cambridge, England.
154. Martyn, D. F. (1947–48). Atmospheric tides in the ionosphere. *Proc. Roy. Soc. (London)* **A189,** 241; **A190,** 273; **A194,** 429, 445.
155. Pekeris, C. L. (1937). Atmospheric oscillations, *Proc. Roy. Soc. (London)* **A158,** 650.
156. Taylor, G. I. (1936). The oscillations of the atmosphere. *Proc. Roy. Soc. (London)* **A156,** 318.
157. Weekes, K., and Wilkes, M. V. (1947). Atmospheric oscillations and the resonance theory. *Proc. Roy. Soc. (London)* **A192,** 80.
158. Wilkes, M. V. (1949). Oscillations of the Earth's Atmosphere. Cambridge Univ. Press, Cambridge, England.
159. Chapman, S., and Tschu, K. K. (1948). The lunar atmospheric tide at twenty-seven stations widely distributed over the globe. *Proc. Roy. Soc. (London)* **A195,** 310.
160. Appleton, E. V., and Beynon, W. J. G. (1948). Lunar tidal oscillations in the ionosphere. *Nature* **162,** 486.
161. Bartels, J., and Johnston, H. F. (1940). Geomagnetic tides in horizontal intensity at Huancayo. *Terr. Magn.* **45,** 269, 485.
162. Chapman, S., and Miller, J. C. P. (1940). The statistical determination of lunar daily variations in geomagnetic and meteorological elements. *Monthly No. Roy. Ast. Soc. Geophy. Suppl.* **4,** 649.
163. Tschu, K. K. (1949). On the practical determination of lunar and luni-solar daily variations in certain geophysical data. *Australian J. Sci. Res.* **A2,** 1.
164. Bagge, E. (1943). Die Bedeutung der Ionendiffusion für den Aufbau der Ionosphäre. *Phys. Z.* **44,** 163.
165. Chapman, S., and Cowling, T. G. (1939). The Mathematical Theory of Non-Uniform Gases. Cambridge Univ. Press, Cambridge, England.
166. Ferraro, V. C. A. (1945). Diffusion of ions in the ionosphere. *Terr. Magn.* **50,** 215.
167. Ferraro, V. C. A. (1946). On diffusion in the ionosphere. *Terr. Magn.* **51,** 427.

168. Jaeger, J. C. (1945). Note on diffusion in the ionosphere. *Proc. Phys. Soc. (London)* **57**, 519.
169. Johnson, M. H., and Hulburt, E. O. (1950). Diffusion in the ionosphere. *Phys. Rev.* **79**, 802.
170. Kirkpatrick, C. B. (1948). The influence of vertical ionic drift on a "Chapman region." *Australian J. Sci. Res.* **A1**, 421.
171. Lovell, A. C. B. (1950). Meteor ionization in the upper atmosphere. *Science Progress* **38**, 22.
172. Seeliger, R. (1948). Electron diffusion in the ionosphere. *Ann. d. Phys.* **3**, 297.
173. Senftleben, H., and Gladisch, H. (1947). Zur Frage der Einwirkung elektrischer Felder auf den Wärmeübergang in Gasen. *Naturwissenschaften* **34**, 187.
174. Epstein, P. S. (1932). Über Gasentmischung in der Atmosphäre. *Gerl. Beitr. Geophys.* **35**, 153.
175. Lettau, H. (1944). Atmosphärische Turbulenz. J. W. Edwards, Ann Arbor, Michigan.
176. Lettau, H. (1948). Zur Theorie der partiellen Gas Entmischung in der Atmosphäre. *Meteorol. Runds.* **1**, 5, 65.
177. Mitra, S. K., and Rakshit, H. (1938). Distribution of the constituent gases and their pressures in the upper atmosphere. *Ind. J. Phys.* **12**, 47.

Estuarine Hydrography

D. W. PRITCHARD

Chesapeake Bay Institute, The Johns Hopkins University, Baltimore, Maryland

CONTENTS

	Page
1. Introduction	243
2. Classification of Estuaries	244
2.1. General Definition of an Estuary	244
2.2. Classification of Estuaries in Terms of Fresh-Water Inflow and Evaporation	245
2.3. Classification of Estuaries in Terms of Geomorphological Structure	245
3. The Physical Structure and Circulation Pattern in Coastal Plain Estuaries	247
4. The Physical Structure and Circulation Pattern in Fiord Estuaries	253
5. The Bar-Built Estuaries	254
6. Theoretical Studies of the Dynamics of Estuarine Circulation	256
6.1. Stommel's Theoretical Study of the Dynamics of a Deep Fiord Estuary	256
6.2. Cameron's Theoretical Study of the Dynamics of a Deep Fiord Estuary	260
6.3. The Dynamics of Coastal Plain Estuaries	262
7. The Flushing of Tidal Estuaries	268
7.1. Flushing Parameters Determined from the Measured Concentration of Fresh Water	269
7.2. The Tidal Prism Concepts in Estuarine Flushing	270
7.3. A Mixing Length Theory of Tidal Flushing	272
7.4. A Quantitative Study of the Salt Balance in a Coastal Plain Estuary	274
List of Symbols	278
References	279

1. INTRODUCTION

The prime emphasis in the field of physical oceanography during most of its history has been directed toward an understanding of the structure and circulation of the open ocean. Until recently, relatively little work has been done on inshore waters, particularly on bays and estuaries which are actually among the most important areas for fisheries production as well as other human use problems. However, the last few years have seen a considerable increase in the effort expended on studies of inshore regions, particularly in the United States and Canada. During the first forty years of this century a relatively greater proportion of the work by marine investigators in Europe was directed towards a study of estuaries than was the case of investigators on this continent. Although the

findings of these earlier European investigations will be included in the general discussion of estuaries, this article will deal primarily with the results of the investigations carried out during the past few years by marine scientists on this continent.

Our knowledge of the physical structure and circulation in estuaries is far from complete. However, with the increased emphasis on the subject, man's understanding of the physical processes in estuaries is rapidly growing, and there is being created a body of concepts which should, within the next decade, advance this important phase of the science of Physical Oceanography into its proper position.

Because of the complexities of shore-line processes, no definite agreement exists in the field as to even the proper definition of what constitutes an estuary. These same complexities result in a number of definite estuarine types, classification of which has not as yet reached general acceptance.

A set of definitions and classifications, based upon our present knowledge of the physical hydrography of estuaries, is given below in order to facilitate the presentation of the material for this paper.

An effort is made here to group under the general heading of estuaries all semi-enclosed coastal bodies of water which present the same general physical problems. The two primary physical factors which serve to classify a region as estuarine are: (1) the mixing of water whose physical and chemical properties are peculiar to the estuarine region, with water of open sea origin; and (2) the relatively large effect of bottom and/or lateral boundaries in defining the circulation pattern.

2. Classification of Estuaries

2.1. General Definition of an Estuary

Certain investigators have, within the scope of their own studies, set down descriptions to serve as definitions of estuaries. Thus, Finch and Trewartha [1] state: "Embayments resulting from submergence are called drowned valleys or estuaries." Ketchum [2], on the other hand, has taken an entirely different approach. In defining an estuary for his studies of tidal flushing, he states that an estuary is any region in which sea water is measurably diluted by land water drainage. These definitions appear in part too limited and in part too broad. In our consideration of estuaries we should not limit ourselves to drowned river valleys. On the other hand, Ketchum's proposed definition would include open coast along which there is measurable dilution of sea water by land water drainage. This broad field of coastal oceanography should not be

included in the general classification of estuaries because of the considerable difference in boundary conditions, and the addition of many physical problems not present within semi-enclosed coastal features. Other investigations are under way in embayments in which the evaporation exceeds the precipitation and fresh water runoff. During the past few years the tendency has been to group these investigations within the general field of estuarine studies.

For the purpose of this discussion the following definition will be used: *An estuary is a semi-enclosed coastal body of water having a free connection with the open sea and containing a measurable quantity of sea salt.* The salinity of an estuary may be less than, equal to, or greater than the salinity of the open ocean, depending upon the relationship between fresh water inflow and evaporation. An estuary is essentially a coastal feature and for this reason certain large bodies of water which might exhibit estuarine features such as the Mediterranean Sea and the Gulf of Mexico are excluded from this definition.

Since any attempt to present a definition of this type in a single, compact sentence is apt to lead to some uncertainties, the several statements immediately preceding and immediately following the definition should be considered as amplifications of its meaning.

2.2. Classification of Estuaries in Terms of Fresh-Water Inflow and Evaporation

Estuaries may be divided into two large groups depending upon the relationship between fresh-water inflow and evaporation. Historically, the term "estuary" has been applied to coastal indentures in which there is a measurable dilution of sea water by land drainage. To this group we will apply the term *positive* estuaries. A second class of estuaries exists in which the evaporation exceeds the land drainage plus the precipitation. In this type, in which there is found a mixture of high salinity estuarine water and sea water, we will apply the general term *inverse* estuary. There may also exist a group of estuaries in which neither the fresh-water inflow nor the evaporation dominates. This class will be termed *neutral* estuaries. Very little work has been done on the neutral estuary, and it will not be considered further in this treatment.

2.3. Classification of Estuaries in Terms of Geomorphological Structure

A further classification of estuaries can be made on the basis of geomorphological structure. The most intense study of estuaries is now being made on a class which may be called coastal plain estuaries. These estuaries have been formed by drowning of former river valleys, either

from subsidence of the land or from a rise in sea level. Thus they are usually an elongated indenture of the coast line with the river flowing into the upper end. A typical coastal plain estuary is a rather shallow body of water often with a dendritic shore line. Most of the eastern shore of North America is characterized by embayments of this class, such as the Chesapeake Bay with its tributary estuaries, the Delaware Bay, the lower Hudson River, and the St. Lawrence River. Most coastal plain estuaries are of a positive nature, that is, the precipitation and fresh-water runoff exceed the evaporation. However, in certain cases, principally where the rivers supplying the main source of fresh water have been diverted to another outlet, coastal plain estuaries of the inverse type do exist.

A second class of estuary which has received considerable attention from marine investigators is the deep-basin type exemplified by the glacial cut fiords of the Norwegian coast and of the Canadian Pacific coast. These estuaries are elongated indentures of the coast line containing a relatively deep basin with a shallow sill at the mouth. Fresh water inflow exceeds evaporation in the majority of the estuaries in this class. However, a few inverse estuaries of this type are found in arid regions.

A third large group of estuaries results from the development of an offshore bar on a shore line of low relief and shallow water. Usually a very narrow channel exists between the open sea and the estuary, or sound, as it is frequently called. These bar-built estuaries may have a river leading into them, in which case they are generally of the positive type, with the fresh-water runoff exceeding the evaporation. Quite frequently, however, bar-built estuaries exist in arid regions where the evaporation exceeds the fresh water inflow and the estuarine waters are then of very high salinity.

Stommel [3] has suggested a classification scheme for estuaries based upon the predominant physical causes of movement and mixing of water in the estuary. These predominant physical causes of mixing may be either tidal, meteorological (wind), or river flow. In the majority of the coastal plain estuaries of the Atlantic seaboard the predominant cause of movement and mixing appears to be the tide, upon which is superimposed a weak river flow. In many of the bar-built estuaries or sounds, the movement and mixing of waters appear to depend primarily on the wind. Finally, as pointed out by Stommel, the position of the wedge of salt water which intrudes into the mouth of the Mississippi River depends entirely upon the flow of the river.

In many estuaries no single cause of movement and mixing predominates. In many bar-built estuaries the tidal motion will dominate near

the channel connecting the estuary with the ocean. Well within the sound, however, both wind and tide may contribute appreciably to the motion and mixing.

This sub-classification suggested by Stommel appears to be well suited for studies of the physical hydrography of estuaries. Each of the above major classes of estuaries would then be broken down according to the dominant physical cause of the movement and mixing of the estuarine water.

The recent studies of estuaries have been confined primarily to four classes of estuaries. The greatest effort in this country has been directed towards a study of the positive coastal plain estuaries of the Atlantic seaboard. Considerable study has been made by Canadian investigators of the basin or fiord estuaries of the Canadian Pacific coast. These estuaries are also of the positive type. In both the coastal plain estuaries of the Atlantic seaboard and the Canadian fiords the predominant physical cause of mixing appears to be the tides. The physical hydrography of certain of the bar-built estuaries along the Gulf of Mexico recently received increased interest. These recent studies along the Texas Coast have included both the positive and the inverse types. In these bar-built estuaries the wind appears to be the dominant physical cause for the movements and mixing, though frequently tidal motion is also important.

3. The Physical Structure and Circulation Pattern in Coastal Plain Estuaries

The majority of the world's harbors occur in estuaries which have been formed by the drowning of the seaward end of river valleys. Extensive fishing grounds usually exist in these coastal plain estuaries. The largest estuary of this type found in the United States is the Chesapeake Bay. For the past two years an extensive investigation of this Bay and its tributary estuaries has been carried on by the Chesapeake Bay Institute of The Johns Hopkins University. The Woods Hole Oceanographic Institution has been engaged in investigations of several coastal plain estuaries in the New England area.

As pointed out above, coastal plain estuaries are relatively shallow bodies of water. Thus 50% of the Chesapeake Bay estuarine system is less than 20 feet deep, and only 8% is deeper than 60 feet. Similar depth distributions are found among the major estuaries of the Atlantic Coast, and even shallower mean depths occur in the smaller ones.

The intensive investigation of the Chesapeake Bay and its tributary estuaries, now under way, has provided a basis for the description of the physical structure and circulation pattern in the coastal plain estuaries in

mid-latitudes. The following discussion represents a summary of these findings.

The salinity increases from zero at the head of the estuary to that of sea water at the mouth. The isohalines are not perpendicular to the

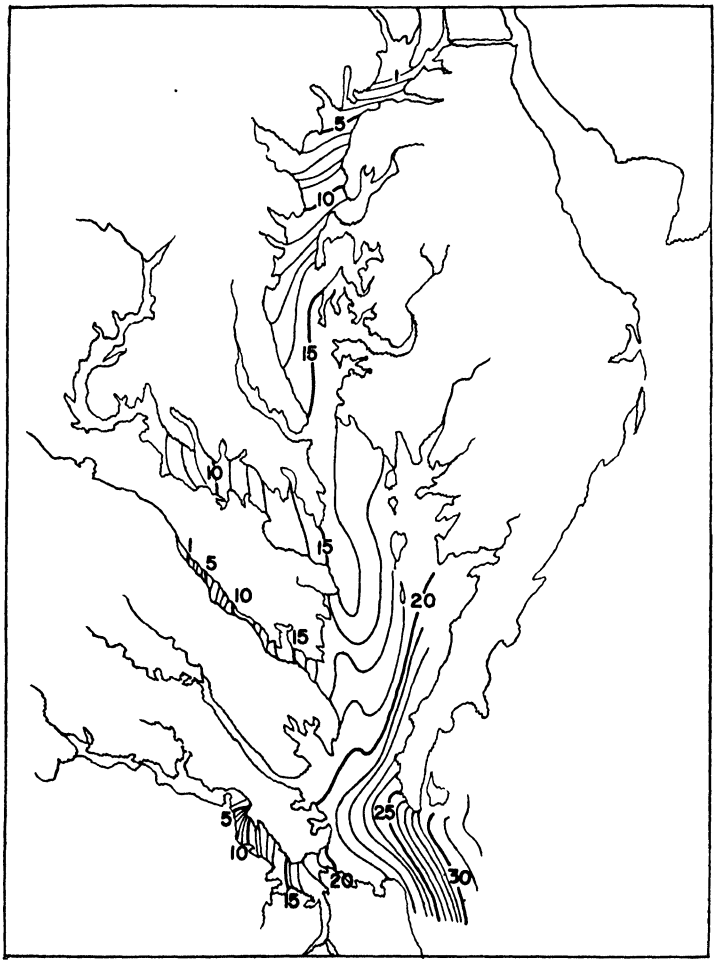


FIG. 1. Typical surface salinity distribution, Chesapeake Bay.

longitudinal axis of the estuary, but rather run obliquely across the estuary with slightly lower salinities on the right-hand side than on the left when facing toward the mouth of the estuary.

The salinity increases with depth during all seasons. The salinity depth curve has the general shape of an inverse tangent curve. Close to

the bottom and, to a lesser extent, near the surface, the curve departs from a typical tangent shape due to boundary effects.

Figure 1 shows a typical salinity distribution in the Chesapeake Bay. Here the lateral gradient in salinity is probably accentuated by the excess of the fresh water runoff from the western shore of the estuary. However, even in narrow estuaries, in which there is no marked difference between the runoff from the land on either side, extensive measurements reveal a lateral gradient giving a lower salinity on the right-hand side than on the left-hand side. Figure 2 shows typical salinity vs. depth curves in this type estuary.

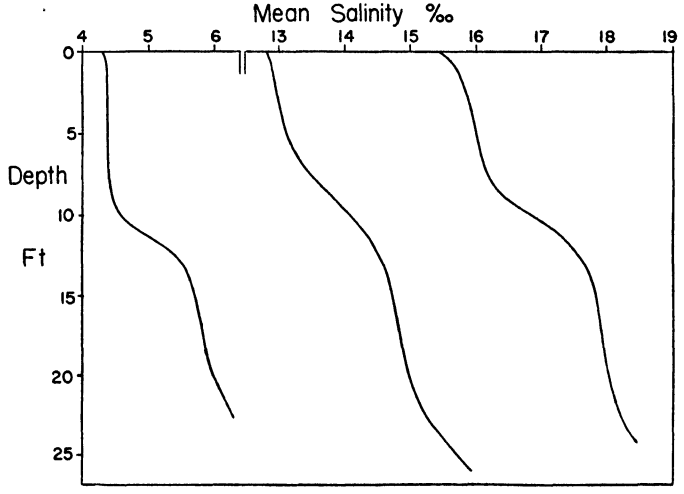


FIG. 2. Typical salinity vs. depth curves, Chesapeake Bay.

In contrast to the fairly regular pattern shown by the horizontal salinity distribution, the temperature distribution over much of the year shows no regular pattern. The horizontal temperature distribution appears to be controlled primarily by local meteorological conditions. In general it can be said that in winter the horizontal temperature distribution shows colder water near the head of the estuary, and warmer water near the ocean. This condition is reversed to some extent during the summer months.

The most apparent water movements are related to tidal motion with tidal currents of about ½ to 3 knots occurring in the Chesapeake Bay system. These tidal currents are oscillatory in character and have superimposed upon them a net current system related to the dynamic structure of the estuary itself.

Current observations taken over one or more tidal cycles reveal that from the standpoint of physical structure and circulation the estuary

may be considered as composed of two layers. In an upper layer there is a net horizontal flow down the estuary towards the ocean. In the lower layer there is a net horizontal flow up the estuary towards the river. Figure 3 shows a series of typical net horizontal velocity profiles. The mean velocity in each of the two layers is approximately one-fifth the magnitude of the maximum tidal current.

The boundary between the upper layer with its net down-estuary flow and the lower layer with its net up-estuary flow is called a *surface of no net*

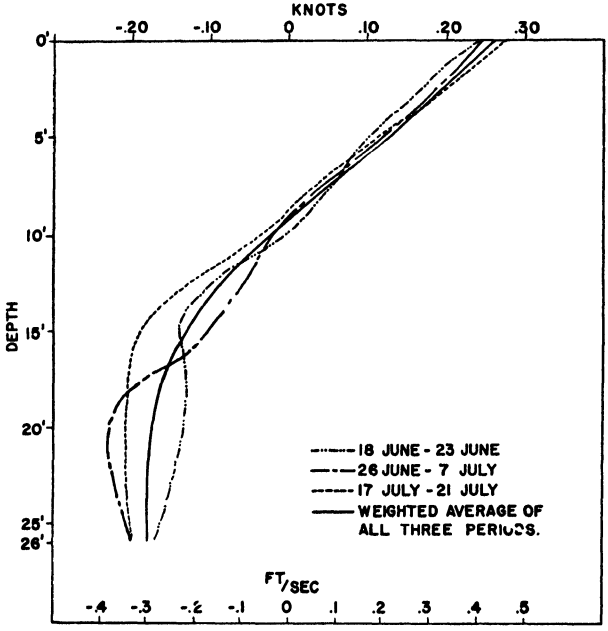


FIG. 3. Net velocity vs. depth curves for three periods in the James River estuary.

motion. This surface is observed to vary in depth, both longitudinally and laterally. Figure 4 shows a section across the James River estuary with the typical lateral slope to the surface of no net motion indicated. This surface slopes downward from the left-hand side of the estuary to the right, and hence the upper layer with its net down-estuary flow is deeper on the right-hand side where the salinity is low than on the left-hand side where the salinity is high.

The boundary between the two layers is close to, but does not necessarily coincide with, the inflection point on the salinity-depth curve. The upper layer is thus of lower salinity than the lower layer. The volume of flow in the upper layer must exceed the volume of flow in the lower layer by an amount equal to the inflow of fresh water from the river.

Though the salinity distribution shows a seasonal variation, it may be considered to be in a steady state during any particular season. Since the upper layer is transporting seaward a net amount of fresh water equal to the fresh water inflow, there must be a flow of salt into the upper layer to maintain the salinity distribution. This is accomplished by a net transfer of water of relatively higher salinity from the lower to the upper

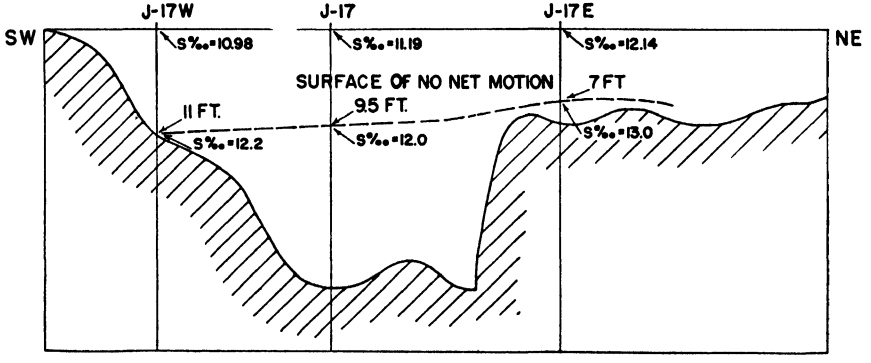


FIG. 4. Cross-section in the James River estuary showing typical slope to surface of no net motion.

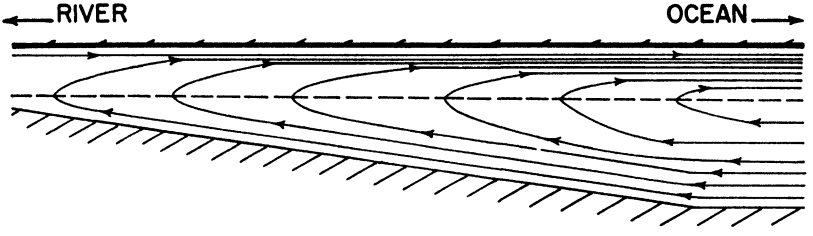


FIG. 5. Schematic presentation of volume transport lines in a longitudinal section taken along axis of estuary.

layer. There must therefore be a negative vertical velocity across the boundary between the two layers. The volume of flow in the upper layer increases towards the mouth, and the volume of flow in the lower layer decreases towards the head of the estuary. A schematic presentation of the volume transport taken in a longitudinal section down the central axis of the estuary is shown in Fig. 5.

The characteristic lateral gradient in salinity and the characteristic slope in the surface of no net motion suggest the influences of the earth's rotation. It has been suggested that this lateral salinity distribution results in a mean lateral pressure force which is in the main balanced by the Coriolis force related to the mean horizontal motion. Indeed, exten-

sive observations at three cross-sections in the James River estuary have shown that more than 75% of the Coriolis force associated with observed mean motion is balanced by the lateral pressure force related to the lateral salinity gradient.

These investigations have led to a proposed dynamic structure for this type of estuary. The surface of the estuary must have a mean slope downward towards the mouth, and also downward from the right side of the estuary towards the left. These slopes of the constant pressure surfaces decrease in magnitude with depth to some intermediate depth which corresponds to a level pressure surfaces. Below this level surface the pressure surfaces slope downward from the mouth of the estuary towards the head and also downward from the left side to the right side.

Consider a left-handed coordinate system with the origin at the surface in the fresh water at the head of the estuary. The x_1 coordinate is directed down the estuary along the central axis. The x_2 coordinate points vertically downward. The lateral coordinate is x_3, and is directed horizontally from the center of the estuary toward the right-hand shore. Assuming no mean lateral motion, the horizontal components of the equation of motion may be written

$$\bar{v}_1 \frac{\partial \bar{v}_1}{\partial x_1} + \bar{v}_2 \frac{\partial \bar{v}_1}{\partial x_2} = -\alpha \frac{\partial p}{\partial x_1} + m_1 \tag{1}$$

$$0 = -\alpha \frac{\partial p}{\partial x_3} + f\bar{v}_1 + m_3 \tag{2}$$

where:

$\bar{v}_1$ = mean horizontal velocity along x_1
$\bar{v}_2$ = mean vertical velocity
α = specific volume
p = mean pressure
m_1 = x_1 component of the frictional stress
m_3 = x_3 component of the frictional stress
f = Coriolis parameter

The first equation was solved for a special type of stream function by Cameron [4]. This equation deals with the field acceleration terms, and shows that the longitudinal pressure gradient is related to these accelerations. The second equation expresses the balance between the lateral pressure force, the Coriolis force related to the mean horizontal motion and a frictional term. From an analysis by the author of the observed data from the James River estuary, it appears that the lateral pressure force, which is associated with the lateral salinity gradient, is required by the equation of motion to balance the major part of Coriolis force resulting from the observed mean horizontal motion. The frictional term in the

second equation amounts to less than one-quarter of the value of the other terms.

In some coastal plain estuaries, particularly those with depths of less than 20 feet, vertical stratification appears to be lacking and the two-layer circulation pattern is not well developed. Ketchum [5] has discussed this type of estuary in his study of the Raritan River. He has suggested that the salinity balance within the estuary is maintained by horizontal mixing due to the oscillatory tidal motion. This type of estuary will be discussed more fully in the section dealing with flushing studies.

4. The Physical Structure and Circulation Pattern in Fiord Estuaries

Several physical investigations have been carried out during the past half-century in the Norwegian fiords. This European work has supplied the background of knowledge for the recent investigations of the fiords along the Canadian Pacific coast. These recent studies have been reported by Tully [6] and Cameron [4], and have been carried out principally by personnel of the Pacific Biological Station, Nanaimo, B. C.

The following description of the physical structure and circulation pattern in a fiord estuary is based principally on Tully [6]. The published work deals primarily with Alberni Inlet, a fiord estuary typical of the British Columbian and Alaskan Coasts.

This fiord estuary is an elongated, U-shaped trough extending inland from the Pacific Ocean approximately 35 miles. In general the sides are precipitous and rocky. The sill depth at the mouth is approximately 120 feet, while the basin which occupies much of the length of the fiord has a maximum depth of over 1000 feet. Much of the estuary is over 600 feet in depth. The main supply of fresh water inflow is from the Somass River at the head of the estuary.

Above a depth of about 33 feet the salinity decreases steadily from the mouth toward the head. Below 33 feet the waters of the fiord are approximately continuous in salt content with the open ocean waters. Thus it would appear that actual estuarine conditions, that is, the mixing between sea water and fresh water from land drainage, occur only in the upper 30 to 40 feet.

Tully observed a very sharp salinity gradient normally occurring at about 13 feet. This sharp gradient appeared to mark the boundary between a surface layer in which the net motion was directed down the estuary toward the mouth, and a deeper layer in which the net motion was directed up the estuary towards the head. Tully concluded that the region of the fiord below the sharp salinity gradient could itself be

considered as two layers. He assumed that the majority of the net up-estuary motion was confined to a middle zone occurring between the depth of the sharp salinity gradient and a depth of 33 feet. The deep zone below 33 feet was considered to have vanishingly small velocities.

The circulation and structure of the upper and middle zones in this fiord estuary appear similar to the two layers found in the Chesapeake Bay and its tributaries. On the other hand the great volume of the deep zone present in the fiord estuary affords a source of sea water for mixing with the upper two layers that is entirely absent in coastal plain estuaries. Consequently, the longitudinal salinity gradient in the middle zone of the fiord is very small compared to the seaward increase of salinity found in the upper zone.

Tully observed some very interesting relationships between the estuarine properties and both the tidal motions and the river discharge. These relationships and the implications suggested by them are listed below:

(1) A nearly linear relationship exists between tidal velocities and the increase in sea water per unit volume of fresh water in the upper layer. This observation confirms Tully's assumption that the rate of mixing, which leads to the seaward increase in salinity in the upper zone, is related primarily to tidal action.

(2) The depth of the upper zone decreases with increasing river discharge up to a critical value of the discharge. Beyond this critical value the depth of the upper zone increases with increasing discharge.

(3) Within the range of river discharge below the critical value mentioned in (2), there is also a decrease in the mean salinity of the upper zone. For values of the river discharge above this critical value, the mean salinity in the upper layer increases with increasing river discharge. These last two relationships indicate that the river discharge must exert some influence on the rate of mixing, though evidently tidal action remains the predominant cause of the mixing.

5. The Bar-Built Estuaries

Collier and Hedgpeth [7] have recently discussed the hydrography of certain bar-built estuaries along the Texas Coast. This treatment constitutes the most intense investigation to date of the physical hydrography of this class of estuaries. Of extreme interest is the fact that the system of estuaries studied by Collier and Hedgpeth includes an inverse estuary next to, and in partial communication with, one of the several positive estuaries enclosed by the same offshore bar.

Characteristic features of bar-built estuaries are: (1) a relatively small channel connecting the estuary with the ocean, and (2) shallow depths

within the estuary. Because of the narrowness of the channel, the tidal velocities are usually large in and near it, but are relatively small within the estuary. But, as a result of the shallow depths within the bar-built estuary, even these small tidal velocities contribute to the mixing between the sea water and the land drainage. This fact was clearly demonstrated by Collier and Hedgpeth. They pointed out that there are considerable differences in the salinity distribution during periods of tropical tides and periods of equatorial tides, and that the longer period changes in sea level (there is a considerable seasonal variation in the mean tide level) are reflected in the salinity exchange.

No direct measurements of the net circulation pattern within bar-built estuaries have been reported in the literature. Collier and Hedgpeth discussed the circulation pattern at various stages of the tide as inferred from a study of the salinity distribution. However, this attempt was not sufficiently comprehensive to indicate the manner in which the exchange between the estuarine water and the ocean water entering through the channel actually takes place. There was some indication from their data that at times there is a net movement of the more saline waters up the estuary along the bottom, thus establishing the two-layer system characteristic of the coastal plain estuaries. In general, however, the horizontal mixing related to the tidal motion and to the wind appears to control the salinity exchange, with the net circulation confined to a net movement down-estuary on each ebb tide that exceeds the movement up-estuary on each flood tide sufficiently to remove a volume of fresh water equal to the fresh water inflow into the estuary during the tidal period.

The study of Laguna Madre by Collier and Hedgpeth is of singular interest since in this estuary the evaporation exceeds precipitation and land drainage. The connection with the ocean is through the lower part of Corpus Christi Bay, and the exchange between Laguna Madre and Corpus Christi Bay is greatly restricted due to the shallowness of the tidal flats separating these two bodies of water.

As a result of the excess evaporation and the restricted communication with Corpus Christi Bay, salinities within the Laguna at times exceed 100 parts per thousand and salinities of 60 to 80 parts per thousand are common. The seasonal changes in mean tide level are very significantly related to the surface salinities of Laguna Madre, and it appears evident from the analysis of Collier and Hedgpeth that significant exchanges between the high salinity water of the Laguna and the moderate salinity water (26‰) of lower Corpus Christi Bay occur only during periods when the mean tide level is above normal. From the standpoint of the salinity exchange, quoting Collier and Hedgpeth [7], " . . . the water level

variation coincident with the cycle of tropical and equatorial tides is probably the most important so far as lagoons (of this type) are concerned. It is these variations which bring about the largest exchange of water between lagoon and gulf, and which alternately expose and flood the greatest area of tidal flat and marsh."

The net circulation pattern in inverse estuaries of this type is apparently the inverse of the typical pattern discussed for positive coastal plain estuaries. The lower salinity water from outside the inverse estuary or lagoon can enter only on the surface, because of the higher density of the lagoon water. An outflow of these dense lagoon waters occurs along the bottom when the depth of the inlet allows such exchange. This type of circulation apparently occurs in the Laguna Madre only during periods when the mean tide level is above normal, though in some inverse estuaries, with a less restricted channel to the ocean, such a circulation pattern may be relatively steady.

6. Theoretical Studies of the Dynamics of Estuarine Circulation

There have been two recent theoretical studies of the dynamics of estuaries. Both of these studies have been concerned with deep tidal fiords, which appear to be the simplest and most clear-cut type of estuary, as bottom effects and Coriolis terms may be neglected. The results of these two investigations are in many ways similar, though the theoretical models employed differ considerably.

In the following discussions the coordinate system employed will be that described in Section 5.

6.1. Stommel's Theoretical Study of the Dynamics of a Deep Fiord Estuary

Stommel [3] has treated a simple steady state model of a deep fiord estuary which is considered as being composed of two layers. The lower layer, which occupies the greater portion of the depth of the estuary, is filled with ocean water of constant density ρ_l from a depth $x_2 = \zeta_l(x_1)$ to the bottom. Floating on the top of the layer is a thin sheet of brackish water of density $\rho_u(x_1)$, with a free surface at $x_2 = \zeta_u(x_1)$.

Making use of Tully's observations that the deep zone (below 33 feet) in Alberni Inlet showed no measurable dilution from full sea water, and of Keulegan's [8] description of tests on the interfacial mixing in stratified flows made in flumes, Stommel assumes that the interface between the two layers permits upward movement of deep water into the top layer, where it is mixed, but permits no mixing downwards. Thus, though ρ_l is independent of position, $\rho_u(x_1) \rightarrow \rho_l$ as $x_1 \rightarrow$ infinity. The mixing is considered to be fixed, independent of the mean flow, and the cause of

the mixing is not specified beyond the statement that it may be due to winds, tidal currents, or the shear developed at the interface.

Stommel also assumes that the depth of the top layer, which he designates as $D = \zeta_l - \zeta_u$, is very much less than that of the bottom layer. Within the top layer mixing is considered to be so strong that the density $\rho_u(x_1)$ and the mean velocity $v_u(x_1)$ are independent of the depth x_2. There is a vertical velocity v_m of mixing of deep water into the upper layer, while the horizontal velocity in the deep layer is considered to be vanishingly small.

From a consideration of mass continuity, under the above assumptions, the rate of change of mass flux per unit width of the top layer with horizontal distance is given by

$$\frac{d}{dx_1}(D\rho_u v_u) = v_m \rho_l \tag{1}$$

Using the concept of conservation of salt, and assuming a simple linear relationship between density and salinity, the vertical velocity across the interface is further related to the thickness of and the horizontal velocity in the upper layer by the relationship

$$\frac{d}{dx_1}(Dv_u) = v_m \tag{2}$$

Stommel obtains a third relationship between the velocity of the upper layer, v_u, the vertical velocity at the interface, v_m, the thickness of the upper layer, D, and the density of the two layers, ρ_u and ρ_l, from the following considerations. The assumption is made that because of the vanishingly small velocities in the deep layer, the horizontal pressure gradient at any depth within this layer must vanish. Again, since the horizontal velocities in the deep layer are assumed to be negligible, the flux of momentum in the upper layer is not added to by the water mixed up from the lower layer. This momentum flux does vary with the longitudinal coordinate x_1, however, as a result of the variations of the pressure with x_1. Stommel has expressed these pressure variations in terms of the density of the upper layer, the thickness of the upper layer, and the thickness of the lower layer. From a combination of these concepts, the expression

$$\frac{d}{dx_1}\left[D\rho_u\left(v_u^2 + \frac{gD}{2}\gamma\right)\right] = 0 \tag{3}$$

where $\gamma = \rho_l - \frac{\rho_u}{\rho_l}$, is obtained.

These three equations are independent relationships between the variables D, ρ_u, v_u, and v_m.

Introducing the expression

$$T = D\rho_u v_u \tag{4}$$

denoting the transport of the upper layer, Stommel obtains a second form of equation (3). Thus

$$\frac{dD}{dT} = \frac{1}{\sqrt[3]{bk}}\left[\frac{2b-1}{b-2}\right] \tag{5}$$

where $b = \frac{v_u^3}{k}$ and $k = g\gamma_0 \frac{T_0}{2}$.

Here the subscript "$_0$" indicates the values of γ and T at the head of the estuary where $x_1 = 0$.

Stommel has computed the numerical values of $\frac{dD}{dT}$ against b, and shows that there is a range of b: $0.5 < b < 2.0$ for which the top layer decreases with increasing transport. Since the value of b increases rapidly after a certain value of the transport, Stommel concludes that his two-layer model breaks down after $b \to 2.0$.

Stommel points out that this point of breakdown of the model occurs when the velocity of the upper layer is equal to the velocity of an internal wave at the interface. As will be shown later, Cameron [4] obtains a similar "critical" velocity at the surface, though his theoretical approach is considerably different from Stommel's.

In his discussion of the probable nature of the mixing process of deep water into the surface layers, Stommel makes use of some experiments described by Keulegan [8]. In these tests, Keulegan showed that when light water flows over a pool of resting water at certain velocities, waves appear on the interface between the two layers of different density. If the velocity of the lighter water is increased above a certain critical value, the crests of these waves break off and diffuse water from the lower layer into the lighter fluid. There is apparently no corresponding mixing of light water downward into the lower layer.

Keulegan obtained the empirical relationship between the amount of upward mixing and the velocity of the upper layer of light water. This expression, which has the form

$$v_m = C(v_l - 1.15v') \tag{6}$$

where v' is the critical velocity at which mixing first appears and v_m is the mean velocity of mixing of the deep water into the upper layer, applies over a wide range of relative densities. From the flume experiments Keulegan established that C is a constant with a value of 3.5×10^{-4}.

Assuming that the velocities encountered in a natural estuary are

much greater than the critical velocity v', Stommel combined Keulegan's expression (6) with equations (1) and (3) and solved for the distribution of density in the range where D remains virtually constant. The expression he obtained is

$$\gamma = \gamma_0 e^{-c\frac{x_1}{D}} \tag{7}$$

Stommel compared the actual observed horizontal density distribution of the surface layers in Alberni Inlet with the distribution computed from equation (7), using Keulegan's constant C and a depth of the mixed layer of 23 feet. The fit between the theoretical curve and the observational points is quite remarkable, considering the relatively simple theoretical model employed.

Perhaps the part of Stommel's paper most difficult to understand clearly is the development of the equation for the longitudinal rate of change of momentum flux in the upper layer. The statement is made that this flux of momentum "is not added to by the water mixed up from below because by hypothesis this water has vanishingly small horizontal velocity." It would appear, however, that this vertical flux of water of zero horizontal velocity must contribute a retarding force, much like an eddy stress, and as a result a further term representing this retarding force should appear in equation (3).

Stommel has made no statement concerning possible frictional terms, though they must be present if his model is to be internally consistent. To see this, let us consider the horizontal component of the equation of motion for the upper layer:

$$\rho_u v_u \frac{\partial v_u}{\partial x_1} = -\frac{\partial p}{\partial x_1} + \rho_u m_1 \tag{8}$$

where m_1 represents an eddy frictional term. The field acceleration term involving $\frac{\partial v_u}{\partial x_2}$ is not present since v_u is constant in the upper layer.

If the above expression and the hydrostatic equation are cross-differentiated and added together, the following equation results:

$$\frac{\partial}{\partial x_2}\left(\rho_u v_u \frac{\partial v_u}{\partial x_1}\right) = -g\frac{\partial \rho_u}{\partial x_1} + \frac{\partial}{\partial x_2}(\rho_u m_1) \tag{9}$$

But Stommel has assumed, as a basic premise in his treatment, that

$$\frac{\partial \rho_u}{\partial x_2} = \frac{\partial v_u}{\partial x_2} = 0,$$

and hence

$$0 = -g\frac{\partial \rho_u}{\partial x_1} + \frac{\partial}{\partial x_2}(\rho_u m_1) \tag{10}$$

From this expression it is seen that if there is to be a longitudinal gradient of density, there must also be a frictional term. One of the results of Stommel's development is an expression for the horizontal variation in density, and hence in this model there must be eddy frictional forces.

Comments on the reality of Stommel's model will be reserved until the end of the next section.

6.2. Cameron's Theoretical Study of the Dynamics of a Deep Fiord Estuary

Cameron [4] undertook a theoretical attack on the dynamics of a deep fiord estuary making use of a model of somewhat greater complexity than that studied by Stommel. Like Stommel, Cameron reduced the problem to a two-dimensional one by assuming lateral homogeneity. However, rather than considering simply two layers each of uniform density, he assumes that the horizontal velocity, the vertical velocity, and the density are continuous functions of depth.

Assuming steady state, but retaining the field acceleration terms and a frictional term, Cameron combines the horizontal and vertical components of the equation of motion to obtain the expression

$$(1) \qquad \frac{\partial}{\partial x_2}\left[\rho\left(v_1\frac{\partial v_1}{\partial x_1}+v_2\frac{\partial v_1}{\partial x_2}\right)\right]=-g\frac{\partial \rho}{\partial x_1}+\frac{\partial^2}{\partial {x_2}^2}\left(A_v\frac{\partial v_1}{\partial x_2}\right)$$

where v_1 is the horizontal component of the velocity, v_2 is the vertical component of the velocity, and A_v is the eddy coefficient of viscosity associated with vertical shear.

Cameron introduces a stream function which satisfies the observed and inferred velocity field. The vertical variations of the horizontal velocity and of the stream function assumed in this model are shown in Fig. 6. Here v_s is the surface velocity, and k is a constant of dimension L^{-1}. The vertical velocity v_2 has the same relative variation with depth as the stream function, ψ.

In a relatively shallow surface layer the horizontal velocity is directed down the estuary and decreases from a maximum at the surface to zero at what might be called the depth of no net horizontal motion. Below this depth the horizontal motion is directed towards the head of the fiord. At a depth of about twice the depth of no net motion the maximum negative, or up-estuary velocity, is found. In the deeper waters the velocity approaches zero asymptotically.

The horizontal variations in velocity can be described completely in terms of the horizontal variation of the surface velocity. The depth of no net motion is considered as constant.

Cameron defines a function $\theta = 1 - \frac{\rho}{\rho_s}$ which he calls the anomaly

ratio. Here ρ_s is the density of undiluted sea water. By introducing his assumed velocity functions into equation (1), and by making use of

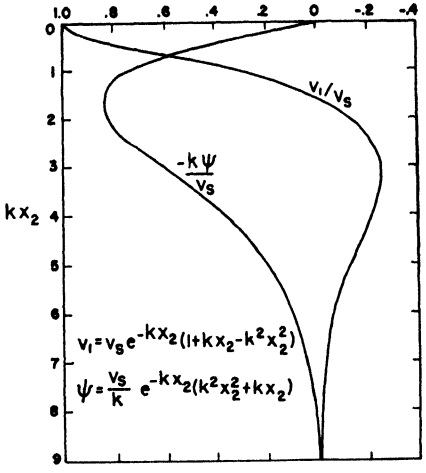


FIG. 6. Stream function and velocity profile assumed by Cameron in theoretical model of a fiord estuary.

the equation of continuity, Cameron develops an equation expressing the horizontal gradient of the anomaly ratio in terms of the surface velocity, the eddy coefficient of viscosity, and the depth. Thus:

$$\frac{\partial \theta}{\partial x_1} = f(v_s, A_v, x_2, k) \tag{2}$$

Here k is again the constant which defines the depth of no net motion and the detailed character of the exponential variation of the velocity with depth.

The solution of equation (2) is expressed as the sum of three functions, which Cameron calls the acceleration, friction, and critical ratios. Based upon terms included in each of these functions, Cameron concluded that the horizontal pressure gradient field associated with the first function is responsible for the field acceleration; and that associated with the second function is responsible for balancing eddy friction. No horizontal pressure gradient field is associated with the critical ratio function.

The introduction into Cameron's solution of continuity in fresh-water transport down-estuary leads to a concept of a critical velocity. This critical, or maximum, velocity would occur at the position where the pressure gradients, associated with the decrease in fresh-water concentration seaward, are sufficient only to produce the accelerations required to maintain a constant fresh water transport.

Since part of the pressure gradient field is required to balance the

eddy friction, this critical velocity could not be attained without a breakdown of the model. However, the concept of the critical velocity allows Cameron to express non-dimensionally the horizontal variation of the surface velocity and the two-dimensional field of the stream function.

The combination of the acceleration, friction, and the critical ratio functions gives the distribution of density, expressed in terms of

$$\theta = 1 - \frac{\rho}{\rho_s}$$

The density field thus determined by Cameron compares favorably with the observed density field found in fiord estuaries.

Certain questions arise in an analysis of Cameron's attack on this problem. The most critical of these questions is concerned with the form of the frictional term. It was assumed that this term is related to the shear of the horizontal velocity. This velocity is actually the net, or non-tidal velocity, and since the oscillatory tidal currents are larger than the non-tidal currents, there is some question as to whether the frictional term might be related to the tidal currents rather than to the net velocity field. Furthermore, Cameron found it necessary to assume a constant eddy coefficient of viscosity.

In general Cameron's model resembles the actual fiord estuary much more closely than does Stommel's. The non-tidal component of currents observed in Alberni Inlet have a vertical variation which closely satisfies the analytical function assumed by Cameron. Moreover, Cameron was able to obtain both the longitudinal and vertical distribution of density, while Stommel dealt with only the horizontal density variations.

Perhaps the greatest disadvantage of Cameron's study as compared to Stommel's is that the former assumed a constant depth of no net motion. Stommel's solution, giving an actual decrease in the depth of his surface layer, appears to be verified in field studies by Tully [6].

Though these two studies are based upon models involving considerably different assumptions, each resulted in determining a critical velocity above which the model breaks down. In several recent unpublished discussions Tully has stated that field observations indicate a breakdown of the estuarine structure of the surface layers near the mouth of certain fiords along the coast of British Columbia. Apparently the critical transport is approached in these fiords at or near the mouth. The agreement of these two studies on this point is of considerable interest.

6.3. The Dynamics of Coastal Plain Estuaries

There have been no published theoretical studies of relatively shallow estuaries. These systems are somewhat more complicated than the deep

fiords because of the effects of shallow depths and sloping sides on the motion and mixing.

In an as yet unpublished investigation the author has approached the problem of the dynamics of estuarine circulation, as applied to the coastal plain estuaries of the Chesapeake Bay system, from a different point of view than that taken by Stommel and by Cameron. This approach involves setting up as complete a set of equations as possible, and then using field data to evaluate the significance of the various measurable terms and to numerically compute from the equations, using the measurable quantities, the steady state distribution of those factors which are not readily measurable.

The first part of this study is now nearing completion and deals primarily with the relatively narrow James River estuary. Extensive measurements of temperature, salinity, and current velocity were made in this estuary in the summer of 1950. These measurements have been used in a semi-empirical study of the factors controlling the salt balance, and of the dynamics of the estuary.

Neglecting molecular stresses, the longitudinal component of the instantaneous equation of motion is

$$(1) \qquad \frac{\partial v_1}{\partial t} + v_1 \frac{\partial v_1}{\partial x_1} + v_2 \frac{\partial v_1}{\partial x_2} + v_3 \frac{\partial v_1}{\partial x_3} = -\alpha \frac{\partial p}{\partial x_1} + f v_3$$

This equation cannot be treated in the form presented, since it is not possible to obtain measurements of the instantaneous pressure gradients and velocities. At our present state of knowledge, it is more fruitful to study the mean values. Also, it becomes necessary to make certain simplifying assumptions regarding the lateral velocity component.

The following simplifications are made in order to apply this equation to the type of estuary under investigation:

(1) Only the mean values of the variables in the lateral direction will be considered.

(2) The lateral velocity component is related to the geometry of the side boundaries. Thus, in a straight-sided U-shaped estuary there would be no lateral velocity term. In an estuary which had a V-shaped cross-section, the lateral velocity would be directly related to the vertical velocity and to the slope of the sides. The cross-sections of a typical coastal plain estuary are more or less V-shaped.

(3) Steady state exists in regard to a time mean taken over one or more tidal cycles.

The instantaneous velocity is replaced by the sum of a mean velocity and a random velocity. Thus $v_i = \bar{v}_i + v_i'$. Applying the above three

conditions and taking the time mean of equation (1), we have

$$(2)\quad \bar{v}_1 \frac{\partial \bar{v}_1}{\partial x_1} + \bar{v}_2 \frac{\partial \bar{v}_1}{\partial x_2} = -\alpha \frac{\partial \bar{p}}{\partial x_1} - \frac{\partial}{\partial x_1} <v_1'v_1'> - \frac{\partial}{\partial x_2} <v_2'v_1'> - <v_2'v_1'> \frac{1}{w} \frac{\partial w}{\partial x_2}$$

where w is the width of the estuary at depth x_2.

Terms of the type $<v_i'v_j'>$ represent Reynold's or eddy stresses. In oceanography they have generally been replaced by friction terms of the type $-A_i \frac{\partial \bar{v}_j}{\partial x_i}$. Thus equation (2) is essentially the equation treated by Cameron [4] where $\frac{\partial w}{\partial x_2} = 0$ (i.e., in a fiord the sides are nearly perpendicular), and horizontal stress term $\frac{\partial}{\partial x_1} <v_1'v_1'>$ was neglected, and where $<v_2'v_1'>$ was replaced by $-A_2 \frac{\partial \bar{v}_1}{\partial x_2}$.

In a parallel study of the salt balance in the James River estuary, which will be discussed in the next chapter, it was found that the horizontal random flux of salt $<v_1's'>$ was negligible. Assuming, by analogy, that the horizontal eddy stress term $\frac{\partial}{\partial x_1} <v_1'v_1'>$ is also negligible, we can solve for the vertical friction component $<v_2'v_1'>$ in equation (2) in terms of measurable quantities. Thus

$$(3)\quad <v_2'v_1'> = -\frac{1}{w}\left[\int w\left(\bar{v}_1 \frac{\partial \bar{v}_1}{\partial x_1} + \bar{v}_2 \frac{\partial \bar{v}_1}{\partial x_2} + \alpha \frac{\partial p}{\partial x_1}\right) dx_2 + C\right]$$

The variables w, and $\bar{v}_1$ can be obtained directly from measurements. As will be shown in the next section, the mean vertical velocity, v_2, can be computed from observations of the horizontal velocity. Temperature and salinity observations can be employed to evaluate the relative distribution of the pressure gradient $\alpha \frac{\partial p}{\partial x_1}$. The absolute distribution of $\alpha \frac{\partial p}{\partial x_1}$, with depth is not obtainable unless the depth of the level pressure surface is known. Thus equation (3) contains two unknown constants, of which one is a function of the depth of the level pressure surface and the other is the constant of integration.

$<v_2'v_1'>$ must be zero at the surface and at the bottom. With these two boundary values it is possible to evaluate equation (3) numerically, and in doing so to determine the absolute longitudinal slope of the pressure surfaces.

Making use of the extensive observations obtained in the James River estuary, the integration on the right side of equation (3) was performed numerically at three sections. The depth of the level pressure surface, as determined from the boundary conditions placed upon $<v_2'v_1'>$, was in all cases closely associated with the observed depth of no net motion, being from $\frac{1}{2}$ to 2 feet deeper than the latter. The variation of $<v_2'v_1'>$ with depth is shown graphically in Fig. 7, for a typical station. As will

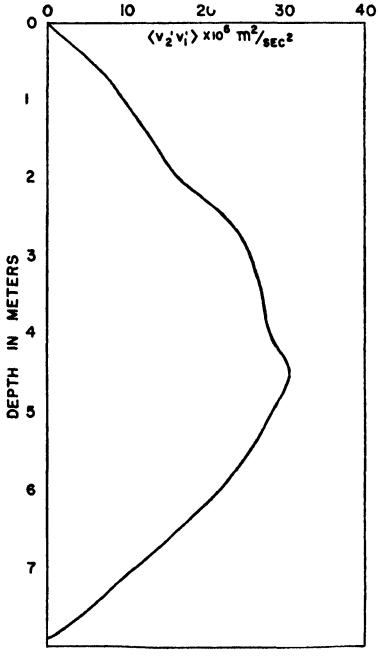


FIG. 7. Variation of the vertical random flux of momentum $<v_2'v_1'>$ with depth in the James River estuary.

be shown later, this variation with depth closely resembles the vertical variation in the random flux of salt $<v_2's'>$. There is some indication that the mean value of this eddy stress term is a function of tidal velocity.

Neglecting the horizontal random terms, the mean lateral component of the equation of motion may be written as

$$0 = -\alpha \frac{\partial \bar{p}}{\partial x_3} + f\bar{v}_1 - \frac{\partial}{\partial x_2} <v_2'v_3'> \tag{4}$$

where f is the Coriolis parameter. All the field acceleration terms involve the lateral velocity $\bar{v}_3$, and are assumed to be zero when averaged laterally across the estuary.

The mean longitudinal velocity $\bar{v}_1$ can be obtained from field observations. Temperature and salinity data allow the determination of the relative vertical distribution of the lateral pressure gradient. When equation (4) is solved for the eddy stress term $<v_2'v_3'>$ there will again be two unknown constants, one of which is related to the depth of the level pressure surface, the other being the constant of integration.

$<v_2'v_3'>$ must equal zero at both the top and bottom boundary, and hence it is possible to evaluate equation (4) numerically in the same manner as equation (3) was evaluated.

The depth at which $\frac{\partial p}{\partial x_1} = 0$ does not necessarily coincide with the depth at which $\frac{\partial p}{\partial x_3} = 0$. However, computations based on the observations in the James River estuary show that the depth of the pressure surface which is level with respect to the lateral coordinate closely corresponds to the depth of the pressure surface which is level with respect to the longitudinal coordinate.

The field acceleration terms in the longitudinal component of the equation of motion are relatively small in the James River estuary and the longitudinal pressure gradient is required primarily to balance the eddy stress term. On the other hand, the term involving the eddy stress in the lateral component of the equation of motion is relatively small, and the lateral pressure gradient is required primarily to balance the Coriolis force related to the mean horizontal motion.

It should be noted that if the width of the estuary increased appreciably in the longitudinal direction, there would be an added term in equation (1) related to this variation in width with x_1. Part of the longitudinal pressure gradient would then be associated with this divergence of the sides of the estuary.

Such a study as the one just described for the James River provides a relatively clear understanding of the relationships between the observed physical structure and the dynamics of the circulation in an estuary of this type. Let us now summarize the observed distribution of physical properties and the related dynamic structure.

The salinity increases with distance from the head of the estuary, where fresh river water enters. This horizontal gradient varies only slightly with depth. The salinity also increases with depth, there being a sharp salinity gradient at about mid-depth. (In the James River estuary the channel depth is about 30 feet, and the mean depth about 20 feet. The halocline, or layer of rapidly increasing salinity, was observed at between 8 and 15 feet.) The layer above the halocline has an observed net motion down the estuary while the relatively higher

salinity layer below the halocline has a net motion directed towards the head of the estuary. The observed depth of no net horizontal motion occurs close to, but does not necessarily coincide with the exact depth of the halocline.

Associated with the longitudinal increase of salinity from the head towards the mouth of the estuary, there is an increase in the volume transport. The volume of flow in the upper layer is increased at the expense of the volume of flow in the lower layer. Thus there is a net vertical flow directed upward, and having a maximum value near the depth of no net horizontal motion.

A pressure surface which is level with respect to the longitudinal coordinate occurs just below the depth of no net motion. Above this reference level the pressure surfaces slope downward in the positive x_1 direction, the upper surface having the largest slope. Below the reference surface the pressure surfaces slope downward in the negative x_1 direction, the slope increasing with depth.

Located just below the depth of no net motion is a pressure surface which is level with respect to the lateral coordinate. Above this reference surface the pressure surfaces slope downwards from the right side of the estuary toward the left side. Below the reference surface the pressure surfaces slope downward from the left side toward the right side of the estuary. Observations indicate that these two reference surfaces are separated by only a few feet. Indeed, considering the possible errors in observation of the pertinent variables and the errors in numerical integration of the equations used in determining these reference surfaces, they may actually be the same pressure surface.

The major portion of the longitudinal pressure gradient is required to balance the longitudinal component of the eddy stress. The remaining part of this pressure gradient is associated with the increase in volume transport down-estuary. This increase in volume transport is reflected either in a field acceleration term or in a term related to the divergence of the sides of the estuary.

The major portion of the lateral pressure gradient is required to balance the Coriolis force related to the mean longitudinal motion. The remaining part of this pressure gradient is associated with the lateral component of the eddy stress.

The eddy stress components $<v_2'v_1'>$ and $<v_2'v_3'>$ both have a bow-shaped vertical distribution, being zero at the upper and lower boundaries, and having a maximum value at some mid-depth. There is some indication that the magnitude of each of these terms is related to the tidal velocity.

It is hoped that studies of this type will lead to a more intelligent

choice of the type of theoretical model which might be employed in a study of estuarine circulation. Too often theoretical attacks are undertaken without sufficient knowledge of the actual structure. In such cases the mathematical simplifications necessary in the development of the theoretical model frequently involve the neglect of terms which may be important, and the inclusion of terms which may be unimportant.

7. The Flushing of Tidal Estuaries

In a positive type of estuary the fresh water which enters from the river will be mixed with the sea water and through processes of advection and diffusion will be distributed throughout the estuary. Because a mean state of volume equilibrium must be maintained, fresh water must pass out of the estuary during any given time period to an amount which is equal to the amount of fresh water introduced from the river during that time period. The river water introduced during, say, a particular tidal cycle will be mixed with the estuarine waters and will, at a rate governed by the natural processes of circulation and mixing, pass through the estuary and out into the ocean.

For purposes of discussion we will identify the fresh water in terms of the period of time since it left the river proper to enter the estuary. The age of any particular increment of river water within the estuary will be defined as the number of tidal periods since that particular increment entered the estuarine region from the river.

The estuary as a whole will contain river water of various ages. At any point in the estuary there will be a certain period of time before any part of the river water entering during any particular tidal cycle can arrive at that particular point. Hence each increment of the estuary will contain river water of all ages greater than a certain minimum age. This minimum age will be a function of the position of the increment.

Within any increment of the estuary the concentration of river water which entered the estuary during a particular tidal cycle will, after the initial period of time required for the first infinitesimal trace to reach the increment, increase to a maximum and then decrease asymptotically to zero. Theoretically the actual concentration of this particular river water in any increment of the estuary will never reach zero, but as its age increases the concentration in each increment of the estuary will ultimately become infinitesimally small.

Studies which deal with the rate at which the concentration of river water or of any dissolved or suspended pollutant within the estuary is decreased are called "flushing" studies. Dissolved or suspended pollutants will be distributed in and flushed from the estuary by the same processes which control the exchange of river water and sea water. For

this reason flushing studies, which have taken the foremost position among estuarine problems within the last two years, have dealt primarily with the distribution of fresh water, or conversely, of salinity.

Recent papers on the flushing problem have been published by Tully, by Ketchum, and by Stommel and Arons. The author has been engaged in a study of the processes which control the flushing from a coastal plain estuary, and the initial results of this as yet unpublished work will also be discussed here.

7.1. Flushing Parameters Determined from the Measured Concentration of Fresh Water

Both Tully [6] and Ketchum [5] have discussed the use of the concentration of fresh water within the estuary in determining the probable distribution of a pollutant. These treatments involve the concept of a *base* salinity as the salinity of the undiluted sea water which is mixing with the fresh water to produce estuarine water.

The concentration of river water within any increment of the estuary is determined by the relationship

$$C_F = \frac{S_b - S}{S_b} \tag{1}$$

where S_b is the base salinity of the sea water and S is the observed salinity of the estuarine water within the increment. This relationship is based on the assumption that the increment of estuarine water is a simple mixture of river water and sea water of a known base salinity.

Both Tully and Ketchum were concerned with the flushing of a pollutant which could be considered as having the same distribution as the fresh water. Ketchum defined a "flushing time" of a segment of the estuary as the *average* time required for the river water, with its contained pollution, to move through the segment. The displacement factor employed by Tully is essentially the inverse of the flushing time expressed in number of tidal cycles.

The flushing time for any segment of the estuary would be the total volume of river water within the segment, determined by the volume integral of equation (1), divided by the daily river flow. Thus, for a given river flow and observed salinity distribution, the average time it would take a pollutant introduced into the estuary to pass through the estuary can be determined. Also the mean concentration of the pollution within any segment of the estuary is given by a relationship between the initial concentration and volume of the pollution, the volume of river flow, and the concentration of river water within the segment.

The choice of the value of the base salinity is critical in the studies

carried out by Ketchum and by Tully. It is a particularly difficult task when treating an estuary which is tributary to a second larger estuarine system. The method provides only a gross picture of the flushing of a pollutant introduced with the river water in the estuary.

7.2. The Tidal Prism Concepts in Estuarine Flushing

Sanitary engineers have for some time used the concept of the *tidal prism* in studies of the flushing of pollutants from harbors. Phelps and Velz [9] discussed the use of this concept in an analysis of pollution in New York Harbor.

The assumption is made that the water brought into the estuary with the flood tide is completely mixed with the polluted estuarine waters. Since the ebbing tide would carry out a volume of water exceeding the volume of water brought in by the flood tide by an amount equal to the inflow of river water during the tidal period, there would be a proportion of the pollutant flushed out of the estuary during each tidal cycle which would be related to the ratio between the total tidal prism (i.e., the volume of water required to produce the observed rise in water level on the flooding tide) and the total volume of water in the estuary.

The fact that complete mixing of the waters within the estuary on each tide could not occur is evident from the observed salinity distributions. Ketchum [2] has presented an improvement in the tidal prism concept which alleviates this shortcoming to some extent. He considers the estuary divided into successive volume segments, the boundaries of which are defined in terms of the mean tidal oscillations. Within each of these segments it is assumed that there is complete mixing at high tide, and hence the tidal prism concept for the complete estuary is reduced to a local intertidal volume concept. Ketchum considers that a segment defined by the average length of the tidal excursion is the largest possible segment in which mixing by tidal action could be considered as complete.

A steady state distribution of fresh and salt water within the estuary is assumed. To maintain this steady state there must be no *net* exchange of salt across each complete cross-section of the estuary during a tidal cycle; but during the same period there must move seaward a volume of fresh water equal to the volume introduced by the river.

The inner end of the estuary is defined by Ketchum as the section above which the volume required to raise the level of the water from low to high water mark is equal to the volume contributed by the river during a tidal cycle. The area of the river above this section and up to the furthest point in the river at which there is a tidal rise and fall is defined as segment O. Each successive segment is defined so that the distance between the upper, or riverward, boundary and the outer, or seaward

boundary is equal to the average excursion of a particle of water on the flooding tide. The length of the estuary would be divided into such segments if, according to Ketchum, the high tide volume of each segment is equal to the low tide volume in the adjacent seaward one. Actually, such a relationship would hold only for estuaries in which the tidal wave is essentially a standing wave. In a large estuary, where the tidal wave has largely progressive wave features, some other means of defining the successive segments would be required.

Designating the volume of river water introduced during each tidal cycle as R, the local intertidal volume as P, and the low tide volume of each segment as V, the above concepts may be briefly stated as follows:

1. The inner end of the estuary is defined as the section above which $P_0 = R$

2. The limits of each successive volume segment (n) are defined such that

$$V_n = V_0 + R + \sum_{1}^{n-1} P \tag{1}$$

Within each of these segments it is assumed that the water is completely mixed at high tide. Therefore, the proportion of water removed on the ebb tide will be given by the ratio between the local intertidal volume and the high tide volume of the segment. This same proportion of river water or of any dissolved or suspended material in the water will be removed by the ebb tide. Ketchum defines an exchange ratio for each segment, then, which has the form:

$$r_n = \frac{P_n}{P_n + V_n} \tag{2}$$

A volume of river water R is received in each segment on each tidal cycle. Though the volume of river water leaving on the next tide will also be R, it will not be composed of the same river water mixture which entered on the previous tide. If R_1 designates the river water arriving at a particular segment on the current tidal cycle, then $r_n R_1$ will leave the segment on the next ebb, and $(1 - r_n)R_1$ will be left behind. The river water which leaves a segment on each ebb tide will be composed of a portion r of the river water mixture accumulated during many tidal cycles.

Ketchum shows that for a constant river flow and after a large number of tidal cycles the total volume of river water Q_n accumulated within any volume segment is given by:

$$Q_n = \frac{R}{r_n} \tag{3}$$

It is possible, from tidal heights, river flow, and the topography of the estuary to divide the estuary into volume segments as defined by equation (1). From these same data the exchange ratio r can be computed from equation (2) for each segment. By means of equation (3) the concentration of river water in each volume segment can be determined. The concentration of any pollutant which entered the estuary with the river water may also be computed.

Ketchum has obtained excellent agreement between the computed distribution of fresh water and the observed mean distribution of fresh water in three different estuaries. There are, however, certain features of this theory which will require further consideration before the procedure could be applied with confidence to all types of positive estuaries.

One serious question arises in regard to the assumption of complete vertical mixing. Ketchum states that "incomplete vertical mixing can be detected readily if salinity observations are available, since water diluted by river effluent will be limited to the upper layers of the water column. Only the mixed volume of water should be considered, in such a case. . . . " Even though salinity observations in many estuaries show that vertical mixing is far from complete, yet river water will occur in appreciable quantities in both the upper, less saline layer and the lower, more saline water. The lower layer with its net up-stream motion will actually be carrying river water up the estuary. Ketchum does not consider this case, which includes, in fact, a considerable number of the coastal plain estuaries.

This development by Ketchum is of considerable value to engineers, since it constitutes a great improvement over the tidal prism technique previously employed in studies of harbor pollution.

7.3. A Mixing Length Theory of Tidal Flushing

Arons and Stommel [10] have presented a theoretical attack on flushing based on Ketchum's fundamental idea that the elemental mixing volume is bounded by the length of the tidal excursion. Their assumptions succeed in reducing the problem of tidal flushing to a simple one-dimensional case.

The model estuary studied is considered to have uniform width w, depth H, and length L. The origin is placed at the head of the estuary where the salinity is zero. The single coordinate, x, is directed positively downstream. The salinity at the mouth of the estuary, where $x = L$, is maintained at the salinity of the open ocean $s = \sigma$.

The tidal rise and fall is assumed to be simultaneous and uniform over the entire channel, with an amplitude of ζ_0, and a tidal current velocity $\mathbf{U}_0 = \frac{\zeta_0 \omega x}{H}$, where ω is the angular frequency of the tide.

For a steady state the equation describing the mean salinity distribution for this simple one-dimensional model is

$$u\frac{\partial s}{\partial x} = \frac{\partial}{\partial x}\left\{A\frac{\partial s}{\partial x}\right\} \tag{1}$$

where s is the time mean salinity at any point x, u is the time mean velocity at x, which is assumed to be equal to that portion of the flow due to the river only, and A is the eddy diffusivity of salt along the x axis.

The eddy diffusivity, A, has classically been considered to be related to a characteristic mixing velocity and to a characteristic mixing length. Arons and Stommel have considered that the characteristic velocity is the amplitude of the tidal velocity, $\mathbf{U}_0$, and that the characteristic length is the total excursion of a particle due to the tides, being given by $2\xi_0 = 2\frac{\zeta_0 x}{H}$. Thus

$$A = 2B\xi_0\mathbf{U}_0 = \frac{2B\zeta_0^2\omega}{H^2}x^2 \tag{2}$$

where B is a non-dimensional constant of proportionality. The second form for A is obtained by substituting for the tidal velocity $\mathbf{U}_0$ and the tidal displacement ξ_0. With this function of x substituted for A in equation (1), Arons and Stommel obtain a solution for the distribution of salinity within the estuary in the form

$$\frac{s}{\sigma} = e^{F\left(1-\frac{1}{\lambda}\right)} \tag{3}$$

where $\lambda = \frac{x}{L}$ is the non-dimensional length parameter, and

$$F = \frac{aH^2}{2B\zeta_0^2\omega L} \tag{4}$$

is a non-dimensional parameter which they term the "flushing number." In this expression for F, a is the mean velocity in the estuary due to the river and is equal to $\frac{D}{wH}$ where D is the river discharge in volume per unit time.

Arons and Stommel presented a family of curves representing the solution of equation (3) for various values of the flushing number, F. The coordinates for these curves were $\frac{s}{\sigma}$ and λ, and hence represent the distribution of salinity along the estuary for various values of F.

Empirical data from the Raritan River estuary in New Jersey and the Alberni Inlet in British Columbia were plotted on the same graph. The

Alberni Inlet data fell closely along the curve for $F = 0.3$ and the Raritan River data fell along the curve $F = 0.8$. Since the model employed was a highly simplified one, it is somewhat encouraging to find that empirical data fit the family of curves with such good agreement.

Equation (4) expresses the flushing number in terms of the river flow, the dimensions of the estuary, the tidal characteristics, and the proportionality factor B. With the exception of B, these factors can be readily determined for any subject estuary. The proportionality factor B was introduced as a constant, and hence could be computed from one known value of F.

Since the two sets of empirical data plotted along two different curves of the family of solutions to equation (3), two determinations of B could be made. Arons and Stommel attempted to make such calculations, and found that the proportionality factor B varied by an order of magnitude for the two cases. This would indicate that though the shape of the theoretical curves is in good agreement with the observations, a calculated value of the flushing number based on equation (4) could not be used as an index of tidal flushing.

7.4. A Quantitative Study of the Salt Balance in a Coastal Plain Estuary

The theoretical treatment of the salinity distribution in a positive estuary by Arons and Stommel, discussed in the previous section, was based on a simplification of the salt balance equation. The terms retained were those suggested by the success of Ketchum's empirically developed theory of the exchange of salt and fresh water in a tidal estuary.

In the coastal plain estuaries of the type found in the Chesapeake Bay system, the processes which are primarily responsible for maintaining the salt balance appear to be different than those proposed by Arons and Stommel for an estuary which is vertically homogeneous. In the Chesapeake Bay and its tributary estuaries the lower layer is of higher salinity than the upper layer, and has a net motion up the estuary, as compared with the down-estuary flow of the surface waters. The author has been engaged in a study of the processes which control the exchange of fresh and salt water in such an estuary. The results of the first part of this investigation, which has dealt with conditions in the James River estuary, are of sufficient significance to be reported here.

The coordinate system to be employed is that described in Section 3. x_i will be used to designate the three coordinate axes, x_1 (longitudinal), x_2 (vertical), and x_3 (lateral). v_i then is the vector velocity with the three components, v_1, v_2, and v_3.

Neglecting molecular diffusions, the instantaneous local rate of change of the salt concentration, s, is given by

$$\frac{\partial s}{\partial t} = -\frac{\partial (v_i s)}{\partial x_i} \tag{1}$$

Introducing the sum of a mean and a random term for the instantaneous values of velocity and salinity, and taking the time mean of equation (1), we have

$$\left\langle \frac{\partial s}{\partial t} \right\rangle = -\bar{v}_i \frac{\partial \bar{s}}{\partial x_i} - \frac{\partial}{\partial x_i} \langle v_i' s' \rangle \tag{2}$$

This equation states that the time mean local change of concentration results from two types of terms, the first related to the mean advection, and the second related to the random flux. Oceanographers have called this latter process "eddy diffusion," and have replaced the mean products $\langle v_i' s' \rangle$ by the so-called diffusion terms having the general form $-A_i \frac{\partial \bar{s}}{\partial x_i}$ where A_i represents the eddy coefficient of diffusion along the three coordinate axes. The basic equation used by Arons and Stommel made use of only the longitudinal advective term and the longitudinal diffusion term.

There is some question as to the physical reality of introducing the concept of eddy diffusivity. Some investigators feel that there is no basis for using terms of the type $-A_i \frac{\partial \bar{s}}{\partial x_i}$. In any case, little is known of the actual character of the diffusivity, and the result of replacing $\langle v_i' s' \rangle$ with $-A_i \frac{\partial \bar{s}}{\partial x_i}$ is merely to replace one group of terms about which we know very little with a second group of terms about which we also know very little. For this reason we will retain the terms $\langle v_i' s' \rangle$.

There are three advection terms and three random flux terms in equation (2). When this equation is applied to a relatively narrow estuary such as the James River, the mean values of the parameters with respect to the lateral coordinate only need be taken, and the only lateral velocity that needs to be considered is that component of the velocity which is related to the variations in width of the estuary. The change in width with longitudinal distance, x_1, is sufficiently small to be neglected. However, the change in width with depth is of considerable importance. It can be shown that under these conditions equation (2) takes the form

$$\left\langle \frac{\partial s}{\partial t} \right\rangle = -\bar{v}_1 \frac{\partial \bar{s}}{\partial x_1} - \bar{v}_2 \frac{\partial \bar{s}}{\partial x_2} - \frac{\partial}{\partial x_1} \langle v_1' s' \rangle - \frac{\partial}{\partial x_2} \langle v_2' s' \rangle - \langle v_2' s' \rangle \frac{1}{w} \frac{\partial w}{\partial x_2} \tag{3}$$

where w is the width of the estuary at depth x_2.

An extensive time series of observation of velocity and of salinity at several cross-sections of an estuary provide empirical values for $\bar{v}_1$, $\frac{\partial \bar{s}}{\partial x_1}$, and $\frac{\partial \bar{s}}{\partial x_2}$ as functions of depth at discrete values of x_1. From a consideration of the volume continuity the mean vertical velocity $\bar{v}_2$ as a function of depth can be obtained from observations of the longitudinal velocity $\bar{v}_1$ at successive positions in the estuary. Thus there remain the random flux terms to be evaluated.

It is possible to apply an integrated form of the salt balance equation to a segment of the estuary in which the boundaries are so adjusted that only the longitudinal (x_1) components need to be considered. Such an application allows the comparison of the longitudinal random flux term $\frac{\partial}{\partial x_1} <v_1's'>$ with the mean longitudinal advective term, which can be computed from measurements.

The last two terms in equation (3) contain the vertical random flux term $<v_2's'>$. The equation may be solved for this term, the solution being an integral equation involving the two advective terms, the longitudinal random flux term, and the width of the estuary. The data obtained in the field studies in the James River were employed in evaluating the magnitude of the various terms in equation (3).

The field program was undertaken during a period which approximated the steady state in salinity distribution. The mean local time variation $\left\langle \frac{\partial s}{\partial t} \right\rangle$ was less than 10^{-7}‰/sec.

The mean longitudinal advective term $\bar{v}_1 \frac{\partial \bar{s}}{\partial x_1}$ was very significant, varying from about $+5 \times 10^{-5}$‰/sec in the surface layers to about -5×10^{-5}‰/sec in the lower layers. In comparison, the longitudinal random flux term $\frac{\partial}{\partial x_1} <v_1's'>$ was negligible, being less than 10^{-7}‰/sec.

The mean vertical advective term $\bar{v}_2 \frac{\partial \bar{s}}{\partial x_2}$ was considerably smaller than the longitudinal term, but was still significant, particularly at middepths where it attained a maximum of about 1×10^{-5}‰/sec per second.

The vertical random flux term $\frac{\partial}{\partial x_2} <v_2's'>$ appears to be of the same order of magnitude as the mean longitudinal term, and hence highly significant. In this analysis of James River data it had values of about -4×10^{-5}‰/sec in the surface layers and approximately $+4 \times 10^{-5}$‰/sec in the deeper layers.

The last term in equation (3), which involves the variation of the

width of the estuary with depth, was of the same order of magnitude as the vertical advective term.

An evaluation of equation (3) was made for three different sections, each of which was studied over three periods of observations. One of the observational periods was taken during neap tides, the second during spring tides, and the third between neap and spring tides.

An analysis of these nine cases revealed that the vertical random flux of salt $<v_2's'>$ has a bow-shaped vertical distribution (see Fig. 8). The

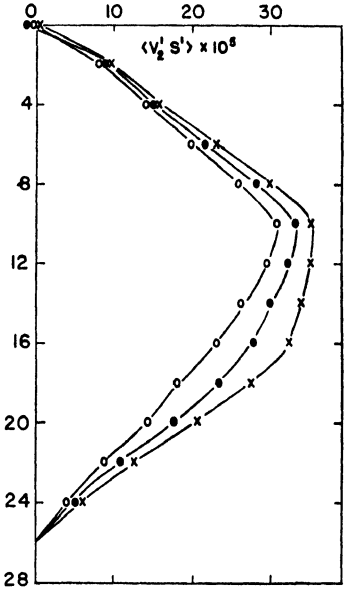


FIG. 8. Variation of the vertical random flux of salt $<v_2's'>$ with depth in the James River estuary for three different periods.

mean vertical magnitude varies both with position and with time, and the time variation appears to be a linear function of tidal velocity. This latter fact is significant since it shows that vertical mixing is related to the tidal velocities rather than the mean velocities. Figure 9 presents graphically the vertical variation of the various terms in equation (3) under steady state conditions, for a typical position in the James River estuary. The longitudinal advective term and the vertical random flux term dominate. Of secondary, but still significant importance are the mean vertical advective term and a term related to the variation of the width of the estuary with depth. The horizontal random flux term is negligible.

Any study of the flushing of pollution from an estuary of this type

must take into consideration the results of this study, since the same processes which control the distribution of salinity would also control the distribution of the pollutant.

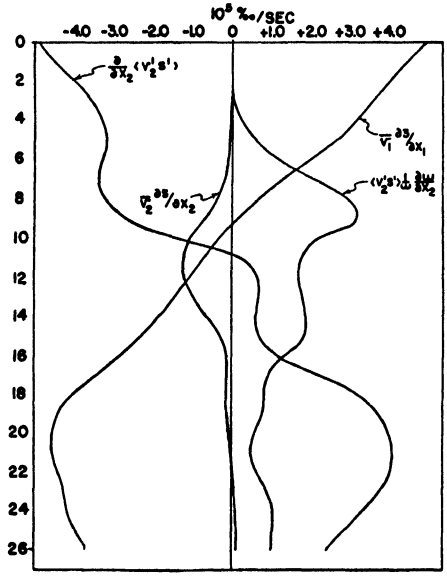


FIG. 9. Variation of the significant terms in equation (3) with depth in the James River estuary.

LIST OF SYMBOLS

x_i tensorial notation for the three coordinate axes
x_1 coordinate axis directed horizontally down the central axis of the estuary, from the origin in fresh water at the head of the estuary towards the sea
x_2 coordinate axis directed vertically downward
x_3 coordinate axis directed laterally across the estuary
v_i vector velocity
v_1, v_2, and v_3 velocity components along the coordinate axes x_1, x_2 and x_3 respectively
α specific volume of the estuarine waters
p pressure
m_1, m_3 frictional stress components along the x_1 and x_3 coordinate axis, respectively
g force of gravity
f Coriolis parameter
ρ density of the estuarine waters
ρ_u density of the upper layer
ρ_l density of the lower layer
ζ_l depth of boundary between upper and lower layer
ζ_u elevation of free surface

D depth of upper layer; also, river discharge in volume per unit time
v_u mean horizontal velocity in upper layer
v_m vertical velocity of mixing of deep water into upper layer
T transport in the upper layer
$b = v_u^3/k$
$k = g\gamma_0 \frac{T_0}{2}$ (after Stommel); also, constant in stream function assumed by Cameron
v' Keulegan's critical velocity
A_v eddy coefficient of viscosity
ψ stream function
θ anomaly ratio, and equals $1 - \rho/\rho_s$
ρ_s density undiluted sea water
$\bar{v}_i$ mean vector velocity
v_i' random vector velocity
$\bar{v}_1, \bar{v}_2, \bar{v}_3$ mean velocity components
v_1', v_2', v_3' random velocity components
$< >$ time mean symbol
w width of estuary
$<v_2'v_1'>, <v_2'v_3'>$ components of the random flux of momentum
C_F concentration of fresh water
s salinity
s_b base salinity of sea water
R volume river water introduced into estuary on each tidal cycle
P local intertidal volume
V low tide volume of particular segments of the estuary
r exchange ratio
Q_n total volume river water accumulated in each segment of estuary
σ salinity of undiluted sea water
ζ_0 tidal amplitude
U_0 amplitude of tidal velocity
ω angular frequency of the tide
H depth of channel
L length of estuary
u velocity
ξ_0 tidal displacement
λ non-dimensional length parameter
F flushing number
B proportionality factor
$<v_2's'>, <v_1's'>$ components of the random flux of salt

References

1. Finch, V. C., and Trewartha, G. T. (1942). Elements of Geography. McGraw-Hill, New York.
2. Ketchum, B. H. (1951). The exchange of fresh and salt water in tidal estuaries. *J. Mar. Res.* **10**, 18–38.
3. Stommel, H. (1951). Recent developments in the study of tidal estuaries. Technical Report, Woods Hole Oceanographic Institution, Ref. No. 51–33.

4. Cameron, W. M. (1951). On the dynamics of inlet circulations. Doctoral Dissertation, Scripps Institution of Oceanography, University of California.
5. Ketchum, B. H. (1950). Hydrographic factors involved in the dispersion of pollutants introduced into tidal waters. *J. Boston. Soc. Civ. Eng.* **37**, No. 3, 296–314.
6. Tully, J. (1949). Oceanography and prediction of pulpmill pollution in Alberni Inlet. *Bull. Fish Res. Board. Canada* **83**, 169 pp.
7. Collier, A., and Hedgpeth, J. W. (1950). An introduction to the hydrography of tidal waters of Texas. *Publ. Inst. Mar. Sci.* **1**, No. 2, 123–194.
8. Keulegan, G. H. (1949). Interfacial instability and mixing in stratified flows. *J. Res. Natl. Bur. Stand.* **43**, PR 2040, 487–500.
9. Phelps, E. B., and Velz, C. J. (1933). Pollution of New York Harbor. *Sewage Works J.* **5**, No. 1, 117–157.
10. Arons, A. B., and Stommel, H. (1951). A mixing length theory of tidal flushing. *Transact. Am. Geophys. Un.* **32**, No. 3, 419–421.

The Earth's Gravitational Field and Its Exploitation

GEORGE PRIOR WOOLLARD

University of Wisconsin, Madison, Wisconsin and Woods Hole Oceanographic Institution, Woods Hole, Massachusetts

CONTENTS

	Page
1. Introduction	281
1.1. Historical Development of Gravitational Theory	281
1.2. The Experimental Determination of "g"	282
1.3. The Mathematical Determination of "g"	284
2. The Exploitation of Gravity	286
2.1. General Facts Concerning Earth's Gravitational Field	286
2.2. Factors Contributing to Gravity Anomalies	288
2.3. Accuracy of Gravity Observations	290
3. The Exploitation of Gravity Measurements in Geodetic Studies	293
3.1. Types of Measurements	293
3.2. Practical Significance of Gravity Measurements in Geodetic Work	296
3.3. National Gravity Base Values	297
3.4. Work at Sea	300
4. Geologic Uses of Gravity Data	301
4.1. Earth Crustal Studies	301
4.2. Geological Exploration	303
4.3. Gravity Studies of the Strength of the Earth	304
Appendix	305
List of Symbols	310
References	310

1. INTRODUCTION

1.1. Historical Development of Gravitational Theory

Our present knowledge of the earth's gravitational field can be said to date from the experiments of Galileo from the leaning tower of Pisa in 1590 when he demonstrated that the rate of acceleration of falling bodies is constant and not a function of mass. The so-called law of gravitational attraction relating the force between two bodies as being directly proportional to the ratio of the product of their masses to the square of the distance between them was not propounded until 1686 when Sir Isaac Newton deduced it from a study of Keppler's empirical laws of motion for the planets.

$$F = \gamma \frac{m_1 m_2}{r^2} \tag{1}$$

F = Force of attraction between masses m_1 and m_2, r = distance apart, and γ = constant whose value depends on the units used. The role of the publicized falling apple actually was a very minor one although if it had fallen from a high limb and hit Newton on the head, the case for its significant contribution would have been considerably strengthened.

The experimental verification of Newton's premise, however, did not take place until a little over a hundred years later when Cavendish in 1798 proved it to be correct using a torsion balance and established for the first time the value of the gravitational constant, γ. Since that time this fundamental number of nature has been established by many investigators, one of the most recent and accurate values being that of Heyl [1] at the National Bureau of Standards in Washington. The present adopted value is $\gamma = 6.668 \times 10^{-8}$ cgs units.

Another of Newton's deductions having a direct bearing upon our knowledge of gravitation was that although the earth appeared to be spherical, it was flattened at the poles because of the centrifugal force of its rotation. At the time that he advanced this idea it had just been demonstrated through measurements of the length of a degree in latitude in northern and southern France that if anything, the opposite was true, i.e. that the equatorial radius of the earth was shorter than the polar radius. This theory of Newton's in direct opposition to what appeared to be "scientific proof," naturally led to some dissension in scientific circles. The upshot was that the French Academy in 1755 sent an expedition to Peru (now Ecuador) near the Equator and another to northern Lapland to measure the length of a degree of latitude in these places and thus determine beyond a reasonable doubt if the earth were spherical or not, and if not, in which direction it was flattened. Newton's theory was vindicated and this so impressed Voltaire that he was led to write, "they who the frozen wastes did roam found that which Newton had foretold who stayed at home."

With the degree of polar flattening indicated by these measurements it was not long before the best mathematical brains of the time were working on the idea of how to deduce the value of gravitational attraction at any place, and the physicists were working on the problem of how best to measure gravitational attraction.

1.2. The Experimental Determination of "g"

The development of the first gravity measuring instrument came about somewhat by accident. A clockmaker in Paris, Jean Richer, built a new pendulum clock for the astronomic observatory in French Guiana.

This clock, while it kept perfect time in Paris, lost 2½ minutes a day in French Guiana. Yet when it was returned to Paris it was found to keep perfect time again. This phenomenon came to the attention of Pierre Bouguer who was a member of the French Academy expedition then outfitting to go to Peru and he had a crude pendulum apparatus built to take along. If Newton were right both in his theory of polar flattening and gravitational attraction varying inversely as the square of the distance, then the pull of gravity should vary with both latitude and elevation.

Apparently there is no record of the difference in gravitational attraction Bouguer was able to measure between France and South America, but he did discover a marked difference in period of his pendulum apparatus at the base of the Andes and at Quito in the high Andes. The difference indicated was much different than would be predicted by theory. It corroborated some earlier measurements of his on the attraction of the plumb bob by the Andes. These indicated that this mountain range was underlain at depth by a deficiency in mass. Naturally, at the time, this evidence of anomalous conditions beneath the mountains was suspected of being in error and it was many years before it was actually substantiated beyond any reasonable doubt. Despite the skepticism regarding Bouguer's measurements attention was called to the utilization of pendulums for measuring gravity. The method looked quite feasible since the mechanical difficulties of making a simple pendulum of known dimensions and near frictionless movement had already been thoroughly studied by the skilled clockmakers of the day. The principal difficulty appeared to lie in accurate measurement of the period of oscillation. With the development of pendulum gravity apparatus it soon became apparent, however, that other considerations such as temperature and barometric effects, the sway of the supports, flexure of the pendulum shaft, air dampening, and wear of the knife edges had to be considered, and with some instruments also changes in the earth's magnetic field. It was also found that reliable absolute gravity values could not be determined with a simple pendulum because the equivalent length (l) of a simple physical pendulum must be expressed in terms of its moment of inertia (k), mass (m), and the distance from the point of rotation to the center of gravity (s), therefore is not a simple measurement of length as called for in the theory of a mathematical simple pendulum whose period (T) is expressed

$$T = 2\pi \sqrt{l/g} \tag{2}$$

For a simple physical pendulum

$$T = 2\pi \sqrt{k/mgs} \tag{3}$$

It was not until the development of invariable length reversible pendulums whose equivalent length can be measured directly as the distance between the two knife edges, that reliable values of "g" were obtained. Without going into details of instrument construction or the method of observation and reduction of data, let it suffice to say that present day absolute measurements of gravity have an accuracy of about 5 parts in 1,000,000. Relative changes in gravity on the other hand can be measured by means of spring gravimeters to 1 part in 1,000,000,000.

1.3. The Mathematical Determination of "g"

The mathematical approach to the problem of determining a gravity value anywhere was not possible until after the work of Cavendish established the gravitational constant and experimental observations of values of gravitational acceleration were feasible. Further, the development of Clairaut's equation relating the theoretical gravitational attraction at the surface of the earth for a given figure and mass for the earth and the centrifugal force of rotation was a necessary prerequisite.

Newton had previously shown that the force acting on a unit mass body m at the surface of the earth was the same as the acceleration that would be caused by gravitational attraction if the body were allowed to fall according to the second law of motion,

$$F = m\alpha \tag{4}$$

where α is the acceleration and m the mass. Substituting for F the gravitational relation,

$$F = \gamma \frac{mM}{r^2} \tag{5}$$

and letting m equal a unit mass, M the mass of the earth (considered concentrated at its center), r the radius of the earth and replacing α with g, the gravitational acceleration is given by

$$g = \frac{F}{m} = \gamma \frac{M}{r^2} \tag{6}$$

When Cavendish experimentally determined γ it was then possible, using experimental values of g, to determine M since the size of the earth was known from geodetic determinations of the length of a degree of latitude. The only missing factor was σ, the density of the earth. However, by approximating the earth's shape as that of a sphere it was possible to determine the density (σ) from

$$g' = \gamma \tfrac{4}{3}\pi r\sigma \tag{7}$$

where g' is the observed acceleration at the surface plus the normal outward component of centrifugal acceleration. This was necessary as equation (6) applies only to a non-rotating earth.

Incidentally the value of $\sigma = 5.5\ g/cm^3$ turned out to be much greater than anticipated since the density of all observable rocks in the earth averages somewhat less than $3.0\ g/cm^3$.

With the mass of the earth and an equatorial value of gravity established it was possible for the first time to actually predict the probable value of gravity at any one place on the earth using the theorem of Clairaut which has the form

$$f + b' = \tfrac{5}{2}c' \tag{8}$$

where f is the geometric polar flattening of the earth which is expressed by

$$f = \frac{a - c}{a} \tag{9}$$

in which a = equatorial radius and c = polar radius.

b' is the gravitational flattening of the earth which is expressed by

$$b' = \frac{g_c - g_a}{g_a} \tag{10}$$

in which g_c = polar gravity value and g_a = equatorial gravity value. b' also is the coefficient in the expression

$$g_0 = g_a(1 + b' \sin^2 \phi) \tag{11}$$

in which ϕ is any latitude and g_0 is the gravity value at that latitude.

c' is the ratio of the equatorial centrifugal acceleration to the equatorial gravitational acceleration at the equator.

$$c' = \omega^2 \frac{a}{g_a} \tag{12}$$

in which ω = angular velocity of rotation and a = equatorial radius.

At present on the assumption that the earth has a shape approaching that of an ellipsoid of revolution whose equatorial radius is 6,356,909 meters and with a value of $f = 1/297.0$, the sea level value of gravity can be calculated at any place from the expression

$$g_0 = 978.049(1 + 0.0052884 \sin^2 \phi - 0.0000059 \sin^2 2\phi) \tag{13}$$

This is the International Gravity Formula in which the constant 978.049 is a statistically determined value for the equatorial sea level value of gravity (g_a) based upon observational data. The constants for

the term $\sin^2 \phi$ incorporate the effect of both geometric flattening and increase in centrifugal force as the poles are approached. The term $0.0000059 \sin^2 2\phi$ is a correction for non-conformity to the spheroidal shape assumed for a rotating body, which reaches a maximum at 45° latitude.

Actually on the basis of gravity anomaly values the earth probably more exactly approaches a triaxial ellipsoid with an elliptical equator whose major axis lies at about 10°W of Greenwich. The difference in equatorial radii has been estimated to be, however, only 150 ± 58 meters. An expression for theoretical sea level gravity considering this term is given by

$$(14) \quad g_0 = 978.046[1 + 0.005296 \sin \phi + 0.0000116 \cos^2 \phi(2\lambda + 10°) - \sin^2 2\phi(0.000007)]$$

in which λ = longitude.

However, since the equatorial ellipticity is so small most computations at present are based on the International Formula.

2. The Exploitation of Gravity

2.1. General Facts Concerning the Earth's Gravitational Field

Before proceeding with a discussion of the ways in which gravitational attraction is now being utilized it may be well to recapitulate the fundamental facts concerning the earth's gravity field since all were not brought out in the Introduction. Briefly they are as follows:

(a) The gravity field vector is peculiar in that the three space components are very unequal. The horizontal components are so small as to be negligible and the vertical component essentially equals the total vector.

(b) The force of gravity, i.e., the pressure exerted in dynes, by 1 gm on its base is numerically equal to the acceleration of gravity measured in the same units. This was shown in the Introduction by equating the force of gravitational attraction (4) $\left(F = \gamma \frac{mM}{r^2}\right)$ to the force in Newton's second law of motion (6) $F = m\alpha$ whereby $\alpha = g = \gamma \frac{M}{r^2}$.

(c) The potential of the gravity field is a scalar quantity whose first negative derivatives with respect to the space coordinates represent the components of gravity.

(d) The gravity potential at the earth's surface can be defined as work performed by a mass of 1 gm falling from space on the earth. Since work is the product of a force and distance the attraction potential for a 1 gm

mass is

$$V = Fr = \left(\gamma \frac{M}{r^2}\right)(r) = \gamma \frac{M}{r} \tag{15}$$

Gravity potential may also be defined as energy potential of a unit body whose weight is mg at an elevation (r) on the surface of the earth. The potential is thus the same considered from either standpoint

$$V = (1)\left(\gamma \frac{M}{r^2}\right)(r) = \gamma \frac{M}{r} \tag{16}$$

(e) For a rotating earth the potential of centrifugal force must be added to that of the attraction potential to obtain the actual potential. This additional potential is expressed by

$$V' = \tfrac{1}{2}\omega^2(x^2 + y^2) \tag{17}$$

in which ω = angular velocity of rotation of the earth = $2\pi/86{,}164$ ft/sec and x and y = space coordinates of the point outside of the rotating earth from the axis of rotation for which the potential is being determined. The total potential is expressed as

$$U = V + V' \tag{18}$$

(f) For any point outside a rotating mass the potential function is finite and continuous and follows Laplace's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} - 2\omega^2 = 0 \tag{19}$$

(g) Points of the same potential may be connected to form an equipotential surface. Such a surface constitutes a "level" surface which everywhere is at right angles to the force of gravity. There is no force component along an equipotential surface.

(h) An equipotential surface does not constitute a surface of equal gravity and the force of gravity may change along any equipotential surface. For example, the ocean surface approximates an equipotential surface everywhere perpendicular to the force of gravity, but the gravitational attraction over the ocean is quite variable.

(i) It is not possible to compensate for the earth's gravitational field as is the case in magnetic measurements. As a consequence, in order to exploit the earth's gravitational field for purposes other than in straight problems in physics involving falling bodies, it is necessary to determine anomalies; that is, differences between observed values and those computed for what the gravitational attraction should be at the point of observation.

2.2. Factors Contributing to Gravity Anomalies

Any measured value of gravity is a composite of many mass effects and these, approximately in the order of their importance, are as follows:

1. The shape and size of the earth as a whole.
2. The rotation of the earth.
3. Changes in elevation.
4. Horizontal mass discontinuities caused by changes in the thickness of the outer crustal layers of the earth.
5. Horizontal changes in mass caused by changes in density of rocks in the crystalline complex beneath the surface sediments.
6. Horizontal changes in mass caused by changes in the density of the surface and near surface rocks.
7. Changes in topographic configuration of the buried crystalline rock surface.
8. Changes in the relief of the surrounding surface topography.
9. Response of the earth to tidal forces and changes in barometric pressure.

A better appreciation of the relative importance of the above effects can be gained from a comparison of the gravitational effects they produce. Roughly the acceleration of gravity at the surface of the earth is taken as 980 cm/sec^2 for most problems in physics involving falling bodies. More exactly its sea level value varies from approximately 978.049 cm/sec^2 at the equator to 985.221 cm/sec^2 at the poles. This change incorporates both the effect of change in shape of the earth and the changes in centrifugal force and represents the major changes in gravitational force present. The effect of all other factors is not apt to exceed 1 cm/sec^2.

Because of the small magnitude of the gravitational effects produced by all the other factors listed as influencing the force of gravity a different set of units from the standard cgs values has been adopted for working with gravity anomalies. The principal unit is the gal, named in honor of Galileo, which is the equivalent of an acceleration of 1 cm/sec^2 or a force of 1 dyne. The common units used are the milligal (mgal) = 0.001 gal and the gravity unit = 0.1 mgal. Most gravity anomalies in geodetic work are expressed in mgals and those in geologic exploration work expressed in gravity units. An error in position of 1 mile will have an effect of approximately 1 mgal (0.001 cm/sec^2) in the anomaly.

The effect of changes in elevation amounts to 0.3086 mgal/meter or 0.09406 mgal/foot. The sign of the correction is (−) for increases in elevation above the reference level and (+) for elevation below the reference level. Actually the effect of changes in elevation is less than

that indicated because of the mass effect of the material included between the point of observation and the reference level.

This mass effect is expressed by $2\pi\gamma\sigma h$, the gravitational effect of a semi-infinite slab of h thickness and σ density. In metric units it is 0.04185σ mgals/meter and in English units 0.01276σ mgal/foot. The sign of this correction is always opposite to that for the elevation correction alone. The net effect for both elevation and included mass for an increase in elevation of 1 foot assuming normal crustal rocks with a density of 2.67 gm/cm^3 is -0.06 mgal.

The effect of a change in crustal structure in the outer portions of the earth shows up in the anomalies obtained when computed values of gravity allowing for position and elevation plus the mass above sea level are compared with observed values. It is found for example, that there is an apparent thickening of the low density outer crustal layer beneath the major mountain ranges. This results in anomalies in these areas as great as -400 mgal (0.4 gal). These findings substantiate the indications discovered by Bouguer in the Andes in 1740 and 1755, and anomalies so determined are termed Bouguer anomalies in honor of the original discoverer. In general it can be said that there is a direct relation between the height of land and the magnitude of the Bouguer anomalies.

The gravitational effect of horizontal changes in lithology within the crystalline rock complex beneath the sediments seldom amounts to as much as 100 mgal and usually is much less. The effect for example of a volcanic pipe two miles in diameter and of considerable depth may not exceed 30 mgal.

The effect of buried topography seldom amounts to 20 mgals and usually is less than 5 mgal. The Nemeha Ridge, a buried granite ridge for example with over 5000 foot relief in Kansas, has a gravitational effect of only 2 mgal.

The effect of surface and near-surface changes in rock density usually amounts to less than 10 mgal. In general, these changes are removed in computing Bouguer anomalies by changing the density value (σ) in the mass correction term.

The effect of topographic irregularities will depend upon the degree of relief present but more particularly on the actual relief at the point of observation. Just a matter of a couple of hundred feet difference in position may make a considerable difference in the gravitational effect of the surrounding topography. A useful rule of thumb is that if the angle between the horizontal and the top or bottom of the adjacent topographic feature is less than 10° the topographic effect will not exceed 1 mgal. Corrections for observations on a precipitous slope, however, may amount to 20 mgal or more. Regardless of whether the adjacent

topographic feature is a hill or valley, the sign of this correction is always negative. If a hill rises above the point of observation, its mass effect counteracts that of the earth as a whole. If a valley is present, there is a deficiency in mass attraction from that computed in the simple Bouguer anomaly reduction without regard to terrain.

Earth tide effects are usually less than 0.2 mgal, but on spring tides with the sun and moon aligned the effect may reach 0.3 mgal. Corrections are computed using suitable astronomic tables and charts. Barometric effects seldom reach 0.05 mgal and, although they are measurable and could be corrected for, are neglected in current methods of gravity reduction.

If it were possible to obtain everywhere absolute values of gravity accurate to 0.01 mgal, the above discussion of the relative importance of different factors influencing the value of gravity would define the extent to which gravity measurements could be exploited for different purposes.

Actually it is not possible to achieve an absolute accuracy of 1 mgal at the present time, and although this does not hinder many studies dependent upon relative changes in gravity, it does limit others. Similarly the accuracy with which position or elevation is known may constitute the limiting factor for some studies, and in others it may be the density value of the local geologic formations.

2.3. Accuracy of Gravity Observations

As indicated earlier, determinations of absolute gravity on the basis of the most recent determinations by Heyl and Cook [2] at the National Bureau of Standards in Washington, D. C. and by Clark [3] at the National Physical Laboratory at Teddington, England, do not appear to be better than 5 mgal. This is based on relative measurements between the two stations by Browne and Bullard [4] and the writer [5]. Relative measurements of changes in gravity, however, have been reported by Slichter [6] to 0.001 mgal in connection with earth tide studies. Between these two limits of accuracy are those obtained using invariable length pendulums for measuring relative changes in gravity which on the average are good to about ± 0.5 mgal. In such measurements pendulums are swung at a base having a value of gravity equal to g_1 and the period equal to T_1 is recorded. The change in gravity (Δg) to some other point, where the period of the pendulums is T_2, is given by the expression

$$\Delta g = g_1 \frac{T_1^2 - T_2^2}{T_2^2} \tag{20}$$

The accuracy of this method is limited by the precision with which the period can be determined and the evaluation or elimination of

extraneous effects such as sway of the supports, temperature changes, air dampening, changes in magnetic field, etc.

The spring type gravimeters, while having high sensitivity, cannot be used for determining absolute gravity since what is measured is either a spring elongation or the restoring force on a spring to bring a suspended mass back to a fixed (null) position.

In either case the spring elongation produced is a function of a change in gravity, Δg, rather than the actual gravitational attraction g.

The principle of operation of such instruments is that the change in gravitational attraction on a mass (m) suspended on a spring causes a change in elongation (Δd) which can be related to the spring constant (C) by the expression

$$d = mg/C \tag{21}$$

in which d is the total elongation of the spring.

The period (T) of this same system vibrating in simple harmonic motion is expressed by

$$T = 2\pi \sqrt{m/C} \tag{22}$$

From these two expressions it is seen that

$$\Delta d = \Delta g \frac{T^2}{4\pi^2} \tag{23}$$

In unstable (null reading) type gravimeters there is an element that contributes a force in the same direction as that of the gravitational force being measured and a restoring force element. In such instruments the period (T') is given by

$$T' = 2\pi \sqrt{m/C - C'} \tag{24}$$

in which C is the spring constant for the restoring force and C' the constant for the force having the opposite direction. Since the net restoring force is $C - C'$, it is obvious that as the algebraic sum of the two approaches zero the period will approach infinity and also the sensitivity of the instrument.

Because of their high sensitivity spring gravimeters are subject to reading errors owing to changes in temperature, barometric pressure, seismic disturbances and, with some types, changes in magnetic field. Readings are also subject to variation with time (drift) which may or may not be uniform. By thermostating the instrument at some value above that normally reached during the day, as 115°F, temperature effects can in large measure be eliminated. Similarly by sealing the

moving element in a partial vacuum barometric effects can be eliminated. The use of nonmagnetic materials as bronze, German silver or quartz eliminates the effect of change in magnetic field. The effects of "drift," however, must be determined by periodic repeat measurements at a base station and this constitutes one of the principal disadvantages of using this type of instrument for other than studies of local changes in gravitational attraction. Another limiting factor is the limited range of such instruments, usually 50 mgal. In order to obtain a greater range it is necessary to sacrifice sensitivity as indicated in equation (24). Another factor which is introduced when the entire length of a spring is being used is the possibility of non-linearity whereby the scale constant will not be uniform throughout.

In the use of high-range gravimeters, such as have been developed for geodetic studies where the range is 3000 mgal, one of the principal problems is calibrating the instrument with sufficient accuracy to obtain a reading reliable to 1 mgal over its entire range. The usual calibration, based on a series of repeat-measurements between bases having a difference of less than 100 mgal, is good only to 1 part in 2000 at best. A better method of calibration would obviously be to check such an instrument against relative pendulum measurements covering the entire range of the instrument. Unfortunately there is no line of pendulum measurements having an accuracy of as much as ± 1 mgal that covers a large enough change of gravity for this purpose. To remedy this situation the writer started a program last year for establishing such a series of measurements. This work is being done under the auspices of the Geophysics Research Division, Air Force Cambridge Research Center, and will result in a series of gravity bases between Fairbanks, Alaska and Mexico City, Mexico. Quartz pendulums developed by the Gulf Research & Development Co. are being used and gravity bases are being established at the principal airports along the route. When completed it will be possible by using air transportation to check the calibration of any gravimeter over its entire range up to 4000 mgal inside of a week.

With a similar purpose in mind a series of pendulum stations covering approximately 2500 mgal change in gravity and following a more direct north-south route between Brownsville, Texas, and La Pas, Manitoba, Canada, was established in 1950 by the United States Coast and Geodetic Survey using the Brown pendulum apparatus developed by that organization. Although these pendulum observations are not as accurate as might be desired, they have been adjusted through a series of gravimeter observations made at the same time to a probable accuracy of ± 0.7 mgal and, at present, constitute the only calibration range for adequately checking high range gravimeters in existence.

3. The Exploitation of Gravity Measurements in Geodetic Studies

3.1. Types of Measurements

Since the largest and most easily measured changes in gravity are related to the change in shape of the earth it was natural that geodesists were the first to attempt to use gravity for purposes other than studying falling bodies.

From the latter part of the 19th century to the present gravity measurements have been an integral part of the geodetic program of most countries. These measurements give information for determining—

1. The actual shape of the earth (the geoid) and its departure from the assumed mathematical shape.

2. The degree of earth curvature present in any area. This information is necessary for the adjustment of triangulation networks with astronomic determinations of position.

3. The areas that are characterized by large departures of the vertical effects (differences between ground triangulation positions and astronomic position) caused by local anomalous masses which warp the equipotential surface defining the geoid.

The development of methods whereby it is possible to compute directly from gravity anomalies the deflections of the vertical to be expected at any place as well as undulations of the geoid makes it possible for the first time to include the oceanic areas in such geodetic studies. Incidentally, these methods also materially reduce the cost of many geodetic studies since gravity surveys are relatively inexpensive as compared to triangulation surveys.

In contrast to the Bouguer anomalies previously discussed, a so-called free air anomaly is used in most geodetic work. In this anomaly the correction for the mass effect of the material between the point of observation and the reference datum is omitted, and the anomaly is expressed by

$$F_a = g - [g_0 - k_h - t] \tag{25}$$

in which F_a = free air anomaly
g = observed gravity
g_0 = theoretical sea level gravity from the International formula
k_h = elevation correction
t = correction for the surrounding terrain

The Bouguer anomaly is not used since it is found that the correlation of Bouguer anomaly values with elevation is so pronounced that it tends to

obscure all other effects. This is shown in Table I in which average Bouguer anomaly values for different elevations are listed.

In order to remove the correlation between the anomaly values and elevation it is necessary to introduce another correction for the indicated mass deficiency at depth. This is known as the isostatic correction and its inclusion reduces the new anomaly values, termed isostatic anomalies, to

TABLE I

Elevation meters	Average Bouguer anomalies* mgals
0 to 400	−0.010
401 to 800	−0.049
801 to 1,200	−0.104
1,201 to 1,600	−0.137
1,601 to 2,000	−0.176
2,001 to 2,500	−0.194
Over 2,500	−0.220

* Based on samples of 50 in each elevation group.

minimum values which show no dependence upon elevation. Because they are independent of elevation, isostatic anomalies are regarded as ideal for geodetic studies. In level plains and plateaus though, it turns out that the magnitude of the isostatic correction is essentially equal to the mass correction and as the two are of opposite sign the isostatic anomaly is numerically much the same as the free air anomaly. This relation can best be seen in Table II which lists representative anomaly values for different types of terrain.

It will be noted from Table II that whereas there is no correlation between the Free Air anomalies and elevation in the areas of low relief, there is a definite correlation between these anomalies and elevation in areas of high relief. This suggests that any mass compensation at depth is on a regional scale rather than a local one and this is further substantiated by the sign of the free air anomalies. Positive values are obtained in areas of mountain peaks and negative values in the valley areas.

The fact that in all cases the correlation between the Bouguer anomalies and elevation is removed and the anomalies reduced to very small values by introducing a correction for a deficiency in mass at depth (compensation) that is proportional to the elevation strongly supports the theory of Isostasy. This theory in brief states that above some depth beneath the surface all crustal columns exert the same pressure, i.e. have the same mass. The apparent general application of this concept both

in continental and oceanic areas furnishes the principal support for the adoption of isostatic anomalies for geodetic studies in all areas.

Two methods have been evolved for arriving at the magnitude of the isostatic correction. The simplest from the standpoint of computation is based on the Pratt or "dough" theory of isostasy which assumes all crustal columns have the same mass above −113.7 km. The sea level crustal column extending down to −113.7 km with a density of 2.67 gm/cm^3 is taken as a standard. Any column with an elevation above

TABLE II

Type area	Place	Elevation meters	Gravity anomalies (mgal) Free air	Bouguer	Isostatic
High plateau	Guymon, Okl.	949	− 20	−121	−19
	Bridgeport, Nebr.	1114	− 14	−139	− 9
	New Castle, Wyo.	1328	+ 31	−112	+28
Low plains	Texarkana, Ark.	99	+ 9	− 2	+ 8
	Alexandria, La.	24	− 1	− 4	− 4
	Memphis, Tenn.	80	+ 12	+ 4	+10
Mountain peaks	Pikes Peak, Colo.	4293	+206	−215	+19
	Cloudland, Tenn.	1890	+131	− 55	+ 1
	Medicine Mt. Wyo.	2778	+191	−111	+67
Deep valleys	Grand Canyon, Ariz.	849	−108	−184	−12
	Grand Junction, Colo.	1398	− 29	−184	+22
	Smith Gap, Pa.	472	− 6	− 52	−43

sea level, therefore, has a proportionally lower density depending upon its elevation. The correction is figured on the basis of the gravitational attraction of a vertical cylindrical column extending upward from −113.7 km with a density value determined by the proportional change in elevation of the actual column above the standard sea level column.

An alternate method of determining the isostatic correction is based upon the idea that all crustal columns have the same density and that differences in elevation reflect differences in the thickness of the outer crust which is considered to be in hydrostatic equilibrium and floating in a denser subcrustal stratum. In other words the situation is analogous to blocks of ice floating in water. The greater the thickness of any block, the higher the freeboard above water and also the greater the submerged portion below water level. This theory is known as the Airy or the "roots of mountains" theory. In practice it makes little difference which

theory is assumed in determining the isostatic correction despite the marked difference in physical conditions assumed. The explanation of this result appears to lie in the fact that under either theory compensation is accomplished above the same depth level.

The expression for the isostatic anomaly incorporating all corrections is given by

$$I_a = g - [g_0 - k_h + (B_h - t) - I] \tag{26}$$

in which k_h = elevation correction
B_h = mass correction above sea level
t = terrain correction
I = isostatic correction

The actual application of the isostatic correction as applied by the United States Coast and Geodetic Survey on the basis of local compensation, however, is a laborious process since changes in compensation for changes in elevation of all the surrounding terrain must be evaluated as well as the isostatic compensation immediately under the point of observation.

3.2. Practical Significance of Gravity Measurements in Geodetic Work

Departure of the vertical effects due to local warping of the geoid may result in errors as large as 1 minute in position. Such differences, aside from the problem they present in the adjustment of triangulation surveys, materially affect the accuracy of all position fixing schemes utilizing base lines and three station reception. While less critical, it also affects the accuracy with which earthquake epicenters can be located. Another consideration is the effect on precise leveling. Since in general the geoid rises in topographically high areas, appreciable errors in precise leveling are apt to result from the fact that the level bubble which is always aligned in the geoid equipotential surface will follow its configuration. In the Rocky Mountains this effect is estimated to be as much as 40 meters. The effect on gravity anomalies in such areas is obvious. From an engineering standpoint the seriousness of the effect will depend upon the type of project, and would be most significant in any project involving a head of water.

From the standpoint of cost there is no question about geodetic measurements based on gravity anomaly surveys being both cheaper and quicker than standard triangulation methods. From the standpoint of accuracy the agreement determined by the United States Coast and Geodetic Survey [7] has shown that gravimetrically reduced astronomic data agree within 1 second of arc with values determined by standard astronomical triangulation methods.

The advantage in having a quick method of ascertaining local departures of the vertical in remote, incompletely surveyed areas also comes into international relations particularly where international boundaries are involved. For example the boundary between the United States and Canada west of the Great Lakes was determined astronomically and on maps appears to be straight and to follow a parallel of latitude. Subsequent triangulation surveys, however, show that this boundary actually is as crooked as the proverbial snake. If it were straightened out, many Canadians would be residing illegally in the United States and vice versa, and presumably would have to move.

Confusion is also caused by different governments or organizations within a government issuing maps based on different geodetic datum points influenced by different departure of the vertical effects. For example three different sets of coordinates may be obtained for the same point in the city of Honolulu depending upon whether maps of the United States Hydrographic Office, the Coast and Geodetic Survey or the Geological Survey are used. In an area like Alaska which is not likely to be surveyed completely for a long time there are seven different geodetic datum bases in use, and until these are adjusted there is certain to be local confusion where surveys based on different datum points meet.

With the development of long range position location schemes, e.g., in SOFAR, which utilize base lines of transoceanic length, the fact that different countries base their maps on entirely different spheroids of reference as well as different geodetic datum bases will also affect the accuracy of the positions determined.

3.3. National Gravity Base Values

The principal difficulty in using gravity work, particularly for the study of the geoid as a whole, lies in getting all data on absolute datum. As indicated earlier absolute values of gravity can now be determined with an accuracy of 5 mgals. However, as shown in Table III, independently-determined absolute values of gravity over the past 60 years at different places vary greatly when the values are compared by relative measurements to a single station.

To eliminate confusion as to what is the absolute gravity datum the value determined at the Geodetic Institute at Potsdam has been adopted by international agreement as a base value, and all values are now referred to it. Since its adoption the more recent determinations of absolute gravity at Washington and at Teddington indicated it may be about 18 mgal in error. From a practical standpoint this error is not too significant since it is used only as a datum value. Of greater significance is the accuracy with which the relative gravity measurements between

TABLE III

Station	Absolute value: Directly established (gal)	Absolute value: Relative to Potsdam (gal)	Differences (mgal)
Potsdam, Germany	981.274		
Vienna, Austria	980.862	980.853	+ 9
Paris, France	980.970	980.944	+26
Rome, Italy	980.343	980.347	− 4
Madrid, Spain	979.977	979.981	− 4
Teddington, England	981.181	981.194	−13
Washington, D.C.	980.080	980.100	−20

the various national gravity bases and Potsdam are established. Since all gravity surveys in any one country are based on a single national base, the accuracy with which that base is established influences all subsequent work in that country and influences the results obtained for any world-wide geodetic study based on data from several countries. An example of variations obtained is furnished by data for Denmark. Copenhagen has had more direct ties made with Potsdam than any other place and the results vary as shown in Table IV.

TABLE IV

Observer	Value relative to Potsdam (gal)	Difference relative to mean (mgal)
1930 (Schmehl)	981.5556	−3.0
1930 (Andersen)	.5621	+3.5
1935 (Andersen)	.5606	+2.0
1935 (Brockamp)	.5577	−0.9
Weighted mean adjusted value (Morelli) [8]	.5586	
Weighted mean adjusted value (Hirvonen) [9]	.5575	
Average adjusted value	.5586	

Where values are based on indirect ties to Potsdam the range in values is apt to be even greater. Examples are Paris, France, a station which checks well with others and Dehra Dun, India, a station which does not. These are shown in Table V.

Statistically the weighted mean values, where there are a number of observations, should be essentially correct. Actually repeat measure-

TABLE V

Observer	Value relative to Potsdam	Difference relative to mean (mgal)
Paris 1900 Putnam from Potsdam	980.942	− 2.5
1900 Haid from Karlsruhe, Germany	.933	−11.5
1926 Vening-Meinesz from DeBilt, Holland	.9428	− 1.7
1932 Holweck from Binningen, Holland	.9466	+ 2.1
1933 Nörlund from Potsdam, Germany	.9439	− 0.6
1932 Nörlund from Binningen, Holland	.9450	+ 0.5
1934 Holweck from DeBilt, Holland	.9456	+ 1.1
1934 Holweck from Binningen, Holland	.9438	− 0.7
1935 Lejay from Potsdam, Germany	.9431	− 2.4
1935 Lejay from Binningen, Holland	.9406	− 3.9
Weighted mean adjusted value (Morelli) [8]	980.9454	
Weighted mean adjusted value (Hirvonen) [9]	.9435	
Average adjusted value	.9445	
Dehra Dun		
1901 Alessio from Potsdam via Colaba	979.059	−18.0
1904 Lenox-Conynghan from Kew, England	.063	−14.0
1905 Hecker from Potsdam via Jalpaiquri	.065	−12.0
1913 Alessio from Genoa, Italy	.079	+ 2.0
1924 Cowie from Kew, England	.054	−23.0
1927 Glennie from Cambridge, England	.072	− 5.0
1929 Glennie and Cowie from Kew	.068	− 9.0
1929 Spoleto from Genoa, Italy	.069	− 8.0
1929 Vening-Meinesz from DeBilt via Colombo	.075	− 2.0
1932 Lejay from Potsdam via Colombo	.085	+ 8.0
1939 Browne and Glennie from Cambridge	.056	−21.0
Weighted mean adjusted value (Morelli) [8]	.073	
Weighted Mean adjusted value (Hirvonen) [9]	.081	
Average adjusted value	.077	

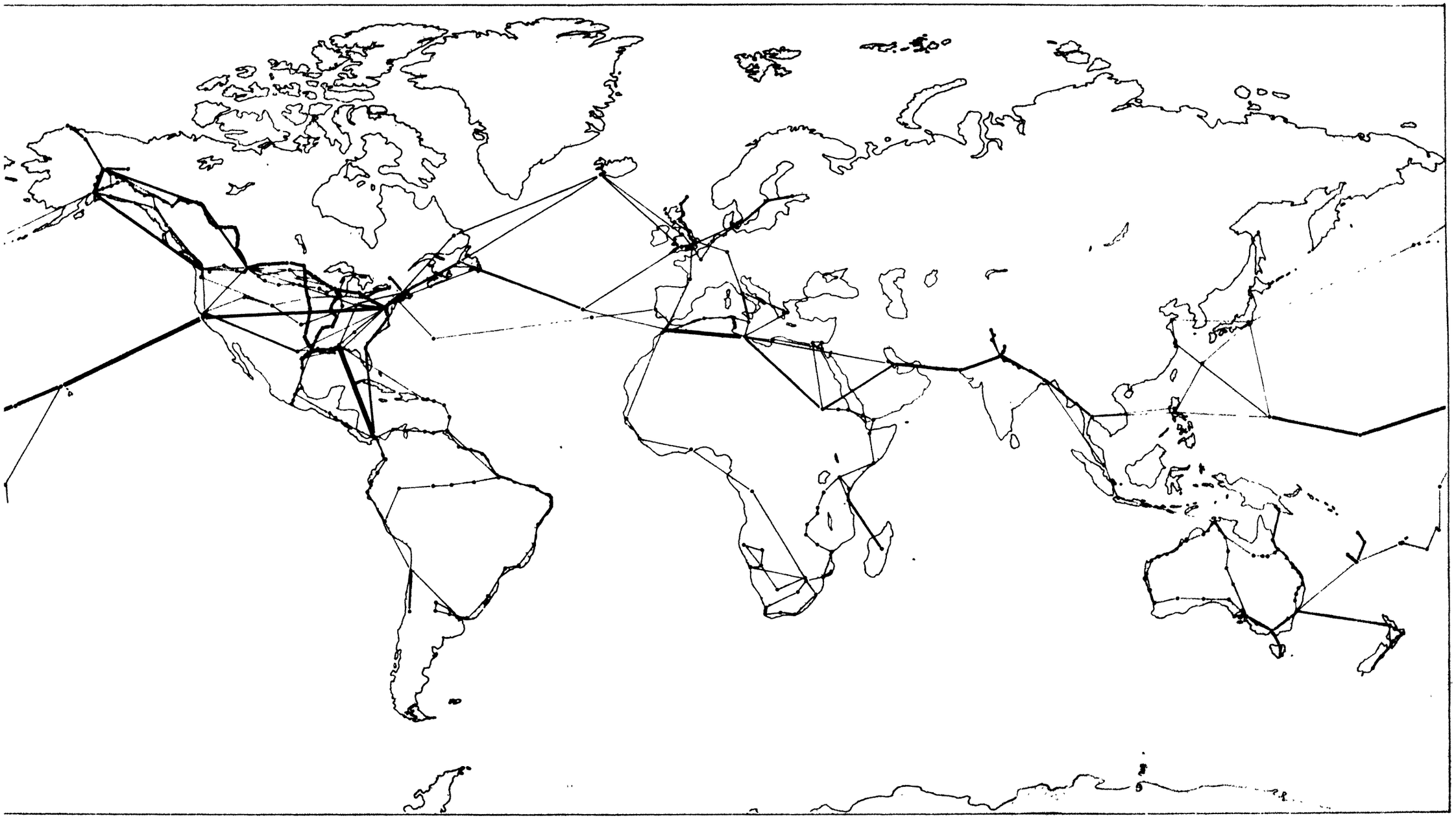
ments by the writer using improved gravity apparatus show that whereas this is true for Paris, in the case of Dehra Dun the correct value differs 14 mgal from the weighted mean. That this difference is real was verified this past year by another series of observations using a different gravity apparatus.

The establishment of a world network of gravity bases on a common datum is still not realized. However, under the auspices of the Office of Naval Research and the Woods Hole Oceanographic Institution the writer has been engaged since 1948 in checking the principal gravity bases of the world and establishing auxiliary bases and stations in areas not yet covered by gravity surveys. Because of international complications it is, of course, not possible to actually cover the world at this time. There is, however, much work to be done in those parts of the world that are easily accessible. The development of the network of bases established to date is shown in Fig. 1. This work has been carried out using a special quartz spring gravimeter having a range of about 5,000 mgal which is temperature-compensated that was developed for the project by the Houston Technical Laboratory of Houston, Texas. The observations on the average appear to be good to about 1 mgal over the 4000 mgal change in gravity covered to date. Values for the principal bases reoccupied and new bases established are given in the Appendix.

3.4. Work at Sea

Until Vening-Meinesz of Holland developed the first gravity apparatus that could be used at sea and give reliable results, there were no gravity data outside of the land areas. Since his apparatus must be used on a submerged submarine the amount of data in the oceanic areas has been limited and up to 1947 consisted of essentially a traverse around the world plus three other traverses in the Atlantic Ocean by Vening-Meinesz [10], and local studies in the East Indies, the Mediterranean Sea, the West Indies, and Japanese Islands by various groups. However, under the Office of Naval Research and with the cooperation of the United States Navy, a group working under Prof. Maurice Ewing of Columbia University, New York, is extending the studies of the gravity field in the oceans using Vening-Meinesz gravity apparatus. To date several traverses across the Pacific Ocean with some short traverses off the Atlantic Coast of North America have been made. Similar measurements by B. L. Cooper of Cambridge University, England, are also in progress in the Mediterranean.

In order to intergrate the gravity program at sea with those on land, a part of the program for establishing gravity bases around the world includes observations at the docksid stations set in the submarine



WORLD NETWORK OF GRAVIMETER BASES
1950

Fig. 1.

program. Other data being tied into the general network are oil company surveys in various parts of the world that have been made in connection with geologic prospecting. Within a relatively short time it appears a sufficient amount of data will be assembled on a common datum to permit many geodetic and geologic studies to be undertaken that previously were not possible.

4. Geologic Uses of Gravity Data

4.1. Earth Crustal Studies

As indicated by anomalous departure of the vertical relations and the marked negative Bouguer gravity anomalies obtained in mountain areas, geologic factors have a marked effect upon gravity values. Under the Pratt theory of isostasy an actual change in density beneath such areas is visualized and the mountains stand high because they are light. Although most geologists do not support this concept, recently Bucher [11] expressed the thought that the continents themselves may have developed from the ocean floor by thermal and chemical processes involving large changes in volume and density.

On the basis of evidence of downward warping in mountain regions many geologists feel that the concept embodied in the Airy hypothesis of isostasy is supported; that is, the crustal blocks of the earth are essentially in hydrostatic equilibrium and the elevation is a function of block thickness rather than change in density. This theory also appears to be supported by seismic studies both in North America and Europe which indicate the presence of low velocity "roots" beneath the mountains whose depths agree very closely with those calculated from Bouguer anomaly values. However, seismic data also suggest that there is a very real density difference between the rocks composing the continents and the floor of the oceanic basins. The seismic velocity found for the rocks beneath the ocean sediments is about the same as that found at a depth of 10 km in the continents. Other evidence indicating a physical difference in the geology of the continents and oceanic basins are the Bouguer anomalies. In the ocean basins after allowing for the low density layer of sea water present the Bouguer anomaly values are found to be about 300 mgal higher than on the continents. Another difference is the depth of the Mohorovičič discontinuity which lies at a depth of about 35 km in the continents but on the basis of recent seismic studies [12, 13] occurs at only 5 km below the bottom of the ocean. The situation is analogous to what would be obtained for a group of thick but variable white pine "continental" blocks and thinner but variable oak "oceanic" blocks floating in water. Thus there appear to be physical features embodied

in both concepts of isostatic compensation corresponding to the situation in nature.

Although gravity surveys that make use of Bouguer anomalies offer a method of studying these differences in crustal structure, there is a serious drawback to their use for this purpose. Because an infinite number of mass distributions in space can create the same gravitational effect, no unique solution as to the actual subsurface mass distribution can be determined from gravity data alone. Because of this limitation the principal uses of gravity data in studying crustal structure have been in extending knowledge where the key to the structure present has been established from seismic studies and in making qualitative investigations as to where there are differences in crustal structure. From such studies it is found for instance that most young mountain ranges as the Rocky Mountains, the Alps, the Andes, and the Himalayas have a mass deficiency at depth which results in a Bouguer anomaly of about −300 mgal, and on the assumption of isostasy this deficiency almost exactly compensates for the mass of the mountains above sea level. On the other hand, old eroded mountain ranges as the Appalachian Mountains have a much smaller deficiency in mass associated with them and although the Bouguer anomaly value may be only −100 mgal in these areas, this mass deficiency may be much greater than is required for isostatic compensation. As a consequence isostatic anomalies of approximately −50 mgal may be obtained. The inference from these relations is that erosion and reduction in elevation in these areas have progressed faster than the readjustment of the compensating mass at depth. In areas of normal faulting some mountains appear to have only a partial compensating "root," and they are characterized by positive isostatic anomalies. The inference is that these mountains were formed by an entirely different mechanism from the major folded mountain ranges and actually constitute an excess load on the crust. Mountains of this type would be those resulting from block faulting whereby a block bounded by shear planes is elevated by lateral compressive forces. Another type of mountain having similar relations is a volcano. In this case the elevation results from the accretion of flow material piling up on the original ground surface around a central vent.

In some areas, such as penisular India, there appears to be a marked thinning of the upper granitic crust manifested by positive anomalies which do not correlate with any surface features. The belt of positive anomalies in India extends completely across the peninsula, and has a width of about 300 miles and an average isostatic anomaly of about 20 mgal. Gravity surveys covering the oceanic "deeps" and island arcs show that these areas are characterized by parallel belts of positive and

negative anomalies which have been interpreted as resulting from downbuckling of the earth's outer crustal layer.

The most surprising thing is that isostasty does appear to be essentially complete over most of the earth despite the above exceptions and regardless of whether the observations are made on the continents or at sea. This subsurface adjustment in the geologic structure of the crust to maintain a constant pressure above about 100 km below sea level everywhere is one of the most remarkable major natural phenomena of the earth.

4.2. Geological Exploration

From the standpoint of the economic exploitation of the earth's gravitational field in the search for mineral and oil deposits, the quantities that must be measured and studied are extremely small. Oil cannot be searched for directly using gravity and must be sought by searching for some geologic trap resulting from folding, faulting, differential compaction of the sediments or change in lithologic facies. Frequently the entire gravity anomaly for such a trap will not exceed 2 mgal, and this effect may be superimposed upon the regional gradient of a large crustal feature involving perhaps a 100 mgal. As a consequence it is not sufficient to take the Bouguer anomaly values and plot them. It would be like looking for a wart on the contour map of an elephant. Neither will isostatic anomalies suffice which remove the bulk of the Bouguer anomaly because the geologic assumptions on which the isostatic correction is computed may or may not actually exist in the area under consideration. As a consequence the large anomaly effects are removed by arbitrary methods in order to isolate the small residual values of potential economic importance. Many methods have been devised for doing this, and they vary from the subtraction of smoothed values representing the regional effects to the determination of second derivative values. The actual method applied will vary with the size of the residuals in the area and the individuals making the reduction.

In areas such as Arabia the gravity anomalies associated with dense folded limestone which are apt to form oil traps may be relatively large and amount to 10 or more milligals. In the Gulf Coast region of the United States where oil traps are formed by salt domes rising from depths like factory chimneys which pierce and drag upward the surrounding sediments, the gravity anomalies are usually less than 5 mgal. In the high plains of Wyoming and Montana the gravity anomalies associated with slight faulting and folding may be less than 1 mgal. In the search for metallic ore deposits very small anomalies are also the rule. As a result, not only are high precision gravimeters which have an accuracy of 0.01

mgal required in prospecting but also detailed corrections for the surrounding terrain and earth tides must be made.

In contrast to geodetic gravity surveys in which a station spacing of 10 km is sufficient except in the immediate vicinity of astronomic stations in departure of the vertical studies, gravity surveys for mineral exploration require a very high density of stations. In prospecting for oil, stations may be located from 0.25 to 1 mile apart. In prospecting for metallic ores observations may be made at intervals of as little as 100 ft. Another difference is that because most oil traps or ore deposits as a rule only involve local areas, it is not necessary that the data be put on an absolute datum. Many surveys as a result are based on arbitrary datum values which make these data useless for geodetic purposes until they have been tied to absolute controlled surveys. Similarly non-linearity in the response of instruments is not as critical in exploration work as in geodetic work because of the small changes in overall gravity covered by any one survey. Likewise the accuracy of calibration is not as important as in geodetic surveys and a calibration of 1 part in 1000 is usually accurate enough. As a consequence of these differences in requirements, it is sometimes difficult to use gravity data taken for prospecting purposes for geodetic purposes. For example it was found that one series of gravity observations carried out locally to 0.01 mgal in oil exploration work had an accumulative error of 8 mgal in 300 miles of north-south travel because of the above effects.

4.3. Gravity Studies of the Strength of the Earth

As previously mentioned, corrections for earth tides effects now often are applied to gravity data being used for prospecting purposes. Studies of crustal response to tidal influences can also be used to study the strength of the earth. Two years ago, under the auspices of the Shell Oil Co., simultaneous gravity observations were made at 15 minute intervals over a 2 week period at a network of stations covering most of the earth to study not only the degree of response of different geologic structures but also the lag in response. More recently Slichter [6] has been making measurements in various areas with highly sensitive gravity apparatus that records not only earth tide effects but also the effect of changes in barometric pressure.

Another type of investigation pertaining to the strength of the earth's crust includes gravity surveys in areas as Fennoscandia and Canada, which in recent geologic times were subject to heavy ice loading. The ice in both areas is now gone and the areas after having been depressed by the ice load are now rising. This rise is indicated by changes in sea level as shown by tide gage records and also such other phenomena as raised beaches and strand lines. Gravity surveys show that both areas are

characterized by negative isostatic gravity anomalies and thus suggest that the areas are still depressed below the natural isostatic equilibrium position and probably will continue to rise for some time. Similarly, gravity anomaly values around volcanic islands suggest that the superimposed load represented by the accretion of volcanic flow material has caused a downward flexure of the crust which is spread out over a considerable distance. Such studies, therefore, indicate that the earth's crust is not rigid but rather acts like an elastic body.

If the earth's crust does behave elastically, it obviously must be in a constant state of flexural adjustment in response to changes in surficial loading induced by erosional agents, volcanism, melting ice, etc., and as a result the force of gravity at any one place will not be constant. As has already been indicated there is a periodic change in gravity with the lunar cycle due to the response of the earth's crust to tidal forces resulting in a maximum variation of about 0.3 mgal. The change in gravity to be expected as a result of long term changes in terrestrial loading, however, may be relatively large. For example from both tide gage studies and precise leveling surveys it is known that southern Finland is rising at a rate of about 5 mm/year and the Gulf of Bothnia about 11 mm/year. From a study of the height of raised beaches in the area there has been at least 275 meters uplift since the removal of the ice cap which once covered the area, and on the basis of the present isostatic anomaly of about -14 mgal the area may continue to rise another 100 meters if it is assumed the observed anomaly is due entirely to crustal displacement. If isostatic compensation were accomplished instantaneously with change in elevation, the gravity value would remain essentially constant. However, there is every reason to believe that there is an appreciable time lag in achieving compensation for any change in elevation, as evidenced by the relatively large isostatic anomalies found in old mountain range areas. Therefore, in any area which is not in isostatic equilibrium, it may be expected that the gravity values will change with time and approach a constant value as isostatic equilibrium is achieved.

As information on the rate of change in elevation and gravity accumulate and results of current studies on the response of the earth at various places to dial influences become available, it will be possible to learn more about the viscosity of the earth's outer crust. This in turn will lead to a better understanding of the subsurface transfer of mass in response to changes in external loading, internal thermal convection and possible other causes which all undoubtly enter into the phenomenon of isostasy.

APPENDIX

In Table VI all values are given relative to the absolute gravity base at Potsdam, Germany, whose adopted value is 980.274 gal.

TABLE VI. Comparative gravity values at principal bases.

Station	Value in use	Gravity meter value	Instrument	Observer
		NORTH AMERICA		
		Canada		
Ottawa, Ontario				
Dominion Observatory	980.6220	980.6215	W10b	Woollard
		Mexico		
Mexico, D. F.				
Tacubaya Meteorological				
Observatory	977.9410	977.9431	W10c	Harding
		Panama		
Balboa				
YMCA Base (USC&GS)	978.2386	978.2410	W10c	Harding
		978.2417	W10e	Harding
		978.2418	W41b	Harding
Submarine "g" Base	978.245	978.2416	W10e	Harding
		978.2420	W41b	Harding
Coco Solo Submarine "g" Base	978.251	978.2590	W10a	Woollard
		978.2585	W10c	Harding
		978.2583	W41b	Harding
		United States		
Washington, D. C.				
Commerce Bldg.	980.1180	980.1190	W10a	Woollard
Smithsonian Institution	980.1190	980.1179	W10a	Woollard
New Jersey Ave. Base	980.1170	980.1145	W10a	Woollard
San Francisco Submarine "g"				
Base, Fort Mason	979.9960	979.9980	W10a	Woollard
		SOUTH AMERICA		
		Argentina		
Buenos Aires				
National Meteorological				
Observatory	979.7050	979.7064	W10c	Harding
La Plata				
Observatory	979.7480	979.7523	W10c	Harding
Cordoba				
Observatory	979.3380	979.3428	W10c	Harding
		Brazil		
Rio de Janiero				
Observatory	978.8048	978.8060	W10c	Harding
		Chile		
Santiago				
Institute Geographico, Militar	979.4293	979.4289	W10c	Harding
		Ecuador		
Quito, Ecuador				
Observatory	977.2790	977.2795	W10c	Harding

TABLE VI. Comparative gravity values at principal bases. (*Continued*)

Station	Value in use	Gravity meter value	Instrument	Observer
		Venezuela		
Caracas				
Cartographia Nacional	978.0664	978.0679	W10c	Harding
		ATLANTIC AREA		
		Puerto Rico		
Mayaquez				
Federal Prison	978.6690	978.6655	W10c	Harding
San Juan				
House of Representatives	978.6760	978.6850	W10c	Harding
		EUROPE		
		Denmark		
Copenhagen				
National Base	981.5581	981.5574	W10b	Woollard
		England		
London				
Greenwich Observatory	981.1882	981.1904	W10b	Woollard
Teddington				
National Physical Laboratory	981.1953	981.1961	W10b	Woollard
		981.1960	W10e	Woollard
Cambridge				
Pendulum House	981.2645	981.2684	W10b	Woollard
		981.2681	W10e	Woollard
		Finland		
Helsinki				
University	981.9158	981.9133	W10b	Woollard
		France		
Paris				
Observatory	981.9440	981.9439	W10b	Woollard
		Holland		
DeBilt				
Meteorological Observatory	981.2679	981.2681	W10b	Woollard
		Italy		
Rome				
Technical College	981.3663	981.3633	W10b	Woollard
		Scotland		
Edinburgh				
Royal Observatory	981.5801	981.5834	W10e	Woollard
		Sweden		
Stockholm				
Riketsallmänna Kartwerke	981.8472	981.8449	W10b	Woollard

TABLE VI. Comparative gravity values at principal bases. (*Continued*)

Station	Value in use	Gravity meter value	Instrument	Observer
		AFRICA		
		Anglo-Egyptian Sudan		
Khartoum (Curry)	978.308	978.3063	W41b	Harding
Gordon College (Munsey)	978.300			
		Egypt		
Cairo (Curry)	979.292	979.2948	W41b	Harding
Helwan Observatory (Lejay)	979.295			
Suez				
Submarine "g" Base	979.334	979.3078	W41b	Harding
		French Morocco		
Averroe				
Geophysical Observatory	979.565	979.5577	W41b	Harding
		French West Africa		
Dakar				
Submarine "g" Base	978.484	978.4846	W41b	Harding
		Kenya		
Nairobi				
Penfolds House (Bullard)	977.5283	977.5303	W41b	Harding
East Shell House (Bullard)	977.5336	977.5375	W41b	Harding
		Tanganyika		
Dar-es-Salaam				
Survey Dept. Base	978.1164	978.1198	W41b	Harding
		Tunisia		
Tunis				
Submarine "g" Base	979.926	979.9152	W41b	Harding
		Union of South Africa		
Capetown				
MacLean Observatory	979.652	979.6537	W41b	Harding
		PACIFIC AREA		
		Hawaii		
Honolulu				
Submarine "g" Base	978.9410	978.9440	W10a	Woollard
		Guam		
Submarine "g" Base	978.5380	978.5410	W10a	Woollard

TABLE VI. Comparative gravity values at principal bases. (*Continued*)

Station	Value in use	Gravity meter value	Instrument	Observer
		New Zealand		
Wellington				
Observatory	980.2620	980.2662	W10e	Muckenfuss
Christ Church				
Geophysical Observatory	980.504	980.5093	W10e	Muckenfuss
Otago University	980.737	980.7426	W10e	Muckenfuss
		Philippine Islands		
Manila				
Observatory (USCGS)	978.372	978.3656	W10b	Woollard
(Lejay)	978.359	978.3642	W10e	Muckenfuss
Submarine "g" Base	978.360	978.3603	W10b	Woollard
		978.3590	W10e	Muckenfuss
Geodetic Survey	978.3670	978.3635	W10b	Woollard
		978.3629	W10e	Muckenfuss
		Japan		
Tokyo				
University	979 799	979.8025	W10b	Woollard
		979.8019	W41b	Muckenfuss
		ASIA		
		Arabia		
Aden				
Submarine "g" Base	978.323	978 3256	W41b	Harding
		India		
Dehra Dun				
National Gravity Base	979.063	979.0654	W10b	Woollard and Gulatee
		979.0645	W10e	Muckenfuss

The gravimeter measurements listed are all relative to the Potsdam gravity value at the National Bureau of Standards in Washington, D. C. in the United States (980.100 gal). This value established by a direct tie with pendulums by Brown [14] in 1936 appears to be correct on the basis of the mean error curve established for the difference in relative gravity measurements and Potsdam gravity values at ten national gravity bases tied directly to Potsdam.

In identifying instruments the alphabetical symbol identifies the kind of instrument and the number subscript the specific instrument. W = Worden gravimeter. In citing the value in use at natural gravity

bases the value given is the average of the adjusted values as determined by Morelli [8] and Hirvonen [9].

List of Symbols

a	equatorial radius
b'	gravitational flattening of earth
B_h	Bouguer (mass) correction
c'	ratio equatorial centrifugal acceleration to equatorial gravitational acceleration at the equator
c	polar radius
C	constant
d	total elongation
f	geometric polar flattening of the earth
F	force
F_a	free air anomaly
g	observed gravity
g_0	sea level gravity at any point
g_a	equatorial sea level gravity
g_c	polar sea level gravity
I	isostatic correction
I_a	isostatic anomaly
k	moment of inertia
k_h	elevation correction
l	length of pendulum
m	mass
M	mass of earth
s	distance from center of rotation to center of gravity of a pendulum
r	distance between two bodies on radius of earth
t	terrain correction
T	period
U	total gravitational potential
V	gravitational attraction potential
V'	gravitational potential due to centrifugal force
x, y, z	spherical coordinates
α	acceleration
γ	gravitational constant
ϕ	latitude
λ	longitude
σ	density
ω	angular velocity

References

1. Heyl, P. R. (1930). A redetermination of the constant of gravitation. *Bur. Stand. J. Res.* **5**, 1243.
2. Heyl, P. R., and Cook, G. S. (1939). The determination of absolute gravity at Washington. *Bur. Stand. J. Res.* **17**, 804.
3. Clark, J. S. (1939). The determination of absolute gravity at Teddington. *Phil. Trans.* **A238**, 65.
4. Browne, B. C., and Bullard, E. C. (1940). Comparison of the acceleration due to gravity at the National Laboratory, Teddington and Bureau of Standards, Washington, D. C. *Proc. Roy. Soc. London* **A175**, 110.

5. Woollard, G. P. (1950). The gravity meter as a geodetic instrument. *Geophysics* **15**, No. 1, 1.
6. Slichter, L. B. Personal communication.
7. United States Coast and Geodetic Survey. (1949). Deflections of the vertical from gravity anomalies. Army Map Serv. Tech. Rept. No. 2.
8. Morelli, C. (1946). Compensaziano della rete internazionale della stazioni di riferimento per le misure di gravita relativa. *Inst. Geofis. Trieste Pub.* **211.**
9. Hirvonen, R. A. (1948). On the establishment of the values of gravity for the National Reference Stations. *Ann. Acad. Sci. Fenn. Helsinki, Geol.-Geog.* **HIII,** 17.
10. Vening-Meinesz, F. A. (1932). Gravity Expeditions at Sea 1923–1930. N.V. Technishe Boekhandel en Drukkerij, J. Waltman, Jr. Delft, Holland.
11. Bucher, W. H. Personal communication.
12. Ewing, W. M., and Worzel, J. L. Personal communication.
13. Raitt, R. W. Personal communication.
14. Brown, E. J. (1936). A direct gravity tie between Washington and Potsdam. *Spec. Pub. U. S. Coast Geod. Surv.* **201.**

Aeromagnetic Surveying

JAMES R. BALSLEY

Geological Survey, U. S. Department of the Interior, Washington, D. C.

CONTENTS

	Page
1. Introduction	314
2. Basic Instrument	314
2.1. Detector Element	315
2.2. Orientation of Detector Element	317
2.3. Installation of Detector Mechanism	320
2.4. Measurement and Recording of Detector Output	320
3. Associated Equipment	322
3.1. Aerial Cameras	322
3.2. Electronic Navigation Aids	322
3.3. Altimeters	322
3.4. Correlating Circuit	322
4. Field Survey Technique	323
4.1. Aircraft	324
4.2. Crew	324
4.3. Operational Procedure	324
4.3.1. Flight Pattern	325
4.3.2. Flight Elevation	325
4.3.3. Magnetic-Control Pattern	325
4.3.4. Recording and Correlation of Data	326
5. Office Compilation of Field Data	326
6. Interpretation of Results	329
7. Results of Aeromagnetic Surveys	329
7.1. Fairfax Quadrangle, Virginia	331
7.2. Northwestern Maine	334
7.3. Bikini Atoll	336
7.4. Upper Peninsula, Michigan	337
7.5. Allard Lake District, Quebec	338
7.6. Adirondack Mountains, New York	339
7.7. Eastern Pennsylvania	341
8. Advantages and Limitations	342
8.1. Speed, Ease, and Economy of Operation	342
8.2. Quality of Results	342
8.3. Instrumentation	343
9. Applicability	344
List of Symbols	344
References	345

* Publication authorized by the Director, U. S. Geological Survey.

1. Introduction

The recent development of the airborne magnetometer is one of the major advances in geophysical exploration. For years, geophysicists have recognized the advantage of aeromagnetic surveys [1] and a few crude measurements had been attempted [2–7], but only under the impetus of war was equipment produced that was adequate to make measurements accurate enough to be used in geophysical exploration. The basic element of the airborne magnetometer—the fluxgate or saturable inductor—is old, and its use as a means of measuring magnetic fields has been recognized since 1936 [8–10]. However, the full utilization of the fluxgate and its incorporation into a functioning geophysical field instrument was not accomplished until 1944, when the cooperative efforts of engineers, physicists, geologists, and geophysicists of several different agencies and private organizations produced equipment that could not have been designed and perfected otherwise [11–17].

Variations in the earth's magnetic field have been detected or measured for several centuries by instruments that operate by means of the mechanical force produced by the earth's field acting on a magnet or electric current. Obviously, such instruments must be used statically and cannot be used to yield accurate measurements in a vehicle in motion. Various methods of making such measurements have been proposed, but the fluxgate instruments have proved the most successful [18, 19]. These instruments operate by means of the relationship between the earth's magnetic field and the permeability of the material used in the detecting element. Because there are no moving parts in this element, the measurement is independent of the motion and acceleration of the vehicle.

Many reports and articles have been published that deal with one or more phases of the development and use of airborne magnetometers. The purpose of this report is to review and summarize those dealing with the instrumentation and use of the equipment, and in particular those dealing with the results that have been obtained.

2. Basic Instrument

The most useful type of airborne magnetometer consists of a detecting mechanism, self-oriented with respect to the earth's field; the electronic oscillators and amplifiers required for its operation; and the equipment used to measure and record its output, the variations of the earth's total magnetic field.

2.1. Detector Element

The heart of the airborne magnetometer is the fluxgate or saturable inductor. This consists of a core of easily saturable ferromagnetic material of high permeability (usually permalloy) with an external winding to which is applied an alternating voltage that produces a magnetic field h that drives the core cyclically through saturation (Fig. 1). If an external magnetic field, H_0, is applied also to the core, the total field, H,

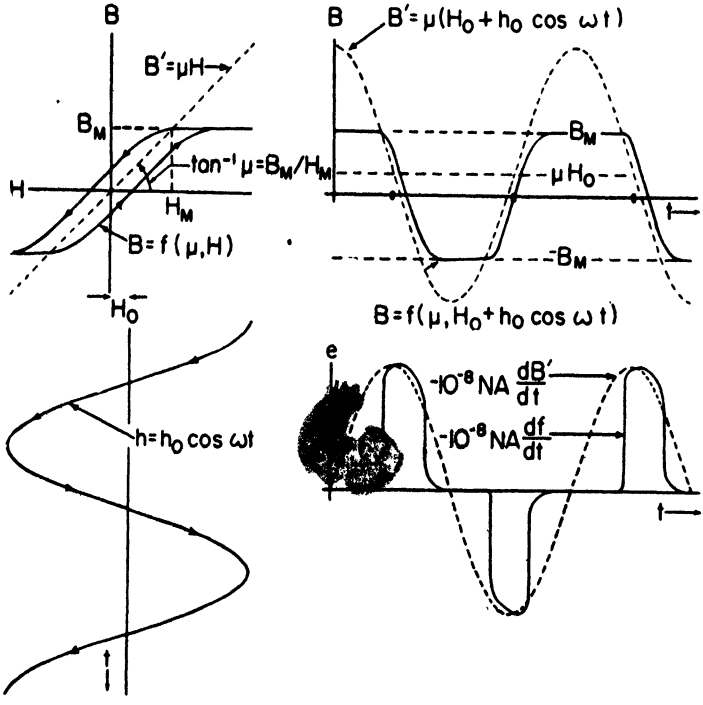


FIG. 1. Schematic diagram of operation of saturable inductor. (By courtesy of the American Geophysical Union [20].)

is increased during one half the cycle and decreased during the other half. If the permeability, μ, of the core is high, the flux density, μH_0, produced in it by the external field is an appreciable part of the flux density at saturation, B_M, although the external field, in this case that of the earth, is low (0.6 gauss).

Owing to this effect of the core, the output electromagnetomotive force e of the coil will be distorted and will have an asymmetrical wave form containing both even and odd harmonics of the driving frequency. Both even and odd harmonics are functions of the external field, H_0, but

only the even harmonics are sensitive to its sign and vanish when it equals zero. Measurement of one or more of the even harmonics of the output emf provides, therefore, a means of measuring the external field [20–22].

Two basic methods have been employed to make such measurements; one uses all the even harmonics [23–25], and the other only one [26–27]. The former has the advantage of using more of the available energy and,

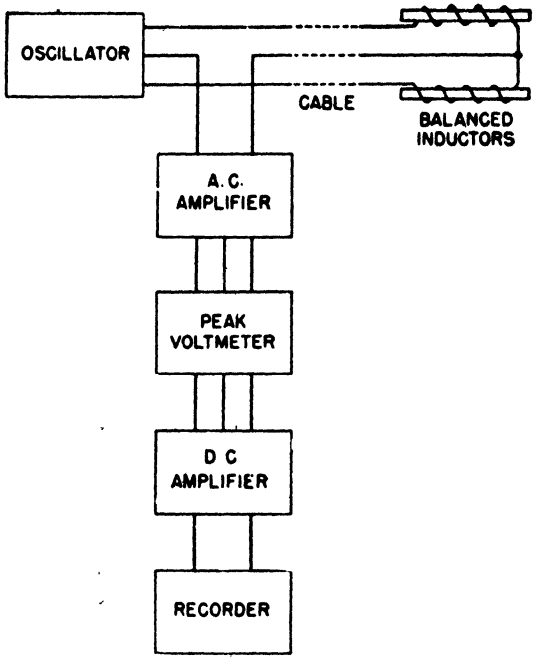


FIG. 2. Balanced inductor circuit. (By courtesy of the American Geophysical Union [20].)

therefore, requires less amplification; the latter, though using a smaller source of energy, has the advantage of using a much simpler and more trouble-free electronic circuit.

The system used to measure all the even harmonics requires a balanced inductor consisting of two identical parallel permalloy cores, magnetically joined at the ends, each with identical coils so arranged that the driving fields are equal in magnitude but opposite in direction. Thus, when the external field opposes the driving field in one coil it adds to the driving field in the other. The coils are electrically connected so that the even harmonics of the output emf add and the odd harmonics oppose. One adaptation of this system is shown in Fig. 2.

The single harmonic system uses a single-element inductor and a narrow band-pass filter that excludes all but the desired frequency of the output emf, usually the second harmonic. Such a system is shown in Fig. 3.

The major portion of the earth's field acting on the detector coil is nullified by a magnetic field produced by means of a steady direct current

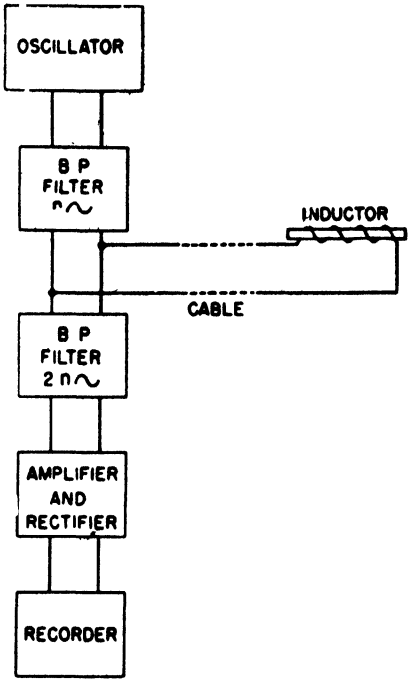


FIG. 3. Second-hamonic inductor circuit. (By courtesy of the American Geophysical Union [20].)

flowing through the coil or through a coaxial secondary winding around it. Thus, instead of measuring one gamma in a field of 60,000, which requires considerable precision, the equipment measures one gamma in a field of 5,000 or less.

2.2. Orientation of Detector Element

Because the saturable inductor measures the external field parallel to its axis, it can be used to measure any of the components of the magnetic field of the earth. Several different types of ground instruments have been constructed that measure one or more components, and several attempts have been made to use a damped pendulum or vertical gyroscope to orient the detector so that the vertical component could be

measured in airplanes or other moving vehicles. Neither the damped pendulum nor gyroscope determines the true vertical in a moving vehicle; the most precise systems now available for use in aircraft are accurate to about 10 minutes of arc, but only when the aircraft is flown at high altitude in smooth air. The error produced by this angular displacement is given by

$$\text{Error} = T[\cos (V - I) - \cos (V' - I)] \tag{1}$$

where T is the total intensity of the earth's magnetic field; I, the magnetic inclination; V, the true vertical or 90°; and V', the erroneous vertical established by the orienting mechanism. An error of 10 minutes of arc in the orienting mechanism gives a magnetic error of 75 gammas if the inclination is 65° and the total intensity 55,000 gammas.

A damped-pendulum system has been used in high-flying airplanes to measure the components of the earth's field so that isogonic charts may be prepared [28–29]. Most exploratory aeromagnetic surveys, however, must be more precise and must be conducted close to the surface of the ground, where turbulent air and frequent maneuvers of the aircraft introduce violent motion that materially increases the error in the orienting system and therefore in the magnetic measurement.

In order to produce an instrument of sufficient accuracy for aeromagnetic surveying, a method of orientation that does not require a geodetic reference is necessary. The simplest method is to use the magnetic field as a reference, although a system with such a reference can measure only the scalar intensity and not the direction of the magnetic vector. If the detector element is oriented parallel to the earth's field, equation (1) reduces to:

$$\text{Error} = T(1 - \cos \phi) \tag{2}$$

where ϕ is the angle between the earth's magnetic field and the axis of the coil. If this angle is 10 minutes and the total intensity 55,000 gammas, the error is 0.23 gamma. Obviously orientation in the direction of the earth's field has the highest inherent accuracy.

Two systems have been developed to accomplish this orientation. In one (Fig. 4), three inductors are placed at mutual right angles [30–36]. The output of each of two inductors is fed to each of two orienting motors, which act to return that element to the position at which it produces no signal, that is, at right angles to the earth's field [37, 38]. These two orienting inductors, therefore, determine a plane at right angles to the earth's field and because the third, or detector, element is also at right angles to this plane, it must be parallel to the earth's field. To correct

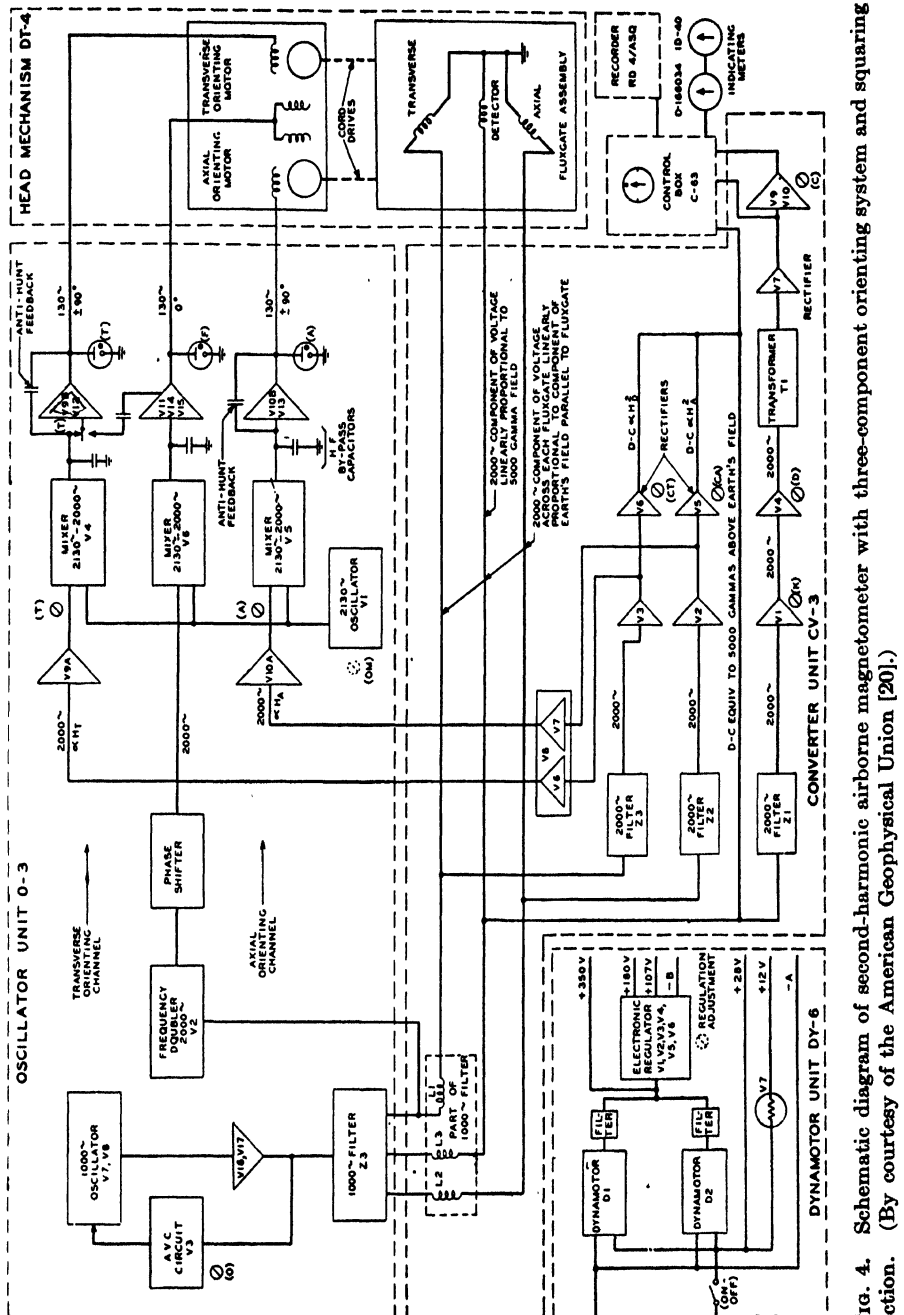


FIG. 4. Schematic diagram of second-harmonic airborne magnetometer with three-component orienting system and squaring correction. (By courtesy of the American Geophysical Union [20].)

for the effect of the small errors of orientation produced by any time lag in the orienting mechanism the output of the two orienting inductors is squared and is used to reduce by the proper amount the nulling field produced by the secondary winding around the detector element. Such a squaring method could be used without orientation to give a measure of the total field, but the requirements that this places on the electronic circuits would be difficult to meet. The system described could be considered either as an oriented detector with a squaring correction or as a partly oriented squaring detector. The squaring correction has been employed in conjunction with the single-even-harmonic system but has not been reported as being used with the all-even-harmonic system.

The other orienting system uses only two inductors—an orienting inductor on a rotating disk and a detector element normal to it [16, 25]. If the rotating disk is not normal to the earth's field an alternating signal is produced by the orienting inductor. The phase relationship of this signal is established by an electronic reference, and the alternating current is used to drive two servo-motors that orient the disk to a position at which no signal is produced by the orienting inductor, that is, normal to the earth's field. No squaring correction is described with this system, but the correction can undoubtedly be made though probably with some difficulty.

2.3. Installation of Detector Mechanism

The detector mechanism must be located so that it is not appreciably affected by the magnetic material of the vehicle. This may be accomplished by removing the detector mechanism from the vehicle and towing it by means of a winch and cable system in a streamlined, bomb-shaped nacelle, called a "bird" [39] if the vehicle is an airplane. The towing cable consists of the necessary electrical conductors and a stress-bearing member of nonmagnetic material, usually phosphor bronze.

Inboard installations have been made in various aircraft, but they require elaborate compensation for the magnetic material in the plane and do not generally utilize the full sensitivity of the instrument [40–41]. Such installations are made on the wing tip or, if the plane does not have a tail wheel, in a nacelle or "stinger" extending from the tail. With the most ideal conditions, a "heading effect" of at least 10 gammas is introduced in the magnetic record by the orientation of the aircraft in the earth's field.

2.4. Measurement and Recording of Detector Output

The two basic methods of producing and measuring the detector output have been previously discussed and are shown in Figs. 2 and 3.

Both require a driving oscillator, usually of 1000 cycles, a linear rectifier and amplifier or differential amplifier, a control box, and a recorder. Figure 5 shows the basic components of the second-harmonic equipment.

In both systems the control box is provided with a range or sensitivity selector-switch that makes it possible to change instantly the full-scale range or sensitivity of the recorder. Five or six sensitivity ranges are available, from 20 gammas per inch to 1000 gammas per inch. The magnetic value of the zero line or base of the recorder chart is changed

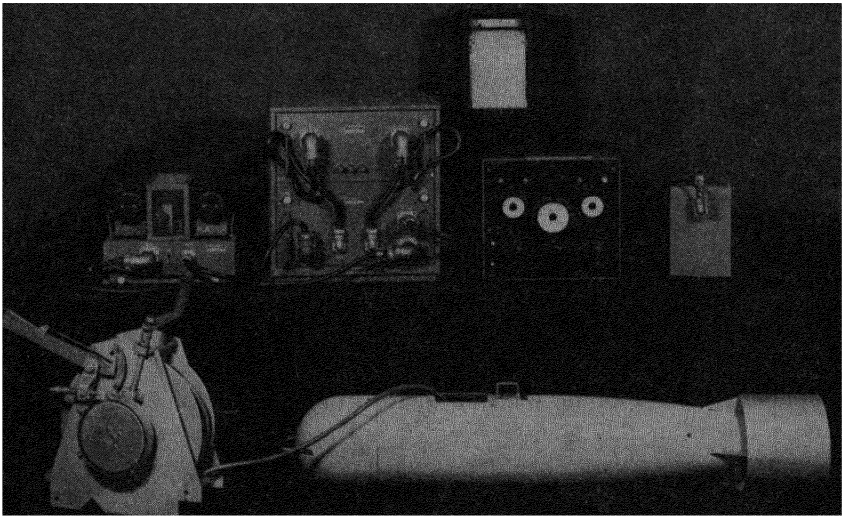


FIG. 5. Airborne magnetometer. Left to right, first row: winch, "bird"; second row: dynamotor, oscillator and converter unit, control box, battery box; top: recorder. (Official U. S. Navy photo.)

instantly, either by hand or automatically, in fixed increments, the size of which may be selected by the operator. Thus an anomaly greater than the set range of the recorder can be recorded either by decreasing the sensitivity of the recorder with the sensitivity selector switch, or by changing the range of the recorder by adding an increment to the magnetic value of the zero line.

The recorder, usually a Leeds & Northrup or an Esterline-Angus, draws a continuous magnetic profile on a roll of graph paper. Several chart speeds are available, and under most conditions one is chosen that gives a horizontal scale approximately equal to the scale at which data are to be compiled. The recorder is equipped with a device that stamps the chart with the settings of the range-selector switch and the magnetic value of the zero line of the chart.

3. Associated Equipment

The airborne magnetometers produce a continuous record of the total magnetic field along the flight path of the aircraft, but these data can not be plotted and used unless they are constantly correlated with the position of the plane in space [11–15, 42–44]. This correlation has been accomplished by several methods.

3.1. Aerial Cameras

The flight path of the plane may be recorded photographically either by taking a series of intermittent pictures with a modified 35 mm. motion-picture camera or by making a continuous photograph with the 35 mm. Sonné strip camera. The Sonné photograph is produced by moving film past a slit at a speed equal to that of the movement of the photographic images of ground points beneath the airplane. The optical path of this camera is generally stabilized by means of gyroscopically controlled mirrors so that the point vertically beneath the plane may be determined.

3.2. Electronic Navigation Aids

In unmapped areas or in regions of water, desert, or featureless terrain the photographic method of location cannot be used; and electronic navigation aids, such as Shoran or Decca, are necessary. In these methods the plane is located by translating into distance the transit time of radio pulses traveling between it and two ground stations at known locations.

Shoran, the only equipment of this type now available for use in aircraft, utilizes high radio frequencies, which have the disadvantage of requiring "line-of-sight" transmission. Therefore, at least two ground stations must be visible from the aircraft at all times. This requirement restricts the use of the method to those areas in which there is level terrain or over which low-level surveys are not required.

3.3. Altimeters

The third dimensions of the position of the aircraft is determined by one or more altimeters, either barometric or radio. The former measure the height above sea level, the latter above the ground. These measurements may be either photographically or continuously recorded.

3.4. Correlating Circuit

All records obtained in the aircraft must be correlated frequently so that the magnetic intensity can be determined at all points along the

flight path. This correlation is accomplished by an electric circuit triggered automatically or by an observer. The circuit actuates pens that mark the various recorder charts, shutters of cameras which photo-

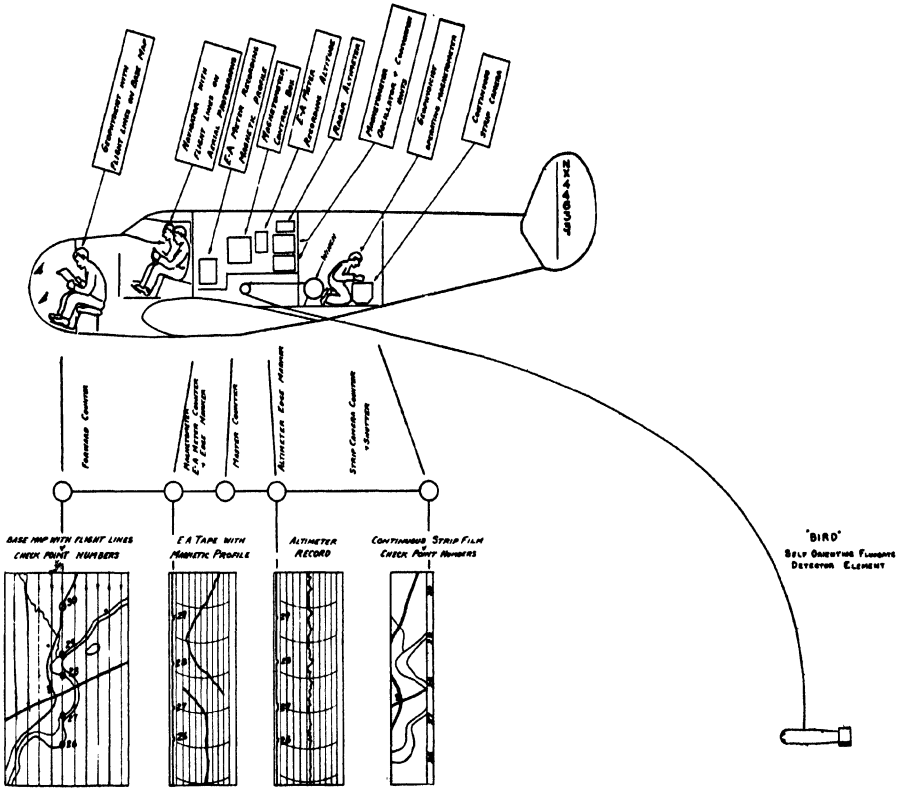


FIG. 6. Sketch of airborne magnetometer and associated equipment installed in AT-11 aircraft showing records obtained and system of correlation. (By courtesy of the U. S. Geological Survey.)

graph the various nonrecording instruments, number counters, and solenoids; these solenoids stamp recorder charts and actuate edge-marking and numbering devices on the Sonné camera (Fig. 6).

4. FIELD SURVEY TECHNIQUE

Given the airborne magnetometer, the geophysicist then has the problem of how to use it most profitably. He must install it in the aircraft best suited to his area of interest, assign to it operating personnel whose combined talents are best adapted to its use, and develop an operating procedure that will yield at lowest cost a magnetic contour

map of sufficient detail and accuracy to delimit the magnetic features of interest to him. There is no single combination of these items that is applicable to all types of surveys [12, 13, 15].

4.1. Aircraft

The basic requirements for the aircraft carrying the magnetometer are that it be safe and economical to operate and that it have a range sufficient for efficient operation and capacity for a crew of three or four and for equipment weighing 300 to 400 pounds, and an additional 200 pounds if Shoran is used.

The smallest fixed-wing aircraft that meets these requirements is that similar to the twin-engine AT-11 Beechcraft and Avro Anson trainers. These are economical to fly, but they have a range of less than 750 miles and therefore are economical to operate only if the project is small or is located very close to an airport. Probably the airplane most useful in this work is the twin-engine Douglas DC-3. It has ample load-carrying and space capacity and a normal range of about 1200 miles, permitting economical operation even when the area to be surveyed is a considerable distance from the nearest airport.

Helicopters have been used because of their ability to fly safely very close to the ground [45]. Although their cost of operation and maintenance is high, their range short, and their load-carrying and space capacity small, they are the only means by which small, detailed, low-level aeromagnetic surveys can be made. These surveys are almost directly comparable with ground surveys in cost and in results obtained.

4.2. Crew

An ideal crew consists of a pilot, copilot, magnetometer operator, and observer or Shoran operator, but space or load restrictions sometimes make it necessary to reduce this number. The pilots should have considerable experience in flying straight and parallel lines, as in photo mapping, and in low-level flying, as in crop-dusting. If Shoran is not used they should be able to interpret readily the maps and photographs used to guide them on the flight lines. The magnetometer operator should be an experienced geophysicist capable not only of operating the magnetometer and making minor repairs on it in flight but of making rough interpretations of the data as they are obtained so that he may change the flight pattern to yield most economically the most useful results.

4.3. Operational Procedure

To produce at low cost an accurate and detailed magnetic map the geophysicist using the airborne magnetometer must: (1) plan his tra-

verses so that he makes the most efficient use of the magnetometer; (2) maintain the level of his survey at the most productive height above the ground; (3) provide a means of removing from the magnetic measurements the effects due to diurnal magnetic variation and drift within the instrument; (4) record and correlate the magnetic and positional data in a form that can be used easily to make magnetic profiles or contour maps [12, 15, 46, 47].

4.3.1. Flight Pattern. For most projects the most efficient pattern is a series of parallel traverses flown at right angles to the trend of the major magnetic anomalies. Because this trend often is not known until the survey is well under way, the traverses are generally plotted at right angles to the geologic grain. The spacing of the traverses is chosen so that a minimum of interpolation of the magnetic data is required between adjacent traverses; thus each of the anomalies measured at the flight elevations must be crossed by at least two traverses.

The pilots are guided on this flight pattern by photographs or maps on which the desired flight lines have been plotted. If Shoran is used, the course is preset by the operator, and the pilot is guided on course by an electronic indicator that shows the position of the plane with respect to the preset course.

4.3.2. Flight Elevation. The intensity of an anomaly produced by a mass of magnetic material decreases rapidly with increasing distance from the source, generally as the second or third power, and the effects of two masses tend to merge at a distance roughly equal to the distance between them [48]. The effects of nongeologic magnetic materials are also subject to the same laws, but as most of these masses are relatively very small, a flight elevation can be chosen that is low enough to give the necessary magnetic detail and is high enough to reduce the effect of nongeologic material to a negligible amount. In the choice of altitude, consideration is also given to safety and ease of operation in the particular terrain involved. At low altitude the manipulation on the aircraft becomes considerably more tiring and hazardous; and because of reduced visibility, the aircraft is more difficult to keep on the prescribed flight path. Experience has shown that a flight elevation of about 1000 feet above the ground is the most reasonable, although an elevation of 500 feet is customary where the fine magnetic detail produced by shallow magnetic materials must be recorded. Where the magnetic materials are deeply buried, as in most oil-producing regions, the flights are generally made at a constant barometric elevation.

4.3.3. Magnetic-Control Pattern. In ground magnetic surveys, the effect of the diurnal variation and instrument drift can be determined by re-occupying a base station or by recording the diurnal variation with a second instrument. These methods are sometimes employed in aero-

magnetic surveying, but because of the large areas involved they are generally not feasible. The same effect is accomplished, however, by flying a series of interconnecting base loops or a series of base lines at right angles to the traverse lines. This base network is internally adjusted and provides a datum to which the traverse data can be adjusted, for each time the traverse crosses a base line the result is the same as though a base station were reoccupied. The network is laid out so that traverses cross a base line every 5 to 20 minutes, depending upon the accuracy required and upon the rate of diurnal variation. In areas where large magnetic features are found, an attempt is made to fly the base lines where the magnetic relief is low; in areas where maps are poor or the terrain featureless, the base lines are flown along easily recognizable linear landmarks such as rivers, railroads, or highways.

4.3.4. Recording and Correlation of Data. The continuous charts of the magnetic intensity and the continuous or intermittent record of the plane's altitude and position are correlated by the circuit previously described. When the continuous Sonné camera is used the observer triggers the circuit when the aircraft passes over an easily identifiable feature. He also makes an approximate plot of the point on a map that is used to aid in the later compilation of the data. With Shoran control an automatic correlation system can be used that is triggered at regular intervals of distance so that office compilation is considerably simplified. A log is kept of the number and direction of the traverse and of the beginning and ending correlating numbers.

5. Office Compilation of Field Data

The basic task undertaken in the office is to plot, in their proper position, the magnetic values determined in the field survey. Because the magnetic data are recorded at a constant rate while the ground speed of the airplane varies, the horizontal scale of the magnetic record is not constant. The track of the airplane must be determined, and the magnetic records plotted to map distances as profiles and adjusted to a common magnetic datum [12, 15, 47].

The first step in compilation is plotting the actual track of the plane. If the photographic method has been used, the positions of the correlation marks indicated by the intermittent or continuous-strip photographs are plotted on the aerial photographs and are transferred from them to the best base map available. If Shoran has been employed, the positions of the correlation marks are plotted by means of their known distance from the determined positions of the two base stations.

The second step in compilation is plotting the magnetic data at a constant horizontal scale by adjusting the correlation marks on the

magnetic record to fit the distances between the correlation marks plotted on the map. This "rectification" may be accomplished on a machine that semi-automatically sets up the proportion between correlation-mark distance on the magnetic record and the map's correlation-mark distance plotted on graph paper. The operator then traces the magnetic profile on the field magnetic record and a new "rectified" profile is scribed on rectilinear graph paper at the desired horizontal and vertical scale.

The third step is adjusting the rectified profiles to a common magnetic datum free from the effects of instrument drift and diurnal variation. The magnetic datum of the traverses is determined by the network of base lines or loops. Each base line or loop is flown first in one direction and immediately in the other. Some of the magnetic profiles when plotted gradually diverge, beginning at the turn-about point and reaching a maximum where the flight started and ended. This maximum divergence is the sum of the diurnal variation and instrumental drift that occurred during the time between the beginning and end of the base-line circuit. It is apportioned so that the two curves coincide, giving a corrected base line which shows the true value of the total magnetic intensity above or below an arbitrarily assigned base. Closure errors of base-line nets usually amount to 1 to 10 gammas per 100 miles of traverse.

After the base lines have been corrected for drift, the points of intersection of the traverses and the base lines are determined from the flight photographs or Shoran plot. At these points, the difference between the magnetic value on the traverse profiles and that on the base-line profile is determined. If the position of the plane were known exactly, the traverses could be adjusted by this simple means to the magnetic datum of the base line as shown in Fig. 7. But under the most favorable circumstances the plane's position can be determined to no better than 50 feet; so these magnetic differences must be averaged to obtain a magnetic datum that is not distorted by positional error. This is done by plotting the differences against time and drawing an average smooth curve through the points. If it is considered necessary, a least-squares adjustment can be made. This curve is the plot of combined instrumental drift and diurnal variation curve against time, with the frequency of observation established by the time intervals between base-line crossings. An arbitrary magnetic datum is chosen, and all the profiles are adjusted to it by removing the effect of diurnal variation and instrumental drift as indicated by the average curve.

The fourth step is magnetic contouring. The contour interval is chosen and intersections of the contour levels and the magnetic profile determined. The positions of these intersections are transferred to the base map, and contours are drawn by connecting points of equal mag-

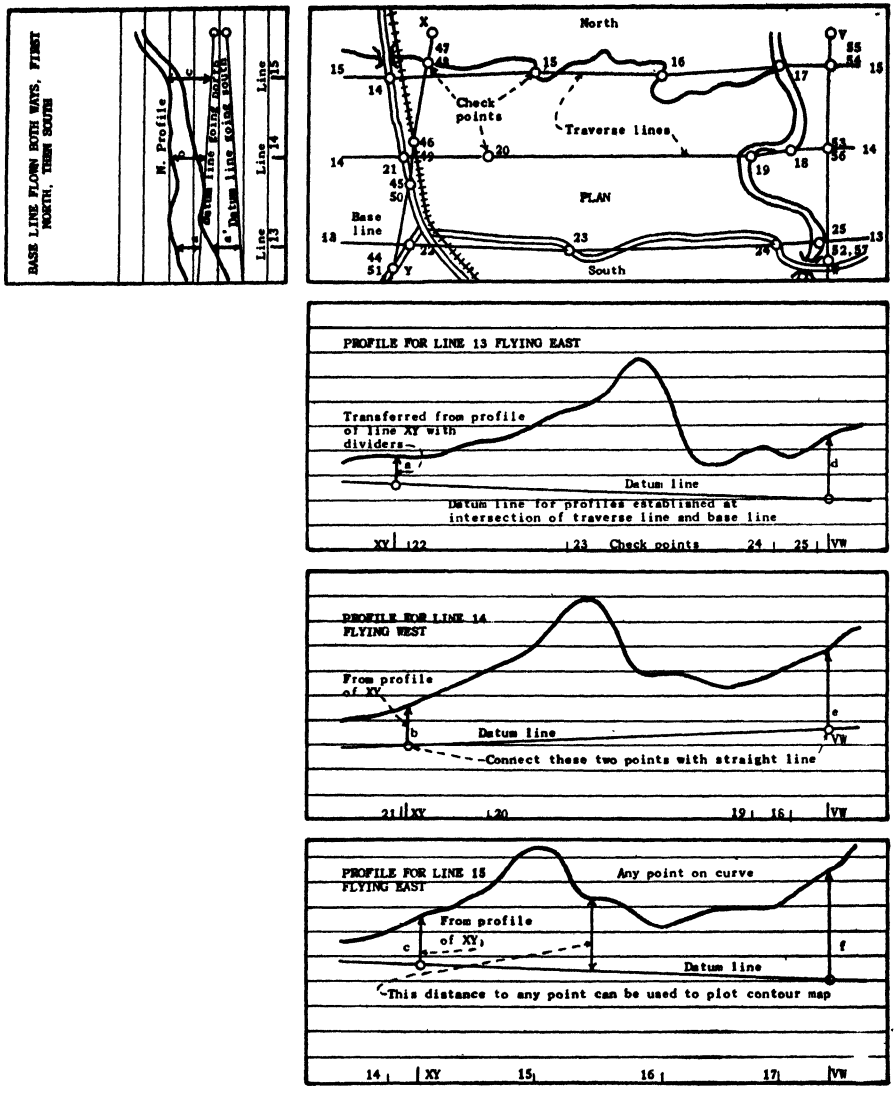


FIG. 7. Diagrammatic sketch of flight pattern of aeromagnetic survey showing method employed for magnetic control. (By courtesy of the *Engineering and Mining Journal* [15].)

netic value. This procedure has the advantage of requiring no interpolation along the traverse lines.

The magnetic contour map prepared by these procedures represents the observed variations in the total-intensity field on a surface described by the flight level of the plane.

Various simplifications of these procedures are possible under favorable circumstances. Where roads or other culture make section corners easily identifiable, the observer, using a gyrostabilized sighting device, can plot accurately the check points and thereby eliminate part of the office compilation required in more difficult terrain. When the magnetic gradient is low and the anomalies broad, a reasonably accurate contour map can sometimes be obtained by plotting the magnetic value measured only at the check points. This eliminates the compilation steps required in "rectification" but at the same time destroys the chief advantage of the airborne over the ground method—that of using the continuous profile to eliminate the interpolation required when individual measurements are plotted. Such simplifications increase the speed and reduce the cost of compilation, factors which must always be balanced against the factor of accuracy.

6. Interpretation of Results

The results of an aeromagnetic survey are compiled into a magnetic contour map or series of magnetic profiles of the same type as those obtained by ground methods, the only difference being that the ground magnetic surveys usually measure variations in either the vertical or the horizontal components of the earth's magnetic field while the aeromagnetic surveys measure variations of the total field. Therefore, the interpretation of these maps and profiles involves the same fundamental theories that for years have been applied to the results of ground surveys. Most of the literature on the analysis of magnetic data deals with the vertical or the horizontal component of the earth's field and is therefore not directly applicable to the study of aeromagnetic data of the total field. However, the necessary modification of the formulas and techniques is not difficult, and some work of this sort has already been published [49–53].

One of the early claims for the airborne magnetometer was that, because of its ability to measure the anomalous magnetic field at more than one level, it could be used to supply information which would permit unique or more accurate determination of the depth of the material producing a magnetic anomaly. Consideration of the potential theory indicates that this hope is unfounded, and the results of multi-level work have been used to prove conclusively that all the essential information of higher-level data is inherent in the low-level data [54]. It has also been shown that if information is desired at a higher level it can be obtained more quickly and at less cost by computation than by flight measurements.

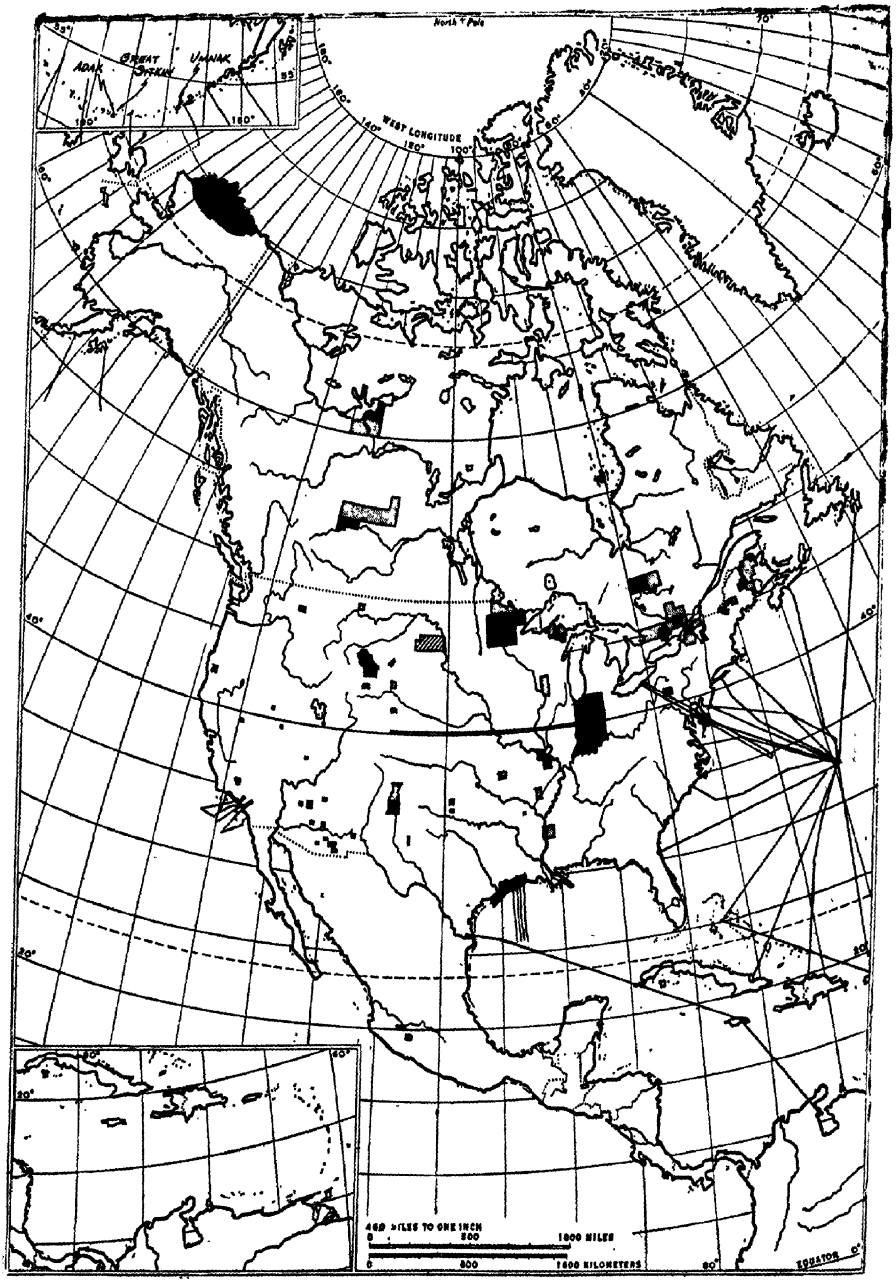


FIG. 8. Map of aeromagnetic surveys available to the public, showing in black those already published, in stipple those surveyed but not yet compiled, in crosshatch that available by purchase from Fairchild Aerial Surveys, Inc., July 1, 1951.

7. Results of Aeromagnetic Surveys

It is estimated that more than a million miles of detailed and accurate aeromagnetic traverses had been surveyed by commercial and government agencies by mid-1951, a little more than seven years since the first survey. This fact alone clearly demonstrates the usefulness and applicability of the method.

Much of the aeromagnetic work has been done by commercial organizations for mining and oil companies, and the results are therefore not generally available to the public; in fact, their existence is frequently unknown. It is known that extensive surveys have been made in at least 30 states of the United States [12, 55–73], Alaska [12, 74], Canada [11, 75, 76], Mexico (see Fig. 8), Venezuela [77, 78] and Bahamas [79–81], Mozambique [82], South Africa [83], U.S.S.R. [84, 85], Pacific Ocean [86, 87], and the eastern Atlantic Ocean [74]. It is unfortunate that the results of much of this work have not been published, but enough are available to permit discussion of the applicability of the method to several specific geologic problems. No attempt will be made to present all the data but rather to review the general conclusions and show typical examples.

7.1. Fairfax Quadrangle, Virginia

Figures 9 and 10 show the results of and geologic aeromagnetic surveys by the U. S. Geological Survey of the Fairfax quadrangle, Virginia. Included in the region are Triassic sedimentary rocks and diabase, granite, and metamorphosed Paleozoic sedimentary rocks and basaltic rock.

The Triassic diabase gives rise to pronounced anomalies, and the magnetic pattern produced by it shows a striking correlation with the geologic map. Exposures are poor in the area, and it is difficult if not impossible to determine from geologic evidence the structure of the larger masses of Triassic diabase whose outcrops are shown to be U-shaped on the geologic map. It is not known whether these masses have the structure of "ring dikes," of southward plunging anticlinal sheets, of northward plunging synclinal sheets, or of spoon-shaped sheets. However, an analysis of the aeromagnetic data shown in Fig. 11 indicates that the masses have a distorted spoon shape, with a vertical western limb and a gently dipping eastern limb. In order to analyze the observed field it was necessary to remove from the data the magnetic ridge apparently produced by a deeply buried large mass of magnetic material diagrammatically shown in the lower right-hand corner of the structure section of Fig. 11.

The Triassic sedimentary rocks are essentially free of magnetic material and are expressed by a very gentle magnetic gradient. The granite, as shown in Fig. 9, apparently has a slightly higher magnetic susceptibility than the Triassic sedimentary rocks for it produces minor

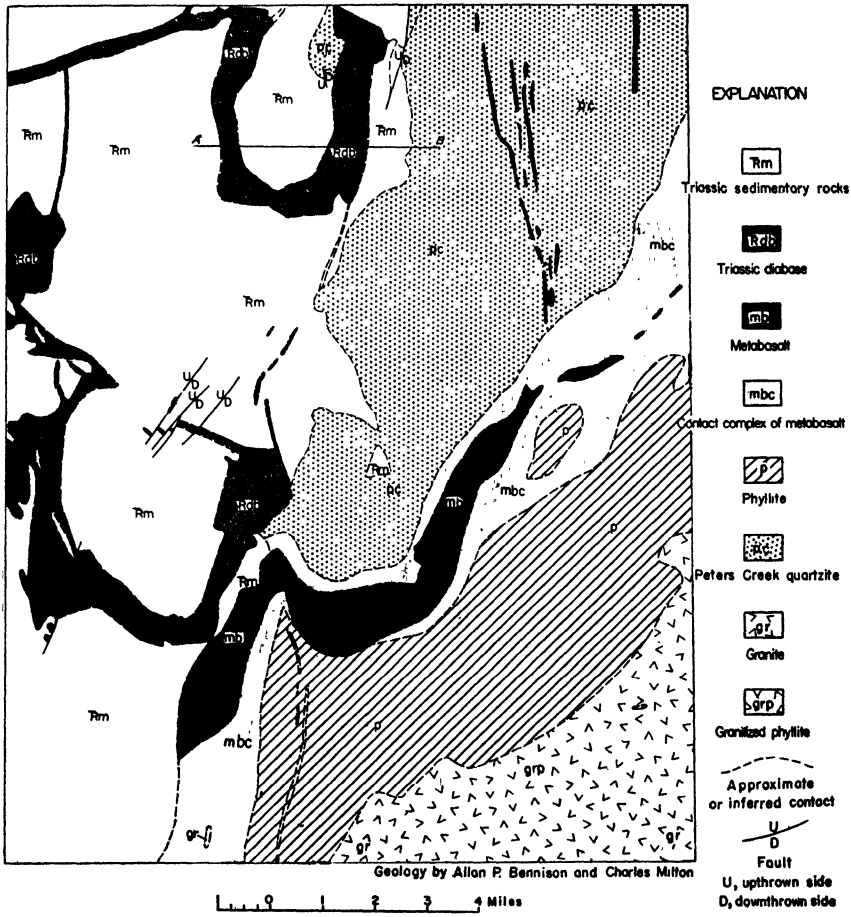
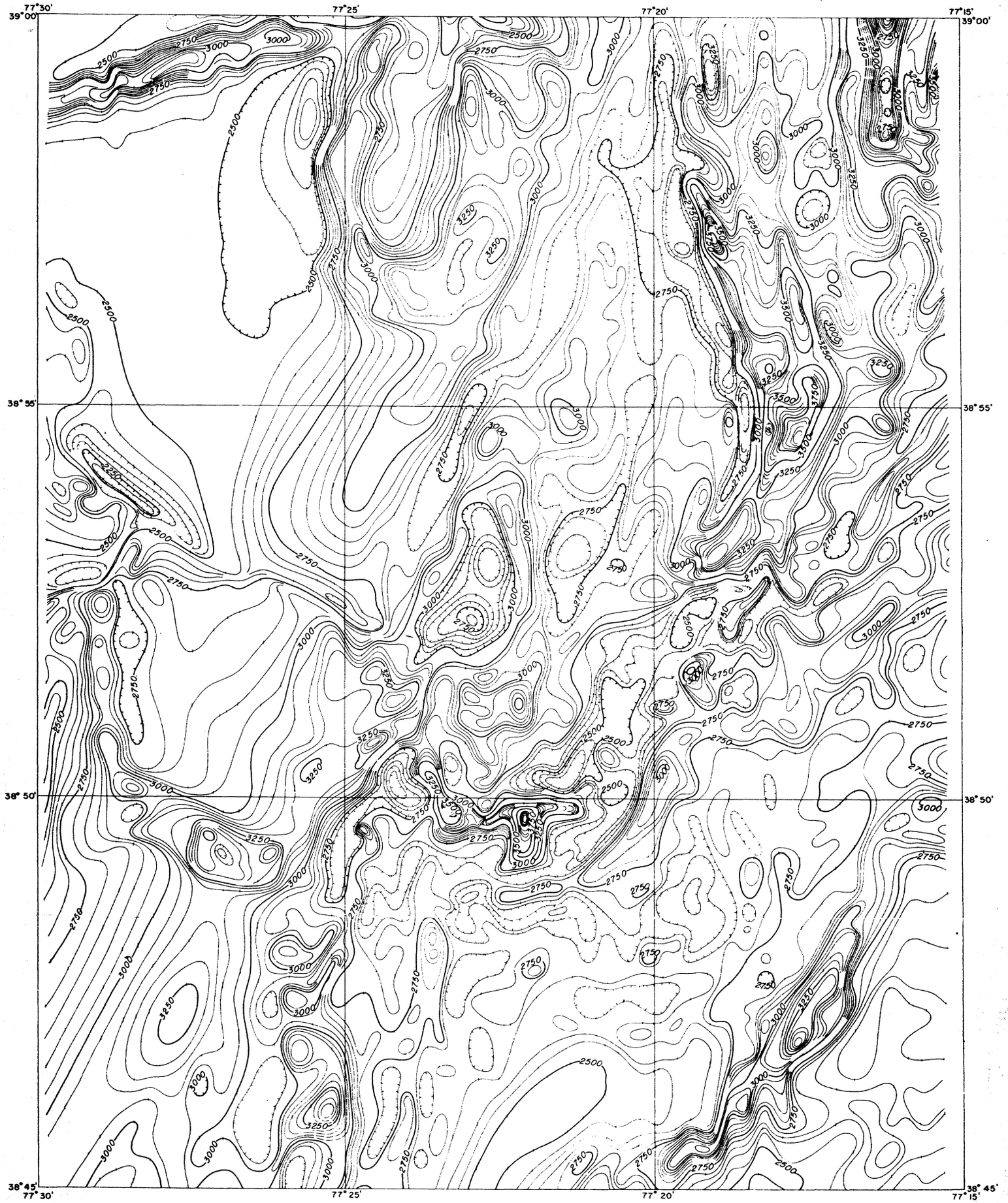


FIG. 9. Geologic map of Fairfax Quadrangle, Virginia. (Data from U. S. Geological Survey.)

broad anomalies with the exception of the one northeast trending large anomaly shown in the lower right-hand corner of the magnetic map (Fig. 10). This anomaly may be caused by a magnetic border facies of the granite, but because exposures are particularly poor in this area, no supporting geologic evidence has been obtained.

The phyllite and schist have, as one would expect, varying but generally low susceptibility, which produces small and irregular anomalies.

Fig. 10. Aeromagnetic map of Fairfax Quadrangle, Virginia. (Data from U. S. Geological Survey.)



BALSLEY, ADVANCES IN GEOPHYSICS, VOL. I

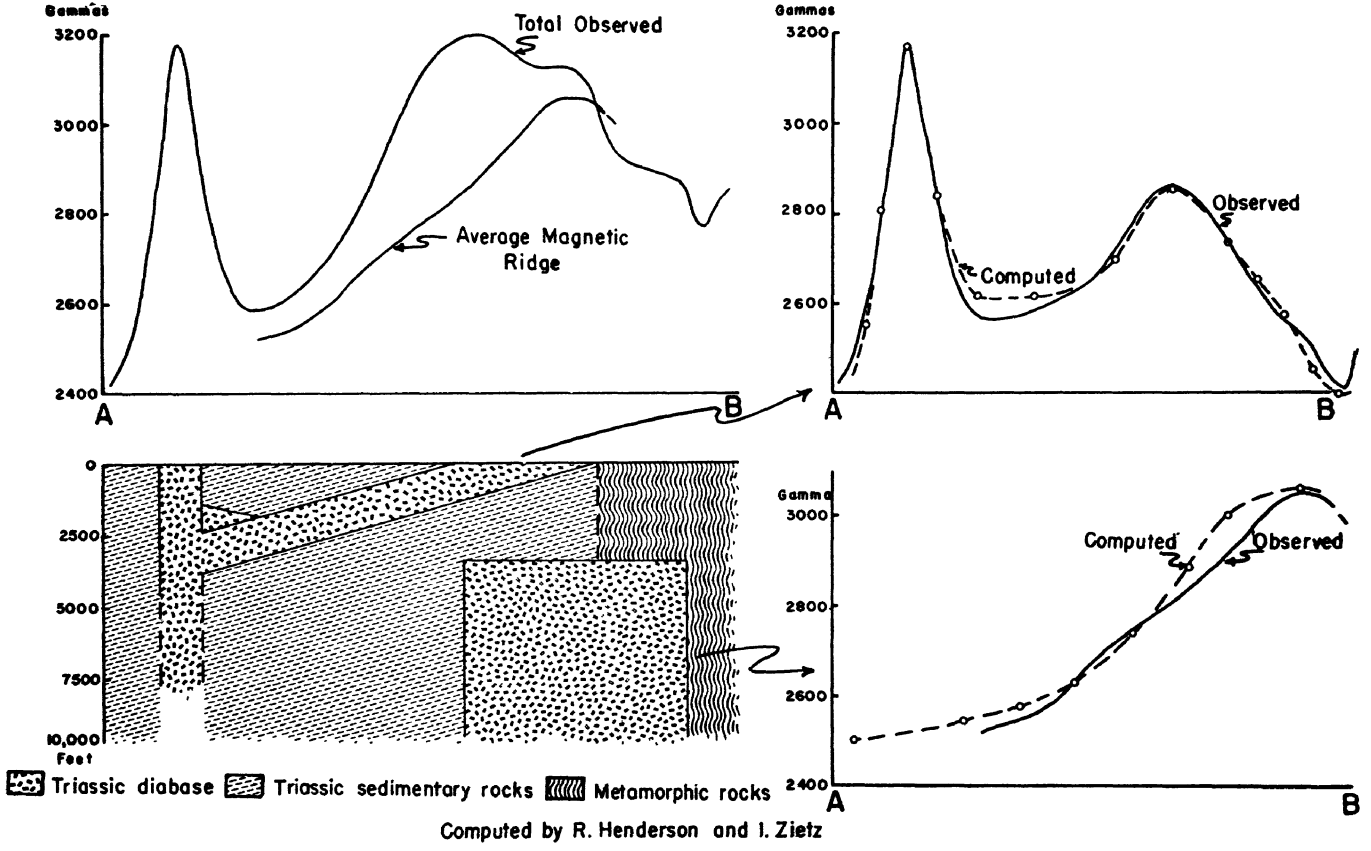


FIG. 11. Analysis of aeromagnetic profile in Fairfax quadrangle, Virginia. (By courtesy of the U. S. Geological Survey.)

The average susceptibility of the schist seems to be slightly higher than that of the phyllite.

The metabasalt produces a distinctive pattern of large anomalies that can be correlated easily with the geologic map. The large anomalies in the upper right-hand quarter of the magnetic map are caused by a swarm of dikes of serpentinized peridotite; it is interesting to note that the dike in the northeastern corner produces a negative anomaly. This is apparently caused by a strong inverse remanent magnetization of the material in the dike. The intense anomaly shown in the central part of the magnetic map is not produced by the structure of the mass of metabasalt but is apparently due to a local increase of the magnetic susceptibility at this point. No bedrock is exposed in this area, but an abundance of magnetite in the soil tends to verify this assumption.

7.2. Northwestern Maine

Figure 12 shows part of an aeromagnetic and geological reconnaissance survey of 2000 square miles in northwestern Maine conducted by the Aero Service Corporation and members of the Geology Departments of Harvard University and the Massachusetts Institute of Technology [60, 61]. The study was undertaken to map masses of ultrabasic rocks that might localize deposits of asbestos, chromite, copper, and talc. As by-products, a geologic map and correlated aeromagnetic survey have been produced.

This correlation is discussed in the following excerpt from the report of Hurley and Thompson [61]:

"In the area underlain by gneiss, the concentrations of magnetite probably tend to follow the warped bedded structures of the gneiss. These structures were not mapped in detail, but there appears to be a principal broad anticline that curves from a northeasterly direction to a northerly one, proceeding north. The principal magnetic trend follows this pattern, with a fair degree of correlation between magnetic contours and strike of bedding. To the west of this strip, the structure shows sharp changes of strike with moderate dips, again paralleled to some extent by the magnetic contour lines. The area has been deeply cut by the drainage. The overall result of the occurrence of magnetite in dissected curved planes is to cause magnetic contour shapes tending to intricate convolutions.

"In contrast with this, the nonlaminated granites, with their more equidimensional concentrations, produce more closed contour patterns of a 'bird's-eye maple' appearance.

"The lack of any sharp gradients in the block of Moose River sediments in the central part of the map is characteristic of this formation

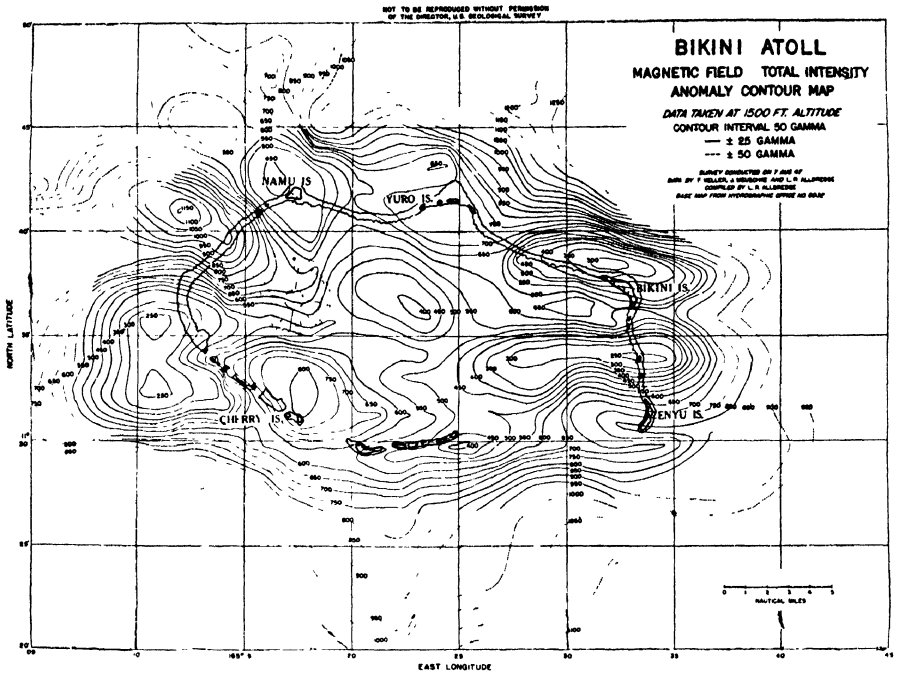


FIG. 13. Aeromagnetic map of Bikini Atoll. (By courtesy of the American Geophysical Union [87].)

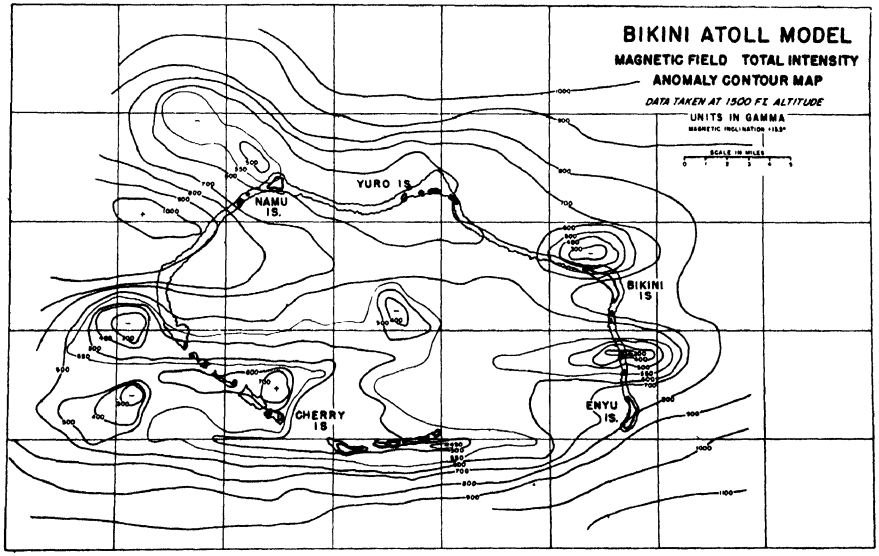


FIG. 14. Magnetic map of model of Bikini Atoll. (By courtesy of the American Geophysical Union [87].)

over the entire area in which these sediments occurred. The magnetic contours conform almost identically with the dikes in the belt of mixed sedimentary and igneous rocks referred to as Siluro-Devonian, and were most useful for plotting the location of the intruded masses beneath the over-burden."

7.3. Bikini Atoll

Figures 13, 14, and 15 show the results of an interesting analysis of an aeromagnetic survey of Bikini Atoll conducted by the U. S. Geological

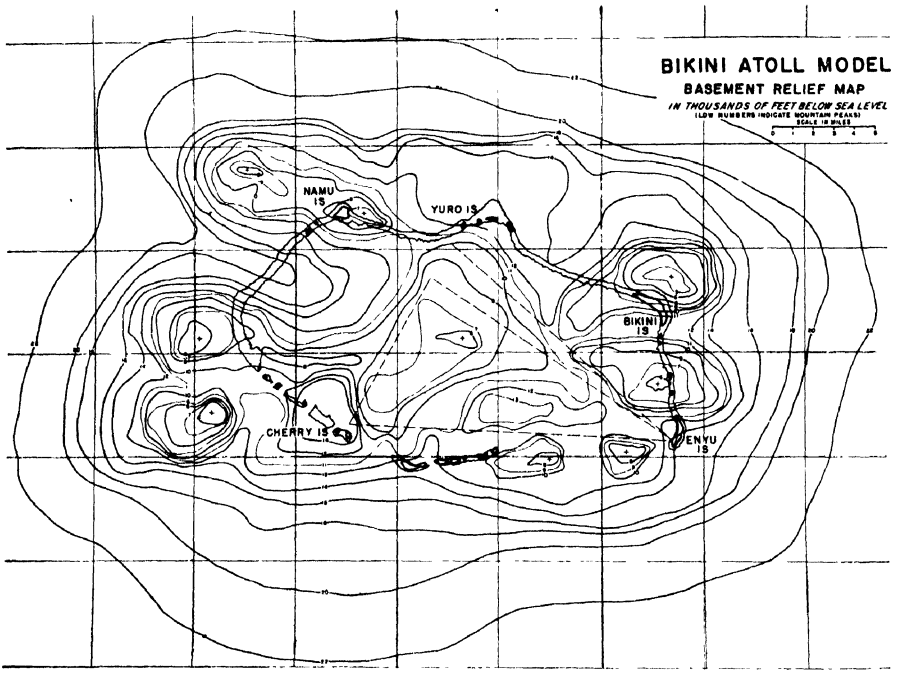


FIG. 15. Relief map of magnetic model of Bikini Atoll. (By courtesy of the American Geophysical Union [87].)

Survey, the Office of Naval Research, and the Naval Ordnance Laboratory [87]. This survey was undertaken to investigate the supposed volcanic foundation of a typical coral atoll. Alldredge and Dichtel assumed a uniform susceptibility and zero permanent magnetization of the magnetic basement rocks; and, using a hand-mixed magnetic clay, they constructed a model of the magnetic basement rocks beneath the atoll. An infinity of such models is possible, but by choosing a susceptibility of 0.008 cgs unit and by using the depth data from three seismic profiles (dashed lines in Fig. 15) it was possible to obtain a unique solu-

tion. Holding the three seismic profiles invariant, a clay model was built and surveyed with a miniature magnetometer. The model was then modified until the miniature magnetometer survey shown in Fig. 14 coincided with the actual survey shown in Fig. 13. The resulting model, Fig. 15, represents, therefore, the relief of the mass of homogeneous basement rocks with uniform susceptibility of 0.008 cgs unit and zero permanent magnetization, which will produce the observed magnetic field.

7.4. Upper Peninsula, Michigan

Figure 16 shows a series of typical north-south profiles obtained from aeromagnetic surveys conducted by the U. S. Geological Survey [62] of

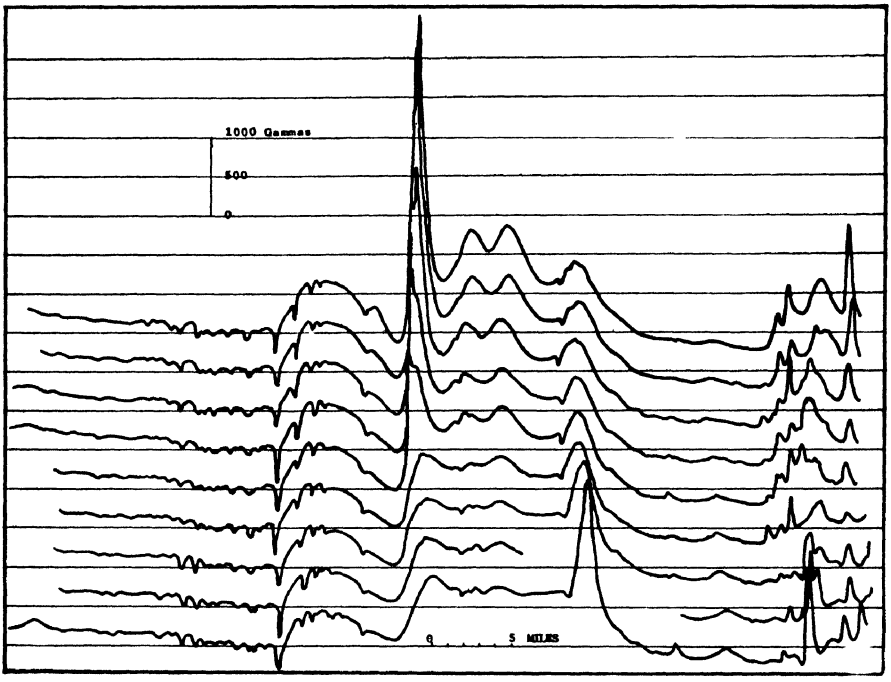


FIG. 16. Typical aeromagnetic profiles in northern Michigan. (By courtesy of the U. S. Geological Survey.)

more than 5200 square miles in the iron-producing region of the Upper Peninsula of Michigan. Thick glacial till covers most of the region and outcrops are sparse, making geologic mapping difficult in the areas between the established mining districts. The aeromagnetic survey was undertaken to supply a framework into which the outcrop and drill-hole information could be fitted. By relating the scattered geologic data to

the observed magnetic trends it is then possible to determine the trend of the geologic structures.

The region is almost entirely underlain by pre-Cambrian rocks; pre-Huronian granite, Lower Huronian or pre-Huronian greenstone, Middle and Upper Huronian slate, graywacke, and iron-formation, and post-Huronian sandstone, conglomerate, and basalt, or diabase.

The most pronounced magnetic anomalies are produced by a magnetic slate 100 to 400 feet stratigraphically above the iron-formation. By tracing the anomaly produced by this formation of the Middle and Upper Huronian sedimentary rocks, it is possible to delineate the geologic structures that incorporate the iron-formation. The series of four anomalies in the center of Fig. 16 are probably produced by this slate and indicate that it has been folded into plunging synclines and anticlines. The sharpness and intensity of the major anomaly indicates that the slate is close to the surface. This information was used to guide a ground party that located one exposure beneath the upper traverse line. This anticline pitches in the direction of the lower profile.

The complex magnetic features of the right end of the profiles reflect the complicated geologic structure of the Iron River mining district. Here both magnetic slate (1000-gamma anomaly, lower right) and greenstone (1000-gamma anomaly, upper right) produce magnetic features of similar appearance.

The series of small sharp negative anomalies in the left of Fig. 16 are produced by a swarm of parallel diabase dikes, post-Huronian in age. These dikes have normal magnetic susceptibility but have strong inverse remanent magnetization that more than compensates for the induced magnetization and produces the negative anomaly. The inverse remanent magnetization of these dikes is remarkably consistent in both direction and intensity and though its cause is not yet known, it is not due to overturning.

7.5. Allard Lake District, Quebec

An aeromagnetic survey totaling 4500 miles of traverse has been made for the Kennecott Copper Co. in an ilmenite-bearing district in Quebec [76]. The survey was undertaken to "outline or delimit the areas of ilmenite-hematite mineralization and in addition to simplify the work of 'dip needle' ground parties attempting to prospect in the more inaccessible and difficult areas" [76]. The ore, coarsely granular ilmenite-hematite, occurs as tabular lenses, dikes, or sills in a large mass of anorthosite and anorthositic gabbro considered to be in the pre-Cambrian Morin series. "In all instances the large massive deposits gave negative anomalies with very sharp gradients, in the magnitude of

3000 to 5000 gammas below the average plateau level. In some instances areas of ilmenite-rich anorthosite also produced negative anomalies, but with less steep gradients. It is of interest to note that in horizontal plan the areal limits of certain anomalies correspond fairly well with the known surface position of the ilmenite bodies. However, since the attitude of the bodies in depth is not known we cannot draw a critical analogy. . . .

"The strong negative anomalies are believed to result from the effect of negative polarization of the ore body itself. Similar negative observations were obtained when dip needle traverses were run over the same occurrences.

"The contact zone between the granite and Morin series produced strong positive anomalies, often in the order of 1–5000 gammas. This is believed due to concentrations of disseminated magnetite in hybrid rocks of the contact zone. The gabbroic facies of the Morin stands out clearly on the aeromagnetic map in rather marked contrast to normal anorthosite. The anorthosite itself exhibits broad magnetic trends with little relief" [76].

7.6. Adirondack Mountains, New York

An aeromagnetic survey of more than 6000 square miles in the northwestern, northern, and southeastern Adirondack Mountains has been completed by the U. S. Geological Survey [66, 67, 69, 72]. This survey was part of a program of geologic mapping in the district to aid the discovery and exploration of deposits of iron ore. Aeromagnetic anomalies indicative of magnetite deposits were located, and dip needle surveys were made of the most promising [68]. By this means, seven deposits of possible economic significance, amounting to 6,000,000 tons of ore indicated by diamond drilling or 14,000,000 tons of inferred ore, were discovered. An additional eleven deposits, amounting to 40,000,000 tons of potential ore, but which are either too small or too low in grade to be of economic importance, were discovered.

It is interesting to note that 68.5 feet of ore was found in the first drill hole bored in the first ore deposit discovered from the air in the Western Hemisphere. This deposit, the Outafit in the Stark quadrangle, New York, is completely covered by a thick overburden of glacial sand and gravel and was discovered by the aeromagnetic survey made in June 1945.

In addition to these results of direct economic value, the aeromagnetic survey has provided information that has greatly aided the program of regional geologic mapping. Much of the area is accessible only with difficulty and many of the valleys are filled by swamps or glacial till.

By using the magnetic data it has been possible to direct the ground geologic mapping to the most critical areas and to trace many of the geologic formations beneath the valley fill. Figures 17 and 18 illustrate this type of problem [69]. The Lyon Mountain granite gneiss (ly) pro-

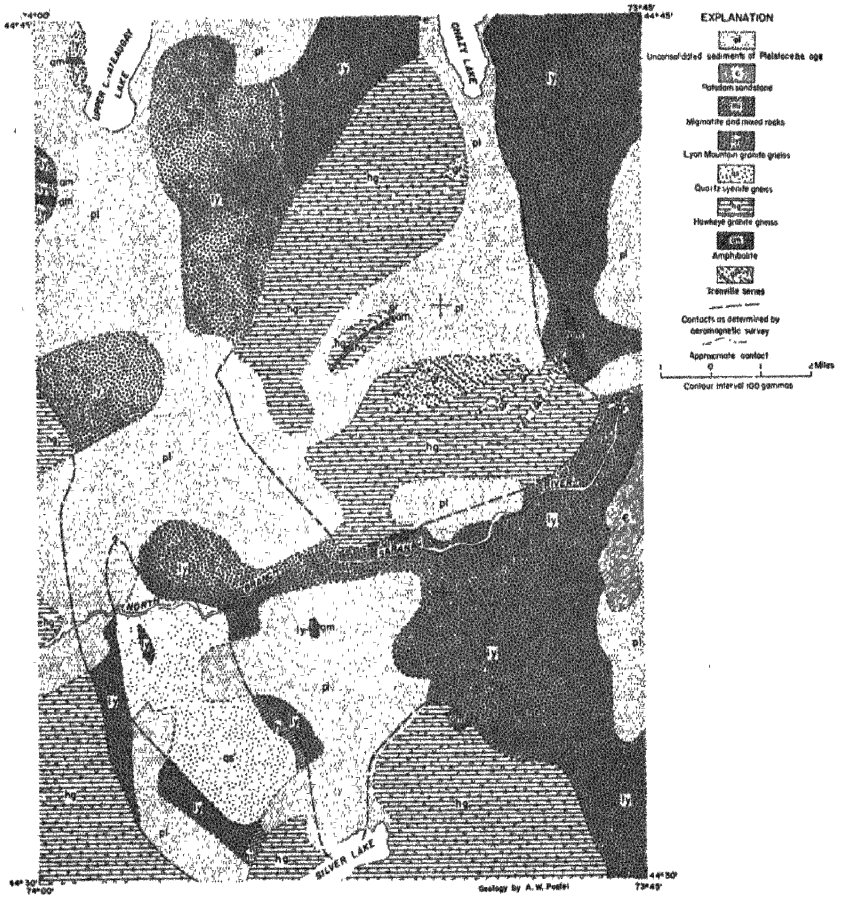


FIG. 17. Geologic map of Lyon Mountain Quadrangle, New York. (Data from U. S. Geological Survey.)

duces strong positive anomalies; and by correlating them with the rock exposures, it is possible to draw reliably the geologic contacts buried beneath the Quaternary valley fill. These "magnetically inferred" contacts are shown by the heavy dashed lines in Fig. 17. Although both the quartz syenite gneiss (qs) and the Hawkeye granite gneiss (hg) give rise to an anomalous field less than that of the Lyon Mountain granite gneiss they may be differentiated because the anomalous field of the

Hawkeye granite gneiss is measurably less than that of the quartz syenite gneiss.

It is not possible here to discuss all the results of the aeromagnetic and geologic surveys of an area as large and as geologically complex as the region surveyed, but a few general conclusions are of interest.

Local variations of the total magnetic intensity are produced by the magnetic properties of masses of rock, but are generally complicated by the shape and structure of the mass. When the size of the rock mass is sufficiently large, however, these effects are limited to its boundaries and it is possible to correlate the observed magnetic intensity directly with the magnetic susceptibility and mineralogic composition of the rock. This correlation has been made for suitable areas in the Adirondacks and it was found that each additional 1% of magnetite content of the rock increased the magnetic susceptibility by 1,000–2,000 cgs units and raised the general level of the observed total magnetic intensity by 500–600 gammas. This rough relationship does not hold true if the magnetite content exceeds 6%.

Negative anomalies in the Adirondacks are invariably associated with a single type of rock—microcline granite gneiss. This rock contains both magnetite and titanhematite (hematite containing up to 10% TiO_2 in solid solution). The relative proportion of these two minerals varies; where the titanhematite-magnetite ratio is high, the anomalies are negative and where it is low, positive. The susceptibility and remanent magnetization of typical specimens were measured in the laboratory. The susceptibility is low and the remanent magnetization inverse in direction where the titanhematite-magnetite ratio is high and the susceptibility increases and the attitude of the remanent magnetization becomes more normal as the titanhematite-magnetite ratio decreases. It seems, therefore, that the titanhematite is inversely magnetized and produces the negative anomalies. This effect opposes that of the induced magnetization of the magnetite so that both add algebraically to produce the anomaly, either positive or negative depending upon the relative proportion of the two minerals.

7.7. Eastern Pennsylvania

Probably the economically most important discovery yet made directly by the airborne magnetometer is that of a very large magnetite ore body in eastern Pennsylvania. The magnetic anomaly associated with this deposit was found in an aeromagnetic survey of 1300 square miles conducted by Aero Service Corp. for a private organization [70]. Unfortunately nothing is known of the occurrence except that the

deposit is very large, is at a depth of 1500 to 3000 feet, and has no surface expression.

8. Advantages and Limitations

The airborne magnetometer described in the foregoing sections has certain advantages and limitations, many of which are indicated by the typical examples of the results of its use [12, 13, 88]. It is an instrument that can be used with great rapidity and little cost to obtain accurate results of a reconnaissance type, but for some types of projects these particular advantages must be considered limitations.

8.1. Speed, Ease, and Economy of Operation

The greatest advantage of the airborne magnetometer is its speed, for a three-or four-man field crew can generally survey more than 100 miles of useful traverse per flying hour, 7,000 to 10,000 miles per month. This excludes the time lost in turns, establishing base lines, and flights to and from the airport. Although this rate is affected slightly by the length of the flight lines, and size of the project and its distance from the airport, the rate is practically independent of the type of terrain, a factor that can change the speed of ground surveys many times. The rate of office compilation depends upon the type of result desired and to some extent upon the quality of the base maps and the intensity of the magnetic anomalies, but to produce a contour map requires generally about one man-hour of office compilation per traverse mile.

The cost of operation is also practically independent of the type of terrain, but it is affected by the same factors that affect the speed of field survey and office compilation [89]. Although commercial organizations quote costs that vary considerably with the company and the project, they will conduct an average survey and prepare a contour map for about $10 per traverse mile. If Shoran is used for location, the cost is generally doubled.

8.2. Quality of Results

The over-all accuracy, that is, the ability to make and duplicate a true magnetic map, is dependent upon both the precision of the magnetic measurement and the accuracy of the position measurement. The airborne magnetometer can measure the difference between two magnetic fields with an accuracy of ± 2 gammas, or ± 1 gamma if particular care is exercised. Because it is possible to make frequent crossings of a base line it is possible to make a more accurate correction for diurnal variation in an aeromagnetic survey than is usually possible in a normal ground survey. For these reasons, the precision of magnetic measurement is

somewhat higher with the airborne magnetometer than with the ground magnetometer.

On the other hand, the accuracy of position measurement in most aeromagnetic surveys can be no better than that of the best available map, and even by using a highly accurate map at a scale of at least 1/31,680 or by using Shoran, the accuracy of position measurement is generally no better than ± 100 feet. This accuracy is considerably less than that obtained by most ground magnetic surveys.

Although the precision of magnetic measurement and accuracy of location are the primary factors affecting the over-all accuracy of a magnetic map, two other factors must be considered when comparing air and ground surveys. The usual ground survey consists of a series of measurements at points with interpolation between, but the air survey consists of a series of line measurements with interpolation in only one direction. This permits a more realistic contouring of the magnetic features along a flight line in an aeromagnetic map. However, because any discrepancies in the magnetic measurement exist between lines rather than between points, as in a ground survey, there is a tendency to develop a false "herringbone" in the magnetic contours, a vague linearity parallel to the direction of flight. These discrepancies have a random distribution and are of the same order of magnitude as those found in ground surveys, but because they are expressed along the length of a line rather than at a point they are more apparent in aeromagnetic contour maps than in the maps obtained by ground methods.

Measurement along a line guarantees that no magnetic feature along a traverse may be missed as between point measurements, and for this reason airborne surveys for some purposes may be more useful than ground surveys. On the other hand, magnetic anomalies attenuate rapidly with distance and tend to merge, so complex magnetic features detected by a detailed ground magnetic survey are not always apparent in an airborne survey made 500 feet above the surface. For the same reason, however, the airborne survey is free of the spurious magnetic features created by most of the nongeologic structures.

The two methods are not directly comparable in quality of results and are not suited to the same type of survey. In general, the quality of their results is similar but the aeromagnetic method is best suited to accurate reconnaissance work and the ground method to accurate detailed work.

8.3. Instrumentation

The greatest disadvantages of the aeromagnetic method are the large capital expenditure required to purchase the necessary equipment and

airplane and the high overhead necessary for their operation. This restricts the use of the method to continued large-scale operation and generally makes its use unfeasible for all but companies whose holdings are of sufficient extent to be surveyed efficiently by this means. In a few instances several companies have formed a cooperative group and have had a survey made of their combined holdings; in one instance Fairchild Aerial Surveys, Inc., made an aeromagnetic survey available by purchase to any buyer. However, the small operator must wait until these practices become more common before he will be able to make use of the airborne magnetometer.

9. Applicability

Th eairborne magnetometer is ideally suited to some magnetic surveys but cannot be used for others. It can be used most productively in relatively flat areas not easily accessible on foot for which good maps and photos are available or in which Shoran or other navigation aids can be easily used. It is ideally suited to nearly all projects that require an accurate reconnaissance map of a large area or whose primary function is prospecting for magnetic anomalies.

The airborne magnetometer is of limited usefulness in mountainous regions for which a detailed and accurate magnetic map is required; in such areas the difficulty of aircraft operation introduces errors into the results. It must be used with caution on any project that involves complex and detailed anomalies produced by shallow magnetic deposits.

Unless a helicopter is used, the magnetometer can seldom be applied economically to projects covering less than 25 square miles or projects requiring accuracy of location to better than ± 50 feet.

In summary, the airborne magnetometer can be used to provide a low-cost accurate survey of large areas, which can then be used to delineate localities for more expensive ground work, both geologic and geophysical. It does not eliminate the need for ground magnetic surveys, but rather relieves the ground magnetometer of the load of reconnaissance work and enables it to be used more productively on detailed work. The development of the airborne magnetometer has not broadened the fundamental science of geophysics, but has placed in the hands of the geophysicist an instrument that can provide him with magnetic maps of tremendously greater coverage in much less time, at less cost, and in some instances of greater accuracy than has heretofore been possible.

List of Symbols

A Cross-sectional area
B Flux density
B_M Flux density at saturation

e Electromagnetomotive force
h Electrically induced magnetic field
h_0 Amplitude of electrically induced magnetic field
H Total magnetic field
H_0 External magnetic field
I Inclination
μ Magnetic permeability
N Number of turns
ϕ Angle between earth's magnetic field and axis of detector coil
t Time
T Total magnetic intensity
V Vertical (90°)
ω Phase angle

References

1. Heiland, C. A. (1935). Geophysical mapping from the air: its possibilities and advantages. *Eng. a. Min. J.*, **136,** 609–610.
2. Grotewahl, M. (1930). Bericht über die Versuchsfahrt des Bidlingsmaierschen Doppelkompasses mit dem Luftschiff Graf Zeppelin. *Terr. Magn.*, **35,** 226–229.
3. Lundberg, H., and Sundberg, K. (1932). Discussion of paper by A. S. Eve, A magnetic method of estimating the height of some buried magnetic bodies. *Am. Inst. Min. Met. Erg. Geophysical Prospecting*, p. 209.
4. Logachev, A. A. (1946). The development and applications of airborne magnetometers in the U. S. S. R. *Geophysics*, **11,** No. 2, 135–147; *Petrol. Eng.* (July 1, 1946), **17,** No. 10, 211–218.
5. Logachev, A. A. (1947). Aeromagnetic surveys in prospecting for iron ore deposits (in Russian). *Vses. Nauch.-Issled. Geol. Inst. Mater., Geofiz.*, No. 11, 3–7.
6. Ramsayer, K. (1941). Die Änderung magnetischer Störgebiete mit der Höhe und ihr Einfluss auf die Flugnavigation. (Variation of magnetic anomalies with altitude and its influence on aerial navigation.) *Beit. angew. Geophys.*, **9,** 65–97.
7. Gebhardt, R. E. (1942). Investigation of height of local magnetic anomaly at Port Snettisham, southeastern Alaska. *Terr. Magn.*, **47,** No. 2, 165–170.
8. Antranikian, Haig (1936). Magnetic field direction and intensity finder. U. S. Pat. 2,047,609. (July 14, 1936.)
9. Aschenbrenner, H., and Goubau, G. (1936). Eine Anordnung zur Registrierung rascher magnetischer Störungen. *Hochfreguenztech. u. Electroakust.*, **47,** 177–181.
10. Elmen, G. W. (1936). Magnetic alloys of iron, nickel, and cobalt. *Bell Systems Tech. J.*, **15,** 113–135.
11. Bailey, R. (1948). Canadian aerial magnetic surveys. *Canadian J. Res.*, **F26,** 523–539.
12. Balsley, J. R. (1946). The airborne magnetometer. *Geophys. Invest.*, Prelim. Rept. 3, 8 pp.: *Petrol. Eng.*, **17** (1946), No. 11, 77–87; No. 12, 104–130.
13. Eckhardt, E. A. (1946). Airborne magnetometer. *Oil a. Gas J.*, **45,** No. 5, 78–79, 91–92.
14. Jensen, H. (1945). Geophysical surveying with the magnetic airborne detector AN/ASQ-3A. *U. S. Naval Ordnance Lab. Rept.* 937, May, 1945, 63 pp., Washington, D. C.
15. Knoerr, A. W. (1946). The airborne magnetometer, a new aid to geophysics. *Eng. a. Min. J.*, **147,** No. 6, 70–75.

16. Muffly, G. (1946). The airborne magnetometer. *Geophysics*, **11**, No. 3, 321–334.
17. Quaile, J. E. (1947). Airborne submarine detector. *Military Eng.*, **39**, No. 258, 166–167.
18. Beers, R. F. (1948). Some problems of magnetometer surveys. *Geophysics*, **13**, No. 3, 495.
19. Jensen, H. (1946). Validity of data from airborne magnetometer. *World Petrol.*, **17**, No. 9, 45.
20. Rumbaugh, L. H., and Alldredge, L. R. (1949). Airborne equipment for geomagnetic measurement. *Trans. Am. Geophys. U.*, **30**, 836–848.
21. Felch, E. P., Means, W. J., Slonczewski, T., Parratt, L. G., Rumbaugh, L. H., and Tickuer, A. J. (1947). Airborne magnetometers. *Elect. Eng.*, **66**, 680–685.
22. U. S. Naval Ordnance Laboratory (1941). Physical principles of the fluxgate and other inductor magnetometers. *Mine Unit Rept.* 263, 44 pp.
23. Vacquier, V. V. (1945). The Gulf absolute magnetometer. *Terr. Magn.*, **50**, No. 2, pp. 91–104.
24. Vacquier, V. V. (1946). Apparatus for responding to magnetic fields. U. S. Pat. 2,406,870. (Sept. 3, 1946.)
25. Wyckoff, R. W. D. (1948). The Gulf airborne magnetometer. *Geophysics*, **13**, No. 2, 182–208.
26. Laird, A. G., and Slonczewski, T. (1947). Magnetic field detector. U. S. Pat. 2,426,622. (Sept. 2, 1947.)
27. Merrill F. G. (1948). Magnetic detector. U. S. Pat. 2,448,613. (Sept. 7, 1948.)
28. Roberts, E. B. (1947). The future use of the airborne magnetometer in general magnetic mapping. *Photogram. Eng.*, **13**, No. 4, 641–643.
29. Schonstedt, E. O., and Irons, H. R. (1949). Airborne magnetometer for measuring the earth's magnetic vector. *Science*, **110**, 377–378.
30. Felch, E. P., Jr., and Slonczewski, T. (1949). Magnetic field strength indicator. U. S. Pat. 2,468,968. (May 3, 1949.)
31. Slonczewski, T. (1949). Magnetic field strength indicator. U. S. Pat. 2,485,931. (Oct. 25, 1949.)
32. Slonczewski, T. (1949). Detection system. U. S. Pat. 2,488,341. (Nov. 15, 1949.)
33. Felch, E. P., Jr., Merrill, F. G., and Slonczewski, T. (1949). Detection system. U. S. Pat. 2,488,389. (Nov. 15, 1949.)
34. Bell Telephone Laboratories (1943). Magnetic airborne detector, development of a magnetic orienting system. *U. S. Office Sci. Res. and Devel. Rept.* **1309**, 72 pp. (Jan. 1943.)
35. U. S. Navy Department (1943). AN/ASQ-3A equipment. *Navy Aeronaut. Spec. M597*, 7 pp. (Nov. 1943.)
36. U. S. War Department, U. S. Navy Department, and Air Council of the United Kingdom (1944). Handbook of Maintenance Instructions for AN/ASQ-3A Equipment. Army-Navy 08-29-14, 193 pp. (June 1944.)
37. Richardson, M. S., and Weid, A. C. (1943). Operation of the universal magnetometer head with linear non-ideal control mechanism. *U. S. Office Sci. Res. and Devel. Rept.* **1930**, 13 pp. (Sept. 1943.)
38. Columbia University (1945). Airborne Instrument Laboratory. Magnetometer suspensions. *U. S. Office Sci. Res. and Devel. Rept.* **5051**, 28 pp. (May 1945.)
39. Smith, P. M. (1944). The towed bird, mechanical details. *Columbia Univ. Airborne Instrument Lab. Rept.*, 32 pp. (Feb. 1944.)

40. Tolles, W. E. (1943). Compensation of induced magnetic field in MAD-equipped aircraft. *U. S. Office Sci. Res. and Devel. Rept.* **1386,** 19 pp. (April 1943.)
41. Tolles, W. E., and Vacquier, V. V. (1944). Compensation of magnetic fields in MAD-equipped aircraft. *U. S. Office Sci. Res. and Devel. Rept.* **4187,** 87 pp. (July 1944.)
42. Burch, J. E. (1947). Cartographic aspects of the airborne magnetometer. *Photogrammetric Eng.*, **13,** No. 4, 633–639.
43. Jensen, H., and Balsley, J. R. (1946). Controlling plane position in aerial magnetic surveying. *Eng. a. Min. J.*, **147,** No. 8, 94–95, 153–154.
44. Klaasse, J. M., Rumbaugh, L. H., and Jensen, H. (1947). Applications of Shoran and photographic techniques to aerial magnetic surveys. *Oil a. Gas J.*, **45,** No. 47, 123.
45. Lundberg, H. (1947). Magnetic surveys with helicopters. *Inst. Min. Met. Bull.*, **488,** 21–28 (July); *Canadian Min. Met. Bull.*, **423,** 392–400 (July); *Rhodesian Min. J.*, **19,** No. 243, 233–237.
46. Jensen, H. (1946). Operational procedure for the airborne magnetometer. *Oil a. Gas J.*, **45,** No. 10, 80–83.
47. Keller, F., Balsley, J. R., and Dempsey, W. J. (1947). Field operations and compilation procedure incidental to the preparation of isomagnetic maps. *Photogrammetric Eng.*, **13,** No. 4, 644–647.
48. Wier, K. L. (1950). Comparisons of some aeromagnetic profiles with ground-magnetic profiles. *Trans. Am. Geophys. Un.*, **31,** No. 2, pt. 1, 191–195.
49. Henderson, R. G., and Zietz, I. (1948). Analysis of total magnetic-intensity anomalies produced by point and line sources. *Geophysics*, **13,** No. 3, 428–436.
50. Vacquier, V. V., Steenland, N., Henderson, R. G., and Zietz, I. (November 1951). Interpretation of aeromagnetic maps. Mem. 47, *Geological Society of America.*
51. Hughes, D. S., and Pondrom, W. L. (1947). Computation of vertical magnetic anomalies from total magnetic field measurements. *Trans. Am. Geophys. Un.*, **28,** No. 2, 193–197.
52. Roman, I. (1946). The resolving power of magnetic observations. *Min. Technol.*, **10,** No. 6, Tech. Paper 2097, 18 pp.
53. Vestine, E. H., and Davids, N. (1945). Analysis and Interpretation of geomagnetic anomalies. *Terr. Magn.*, **50,** No. 1, 1–36.
54. Henderson, R. G., and Zietz, I. (1949). The upward continuation of anomalies in total magnetic intensity fields. *Geophysics*, **14,** No. 4, 517–534.
55. Kastrop, J. E. (1948). Gulf's airborne magnetometer in Florida. *World Oil*, **128,** No. 4, 138–142.
56. Joesting, H. R., Keller, F., and King, Elizabeth. (1949). Geologic implications of aeromagnetic survey of Clearfield-Philipsburg area, central Pennsylvania. *Bull. Am. Assoc. Petrol. Geol.*, **33,** No. 10, 1747–1766.
57. Pirson, S., and Bacon, L. O. (1948). Airborne magnetometer survey in central Pennsylvania. *Penn. State Coll. Bull.*, **42,** No. 10, 55–65.
58. U. S. Geological Survey (1949–1951). Total intensity aeromagnetic maps of 92 counties of Indiana. *U. S. Geol. Surv. Geophys. Invest.* Nos. 7–12, 20–45, 52–76, 82–90, 103–114, and Lake, Newton, Jasper, Benton, LaPorte, Stark, Pulaski, White, Fulton, Posey, Cass, Elkhart, Marshall, and St. Joseph Counties.
59. Joesting, H. R., and Henderson, J. R. (1947). Preliminary report on an experimental aeromagnetic survey in northwestern Indiana: *U. S. Geol. Surv. Geophys. Invest.*, 12 pp., preliminary map 4, pls. 1, 2.

60. Hurley, P. M. (1949). Airborne magnetic survey in Maine. *Eng. a. Mm. J.*, **150**, No. 8, 52–55.
61. Hurley, P. M., and Thompson, J. R. (1950). Airborne magnetometer and geological reconnaissance survey in northwestern Maine. *Bull. Geol. Soc. Am.*, **61**, No. 8, 835–842.
62. Balsley, J. R., James, H. L., and Wier, K. L. (1949). Aeromagnetic survey of parts of Baraga, Iron, and Houghton Counties, Michigan, with preliminary geologic interpretations. *U. S. Geol. Surv. Geophys. Invest.*, preliminary map.
63. U. S. Geological Survey (1950–51). Total intensity aeromagnetic maps of parts of Minnesota. *U. S. Geol. Surv. Geophys. Invest.*, Nos. 46–51, 99–102, and southern Beltrami, Cass, northern Crow Wing, part of Hubbard, western Itasca, western Morrison, eastern Morrison, Todd, and Wadena Counties.
64. U. S. Geological Survey (1950). Total intensity aeromagnetic maps of parts of Missouri. *U. S. Geol. Surv. Geophysical Invest.*, Nos. 13, 14, 77–91, and Coldwater, De Soto, Des Arc, Farmington, Fredericktown, Ironton Richwoods, and St. Clair quadrangles.
65. U. S. Geological Survey. (1950). Total intensity aeromagnetic maps of parts of New Mexico. *U. S. Geol. Surv. Geophysical Invest.*, Nos. 15–18.
66. Hawkes, H. E., Balsley, J. R., and others. (1946). Aeromagnetic map of a part of Oswegatchie quadrangle, St. Lawrence County, N. Y. *Geophysical Invest.*, preliminary aeromagnetic map.
67. Hawkes, H. E., Balsley, J. R., and others. (1946). Aeromagnetic survey at three levels over Benson Mines, St. Lawrence Co., N. Y. *Geophysical Invest.*, preliminary aeromagnetic map.
68. Hawkes, H. E., and Balsley, J. R. (1946). Magnetic exploration for iron ore in northern New York. *U. S. Geol. Surv., Strategic Minerals Invest.*, Prelim. Rept. 3–194,9 pp.
69. Postel, A. W. (1951). Geology of the Clinton magnetite district, New York. *U. S. Geol. Surv. Prof. Pap.*, **237**.
70. Jensen, H. (1951). Aeromagnetic survey helps find new Pennsylvania iron ore bodies. *Eng. a. Min. J.*, **152**, No. 8, 56–59.
71. Jensen, H. (1949). Airborne magnetic profile above 40th parallel, eastern Colorado to western Indiana. *Geophysics*, **14**, No. 1, 57.
72. U. S. Geological Survey, Total intensity aeromagnetic maps on open file. Magnet Cove, Arkansas; Gulf of Mexico; Tri-State area, Missouri, Kansas, and Oklahoma; Big Horn Basin, Wyoming; Great Sitkin, northern Adak, and northeastern Umnak Islands, Alaska; Cranberry Lake, Starke, Russell, and Childwold quadrangles, New York; Coeur d'Alene district, Idaho.
73. Frowe, E. (1948). Exploration in the Gulf of Mexico with the air-borne magnetometer. *Oil a. Gas J.*, **47**, No. 32, 104.
74. Keller, F., Meuschke, J. L., and Alldredge, L. R. A report on Project Volcano aeromagnetic surveys in the Aleutians, Marshall, and Bermuda Islands. To be published in *Trans. Am. Geophys. Un.*
75. Canada Department of Mines and Technical Surveys, Geological Survey of Canada (1951). *Geophys. papers*, 7–69.
76. Bourret, W. (1949). Aeromagnetic survey of the Allard Lake district, Quebec. *Econ. Geol.*, **44**, No. 8, 732–740.
77. Affleck, J. (1948). Aeromagnetometer profile flown from Venezuela to Texas. *World Oil*, **128**, No. 3, pp. 223–228.
78. Due Rojo, A. (1948). La prospeccion magnetica por avion en EE.UU. (Airborne magnetic prospecting in the United States of America.) *Rev. Geofis.*, **7**, No. 26, 211–215.

79. Bemrose, J., Higgblom, J. C., Holt, T. C., Richard, T. C., and Watson, R. J. (1950). Bahamas airborne magnetometer survey. *Geophysics*, **15**, No. 1, 102–109.
80. Jensen, H. (1948). Some technical aspects of Bahamas airborne magnetometer survey. *Geophysics*, **13**, No. 3, 495.
81. Burns, W. W. (1947). Bahamas oil exploration—first major survey with airborne magnetometer. *Petroleo Interamericano*, **5**, No. 12, 40–45.
82. *World Oil* (1949). Aerial surveying speeds Mozambique exploration. **128**, No. 12, 200–202.
83. Weiss, O. (1949). Aerial magnetic survey of the Vredefort Dome in the Union of South Africa. *Min. Eng.*, **1**, No. 12, 433–438.
84. Suslennikov, V. V. (1947). Principal results of the aeromagnetic survey in the Karelian-Finnish S.S.R. (in Russian). *Razvedka Nedr.* **13**, No. 5, 67–71.
85. Katskov, A. I. (1948). Aeromagnetic surveys and their value in geophysical studies (in Russian). *Vses. Nauch-Issled. Geol. Inst. Mater., Geofiz.*, No. 11, 8–11.
86. Alldredge, L. R., and Keller, F. (1949). Preliminary report on magnetic anomalies between Adak, Alaska, and Kwajalein, Marshall Islands: *Trans. Am. Geophys. Un.*, **30**, No. 4, 494–500.
87. Alldredge, L. R., and Dichtel, W. J. (1949). Interpretation of Bikini magnetic data: *Trans. Am. Geophys. Un.*, **30**, No. 6, 831–835.
88. Balsley, J. R. (1951). Techniques and results of aeromagnetic surveying. *Proc. U. N. Sci. Conf. Conservation and Utilization of Resources*, **3**, 8–10.
89. Deegan, C. J. (1948). Economics of aerial magnetics. *Oil a. Gas J.*, **47**, No. 12, 67–71, 77.

Author Index

Numbers in parentheses are reference numbers and are included to assist in locating references in which the authors' names are not mentioned in the text. Numbers in italics indicate the page on which the reference is listed.

Example: Adel, A., 163 (18), *235*, means that this author's article is reference 18 on p. 163 and is listed on p. 235 at the end of the article.

A

Adel, A., 163 (18), *235*
Affleck, J., 329 (77), *348*
Aiken, H. H., 11 (2), 12 (2), 32 (2), 42 (2), *43*
Alessio, 299
Alldredge, L. R., 315 (20), 316 (20), 317 (20), 319 (20), 329 (74), 331 (74, 86, 87), 334 (87), 335 (87), 336 (87), *346*, *348*, *349*
Allen, W. C., 131, 132, *153*
Almond, M., 128 (28), *152*
Andersen, 298
Antranikian, H., 314 (8), *345*
Appleton, E. V., 226 (160), *241*
Archenhold, F. S., 221 (147), *241*
Arons, A. B., 269, 272, 273, 274, *280*
Aschenbrenner, H., 314 (9), *345*

B

Babcock, H. D., 161 (11), 187 (72), 196 (84), *235*, *237*, *238*
Bacher, R. F., 192, *238*
Bacon, L. O., 329 (57), *347*
Bagge, E., 229 (164), *241*
Bailey, V. A., 204 (97), *238*
Baker, J. G., 124
Balsley, J. R., 313, 314 (12), 322 (12, 43), 324 (12), 325 (12, 47), 326 (12, 47), 329 (12, 62, 66, 67, 68), 336 (62), 339 (66, 67, 68), 342 (12, 88), *345*, *347*, *348*, *349*
Barbier, D., 197, *238*
Bartels, J., 158 (A), 226 (161), *234*, *241*
Bates, D. R., 160 (4), 185 (61), 206 (111), 208 (114, 115, 116), 211 (122), 212 (132), *235*, *237*, *239*, *240*
Bayley, R., 314 (11), 322 (11), 329 (11), *345*
Beers, R. F., 314 (18), *346*
Bellamy, J. C., 1, 38 (7), 39 (7), *43*
Bemrose, J., 329 (79), *349*
Berg, H., 219 (139), *240*
Bergeron, T., 89 (2), *116*
Berkely, E. C., 10 (1), 11 (1), 15 (1), 42 (1), *43*
Berkson, J., 50, *83*
Beynon, W. J. G., 177 (39), 219 (39), 226 (160), *236*, *241*
Bhar, J. N., 199 (91), *238*
Biondi, M. A., 212 (133), *240*
Bjerknes, J., 97, *117*
Blackett, P. M. S., 127, *152*
Boffi, J. A., 91 (9), 115 (9), *116*
Bohn, J. L., 132, *153*
Bohr, N., 206 (105), *239*
Bolin, B., 87, 115 (42), *118*
Booker, H. G., 239 (140), *240*
Boothroyd, S., 122 (4), *151*
Bourret, W., 329 (76), 338 (76), 339 (76), *348*
Bradford, F., 60 (27), *84*
Bradt, H. L., 164 (22), *235*
Branscomb, L. M., 187 (65), *237*
Brasefield, C. J., 176 (32), *236*
Briggs, B. H., 219 (141), *240*
Brockamp, 298
Brown, E. J., 309, *311*
Brown, S. C., 212 (133), *240*
Browne, B. C., 290, 299, *310*
Bucher, W. H., 301, *311*
Buckingham, R. A., 212 (132), *240*
Buddhue, J. D., 131, *153*
Bullard, E. C., 290, *310*
Burch, J. E., 322 (42), *347*

Burhop, E. H. S., 206 (112), 211 (112), *239*
Burnight, T. R., 132, *153*
Burns, W. W., 329 (81), *349*

C

Cabannes, J., 187 (73), *237*
Cameron, W. M., 252, 253, 258, 260, 261, 262, 263, 264, *280*
Cauer, H., 160 (8), *235*
Chackett, K. F., 143, *153*, 162 (17), *235*
Chamanlal, C., 127, *152*
Chandrasekhar, S., 197, 199 (88, 89), 208 (117), *238*, *239*
Chapman, R. M., 161 (12), *235*
Chapman, S., 158 (A), 167, 184, 225 (159), 226 (162), 229 (165), *234*, *241*
Charlier, C. V. L., 51, *84*
Charney, J., 94 (14), 96 (19), 108, *117*
Chow, V. T., 55 (16), 66, *84*, *85*
Claassen, H. H., 160 (10), 196 (86), *235*, *238*
Clapp, P. F., 91 (8), *116*
Clark, J. S., 290, *310*
Clopper, C. J., 56, *84*
Cockroft, A. L., 211 (123), *240*
Collier, A., 254, 255, *280*
Conrad, V., 46, *83*
Cook, G. S., 290, *310*
Cook, M. A., 139, *153*
Coster, D., 208 (118), *239*
Court, A., 45, 52 (15), *84*
Cowie, 299
Cowling, T. G., 194 (80), 229 (165), *238*, *241*
Craggs, J. D., 212 (134), *240*
Craig, R., 234
Crary, A. P., 176 (36, 37), *236*
Crozier, W. D., 133, *153*
Curran, S. C., 211 (123), *240*
Currie, B. W., 172 (31), *236*
Cutolo, M., 204 (98), *238*

D

David, F. N., 56 (21), *84*
Davids, N., 329 (53), *347*
Davies, J. G., 128, *152*
Dean, G., 89 (5), 90, 93, *116*
Deegan, C. J., 342 (89), *349*
Defant, A., 107, 109, *118*
Dempsey, W. J., 325 (47), *347*
Denning, W. F., 122, *151*
Dennison, D. M., 194 (81), *238*
Dichtel, W. J., 331 (87), 334 (87), 335 (87), 336 (87), *349*
Dickey, F. P., 161 (14), *235*
Dirac, P. A. M., 199 (92), *238*
Ditchburn, R. W., 206 (119), *239*
Dixson, F. N., 56 (21), *84*
Dobson, G. M. B., 122, 133, 139, *151*
Donaldson, R. J., Jr., 158 (G), 234, *234*
Donavan, R. A., 160 (6), *235*
Doob, J. L., 55 (18), *84*
Dow, W. G., 140, *153*
Due Rojo, A., 329 (78), *348*
Dufay, J., 187 (73), *237*
Durand, D., 76 (36), 79 (36), *85*
Dzerdzevski, B. L., 88 (6), 89 (6), *116*

E

Eady, E. T., 95, 96 (18), 107, 108, 114, *117*, *118*
Eckardt, E. A., 314 (13), 322 (13), 324 (13), 342 (13), *345*
Edgeworth, F. Y., 51, *84*
Eisenhart, C., 56 (21), *84*
Ekman, V. W., 94 (15), *117*
Elkin, W. L., 23, *152*
Ellyett, C. D., 128, 150, *152*, *154*
Elmen, G. W., 314 (10), *345*
Elsasser, W. M., 185 (62), *237*
Epstein, B., 60 (29), *84*
Epstein, P. S., 230 (174), *242*
Ewing, W. M., 300, 301 (12), *311*
Eyfrig, R., 177 (40), *236*
Eyring, H., 139, *153*

F

Farmer, F. T., 204 (99), *238*
Fedynsky, V. V., 148, *154*
Felch, E. P., 316 (21), *346*
Felch, E. P., Jr., 318 (30, 33), *346*
Fennell, P., 177 (47), *236*
Ferraro, V. C. A., 229 (166, 167), *241*
Ferrell, O. P., 177 (48, 49), *236*
Finch, V. C., 244, *279*

Fisk, J. B., 206 (106), *239*
Fjortøft, R., 94 (14), 96 (20), 108, *117*
Fleming, J. A., 158 (B), *234*
Flory, P. J., 208 (120, 121), *239*
Forsyth, P. A., 172 (31), *236*
Frost, S., 15 (3), *43*
Frowe, E., 329 (73), *348*
Fultz, D., 94, *117*
Fundaminsky, A., 211 (122), *239*

G

Gauzit, J., 187 (73), *237*
Gaydon, A. G., 190, 194 (82), *238*
Gebbie, H. A., 161 (13), *235*
Gebhardt, R. E., 314 (7), *345*
Gerson, N. C., 155, 158 (D), 167 (28), 177 (41, 50, 51, 53), 185 (63), 221 (48, 51, 53), *234*, *236*, *237*, *241*
Ginsburg, V. L., 204 (100), *239*
Gladisch, H., 229 (173), *242*
Glemic, 299
Glückauf, E., 160 (2), 161 (16), *235*
Godfrey, G. H., 185 (64), *237*
Goldberg, L., 160 (3, 6), 196 (3), *235*
Gollnow, H., 187 (67, 68, 69), *237*
Goodridge, R. S., 60 (27), *84*
Goubau, G., 314 (9), *345*
Goudsmit, S., 192, *238*
Gowan, E. H., 166 (26), *236*
Greenhow, J. S., 150, *154*
Greenwood, J. A., 76 (36), 79 (36), *85*
Grotewahl, M., 314 (2), *345*
Gulatee, 307
Gumbel, E. J., 50, 60, 64, 76, 79 (36), *83*, *84*, *85*

H

Hadley, G., 88 (1), *116*
Hagstrum, H. D., 211 (124), *240*
Haid, 299
Harding, 304, 305, 306, 307
Harding, W. R., 161 (13), *235*
Harteck, P., 160 (5), *235*
Hartree, D. R., 206 (107)
Hartree, W., 206 (107), *239*
Hastay, M. W., 56 (21), *84*
Havens, R., 140, 142, 145, 146, *153*
Hawkes, H. E., 329 (66, 67, 68), 339 (66, 67, 68), *348*
Hazen, A., 49, *83*
Hecker, 299
Hedgpeth, J. W., 254, 255, *280*
Heiland, C. A., 314 (1), *345*
Henderson, J. R., 329 (59), *347*
Henderson, R. G., 329 (49, 50, 54), 333, *347*
Herlofson, N., 127, 129, *152*
Herzberg, G., 203 (76), *238*
Herzberg, L., 161 (11), 174, 190, *235*
Hey, J. S., 127, 129, *152*
Heyl, P. R., 282, 290, *310*
Higgblom, J. C., 329 (79), *349*
Hilsum, C., 161 (13), *235*
Hirvonen, R. A., 298, 299, *311*
Hoffmeister, C., 121, 149, *151*, *154*, 177 (54), 221 (149), *237*, *241*
Holt, R. B., 212 (135), *240*
Holt, T. C., 329 (79), *349*
Holweck, 299
Hoppe, J., 133, *153*
Hopwood, W., 212 (134), *240*
Hough, S. S., 224 (152), *241*
Houtgast, J., 196 (85), *238*
Howland, B., 212 (135), *240*
Howard, R. R., 206 (108), *239*
Hughes, D. S., 329 (51), *347*
Hulburt, E. O., 148, *154*, 229 (169), *242*
Hulthén, L., 206 (109) *239*
Hurley, P. M., 329 (60, 61), 334 (60, 61), *348*
Hutchings, J. W., 89 (7), *116*
Huxley, L. G. H., 204 (101, 102), *239*

I

Insch, G. M., 211 (123), *240*
Irons, H. R., 318 (29), *346*

J

Jacchia, L., 137, 138, 140, 142, 145, *153*, *154*
Jaeger, J. C., 229 (168), *242*
James, H. L., 329 (62), 336 (62), *348*
Jeffreys, H., 97, *117*
Jensen, H., 314 (14, 19), 322 (14, 43, 44), 325 (46), 329 (70, 77, 80), 341 (70), *345*, *346*, *347*, *348*, *349*
Joesting, H. R., 329 (56, 59), *347*

Johnson, M. H., 229 (169), *242*
Johnson, N. K., 176 (33), *236*
Johnston, H. F., 226 (161), *241*
Johnston, H. L., 208 (121), *239*
Jones, L. M., 143,
Jutsum, P. J., 208 (119), *239*

K

Kahlke, S., 148, *159*
Kalashnikov, A., 131, *152*
Kalinske, A. A., 49 (3), *83*
Kaplan, J., 234
Kastrop, J. E., 329 (55), *347*
Katskov, A. I., 331 (85), *349*
Keller, F., 325 (47), 326 (47), 329 (56, 74), 331 (74, 86), *347*, *348*, *349*
Kendall, M. G., 60, *84*
Ketchum, B. H., 244, 253, 269, 270, 271, 272, *279*, *280*
Keulegan, G. H., 256, 258, 259, *280*
Kimball, B. F., 50 (9), 60 (27), *83*, *84*
King, E., 329 (56), *347*
Kirkpatrick, C. B., 229 (170), *242*
Klaasse, J. M., 322 (44), *347*
Knoerr, A. W., 314 (15), 322 (15), 324 (15), 325 (15), 326 (15), 328 (15), *354*
Kohler, M. A., 60 (24), *84*
Koll, R., 140, 142, 145, 146, *153*
Kopal, Z., 140, 145, *154*
Kotani, M., 211 (131), 212 (138), *240*
Kottler, F., 50, *83*
Krautkrämer, J., 219 (142), *240*
Krautkramer, K., 177 (42), *236*
Krogh, M. E., 160 (77), *235*
Kuiper, G. P., 158 (E), *234*
Kuo, H. L., 96 (21, 22), 108, 109, 111, 112, 114, *117*

L

Lagemann, R. T., 161 (14), *235*
Lagow, H., 140, 142, 145, 146, *153*
Laird, A. G., 316 (26), *346*
Lamb, H., 224 (153), *241*
Lameris, A. J., 208 (118), *239*
Landsberg, H. E., 51, *83*, 131, *153*, 234
Langbein, W. B., 55 (16), *84*
La Seur, N. E., 91 (10), 104, 105, 113 (10), *117*
Leech, J. W., 211 (122), *239*
Leighly, J., 81, *85*
Lejay, 299
Lenox-Conynghan, 299
Lettau, H., 178 (60), 230 (175, 176), 234, *237*, *242*
Liller, W., 130, *152*
Lin, C. C., 112, *118*
Lindemann, F. A., 122, 133, 139, *151*
Link, F., 123, *151*
Linsley, R. K., Jr., 60 (24), *84*
Loeb, L. B., 212 (136), *240*
Logachev, A. A., 314 (4, 5), *345*
Long, R., 94, 96 (13), 97, *117*
Lovell, A. C. B., 127, 128, 129, 130, *152*, 229 (171), *242*
Lundberg, H., 314 (3), 324 (45), *345*, *347*

M

McClure, B. T., 212 (135), *240*
McIntosh, R A., 123, *151*
McKinley, D. W R., 127, 128, 129, 130, *152*
McMath, R. R., 160 (6), *235*
McQueen, J H., 161 (15), *235*
Manning, L. A., 128, 149, 150, *152*, *154*, 177 (56), *237*
Maris, H. B., 133, *153*
Martyn, D. F., 204 (97), 224 (154), 226 (154), *238*, *241*
Massey, F. J., Jr., 56 (21), *84*
Massey, H. S. W., 206 (110, 111, 112), 211 (110, 112, 122, 125), 212 (132), *239*, *240*
Means, W. J., 316 (21), *346*
Meek, J. H., 177 (52), 221 (52), *237*
Meinel, A. B., 172 (30), *236*
Merrill, F. G., 316 (27), 318 (33), *346*
Meuschke, J. L., 329 (74), 331 (74), *348*
Milda, J., 211 (128), *240*
Miller, J. C. P., 226 (162), *241*
Millman, P. M., 124, 125, 126, 129, 130, *152*
Minnaert, M., 196 (85), *238*
Mintz, Y., 89 (5), 90, 93, 99 (28), 101 (28), 110 (28), *116*, *117*
Mises, R. von, 55 (18), *84*
Mitra, S. K., 158 (F), 230 (177), *234*, *242*
Mitra, S. N., 177 (43), 219 (43, 143), *236*, *241*

Mohler, O. C., 160 (6), *235*
Moore, C. E., 190, 194 (83), 196 (84), *237*, *238*
Morelli, C., 298, 299, *311*
Morse, P. M., 212 (137), *240*
Moses, H. E., 163 (21), 199 (21), *235*
Mott, N. F., 211 (125), *240*
Muckorfuss, 307
Muffly, G., 314 (16), 320 (16), *346*
Mulders, G. F. W., 196 (85), *238*
Munro, G. H., 177 (44), *236*
Murgatroyd, R. J., 176 (38), *236*

N

Nadig, F. H., 132, *153*
Namias, J., 91 (8), 105, *116*, *118*
Neumann, J. von, 94 (14), 108 (14), *117*
Newell, H., 145, *153*
Nicholls, R. W., 211 (126), *240*
Nielsen, A. H., 161 (14), *235*
Nörlund, 299

O

Ockenden, C. V., 176 (34), *236*
Öpik, E., 122, 129, 131, 133, 135, 139, 146, *151*, *152*, *154*
Oldenberg, O., 187 (66), *237*
Oliver, N. J., 170 (29), *236*
Olivier, C. P., 121, 147, 148, *151*, *154*, 177 (57, 58), *237*
Oxholm, M. L., 196 (86), *238*

P

Palmén, E., 100, *118*
Paneth, F. A., 143, *153*, 160 (1), 162 (17), *235*
Pannekoek, A., 199 (93), *238*
Panofsky, H. A., 47 (2), *83*
Parratt, L. G., 316 (21), *346*
Parson, S. J., 128, *152*
Paulhus, J. H., 60 (24), *84*
Pearse, R. W. B., 190, *238*
Pearson, K., 51, 56, *83*, *84*
Penndorf, R., 163, (20), 199 (20), 234, *235*
Perkeris, C. L., 204 (103), 224 (155), *239*, *241*
Peters, B., 164 (22), *235*
Peterson, A. M., 128, *152*, 177 (56), *237*
Petrie, W., 172 (31), *236*
Pettersson, H., 132, *153*
Pettit, H. B., 221 (150), *241*
Phelps, E. B., 270, *280*
Phillips, G. J., 219 (141), *240*
Piece, J. A., 127, *152*
Pierce, A. K., 160 (6), *235*
Pirson, S., 329 (57), *347*
Pollak, L. W., 46, *83*
Pondrom, W. L., 329 (51), *347*
Porter, J. G., 123, *151*
Postel, A. W., 329 (69), 339 (69), 340 (69), *348*
Potter, W. D., 60 (28), *84*
Powell, R. W., 50, 60 (7, 27), *83*, *84*
Press, H., 60 (30), *85*
Price, W. C., 190, *238*
Price, W. L., 185 (64), *237*
Priestley, C. H. B., 97, 98, 99, 102, *117*
Pritchard, D. W., 243
Putnam, 299

Q

Quaile, J. E., 314 (17), *346*

R

Ragazzini, J. R., 15 (4), *43*
Raitt, R. W., 301 (13), *311*
Rakshit, H., 163 (19), 230 (177), *235*, *242*
Ramsayer, K., 314 (6), *345*
Ratcliffe, J. A., 204 (99, 102, 104), 219 (140), *238*, *239*, *240*
Reasbeck, P., 143, *153*, 162 (17), *235*
Rex, D. F., 105, 106, *118*
Richard, T. C., 329 (79), *349*
Richardson, J. M., 212 (135), *240*
Richardson, M. S., 318 (37), *346*
Riehl, H., 91 (10), 94 (16), 99, 104, 105, 113, *117*
Rietz, H. L., 55, *84*
Roach, F. E., 197, 221 (150), *238*, *241*
Roberts, E. B., 318 (28), *346*
Roberts, V., 161 (13), *235*
Roman, I., 329 (52), *347*
Rossby, C. G., 89 (3, 4), 95, 104, 109, 110, *116*, *117*, *118*
Rosseland, S., 199 (94), *238*

Rotschi, H., 132, *153*
Rumbaugh, L. H., 315 (20), 316 (20, 21), 317 (20), 319 (20), 322 (44), *346*, *347*

S

Schilling, G. F., 177 (55), *237*
Schmehl, 298
Schonstedt, E. O., 318 (29), *346*
Schüler, H., 187 (67, 68, 69), *237*
Scorer, R. S., 219 (144), *241*
Scrase, F. J., 176 (35), *236*
Seeliger, R., 229 (172), *242*
Seely, B. K., 133, *153*
Sekera, Z., 219 (145), *241*
Senftleben, H., 229 (173), *242*
Shapley, H., 121 (4), *151*
Shaw, I. J., 204 (104), *239*
Shaw, J. H., 160 (4, 9), 161 (12), 196 (86), *235*, *238*
Sheppard, P. A., 166 (27), 177 (27), *236*
Shinn, D. H., 219 (140, 141), *240*
Slichter, L. B., 290, 308, *311*
Slobud, R. L., 160 (7), *235*
Slonczewski, T., 316 (21, 26), 318 (30, 31, 32, 33), *346*
Smith, P. M., 320 (39), *346*
Smith, P. T., 211 (127), *240*
Smith, W. V., 206 (108), *239*
Snedecor, G. W., 56, *84*
Sparrow, C. M., 133, *153*
Spencer, N. W., 140, *153*
Spicer, H. C., 51 (10), *83*
Spitzer, L., 144, *153*
Spoleto, 299
Starr, V. P., 97, 99, 102, *117*
Steenland, N., 329 (50), *347*
Stewart, G. S., 128, 129, *152*
Störmer, C., 221 (151), *241*
Stommel, H., 246, 256, 257, 258, 259, 260, 262, 263, 269, 272, 273, 274, *279*, *280*
Stuckelberg, E. C. G., 212 (137), *240*
Suess, H. E., 160 (5), *235*
Sundberg, K., 314 (3), *345*
Suslennikov, V. V., 331 (84), *349*
Svoboda, A., 15 (5), *43*
Swirles, B., 206 (107), *239*

T

Tate, J. T., 211 (124, 127), *240*
Taylor, G. I., 94, *117*, 224 (156), *241*
Thomas, H. A., Jr., 59, *84*
Thomas, R. N., 135, 138, 139, *153*
Thompson, J. R., 329 (61), 334 (61), *348*
Tickuer, A. J., 316 (21), *346*
Tolles, W. E., 320 (40, 41), *347*
Tonsberg, E., 187 (71), *237*
Trewartha, G. T., 244, *279*
Tschu, K. K., 225 (159), 226 (163), *241*
Tully, J., 253, 256, 262, 269, 270, *280*

U

Underhill, B. B., 158 (G), *234*
Unwin, J. J., 212 (132), *240*
Uspensky, J. V., 55, *84*

V

Vacquier, V. V., 316 (23, 24), 320 (41), 329 (50), *346*, *347*
van de Hulst, H., 131, 132, *153*
Van Dijk, E. W., 208 (118), *239*
Van Mieghem, J., 102, *118*
Vassy, A., 166 (23, 25), *235*
Vassy, E., 166 (23, 25), *235*
Vegard, L., 187 (71), *237*
Velz, C. J., 270, *280*
Vening-Meinesz, F. A., 299, 300, *311*
Venkataraman, K., 127, *152*
Vestine, E. H., 329 (53), *347*
Villard, O. G., Jr., 128, 149, 150, *152*, *154*, 177 (56), *237*
Vyssotsky, A. N., 125, *152*

W

Wakeshima, H., 187 (70), *237*
Wallis, W. A., 56 (21), *84*
Warfield, C. N., 140 (54), 142 (54), *153*
Wasiutyński, J., 219 (146), *241*
Watanabe, M., 211 (128), *240*
Watson, F. G., 121, 131, 132, *151*
Watson, R. J., 329 (79), *349*
Waynick, A. H., 177 (45), *236*
Weekes, K., 177 (46), 224 (157), *236*, *241*
Weid, A. C., 318 (37), *346*
Weiss, O., 331 (83), *349*
Weizel, W., 192, *238*
Wenzel, E. A., 143
Wexler, H., 145, *154*, 166 (24), *235*

Whipple, F. L., 119, 124 (14), 131 (38), 135, 138, 139 (50), 140 (51), 145, *152*, *153*, *154*, 177 (59), *237*
Whipple, G. C., 49, *83*
White, R. M., 99, 102, 110, *117*
Wiborg, B. S., 143, *153*, 162 (17), *235*
Widger, W., Jr., 99, 100, *117*
Wier, K. L., 325 (48), 329 (62), 336 (62), *347*, *348*
Wildt, R., 199 (95), *238*
Wilkes, M. V., 224 (157, 158), *241*
Willett, H. C., 104, *118*
Williams, E. J., 211 (129), *240*
Woollard, G. P., 281, 290 (5), 304, 305, 306, 307, *311*
Woolley, R. v. d. R., 199 (96), *238*
Wu, T. Y., 163 (21), 199 (21), *235*
Wyckoff, R. W. D., 316 (25), 320 (25), *346*

Y

Yamanouchi, T., 206 (113), 211 (130, 131), 212 (138), *239*, *240*
Yeh, T. C., 91 (10), 99, 104, 105, 113 (10), *117*

Z

Zietz, I., 329 (49, 50, 54), 333, *347*
Zoch, R. T., 62 (33), *85*

Subject Index

A

Ablation, 120, 121, 131
Adirondack Mountains, N.Y., 339
Aerobee rocket, 143
Aeromagnetic survey, 329, 331, 334, 336, 337, 338, 341
Afterglow, 126
Airborne magnetometer, 314, 342
Aircraft pressure altimeter, 40
Airglow, 164, 169, 180, 189, 196, 208, 219
  definition, 160
  spectrum, 173, 188
Airy, 295, 301
Alberni Inlet, 253, 256, 259, 273
Allard Lake District, Quebec, 338
Analysis, statistical, 47, 49
Angular momentum, 96, 97, 98, 104
  absolute, 98
Anomalies, Bouguer, 289, 293, 301, 302, 303
  free air, 293
  isostatic, 294, 302, 303, 305
Anomalous propagation, 164
Anticyclones, 103, 104, 110
Argon, 162
Association, 211
Astroballistics, 139
Atmosphere, nomenclature, 158
Atmospheric constituents, 188
  densities, 136, 137
  emissions, 169
  gases, spectra of, 190
  tides, 224
Aurora, 164, 169, 180, 208
  sunlit, 164
Auroral spectrum, 170, 188
Automatic processing, 33

B

Baroclinic model, 113
Barotropic model, 109
Bénard cells, 215
Bikini Atoll, 334
"bird," 320, 321
Blocking action, 105, 106
Body-Doppler, 149
Bolides, 120

C

Calculus of observation, 47
Carbon dioxide, 161, 195
Chemosphere, 158, 162, 168, 170, 178, 228
Chesapeake Bay, 247, 254, 263, 274
Circulation of the atmosphere, 103, 107
  basic principles, 91
Circulation cells, meridional, 96, 99, 100, 107
Clairaut, theorem of, 285
Collisional frequency, 168, 169, 180, 184, 202
Collisional phenomena, 199–212
Comet 1862 III, 121
Computers, analogue, 15, 19
  digital, 15, 19
Computing machine, 48
Confidence band, 68
Conservation of abs. vorticity, 110, 111
Continuity, equation of, 215
Convection, small scale, 93
Coriolis force, 94, 251, 252, 266, 267
Corpus Christi Bay, 255
Cosmic flotsam, 162
Cosmic ray, 163
Cross sections, 204
Cyclones, 103, 104, 110

D

Dalton atmosphere, 229
Depth of no net motion, 260, 267
Density function, see Probability density function
Differential equations, 15

Diffusion, 226
eddy, 214, 226
forced, 228
molecular, 214, 227
Diffusive equilibrium, 229
Diffusive separation, 143, 144
Digit, decimal, 31
Distribution, bimodal, 52
frequency, 49, 51
probability, 49
unimodal, 52
Doppler effect, 149
velocities, 131

E

Eddy diffusion, see Diffusion
diffusivity, 216, 273, 275
motion, 107, 111, 114
transfer, 107
transport see transport
E-layer, see Ionic layers
Electron cloud, 131, 213
column, 127, 131
densities, 164
Electronic counter, 34, 36, 38
Energy absorption, 206
balance, 95, 102, 184, 198
cycle, 103
Estimate of risk, 54
Estuary, 243
bar-built, 246, 254
circulation pattern, 255
classification, 245
coastal plain, 245, 247, 269, 274
dynamic structure, 252
dynamics of, 262
salinity distribution, 248, 251, 273, 274
temperature distribution, 249
tidal motion, 249
velocity profile, 250
deep-basin, 246, 253, 256–263
definitions, 244
dynamics of, 256
fiord, see Deep-basin estuary
Exceedance interval, 54
Excitation, 208
Exosphere, 158
Extreme values, theory of, 63, 71

F

Fairfax Quadrangle, Virginia, 331
Fiord, see Deep-basin estuary
Fireball, 120
Flare, 138
Flushing studies, 268
Forbidden atomic lines, 187
Fourier series, 76
Fraunhofer spectrum, 131

G

gal, 288
Gegenschein, 120
Geomagnetic field, 203, 221
Geostrophic flux, 99
wind, 178
Giacobinid shower, 128, 130, 131
Goldstein-Kaplan bands, 170
Gradient, magnetic, 331
Gravimeter, 283, 284
Gravitational acceleration, 284
attraction, 120, 213, 283, 284
constant, 282
separation, 162
Gravity, see Gravitational acceleration
anomaly, 286, 288
bases, world network of, 300
Formula, International, 285

H

Halocline, 266, 267
Heat engine, 93, 103
Heat transfer, 111
Helium, 162, 163, 170
Helix-tapper-bar-typewriter, 34
High index, 104, 114
Hydrogen, 163, 170, 195
Hydroxyl molecule, 175, 206

I

Index cycle, 104, 114
Inertia cycle, 95
Ionic layers, 167, 180
D, 168, 206, 232
E, 130, 147, 149, 232
E_s (Sporadic E), 168, 178, 219, 231

F, 176
F_1, 232
F_2, 185, 214, 232
G, 168
Isogonic charts, 318
Isostasy, 294, 301, 305

J

James River estuary, 252, 263, 275
Jet, planetary, 89, 91
Jet stream, 91, 105, 111

K

Key punch, 9
Kinetic energy, balance equation of, 102

L

Laguna Madre, 255
Large scale motion of the atmosphere, 91
Leuchtstreifen, 149, 219
Level, mandatory, 5, 30
Level, significant, 5, 30
Lithosphere, 158
Low index, 104, 114
Luminous clouds, 177, 189, 213, 219, 221
Luxembourg effect, 203

M

Mach number, 134, 142
Magnetic contour map, 328, 329
Magnetic field, 131, 150, 180, 213, 214, 291
Magnetic storm, 169, 233
Magnetohydrodynamics, 169, 234
Magnetometer, airborne, 314, 342
  ground, 343
Matching card, 41
Mean deviation, 64, 65
Mean free path, 165, 214
Meinel bands, 170, 175
Memory screen, 38
Mesosphere, 158, 170, 180, 212
Meteor, 120
  brightness, 123, 130, 138
  daylight, 129
  drag coefficient, 135
  fast, 125
  heat transfer, 135, 138, 139
  height, 139
  hyperbolic, 124, 129
  orbits, 128
  origin, 126
    cometary, 126
  radiant point, 122
  showers, 121
  slow, 125
  spectra, 124
    bright-line, 125
    continuous, 124
  sporadic, 121, 126
  techniques of observation, 122
    photographic methods, 123
    radar echo, 127, 128
    radio methods, 127
    visual methods, 122
  trail, 121, 123, 134, 138
  train, 121, 130, 147, 164, 213, 227, 229
    luminosity, 121, 126, 134
  trajectory, 122, 123, 127
  velocity, 123, 128, 130
    angular, 122, 123
    distribution, 124
Meteorites, 120, 126, 139
  micro-, 121, 131
Meteoritic dust, 121, 131
Meteoritic energy, 129
Meteoritic ionization, 129
Meteoritic material, accretion of, 132
Meteoritic orbits, 128
Meteoritic process, 133
Meteoritic streams, 121
Meteoroid, 120, 126, 134, 139
  mass, 134, 135, 136
  nature, 138
Microcline granite gneiss, 341
Microfilming, 17
Milligal, 288
Mohorovičič discontinuity, 301
Momentum, see Angular momentum
  balance, 95, 97, 100, 102, 104, 113
  exchange, 100
  flux, 99
  transport, see Transport
Monsoonal circulation, 115
"monsoon" effect, 132
Morse code, 6

N

NACA Tentative Standard Atmosphere, 140
Neon, 162
New York Harbor, 270
Nitrogen, 162, 195
  atomic, 164, 170, 204
  molecular, 164, 172, 204
Noctilucent clouds, 213, 219
Notations, binary, 32
  digital, 25, 32
  graphical, 26
  numerical, 25
  unitary, 26, 31, 32
Numeroscope, 12

O

Ogive, 49
Oil trap, 303, 304
Ore deposits, 303, 304
Oxygen, atomic, 170, 172, 175, 206, 212
  molecular, 163, 172, 193, 195, 204, 206, 212
Ozone, 144, 163, 193, 195
  layer, 166

P

Parent population, 47
Pastagram, 37
Pendulum gravity apparatus, 283
Perseid, 126
  shower, 121
Photochemical equilibrium, 197
Photographic recorder, 33
Photoionization, 189, 206
Planet, 126
Plasma, 169
Playback, 11, 14, 27, 30, 31, 33, 34
Potsdam, 297, 305
Pratt, 295, 301
Probability density function, 49, 62
Probability function, circular normal, 76
Probability function, cumulative, 49
Probability of non-occurrence, 61
Probability of occurrence, 55
Probability paper, 49, 50
  extreme, 67
Punched cards, 9, 21

R

Radiant, 121
Radiation, 93
Radio fadeout, 168
Radio wave probing, 217
Radio "whistles," 127
Radiosonde observations, 3, 9, 38
Random flux, 275
Raob computation data record, 5, 6
Raritan river, 253, 273, 274
Recombination, 198, 211
Record(s), 4, 9
  isometric geographical, 30
  magnetic tape, 14
  microfilm, 17, 23
  multiple, 36
  observational, 4
  original, 5
  perforated tape, 6
  strip chart, 27
  unitary, 23, 33
  vectorial, 27
Recording tables, 34, 37
Recurrence interval, 54
Return period, 71
  design, 74
Reynolds number, 134
Risk, calculated, 59, 60
River discharge, 254
Rocket, 141, 142, 143, 145, 146, 164
  Aerobee, 143
  pressure data, 142, 143
  V-2, 132
Rotation of the earth, 94

S

Salinity, 266, 269, see also Estuary
  base, 269
Salt balance, 274
  dome, 303
S.C.E.L. Grenade, 142, 143
Seasonal variations, 105, 145, 146
Sodium, 163, 170, 175, 206
  chloride, 163
Solar absorption spectrum, 194
Solar chromospheric eruption, 168
Solar ejecta, 209
Solar flares, 168

Solenoidal field, 115
Sound, 246, see also Bar-built estuary
Sound recorder, 13
Space requirements, 32
Spring Gravimeter, 284, 291
Standard atmosphere, see NACA
Standard deviation, 47, 48, 52, 62, 64
Standard relationship, 40, 41
Stokes' law, 227
Stratosphere, 158
Sudden ionospheric disturbances (SID), 233
Sunspot activity, 167, 169
Super-Schmidt meteor camera, 124, 125, 142
Surface of no net motion, 250
Surface stress, 98

T

Tapes, perforated, 6, 20
  magnetic, 11, 13, 22
Temperature, 182–188
  gas, 185
  ionosphere, 182
  mesosphere, 182
  rotational, 185, 195
Tidal influences, crustal response to, 304
Tidal motion, 221, 249, 254, see also Estuary
Tidal oscillations, 164, 179, 222, 224
Tidal prism, 270
Tides, 213, 221
  ionospheric, 222
Transport of
  angular momentum, 97, 98, 108
  heat, 93
  heat, eddy, 102
  mass, net, 98
  momentum, 95, 100
  relative momentum, 98
  vertical momentum, 101
  vorticity, 110, 111
Troposphere, 91, 105, 158
  upper, 92, 100, 113
Turbulence, 110, 148, 214
  large-scale, 109, 111
  large-scale, horizontal, 107, 109
  small-scale, 100
Turbulent motion, atmospheric, 108
Turbulent regions of the E-layer, 168

U

Upper atmosphere,
  densities, 139, 140
  pressures, 140, 142
  seasonal variations, 145
  temperature-altitude diagram, 143
  temperatures, 139, 140, 142
Upper Peninsula, Michigan, 336

V

Vegard-Kaplan bands, 174, 175, 191
Velocities, horizontal, 92
Velocities, vertical, 92
Velocity of escape, 120
Viscosity, kinematic, 165, 214
Viscosity, magnetic, 224
Vorticity, 109
  absolute, 110
  distribution, 110
  of the basic rotation, 94
  of the earth, 94
  of the relative motion, 94
  transport, 109
  vertical component of the absolute, 108

W

Waves, planetary, 95, 104
Water vapor, 161, 163, 193, 195
Weather Station Charts, 7
Wind, 147, 214, 215
  high altitude, 176
  observation, upper air, 29
  velocity components, 28

Z

Zodiacal light, 120, 132, 197, 231