J.A.Ratcliffe

World
University
Library

Sun, Earth and Radio

An introduction to the ionosphere and magnetosphere

J. A. Ratcliffe

Sun, Earth and Radio

An introduction to the ionosphere and magnetosphere

World University Library

Weidenfeld and Nicolson
5 Winsley Street London W1

Photoset by BAS Printers Limited, Wallop, Hampshire
Manufactured by LIBREX, Italy

I M. Norbury.

World University Library

The World University Library is an international series of books, each of which has been specially commissioned. The authors are leading scientists and scholars from all over the world who, in an age of increasing specialisation, see the need for a broad, up-to-date presentation of their subject. The aim is to provide authoritative introductory books for students which will be of interest also to the general reader. The series is published in Britain, France, Germany, Holland, Italy, Spain, Sweden and the United States.

Contents

Introduction

Introduction

When Sir George Cayley, the pioneer of aviation, wrote of 'The uninterrupted ocean of air which comes to the threshhold of every man's door' he reminded us of something we often forget. The magnitude of this ocean can perhaps be realised by recalling that its total weight is about 5×10^{15} tons (five thousand million million tons). It presses in all directions with a force of 14 lbs on every square inch; a square column of air with sides measuring four inches weighs about as much as a heavy man, and the air under an ordinary table weighs $1\frac{1}{2}$ kilograms; if a large aircraft could travel without the air inside it the saving in weight would allow an extra four passengers to be carried. We ignore this air around us only because we have never been without it; once we remember its existence many questions suggest themselves. To what height does it reach, is it the same at great heights as at the ground, what is there beyond it?

Answers to questions of this kind have been sought by sending experimental equipment to great heights, first with kites, then with balloons and aircraft, and later with rockets and artificial satellites. Before rockets were used only the first 30 kilometres of the atmosphere could be explored directly, and what was above had to be deduced theoretically from observations made on the ground. In 1924 it was found that radio waves, returning like an echo from above, could be used to probe the structure of the atmosphere up to heights of about 300 kilometres. The early investigators derived a particular aesthetic satisfaction from devising and using these indirect radio methods for investigating the high part of the atmosphere that was inaccessible for direct experimentation. Progress was slow, ideas developed gradually, and some were later proved wrong. It began to be realised that solar radiations had a profound influence on the upper atmosphere and that by studying it much could be learnt about the sun itself. After about 1950 space vehicles were used to make experiments *in situ*: the previous knowledge was extended to heights greater than 300 kilometres, and in some

respects revised, and the space between the sun and the earth began to be explored. At every stage results of the scientific investigations were used to improve operational radio communications, which depended for their success on the electrical state of the upper atmosphere.

Chapters 1 and 2 give a brief account of what was known before radio waves were used as experimental tools in this kind of work and describe the steps by which these waves were progressively made to reveal the electrical state of the high atmosphere and how it depends on radiations from the sun. Chapters 3, 4, 5 and 6 show how these investigations have led to an understanding of the atmosphere up to a height of about 1,000 kilometres, and of how it behaves when illuminated by solar radiations. Chapter 7 is concerned with what lies above 1,000 kilometres, and describes how artificial satellites have been used to explore the environment in which the earth and its atmosphere is situated. Chapter 8 describes how the results of the scientific investigations have been applied to improve practical radio communications. Some of the most recent developments in the subjects dealt with here are described in appendix 2.

A study of the present kind rests on a few fundamental ideas, such as those of waves, photons, electrons, and atoms, which, although not difficult to understand, may not be familiar to all readers. They are therefore explained, in simple terms, in appendix 1 and are printed in *italics*.

In tracing the development of an important branch of science in this way, it is tempting to ascribe discoveries and ideas to the individuals who were responsible for them. For the earlier investigations this might have been possible, but as the subject developed workers contributed to it in greater and greater numbers, and it is quite impossible to do justice to all in a short account of this kind. A decision was therefore made not to mention, save with one or two exceptions, any living investigator. The exceptions are those

(Barnett, Breit and Tuve) who performed the first experiments in radio sounding, and Sydney Chapman, whose name is universally associated with the theory of how the sun's radiation can produce layers of electrification in the atmosphere. It would seem foolish not to use his name to describe it. I am particularly sorry that my decision to exclude names prevents my mentioning some of those, like Dr D. F. Martyn and Sir Harrie Massey who have contributed some of the most important ideas discussed here, and with whom I have had continuous and stimulating contacts since the start of this work. Much of this book is concerned with their experiments and theories.

1 Early knowledge of the upper atmosphere and the sun

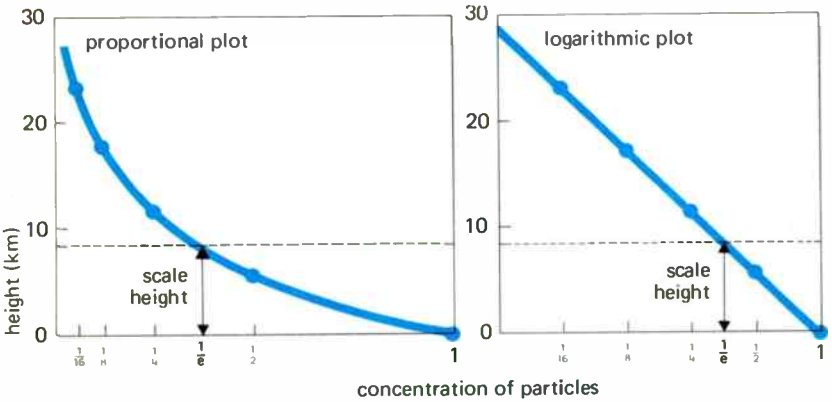
1 Early knowledge of the upper atmosphere and the sun

This book is concerned with the highest part of the earth's atmosphere and beyond, how it has been explored with the help of radio waves and how it is modified and controlled by radiations coming from the sun. Before 1924, when radio waves were first used to explore the upper atmosphere, our knowledge of the region outside the earth was limited to what could be discovered with balloons, which went no higher than about 30 kilometres. Indeed, at that time, we might have known little about what lay at greater heights had it not been that the high atmosphere is ionised and contains free electric charges which give rise to phenomena observable from the ground. Amongst these, the naturally occurring aurora and the easily observed daily changes in the earth's magnetic fields have been studied for more than a hundred years. Observations of these phenomena led to conclusions about the nature of the ionisation and indirectly about the high part of the atmosphere itself. When, in addition to the naturally occurring phenomena, man-made radio waves became available, it was found that this same ionisation controls their travel, so that, by observing their behaviour the existing knowledge could be extended. The deductions from observations of radio waves were, however, sketchy and tentative until, in 1924, they were first used expressly for the purpose of investigating the upper atmospheric ionisation, and from that time onwards our knowledge increased rapidly. Most of this book is devoted to a discussion of these radio wave experiments and what has been deduced from them. This introductory chapter explains what was known about the atmosphere when the radio experiments were first made.

Height distribution of the atmosphere

The atmosphere consists mainly of oxygen and nitrogen held to the earth by the attraction of gravity. The first question that suggests itself is why all the molecules of these gases do not fall to the ground

1·1 The height distribution of gaseous atoms (or molecules) above the earth. The curves refer to a single gas (for example, molecular oxygen) at the same temperature and show how the number of molecules in unit volume (the concentration) decreases with height. The horizontal scale is chosen so that the concentration at the ground is unity. The concentration is halved for each height increment of 6 km so that the decrease is exponential. The increase of height required to reduce the concentration to 1/e (1/2·72) of its original value is called the *scale height*. It is often convenient to plot the logarithm of the concentration against the height, and the graph is then a straight line whose slope gives a measure of the scale height.



and form a very thin layer on the surface. The answer is that if the air were very cold they would do just that, but because it is warm the molecules are in continuous rapid motion, and although they are all pulled downwards by the earth's gravity their heat motions tend to spread them upwards. The way in which they are distributed above the earth is thus the result of two conflicting processes: the tendency to fall to the ground, which depends on their weight, and the tendency to keep moving, which depends on their temperature.

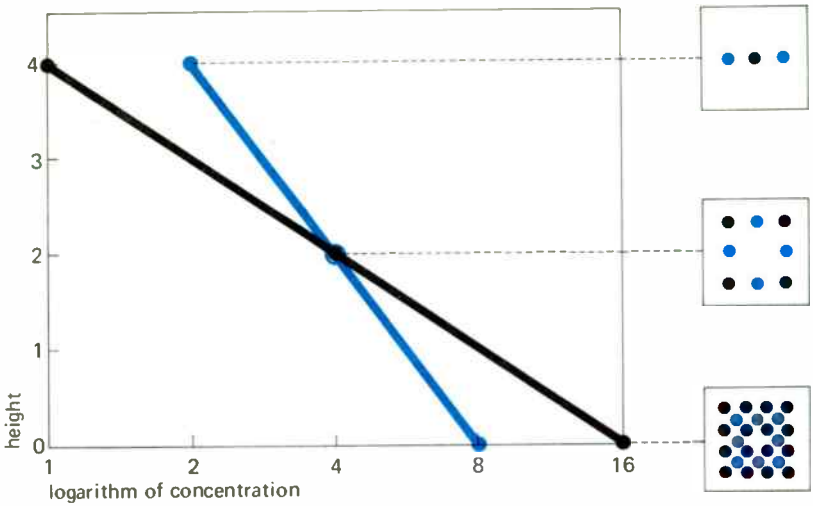
A theoretical investigation shows that, in an atmosphere consisting of a single gas all at one temperature, the concentration of molecules would decrease as we ascended in such a way that, whatever level we started from, a given increase in height would always be accompanied by the same proportional decrease in concentration.

Thus, for example, if the composition and temperature of the atmosphere at all levels were the same as at the ground, the concentration would decrease to one-half of its ground value at a height of 6 kilometres, in the next 6 kilometres it would again be halved, so that at 12 kilometres it would have decreased to one-quarter; in the next 6 kilometres it would be halved again to one-eighth, and so on. If a graph is constructed by plotting the concentration of the molecules against the height, a curve of the kind shown in figure 1·1 is obtained. For our purposes, however, it is often more useful to plot the logarithm of the concentration, for then the graph is a straight line. A height variation of this kind is important in our discussion of the atmosphere; it is called an exponential variation.

In a distribution of this kind the rate of upwards decrease could be described by stating the height interval that would correspond to a decrease of the concentration by half. This would be 6 kilometres in the case discussed. For reasons that we shall not enter into here it is, however, more convenient for theoretical purposes to use the height interval which corresponds to a decrease of $1/2{\cdot}72$ in the concentration.* This height interval is called the scale height and is usually denoted by H; for the air at ground level its magnitude is 8 kilometres. If the results of experiments are plotted logarithmically the scale height can be immediately deduced from the slope of the graph.

Theory shows that the scale height is directly proportional to the temperature and inversely proportional to the molecular weight of the gas. If, therefore, the temperature is increased the scale height is greater and the gas is more extended upwards, so that we have to traverse a greater height interval before the concentration decreases by $1/e$, while if the weight of the molecule is greater, the gas does not extend so far, and the concentration drops to $1/e$ at a smaller height.

* 2·72 is usually denoted by e, and plays a fundamental part in a wide range of mathematical calculations.



The scale height of any gas is a measure of some interesting characteristics. Thus it is the height to which the gas would extend if its pressure were the same at all heights as it is at the ground: a column of air similar throughout to that at the ground would, for example, extend up to a height of 8 kilometres, where it would suddenly come to an end. For this reason the scale height was, at one time, called 'the height of the homogeneous atmosphere'.

The molecules of a gas are in constant motion, their speed depending on their temperature: if a molecule were projected upwards with the average speed corresponding to its temperature, and if it travelled without making collisions, it would reach a height equal to one and a half times the scale height. If it were originally at rest and were dropped from that height, and fell freely, it would arrive at the ground with a speed appropriate to the temperature used in calculating the scale height. Equivalent quantities like these are simple because we have used the fraction $1/e$ in defining the scale height: if a different fraction, such as $\frac{1}{2}$, had been used they would have been more complicated.

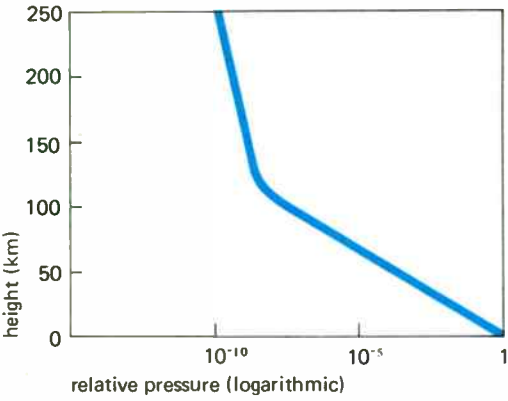
If there were several gases in the atmosphere all with different molecular weights, and if there were no movements tending to mix them together, each would be distributed just as if the others were

1·2 The height distribution of two gases with different scale heights. The weight of the 'blue' molecule is half that of the 'black', so that the scale height of the 'blue' gas is twice that of the 'black'. At the ground the number of 'blue' molecules in unit volume is here supposed to be half that of the 'black' (ratio 8:16), but at height 4 the concentration of 'blue' is twice that of 'black'.

absent. The lightest would have the greatest scale height, it would decrease upwards least rapidly, and even if it formed only a small fraction of the atmosphere at ground level, it would predominate at greater heights. A mixture of gases distributed in this way is said to be 'in diffusive equilibrium'.

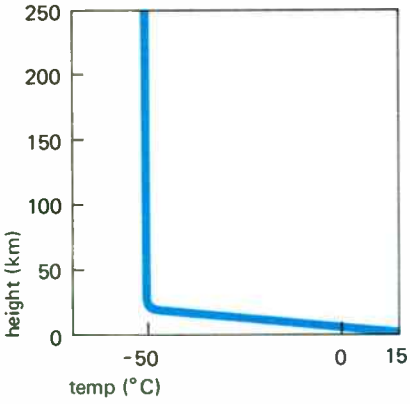
If, however, the gases were mixed by turbulent wind motion they would all have the same height distribution, they would not separate out, and the scale height of the mixture would correspond to the mean of the molecular weights. Thus in ordinary air near the ground the mean molecular weight of the molecules of oxygen and nitrogen corresponds to a scale height of 8 kilometres.

In the early 1920s measurements of atmospheric temperature and pressure were made with balloons up to heights of about 30 kilometres. The temperature was found to decrease steadily upwards from the ground to a height of about 12 kilometres, after which it was independent of height; the pressure was consistent with the idea that the gases were well mixed up to about 20 or 30 kilometres, but at greater heights they were thought to be in diffusive equilibrium so that each varied upwards as it would in the absence of the others. In estimating the proportions of each gas at these greater heights it was remembered that, in addition to the molecular oxygen and nitrogen, the atmosphere at the ground contains a small amount of helium, and that the molecular weight of this gas is only about one-eighth of those of oxygen and nitrogen, so that its scale height is about eight times as great as theirs. Although, at the ground, the proportion of the helium is only about 1 in 100,000 nevertheless because of its larger scale height it will become outstandingly important higher up. With these ideas in mind it was thought that the atmospheric temperature and pressure varied with height as shown in figure 1 3 and that helium predominated at heights above about 100 kilometres. Note that the slope of the 'logarithmic plot' increases by a factor of about eight as it passes from the lower region,

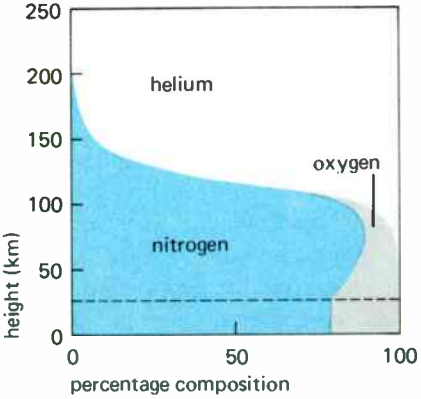


(a)

1·3 In the mid 1920s it was thought that, above about 12 km, the temperature of the air was about −50 °C (b), and that the gases were not mixed above about 20 km. It was then deduced that the pressure and the percentage composition would decrease upwards as shown at (a) and (c).



(b)



(c)

where oxygen and nitrogen are the main constituents, to the upper region where helium is predominant.

At about the time when radio waves were first used to probe the high atmosphere these ideas began to be questioned for reasons unconnected with radio. Three pieces of evidence in particular suggested that previous views about the upper atmosphere required revision: two were based on observations of visible meteors, and of

sound waves emitted from explosions; a third was theoretical and was concerned with the 'photochemistry' of the upper atmosphere.

Meteors, or 'shooting stars', are small particles of matter, about the size of a pin's head, that enter the earth's atmosphere at great speed. The air friction in the lower and denser part of the atmosphere heats them up until they become luminous at heights near 110 kilometres and finally evaporate near 80 kilometres. Theoretical investigations of their heating up and evaporation led to the conclusion that the air density in the region of 100 kilometres must be much greater than had previously been supposed, and it began to be thought that, although the temperature at 12 kilometres was small, it must be greater above, and at about 100 kilometres it must be equal to, or greater than, that at the ground. If this were so then helium would not be more dense than oxygen and nitrogen until heights of 300 or 400 kilometres were reached.

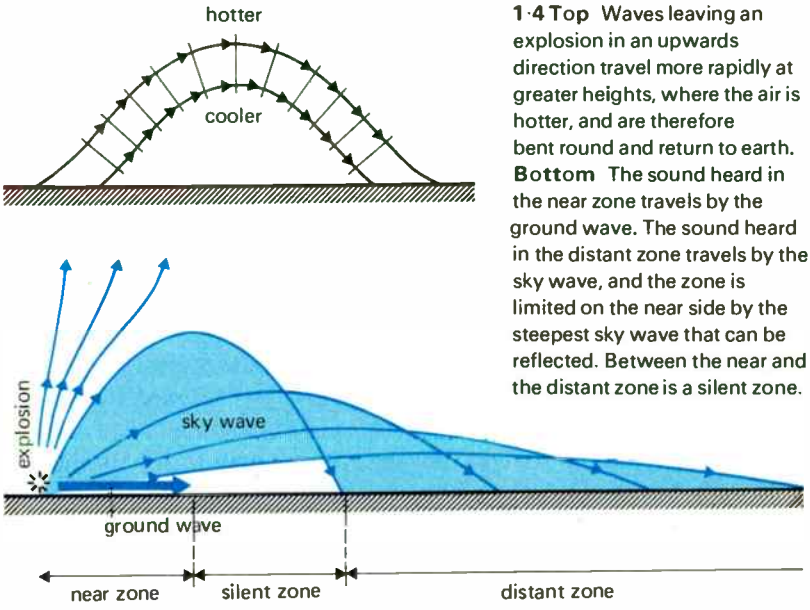
The second relevant observation was made on loud sound waves from explosions. During the First World War it had been noticed that the sound of explosions or intense gunfire was not usually heard at distances around 150 kilometres, although it was audible at greater and smaller distances: there seemed to be a 'zone of silence' between two zones of audibility. The reason for this was that sound is carried through the air by two different waves; a *ground wave* which travels near to the ground, and a *sky wave* which leaves the source in an upward direction, is reflected by the atmosphere at a height of about 30 kilometres, and returns to the ground some hundreds of kilometres from the source. The sound heard in the near zone is carried by the ground wave, that in the far zone by the sky wave, and in the silent zone neither is sufficiently intense to be audible. After the war experiments made to record the sound of loud explosions at several distances led, in the following way, to important conclusions about the temperature of the upper atmosphere.

A sound wave travels more rapidly if the air is hotter so that, if it

leaves the earth at an angle, and the temperature of the air increases upwards, the top of the wave travels faster than the bottom, and finally it bends round and returns to the earth. A wave of this kind travels with little attenuation and is responsible for the sound heard, comparatively loudly, in the 'distant zone'. But not all waves can be reflected: those that leave the ground too steeply travel outwards, into the upper regions of the atmosphere, without being bent sufficiently to return to the ground. The limiting angle of elevation beyond which the waves are not reflected is determined by the temperature of the high atmosphere. Corresponding to this angle there is an inner boundary to the 'distant zone', inside which the sound cannot be heard by way of the sky wave; if it is heard at all it must be by way of the ground wave. But because the ground wave travels in contact with the earth its energy is fairly rapidly dissipated, its strength diminishes, and it can be heard only for a comparatively short distance. The distance within which the ground wave is audible is less than the limiting (nearest) distance at which the sky wave can be heard so that, in between, there is the silent zone where neither is audible.

The sound from experimental explosions was observed at several places at different distances; the time of its arrival and the angle of its descent were measured, and deductions were made about the height from which it had been reflected and the temperature of the air at that height. It was concluded that the temperature at a height of 12 kilometres is less than that at the ground, by about 70°C, but at a height of 30 kilometres it has increased until it is about the same as at the ground, and at still greater heights it exceeds that at the ground.

The observations of meteors and sound waves provided experimental evidence about the density and temperature of the upper atmosphere. At about the same time theorists were considering the way in which ozone was formed from the oxygen of the air by the action of ultra-violet radiation from the sun. Measurements (see



1·4 Top Waves leaving an explosion in an upwards direction travel more rapidly at greater heights, where the air is hotter, and are therefore bent round and return to earth. **Bottom** The sound heard in the near zone travels by the ground wave. The sound heard in the distant zone travels by the sky wave, and the zone is limited on the near side by the steepest sky wave that can be reflected. Between the near and the distant zone is a silent zone.

figure 1·9) had shown that the ultra-violet spectrum of solar radiation is cut off very sharply at wavelengths less than 2,900 Å by the absorbing action of ozone in the atmosphere, and by studying how the intensity of this radiation changes as the sun sets it was shown, about 1925, that the ozone is distributed in a layer some tens of kilometres thick centred at a height of about 25 kilometres. The amount of ozone in this layer is quite small, if it were collected together at the pressure and temperature appropriate to ground level it could all be contained in a layer only 3 millimetres thick. It is because the pressure at the height of 50 kilometres is so small that the thickness of the layer there is so much greater.

Once the layer had been located the question was asked how it came to be formed from the oxygen of the atmosphere. The molecules of ozone consist of three atoms of oxygen and the problem was to find out how they were formed from oxygen molecules containing two atoms. Theory showed that certain wavelengths in the solar ultra-violet radiation disintegrate the molecules to produce oxygen

atoms (O) and that some of these then combine with oxygen molecules (O_2) to produce ozone (O_3) as indicated by the reactions

$$O_2 + \text{ultra-violet} \rightarrow O + O$$
$$O + O_2 \rightarrow O_3$$

Of course not all the atoms combine with molecules; some are left free as atoms, and the surprising conclusion was reached that at heights above about 150 kilometres the molecules are almost completely dissociated so that the oxygen occurs only in the atomic form. The upper atmosphere thus consists of nitrogen molecules oxygen *atoms* and helium.*

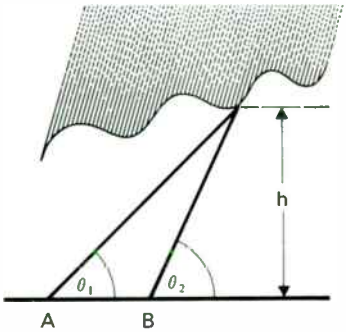
The knowledge of the constituents of the atmosphere obtained in these ways was sketchy and imperfect. It was supplemented by observations of two natural phenomena that occur only because the normal atmosphere is disturbed by influences coming from outside the earth. The first phenomenon was the luminosity in the sky called the aurora, and the second was the variable nature of the earth's magnetic field.

The aurora

It is fortunate, for the purpose of this book, that nature provides a phenomenon in the high atmosphere that is so obvious and striking that it attracted the attention of scientists even as early as 1621. It is seen in the night sky in regions not far from the geographic poles and takes the form of a spectacular display of coloured lights in various complicated forms, often moving and changing rapidly. To those living in middle northern latitudes the luminosity is generally seen towards the north, and since it is often of a reddish colour, like the dawn, it was named the 'aurora borealis', or the 'northern

* It was usually thought that there is no hydrogen in the high atmosphere (compare the modern view on p. 129).

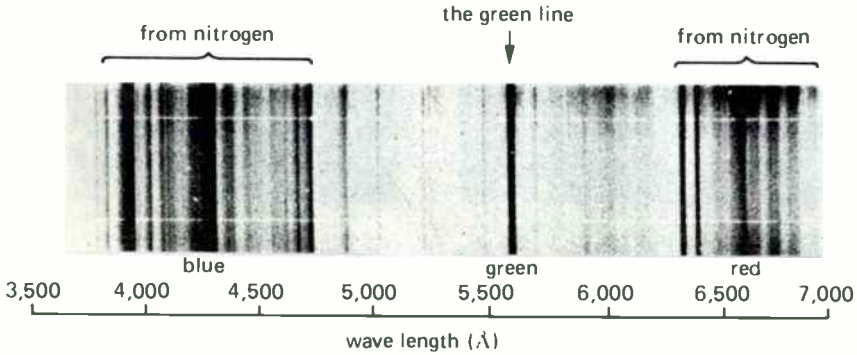
1·5 By measuring the angles of elevation (θ_1 and θ_2) of a particular characteristic of an aurora as seen from two places, A and B, the height (h) of that characteristic can be determined. **1·6 Overleaf** The luminosity in the sky, frequently seen towards the north by those living at fairly high latitudes, often appears like a northern dawn – hence the name *aurora borealis*.



dawn': similar luminosity seen near the south pole is named the 'aurora australis', or the 'southern dawn'. Measurements of the angle of elevation of one particular feature, made simultaneously from several places, have shown that most of the light comes from heights around 100 kilometres. In the early days these observations thus provided the possibility of finding out something about the nature of the atmosphere at heights then inaccessible to man-made equipment.

The light from the aurora is similar to that from an electric discharge in a gas at low pressure, as in a neon lamp, and, since it comes from a height where the pressure is known to be low the question arises whether it, also, might result from an electric discharge. Laboratory studies suggest that light is emitted from a gas when free electrons collide with the atoms or molecules of the gas, and it is natural to suppose that beams of electrons projected from the sun might similarly be responsible for the aurora, provided they could reach the dark side of the earth. Now a stream of electrons is like an electric current and if it approached the earth, which is a huge magnet, it would be deviated so as to impinge equally on the day and the night sides: it would also be concentrated in circular regions

1·7 The spectrum of light from an aurora photographed in 1958. In the early 1920s the parts of the spectrum in the red and the blue were known to be emitted by molecular nitrogen, but the origin of the strong green line was unexplained. It was later shown to come from atomic oxygen.



one round each pole. The aurora is, in fact, most frequently seen (at night) in two roughly circular regions of this kind, so that the hypothesis of excitation by electron streams seems reasonable. This theory of the aurora was accepted in the early 1920s, and it was also realised that the violent changes that are observed in the intensity of the light must correspond to changes in the primary electron stream coming from the sun.

A study of the spectrum of the auroral light revealed lines characteristic of molecular nitrogen, together with a green line which had not previously been obtained from laboratory discharge tubes. For many years the origin of this line remained a mystery, but finally, in 1924, it was demonstrated that it comes from atomic oxygen. This fact is consistent with the inference, drawn later from ozone theory, that at the auroral heights the oxygen is largely in the atomic form.

It is interesting to enquire why this auroral green line was so difficult to produce in the laboratory. A discharge excited by an incident electron beam emits light when an atom first absorbs a certain fixed amount of energy from the beam, and then quickly parts with it in the form of a light photon. Instead of radiating a photon the atom might, however, have passed its energy on to another

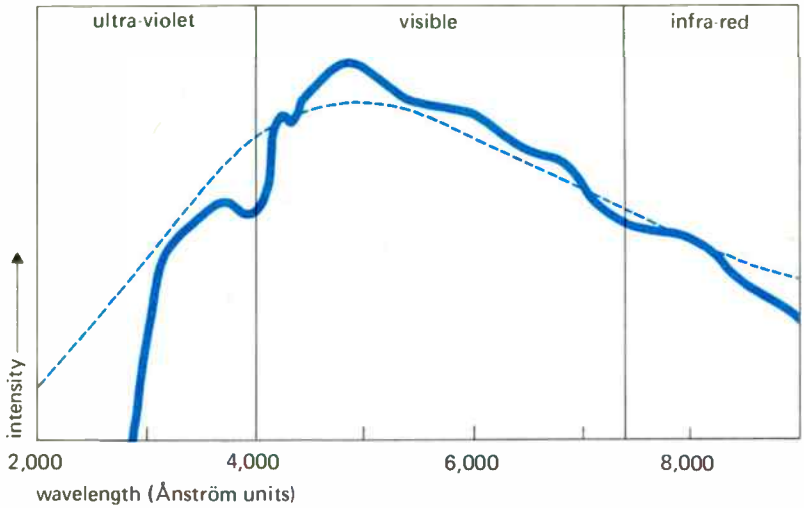
atom or molecule, particularly to one constructed so that it could receive it easily. If during a collision the energy is passed on before the photon is emitted, then there is no light, but if the photon is emitted first there is light. It is all a question of which comes first, the photon emission or the collision.

After excitation by an electron beam, atoms of most gases emit a photon within a short time, of the order of 10^{-8} seconds, much shorter than the time between collisions in the laboratory discharge, so that light is emitted. Oxygen atoms, however, have the peculiarity that, having acquired energy from the electrons, they tend to hold on to it, and they wait a longer time, about 0·5 second, before emitting a 'green-line' photon. During this waiting period, an oxygen atom in a laboratory discharge makes many collisions with other atoms to which it can pass on its energy, long before the time has come when it might have emitted a photon; so there is no green line. In the atmosphere, however, at heights above about 100 kilometres, there are few particles of a kind that can accept the energy from the oxygen atom, and the time between collisions with them is comparable with the delay in the emission of a photon. There is thus a good chance that the photon will be emitted before the energy is handed on in a collision, and it is this photon which forms the auroral green line.

The sun and sunspots

Since the occurrence of the aurora, and the phenomena of geomagnetism to be described in the next section, are found to be closely related to solar events, it is convenient here to describe what was known, in the period before 1924, about the relevant behaviour of the sun.

The distribution of energy in sunlight has been studied for many years and the measurements have been extended at one end into the



infra-red and at the other into the ultra-violet. In the visible part of the spectrum the strength of the radiation varies with the wave length as though the sun were at a temperature of 6,000 degrees,* and this same variation is found to continue at shorter wavelengths into the ultra-violet, until at wavelengths less than about 2,900 Å the atmospheric ozone absorbs the radiation completely. This absorption is, indeed, fortunate for if the shorter wavelengths were to reach the ground they would be lethal to animal and vegetable life.

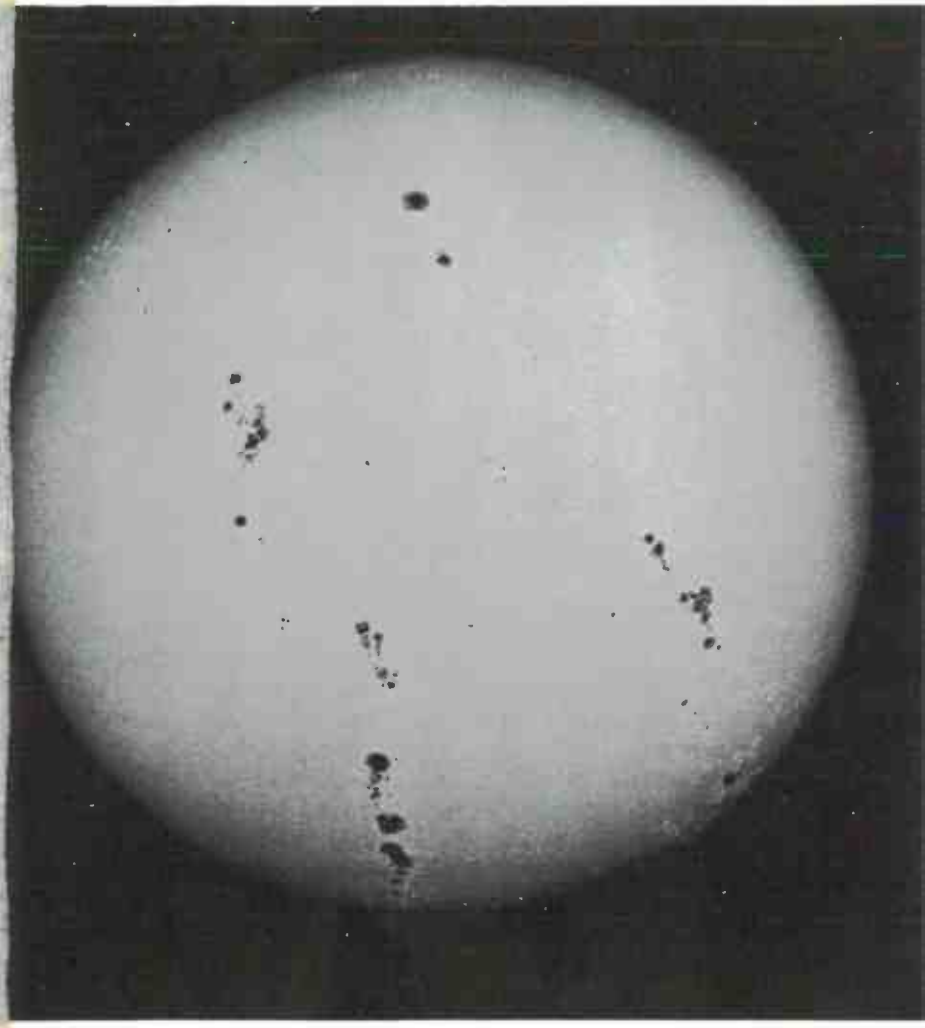
It was usual, in the mid 1920s, to estimate the strength of the radiation at shorter wavelengths on the assumption that the sun continued to radiate, even in the unobservable part of the spectrum, as though its temperature were 6,000 degrees. That would mean that the strength decreased rapidly as the wavelength became shorter, and the strength of x-rays would be very small indeed.

The strength of the solar radiation must be very great near the sun itself and it was suggested by Milne in 1926 that the resulting 'radiation pressure' might, under some circumstances, cause ionised atoms of calcium to be repelled from the sun so that they

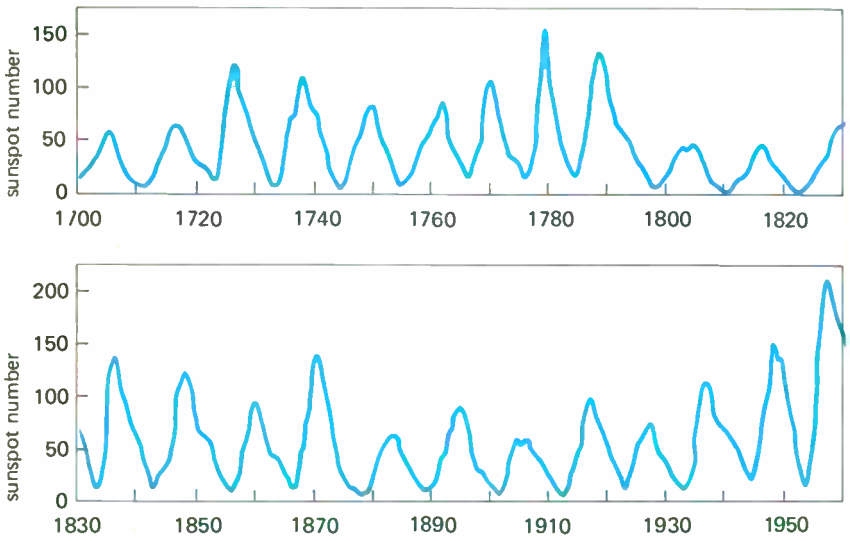
* In this book the word *degrees* referring to temperature means *degrees absolute*. See also the appendix on page 237.

1·8 Left Ground-based observations of the sun's radiation showed that, from the infra-red to the ultra-violet (continuous line), its strength varies as expected from a body at 6,000 degrees (dotted line), but at wavelengths less than 2,900 A the atmosphere absorbs it completely.

1·9 Below A photograph of the sun when there were many spots on its surface.



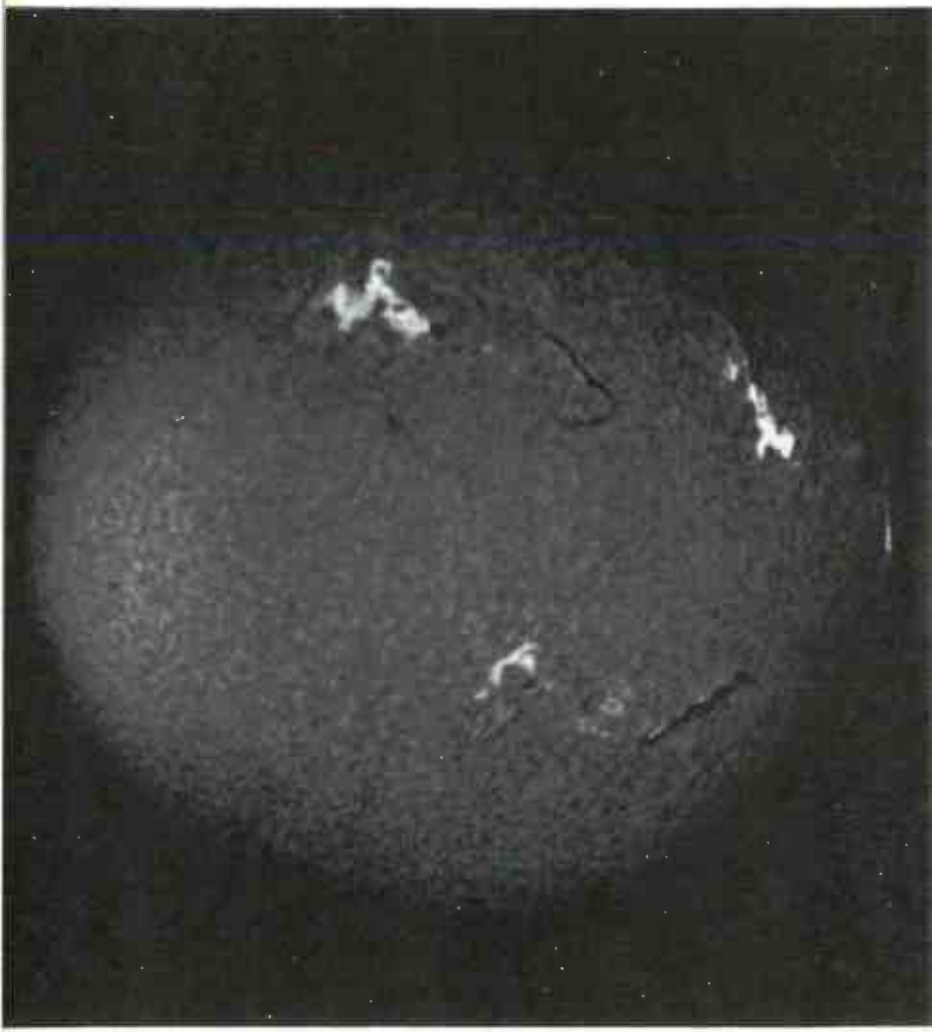
1·10 The *sunspot number* is designed to represent the importance of sunspots by combining information about their number and size. This curve shows how the sunspot number has varied since 1700 and demonstrates the existence of a solar cycle with a period of about eleven years.



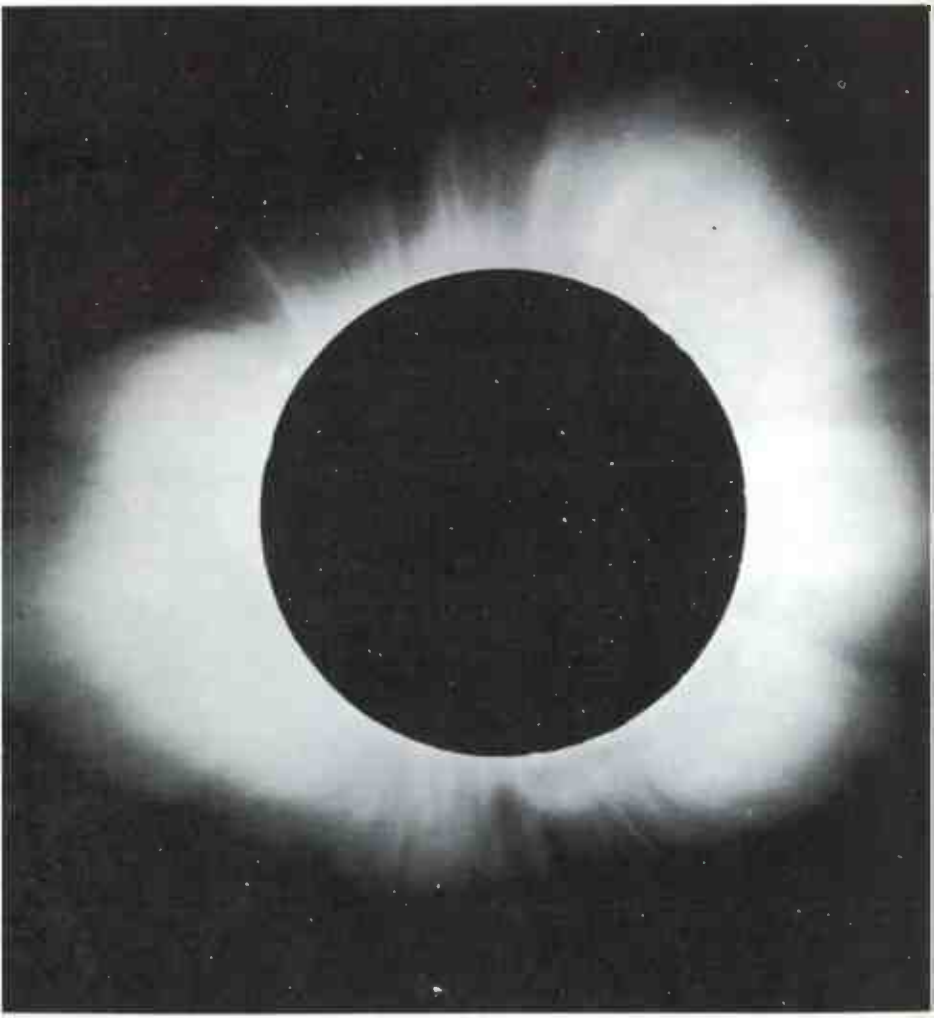
moved towards the earth at about 1,600 kilometres per second.

From time to time dark spots appear on the sun: they are, in fact, very bright and appear dark only because they are seen against the brightness of the solar disc. If one could be separated from the sun and suspended in the sky it would be about one hundred times brighter than the full moon. Observations of sunspots extending over 250 years show that their number waxes and wanes with a more or less regular 'solar cycle' which repeats every eleven years. This cycle is thought to represent some important oscillation in the structure of the sun itself and a corresponding variation in the strength of sunlight has been looked for but none has been found. If a spot persists for long enough it is seen to move across the solar disc in the course of about 13½ days, and sometimes, having reached one side and disappeared, it will reappear on the other side about 13½

1·11 A photograph of the sun taken with the red (H-alpha) light emitted by hydrogen. Three solar flares are visible.



1·12 The luminous corona surrounding the sun can be photographed during a solar eclipse. It is normally invisible compared with the brilliant disc of the sun.



days later. These movements provide evidence that the sun itself rotates with a period of about 27 days.

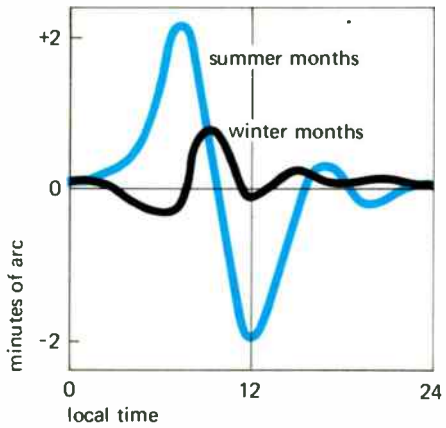
Sometimes the surface of the sun is disturbed by sudden 'storms' which reveal themselves through intense localised bursts of the particular red light, known as H-alpha, emitted by hydrogen. These outbursts are not normally visible against the bright background of the solar disc but can be observed if the H-alpha light is isolated from the rest of the spectrum with the help of a spectrohelioscope. The intensity of the light increases very rapidly, within a minute or less, and then slowly decreases to its normal value over a time of about 30 minutes. The phenomenon, called a *solar flare*, usually occurs near a sunspot. Flares of this kind correspond to the liberation of enormous amounts of energy; they are accompanied by disturbances in the upper atmosphere, some of which are described later (pp. 157 and 198).

The light from the sun comes chiefly from a sharply defined disc, but when it is cut off, during an eclipse, a much more diffuse region of luminosity can be seen extending to distances comparable with the solar radius: it is called the *corona* and is thought to be the source of some important ionising radiations (see p. 150).

Geomagnetism

The aurora provides direct evidence of the state of the upper atmosphere and suggests that it may be profoundly influenced by radiations falling upon it from the sun. Studies of the earth's magnetism (geomagnetism) also suggested to early investigators that incident radiations might ionise the atmosphere at great heights so that it contained charged particles and could carry a current of electricity. These ideas arose from precise measurements which showed that the direction of a compass needle fluctuates in a regular manner throughout the day, and that the nature of the fluctuations is

1·13 Below The direction of a compass needle oscillates in a regular way each day. The curves show the average oscillation observed at Bombay during summer and winter months. The angular movement is quite small: one minute of arc equals 1/60 degree.

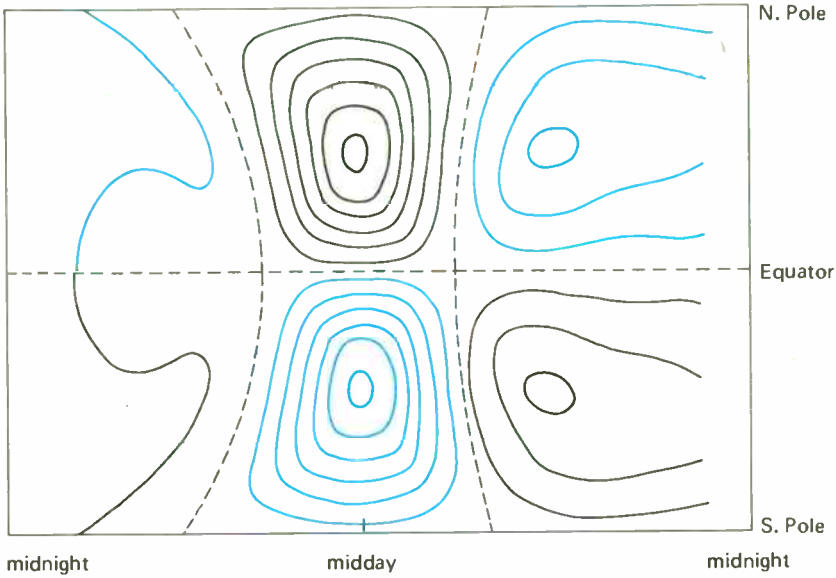


1·14 Right The system of circulating currents in the high atmosphere which causes daily changes in the magnetic field of the kind shown in figure 1·13. Currents flow round the circles in opposite directions for the blue and the black. The map is drawn as seen from the sun, the sub-solar point being marked with an asterisk. 10,000 amps flow between each circle so that, by day, the total current between the centre and the edges of the current system is 60,000 amps. The compass needle oscillates as the earth rotates beneath the currents.

different at different places. Now it is well known that a current of electricity deflects a compass needle, and in 1882 Balfour Stewart suggested that the fluctuations might be caused by electric currents in the high atmosphere. When his suggestion was worked out in detail it appeared that if a huge current system like that of figure 1·14 remained held in position relative to the sun, and if the earth rotated beneath it, the daily magnetic variations could be explained.

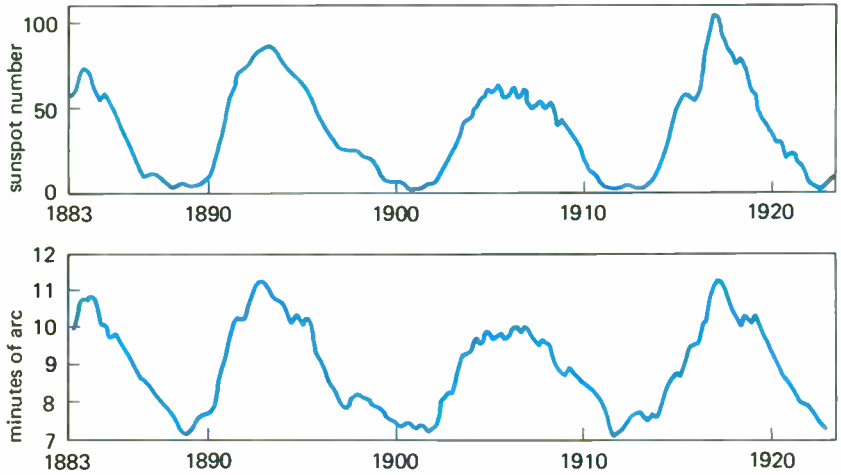
When this suggestion was first made laboratory experiments were showing that, under some conditions, currents could flow through gases provided the pressure was low enough, as it might be in the upper atmosphere; what was required was a battery or dynamo to drive the currents. Stewart suggested that, for the following reasons, this dynamo might be the atmosphere itself.

It is known that a current will flow in an electrical conductor if it is moved across a magnetic field, as when the conducting wires in the rotating armature of a dynamo are swept past the poles of the stationary field magnet. Now if the upper atmosphere is a conductor and if, for any reason, it is moved across the magnetic field of the



earth, it will behave like the armature of a huge dynamo, and electric currents will flow in it. Stewart suggested that the heat of the sun might cause the atmosphere to swing back and forth as day follows night. When the details of this atmospheric dynamo were worked out, it seemed to provide the outline of an explanation for the daily movement of the compass needle.

When the dynamo theory was first suggested it was thought that the upper atmosphere would conduct electricity simply because it was at a low pressure, but towards the end of the nineteenth century it began to be realised that in order to become a conductor a gas must first be ionised by some external agency. The agency that ionised the upper atmosphere would probably come from the sun: it might be x-rays or ultra-violet light, or a stream of particles, and the ionisation produced over the sunlit side of the earth might persist, but with diminished strength, over the night side. The currents flowing in the atmospheric dynamo on the day side would thus be greater than those flowing on the night side, as, indeed, observation had suggested (see figure 1·14).



It was not easy, in the years before the radio experiments of 1924, to decide what sort of solar radiation was responsible for this ionisation. It might consist of photons, like light, ultra-violet, or x-rays, or of moving particles, either charged or uncharged. Charged particles could perhaps be excluded because, like the electrons producing the aurora, they would impinge on the earth mainly round the auroral zones, but there was always the possibility that the stream might consist of a mixture of positive and negative charges which interacted so as to cancel out the guiding effect of the earth's steady magnetic field. A mixed stream of that kind, or a stream of uncharged particles, or radiation consisting of photons, were all reasonable candidates as ionising agents. One thing seemed quite clear; the ionisation was not produced by the electron streams responsible for the aurora, for those impinged on the earth only around the auroral zones, and their intensity varied quite rapidly, whereas the conductivity in the atmospheric dynamo seemed to vary only with a speed corresponding to the normal change from day to night.

There thus grew up a picture of two separate kinds of solar radiation, one reasonably constant and able to ionise the upper atmosphere over the whole of the day side of the earth; and another, thought of as a variable jet of electrons emitted from the sun, striking

1·15 The extent of the daily swing of the compass needle changes through the years in step with the changes in the sunspot number. The curves show the sunspot number and the daily swing between 1883 and 1923.

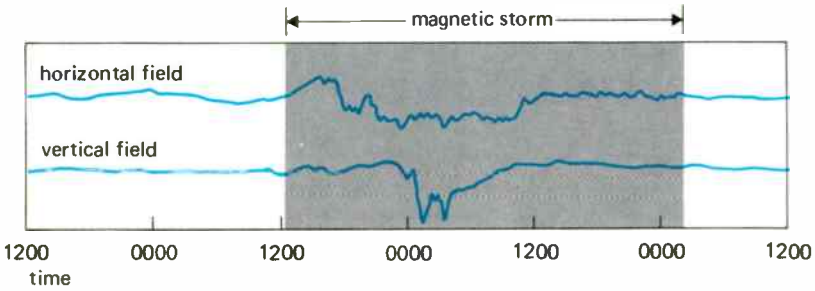
the earth by day and by night but only around the auroral zones, and there producing the aurora.

Observations conducted over a long period had shown that in its daily swing the compass needle oscillated over a range which varied in sympathy with the solar cycle of sunspots, the swing being about 1·8 times greater at sunspot maximum than at sunspot minimum. If the variations in the size of the swing resulted from world-wide changes in the conductivity of the atmosphere the implication would be that the steady solar radiation producing that conductivity itself varied throughout the solar cycle. A variation of this kind was surprising for it was known that there was no corresponding variation in the strength of visible light from the sun.

Another important aspect of geomagnetism had been noticed as early as 1740. On some occasions the regular swing of the compass needle is interrupted; the movements become more pronounced, more rapid, and less regular. These irregularities are called *magnetic storms,* and are most intense near the auroral zones. When a storm of this kind is in progress the aurora is often unusually bright, as if the electron stream from the sun were then more intense. It seems that the electrons which produce the light of the aurora also produce changes in the conductivity of the air and hence alter the flow of the currents. Although such an explanation accounts for the occurrence of magnetic storms, and shows why they are more intense near the auroral zones, it must be revised and elaborated before it can explain all the observations. In particular it cannot explain why, although many storms are most marked near the auroral zones, they are observed, with fair strength, all over the world.

The frequency and intensity of aurorae and of magnetic storms follow the solar cycle closely: the association is, however, only a statistical one and the existence of a particular sunspot cannot, in general, be related to the occurrence of a particular storm or aurora, although in years when there are many spots there are usually many

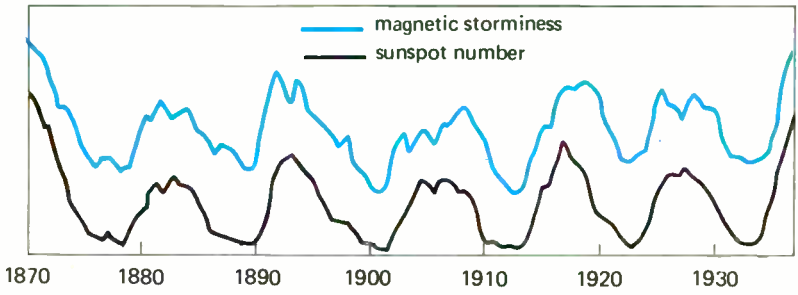
1·16 The strength of the earth's magnetic field undergoes violent changes during a magnetic storm and returns to its normal value only after several hours. The vertical scale on this record is so compressed that the normal diurnal changes are not noticeable.



storms. There is no one-to-one correlation between solar and terrestrial events.

The largest storms are, however, found to occur most often when a very large sunspot is near the centre of the solar disc. This fact seemed to support the suggestion that the storm is caused by a jet of something projected radially outwards from the spot and impinging on the earth. When this relationship between storm and spot was examined more closely it appeared that the storm usually starts, not while the spot is exactly in the centre of the disc, but between one and four days later, as though that was the time taken for the jet of ionising radiation to travel from the sun to the earth. From a knowledge of this time and the distance between sun and earth the speed of the electrons can be calculated, and it is then possible to estimate the latitudes near the poles to which they would be driven by the earth's magnetic field, and the depth to which they would penetrate into the atmosphere. When these calculations were made it became clear that electrons which take one to four days to travel from sun to earth are much too slow to be the direct cause of the aurora, for they would produce an auroral zone much nearer to the pole than the observed one, which is about 25 degrees away, and they would be stopped in the atmosphere at heights greater than 100

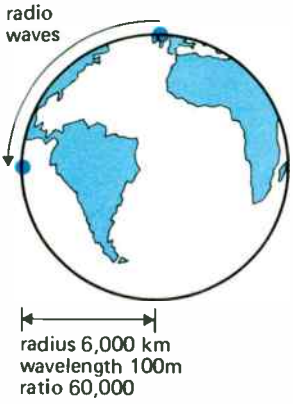
1·17 If the magnetic storminess is averaged month by month and plotted through the years, it is found to follow the solar cycle of sunspots. But there is no clear relation between a single storm and a single spot.

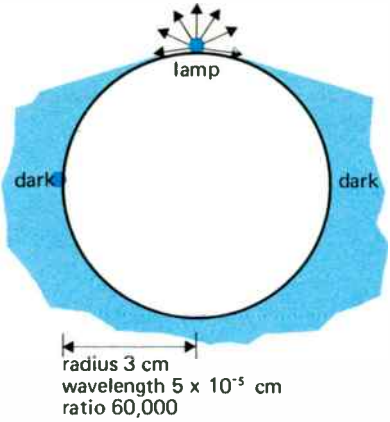


kilometres, where the aurora is known to occur. The speed of electrons capable of producing storms and aurorae must be many times greater. As with the atmospheric dynamo theory of geomagnetism it was clear that the explanation was along the right lines, but that some point of fundamental importance was missing.

The travel of radio waves, and the Heaviside layer

Since the two phenomena, aurora and geomagnetism, so far mentioned, could be observed with the simplest equipment, it is not surprising that they led to early speculations about the electrical state of the upper atmosphere. The third phenomenon that interests us was notice more recently; it concerns the way in which radio waves are transmitted to a distance. In 1901, soon after the invention of radio, Marconi succeeded in transmitting signals from England to America – over the intervening bulge of the Atlantic Ocean. This achievement was difficult to explain, for it was known that radio signals were a form of electromagnetic wave motion like light, the only difference being that their wavelength was much longer; it was also known that waves of that kind travelled nearly, but not quite, in straight lines and the inevitable small amount of





bending, which was produced by the process known as diffraction, was calculated and shown to be much too small to account for their travel round the earth.

After Marconi's experiments new attempts were therefore made to explain what had been accomplished. MacDonald revised the diffraction theory to show that the waves really would be expected to bend the required distance round the earth. This was a remarkable result, and it interested Lord Rayleigh, who immediately wrote a paper to show that MacDonald was wrong and in fact the waves could not be diffracted to anything like that extent. There is a significant passage in Rayleigh's paper, which it is useful to mention to anyone who is too easily blinded by mathematics. He writes:

'Macdonald's results, if they can be accepted, certainly explain Marconi's success; but they appear to me to be open to objection. The first objection that I have to offer is that nothing of this sort is observed in the case of light. The relation of wavelength to diameter of object is about the same in Marconi's phenomenon as when visible light impinges on a sphere one inch in diameter. So far as I am aware no creeping of light into the dark hemisphere through any sensible angle is observed under these conditions even though the sphere is highly polished.' He then proceeds. 'But I shall doubtless be asked whether I have any complaints against the mathematical arguments' and explains in detail why the mathematics was wrong. The interesting point here is that even a mathematician with the

1·18 Far left Radio waves would not be expected to travel from England to America any more easily than light would travel a quarter way round a sphere of radius 3 cm from a point on its surface. Rayleigh showed this by comparing the radii with the wavelengths.
Left Heaviside and Kennelly suggested that a radio wave travelled round the earth after successive reflections from a *Heaviside layer* in the upper atmosphere.

ability of Rayleigh preferred to view the problem in the light of physical common sense rather than through an elaborate piece of mathematics. We should all do well to follow Rayleigh's example.

The other and more correct explanation of the result was more in the nature of a guess than a theory. Almost simultaneously Heaviside in England and Kennelly in America suggested that, in order to bring the waves down at a distance, there must be a reflector overhead, and that it is probably formed by free electrical charges in the upper regions of the atmosphere.

This hypothetical reflecting layer has since been called the Heaviside layer, and much of this book is concerned with the details of its structure. Before we discuss what was known of it before 1924, however, it is interesting to examine a little more closely the suggestions made by Heaviside and Kennelly. Kennelly's suggestion was made in some detail:

'It may be safe to infer that at an elevation of about 80 kilometres, or 50 miles, a rarefaction exists which, at ordinary temperatures, accompanies a conductivity to low-frequency alternating currents about 20 times as great as that of ocean water. ... On waves that are transmitted but a few miles the upper conducting strata of the atmosphere may have but little influence. On waves that are transmitted, however, to distances that are large by comparison with 50 miles, it seems likely that the waves may also find an upper reflecting surface in the conducting rarefied strata of the air. It seems reasonable

to infer that electromagnetic disturbances emitted from wireless sending antennae spread horizontally outwards, and also upwards, until the conducting strata of the atmosphere are encountered, after which the waves will move horizontally outwards in a 50-mile layer between the electrically-reflecting surface of the ocean beneath, and an electrically-reflecting surface, or successive series of surfaces, in the rarefied air above ... As soon as long-distance wireless waves come under the sway of accurate measurement, we may hope to find, from the observed attenuations, data for computing the electrical conditions of the upper atmosphere.'

In comparison with this statement the suggestion of Heaviside was worded remarkably loosely:

'There may possibly be a sufficiently conducting layer in the upper air. If so, the waves will, so to speak, catch on to it more or less. Then the guidance will be by the sea on one side and the upper layer on the other.'

It is thus somewhat surprising that the hypothetical reflector was usually called the Heaviside layer and the name of Kennelly was omitted. The reason for this nomenclature has been explained in the following way. It appears that Heaviside developed his ideas rather more fully and submitted them in an article for publication in *The Electrician*. But unfortunately a referee, to whom it was sent, turned it down and it never appeared in print. Later, in 1912, Eccles wrote a paper on the effect of an ionised layer on radio transmission, and, knowing of Heaviside's rejected article, he went out of his way to call the layer the Heaviside layer, a name which still remains with it.

Let us now return to the idea that the conducting layer is responsible for the long-distance transmission of radio waves. After Marconi's experiment, commercial communication links were set up across the Atlantic and methods of measuring the strengths of the signals were developed. It was then noticed that the strength varied in a regular way throughout the day, the season, and the solar cycle,

and that the regular daily variation was disturbed when a magnetic storm was in progress.

These striking facts suggested that a study of radio waves might lead to new knowledge about the upper atmosphere, and theories were developed to show how an ionised layer reflects radio waves and how its reflecting power depends on the concentration of free charges in it. It then became clear that the daily and seasonal variations of the signal might be explained if the concentration of the charges varies with the inclination of the sun's rays: that the signal strength varies with the solar cycle was particularly interesting, for it seemed to support the evidence, already provided by geomagnetism, that the ionising radiation itself fluctuates with this cycle. The abnormal strength of the radio signal noticed during magnetic storms could be explained in terms of increased ionisation caused by the solar corpuscular stream that produced the storm and the aurora. There was, however, a serious difficulty in this explanation, similar to the one encountered in explaining magnetic storms, for it was known that the aurora occurs only near the auroral zone, whereas radio wave abnormalities are observable all over the world.

The upper atmosphere as understood in the early 1920s

To summarise this chapter it is convenient to describe the upper atmosphere as it was understood in the early 1920s. The air at the ground consisted mainly of oxygen and nitrogen together with a very small proportion of helium. The gases were mixed up to a height of 20 or 30 kilometres, but above that height they separated out, and at great enough heights the lightest, helium, predominated. The temperature decreased in the first 12 or 15 kilometres above ground level but at some indeterminate height it began to increase again until at 100 kilometres it was probably about 300 degrees, twenty degrees greater than at the ground.

On the day side of the earth radiations coming from the sun and falling on this atmosphere produced a layer of ozone at a height of about 25 kilometres and caused the oxygen to exist mainly in the atomic rather than in the molecular form at heights greater than about 150 kilometres. They also ionised the air, at some unknown height, and thereby rendered it conducting so that it reflected radio waves and, through the mechanism of the atmospheric dynamo, caused changes in the geomagnetic field. This solar ionising radiation might be photons of ultra-violet or x-radiation, or might be a stream of particles, charged or uncharged. Whichever it was, its strength waxed and waned with the solar cycle of eleven years.

Sometimes, in addition, streams of electrons were emitted from the sun and, being guided to the auroral zones by the earth's magnetic field, produced the visible aurora and increased the ionisation of the upper atmosphere so that there were disturbances in the smooth variation of magnetism and in radio wave propagation. The frequency of these events varied with the eleven-year solar cycle.

This picture of the upper atmosphere was necessarily somewhat hypothetical. And to justify it, or to extend and modify it, new methods of investigation were necessary. In 1924 radio waves were first used for the express purpose of investigating the electron content of the upper atmosphere and in the words of Kennelly 'wireless waves came under the sway of accurate measurement, and data for computing the electrical condition of the upper atmosphere became available'. In 1951 experimental equipment was sent in space vehicles to make direct measurements at great heights. The rest of this book describes investigations of these two kinds and shows how they have extended and modified the earlier ideas about the upper reaches of the atmosphere.

2 Radio sounding of the upper atmosphere

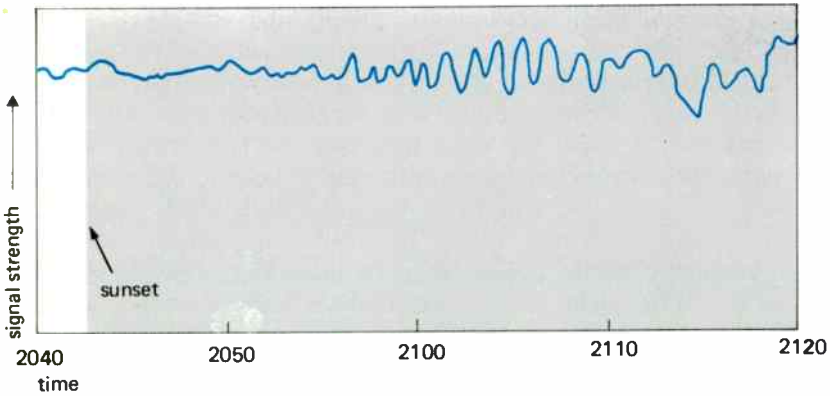
2 Radio sounding of the upper atmosphere

The reflection of radio waves

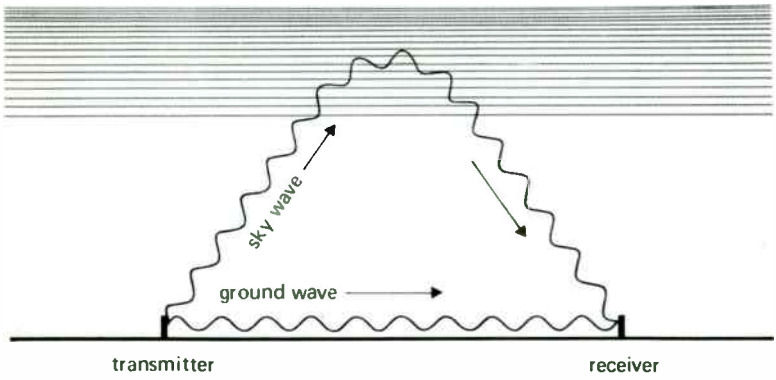
When radio broadcasting started in the early 1920s it was noticed that the strength of the signal received at night over distances of about 100 kilometres often varied over a wide range within a period of a few minutes; sometimes it disappeared completely. This fading of the signal was ascribed to the superposition of two waves; one had travelled over the ground (the ground wave) and the other (the sky wave) had been reflected from the Heaviside layer. Although it had been believed for some time that the layer could reflect transatlantic waves impinging on it very obliquely, there was no evidence that it could also reflect waves travelling nearly vertically, and the suggestion was, at first, no more than a guess. Indeed, if an analogy could be drawn from the behaviour of sound waves (see p. 24), it might be supposed that vertically travelling waves would not be reflected. In 1924, therefore, two sets of workers set out to investigate the matter. Almost simultaneously Appleton and Barnett in England, and Breit and Tuve in America, performed experiments which demonstrated clearly that during darkness a radio wave travelling nearly vertically is in fact reflected from the Heaviside layer.

These were crucial experiments, of far reaching importance, for they suggested that radio waves could be used as a tool to explore the nature of the charged particles in the ionised upper atmosphere. If it proved possible to find the height distribution of the ionisation at different times and at different places, conclusions might be drawn about the atmosphere which had been ionised, and also about the radiation, presumably of solar origin, that ionised it. This chapter gives an account of the different kinds of experiment that were made with these objectives.

2·1 Soon after sunset the fluctuations of the signal from a medium-frequency broadcasting station about 100 km away begin to increase. This diagram shows one of the early records of this fading observed at Cambridge on the broadcast signal from London.



2·2 Fading is caused by the ground wave and the sky wave arriving at the receiver sometimes in step so that they add, and sometimes out of step so that they subtract.



Height of reflection

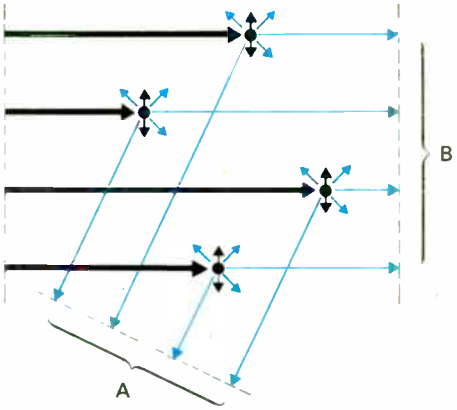
From the time of travel of vertically reflected waves it is possible to deduce the height from which they have been returned if their speed is known. While they are travelling in non-ionised air, low down in the atmosphere, their speed has the well-known value appropriate to empty space, but higher up, when they enter the ionised region, their speed is different, and since nothing was known about the concentration of the ionisation it was not possible to estimate the speeds at different heights. All that could be done was to calculate an *effective height* at which it would be necessary to place a sharply reflecting layer if it were to return the waves with the observed time delay. This effective height was found to be about 100 kilometres; it was believed that the real height of reflection was of the same order, say not more than 20 kilometres different, but there was at first little justification for this belief.

Mechanism of reflection

Once it had been demonstrated that waves are reflected vertically, and the effective height of reflection had been measured, attempts were made to understand in more detail how the reflection was produced. It was supposed that the radio waves set the electrically charged particles of the atmosphere into oscillation and cause them to radiate tiny secondary wavelets. These wavelets travel radially outwards in all directions: the problem was to consider their effect on a receiver. When waves from a single source arrive at a receiver by many different paths they reinforce each other strongly only if the lengths of the paths are all equal. For the purpose of the present discussion a path length must include not only the distance travelled by the wavelet between an oscillating particle and the receiver, but also the distance travelled by the main wave before it sets

the particle into oscillation. Now the atmospheric particles are distributed in random positions, as indicated in figure 2·3, so that, for the wavelets travelling to a receiver situated in most directions, such as A in the figure, the total path length is different for each particle; in the forward direction B, however, the total path (including the part before the particle is reached and the part after the wavelet is radiated) is the same for each particle. In the direction B, therefore, the wavelets add together to produce a strong wave, while in other directions their 'incoherent' combination produces only a wave so weak that it could not be detected until advanced techniques became available after the Second World War. On page 78 we discuss results which have been deduced from these 'incoherent' scattered wavelets, but for the present we shall neglect them, and concentrate our attention on the strong wave re-radiated forwards.

The strong, forward-going, scattered wave is, of course, travelling in the same direction as the original main wave, coming from the transmitter, and it is now necessary to see how they add together. Although they have traversed actual paths which are equal, the wavelets radiated by the oscillating particles are a quarter of a period ahead of the oscillation in the main wave (the phase – see appendix – is advanced by 90°) so that when the two waves are added the oscillation in the combined wave occurs a little earlier than in the original wave (its phase is slightly advanced). The result is that the composite wave appears to have travelled a little faster, and to have arrived earlier. If the concentration of the electrons is greater then the re-radiated wave is greater, so that in the combined wave the oscillation occurs even earlier (the phase is further advanced) and the wave seems to have travelled even more rapidly. A full treatment of the problem shows that the speed depends on the frequency of the wave, and on the mass, charge, and concentration of the charged particles. The change of speed is greater as the frequency is smaller



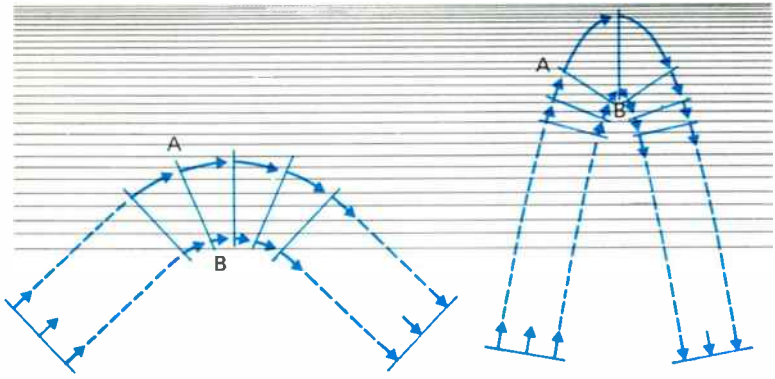
2·3 A radio wave passing through a collection of charged particles. The main wave (black) makes the charged particles oscillate (small black arrows) so that they re-radiate wavelets (blue) in all directions. In a direction such as A the total distance travelled by black plus blue waves is different for each particle and the combined wave is weak. In the forward direction B the total distances are all the same and the combined wave is strong.

and is greater for light charged particles (which can more easily be set into oscillation) than for heavy; light electrons are therefore more effective than heavier ions in speeding up the wave.

Consider next what would happen to a wave impinging obliquely on the ionised upper atmosphere. Suppose, as is likely, that above a certain level the concentration of free charges, and hence the speed of the wave, increases upwards. Different parts of the wave front, travelling upwards obliquely from the ground, would find themselves in places where the charge concentration was different. The greater concentration at greater heights would cause the top of the wave front to travel more rapidly than the bottom so that it would gradually swing round towards the earth to form a reflected wave.* If the original wave travelled more steeply the top of the wave front would have to travel more rapidly to catch up with the bottom, and that would require a greater concentration of charge. If this argument is carried to its conclusion it is clear that even a wave impinging vertically on the Heaviside layer will be reflected provided it can reach a level where the speed of the wave is infinitely great so that it can swing over quickly enough. Since the speed depends in a known way on the mass, charge, and concentration of the electrified particles, and on the radio frequency, it is possible to calculate the

* A similar situation for sound waves is discussed on p. 24.

2·4 When a wave falls obliquely on an atmosphere containing charged particles whose concentration increases upwards, the top (A) of the wave front travels more rapidly than the bottom (B) so that the wave swings over and is reflected. A wave arriving more steeply must travel higher to find enough charged particles to reflect it.

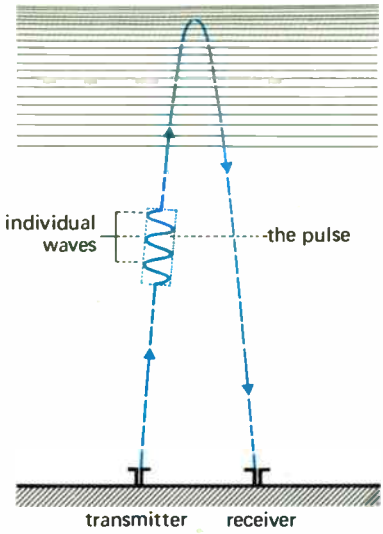


charge concentration required at the reflection level to make the speed infinite, provided the other quantities are known.

Before considering the important information that can be obtained in this way, let us look a little more closely at the conditions that determine reflection.

The travel of a pulse of radio waves

Reflection occurrs when the speed of the wave is infinite, a condition which at first sight might seem a little odd. In considering the matter in more detail, however, we must remember a point that is always important with any kind of wave motion. A steady uninterrupted wave with constant strength would be no use in an experiment designed to measure the height of reflection, for it would simply produce a constant unvarying response in the receiver at the ground, from which nothing could be deduced. To make it useful it must be broken up into small bits, or pulses, whose arrival at the receiver can be timed. This is what Breit and Tuve did. The same effect can be obtained in more roundabout ways; for example,



2·5 A pulse of radio waves used for reflection experiments.

Appleton and Barnett used a wave of constant strength, altered its frequency back and forth continually, and then measured the time that elapsed between the emission and reception of one-and-the-same frequency. Other arrangements can be used, but in all of them some sort of 'label' must be put on the original wave and the arrival of this 'label' at the receiver must be timed. In any kind of signalling it is always necessary to use labelled bits of the wave of this kind – usually called *groups*.

It is known that if the speed of a wave varies with its frequency, as it does in the ionised atmosphere, then the speed of a group is not the same as that of the wave. In an ionised medium, like the one we are considering, the individual waves travel faster than the group, so that when they reach the front they die out, and new ones are continually formed at the rear to replace them. Of course, the individual radio waves are invisible, but the same kind of behaviour is easily observable with the pulse of water waves that spreads out in a circle when a stone is dropped into a pond. Once this pulse has travelled some distance it is seen to consist of a group of waves that moves forward as a whole (with the 'group speed') but if we watch

one wave in the group we see that it moves (with the 'wave speed') faster than the group which contains it, so that ultimately it reaches the front and dies out, meanwhile being replaced by a new wave which appears at the back.

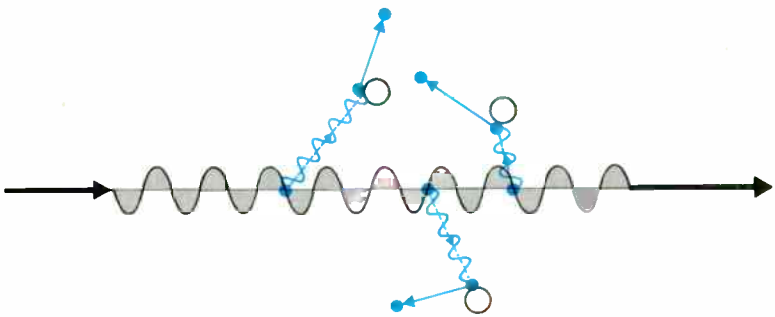
When radio waves travel through ionised air an increase in the concentration of electric charges causes the groups to travel more slowly and the waves more rapidly. When waves penetrate more deeply into the Heaviside layer, therefore, although the waves are speeded up, the group is progressively slowed down, and reflection finally occurs at the level where it comes to rest. It is somewhat easier to visualise the mechanism of reflection in terms of the zero speed of the group, rather than the infinite speed of the wave.

Once the mechanism of reflection was understood it became possible to make deductions about the charged particles at different heights. Before we consider that matter, however, it is convenient to pause and to discuss what determines the strength of the reflected wave.

Strength of the reflected wave and absorption in the layer

The early experiments showed that the waves from broadcasting stations are reflected only at night and that when sunlight reaches the layer at dawn they rapidly become weaker until, one or two hours later, they are undetectable. This weakening could not be attributed to a failure in the reflecting power of the layer, for the arrival of the sun's radiation would be expected to assist the reflection by increasing the concentration of charged particles. It must be that, in its travel to the reflection level and back, the wave gives up some of its energy to the ionised atmosphere around it, so that it is progressively absorbed, and that this absorption is greater when there is more ionisation. The mechanism of the absorption was therefore studied.

2·6 The mechanism of radio wave absorption. The main wave gives energy to the oscillating electrons (blue), which then make collisions with neutral particles (black outline) and bounce off in random directions carrying the energy of oscillation with them so that their temperature is increased.



Reflection has already been explained in terms of the wavelets re-radiated from oscillating charges set in motion by the oncoming wave. If we consider this situation a little more fully we see that energy is fed from the main wave into the charges and is then re-radiated, by way of the wavelets, back into the main wave so as to alter its speed. But that is not the whole story: during their oscillations the charges often collide with the neutral air particles that surround them. After a collision they bounce off, sometimes in one direction and sometimes in another, depending on the precise circumstances of the encounter; the regular oscillations are interrupted, and energy has to be fed in from the main wave to re-start them. The wave thus loses energy every time the oscillating charges make a collision, with the result that it becomes weaker, or is absorbed, as it travels.

The amount of energy absorbed in any particular volume of air

depends on the total number of collisions made by the charged particles in it and hence on the number of charged particles in each cubic centimetre (their concentration) and the number of collisions which each makes per second (its collision frequency). The absorption is determined by the product of these two quantities and because the collision frequency depends on the concentration of neutral particles with which the charged ones can collide, it follows that absorption is greater if the concentration of either the charged particles or the neutral particles is greater. Thus absorption is greater at *times* when the charged particles are more numerous and at *places* where the air is denser so that the concentration of uncharged particles is greater.

The observed decrease in the strength of the reflected waves at sunrise is thus explained as caused by increased ionisation below the level of reflection, at heights where the neutral particles are so numerous that there are many collisions. This low-lying ionisation is first produced by the arrival of the sun's rays at dawn.

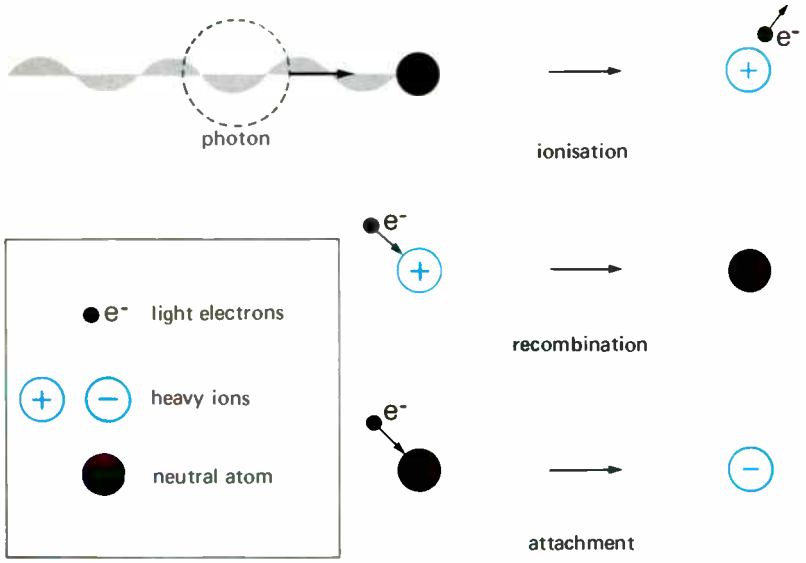
Although most of the absorption occurs low in the atmosphere, where the collision frequency of the electrons is great, a significant amount also occurs near the level where the wave is reflected, for the following reasons. When a wave carrying constant power penetrates to heights where its speed is increased, the strength of its electric field is also increased,* and correspondingly large oscillations are set up in the charged particles. At each collision much energy is thus lost from the wave, which therefore suffers considerable absorption. Now it was stated on page 55 that the speed of the wave becomes very great (ideally infinite) at the level of reflection; the electric field is therefore very great, and there is great absorption at that level, even if the collision frequency is not very large.

* This increase of the wave's electric field is accompanied by a decrease of its magnetic field so that (apart from absorption) the power in the wave, proportional to the product of these two fields, remains constant.

Normally this strong absorption near the reflection level occurs over only a short distance, and its effect on the reflected wave is small. Under some circumstances, however, the radio pulse traverses a comparatively great part of its path with a small group speed so that its total delay time is increased. The absorption is then greater and theory shows that the collision frequency can be deduced if a comparison is made between the increased echo delay and the decreased strength of the pulse. Comparisons of this kind have been used to determine the collision frequency near the peaks of the ionospheric layers discussed on page 119.

It is interesting to enquire what happens to the energy abstracted from the wave. To answer this question we remember, first, that the charged and neutral particles in the atmosphere are in continuous random motion with speeds corresponding to their temperatures. When a charged particle is set into oscillation by a radio wave its energy is slightly increased so that after a collision it 'bounces off' with a little extra speed, as though its temperature were slightly greater, and the process then starts all over again. The result is that the ordered energy of oscillation that the particle receives from the wave is converted, at each collision, into disordered, random energy corresponding to a slight increase in its temperature, and the charged particles are heated up. This heating of the charged particles would continue indefinitely were it not that, ultimately, they pass their heat on to the much more numerous uncharged particles surrounding them. The effective charged particles are light electrons, which hand their energy on to heavy neutral particles only with difficulty: the temperature of the electrons is therefore increased quite substantially when a radio wave that they absorb is sent into the layer. Experiment and theory indicate that the switching on of a powerful broadcasting transmitter can increase the temperature of the electrons at a level of about 100 kilometres by as much as 45 degrees.

2·7 Production and loss of light electrons and heavy ions.
Ionisation: a photon impinging on an atom removes a negative electron and leaves a heavy positive ion.
Recombination: an electron attaches itself to a positive ion so that the charges neutralise each other and a neutral atom is formed.
Attachment: an electron attaches itself to an atom to produce a heavy negative ion.



The concentration of charged particles

Once the height of reflection of the waves had been roughly determined it was possible to estimate the concentrations of the charged particles that reflected them. For this purpose it was necessary to know the mass of the particles, and at first there was doubt about what it might be. It was known that when photons in the solar radiation ionise the air, free electrons and positively charged ions are produced. The electrons do not remain free for long: they either recombine with the positive ions to re-create neutral particles, or they attach themselves to neutral particles to form negative ions. Charged particles with two different masses might therefore be present: electrons, with small mass and negative charge, or ions,

with much greater mass and either a positive or negative charge.

A charged particle has more effect on the travel of a radio wave the smaller its mass, so that if electrons and ions are present in roughly equal numbers the effect of the electrons is predominant. It was therefore reasonable in the first place to suppose that the effective charges were electrons and to calculate their concentration at the level of reflection. It then appeared that there must be about 10^4 electrons per cubic centimetre at the levels near 100 kilometres where the waves were reflected. Other evidence had led to the belief that, at these levels, there are about 10^{12} neutral molecules per cubic centimetre, so that only about one part in 10^8 of the air appeared to be ionised.

Of course, it might be that most of the electrons had become attached to heavier particles to form negative ions, so that positive and negative ions jointly might be the important constituents, in spite of their much greater masses. If that were so they would have to be more numerous in the ratio of the ionic to the electronic masses, that is to say, about 10^4 times more numerous, so that the concentration would be about 10^8 per cubic centimetre, and about one in 10^4 molecules would be ionised. It was important to decide which of these figures was correct and it was realised that the action of the earth's magnetic field could be made to give the answer, as described in the following paragraph.

The magneto-ionic theory

When a charged particle, ion or electron, moves in a magnetic field it travels in a spiral path, simultaneously moving along the field line and rotating round it. The speed of the rotation in the spiral depends on the particle's charge and mass and on the strength of the field, and under the action of the earth's magnetism the rotations of ions and of electrons occur about 100 times and 10^6 times per second

respectively. Now we have seen that when a wave enters the Heaviside layer it sets the charges into oscillation, the energy lost from these oscillations during collisions causes absorption of the wave and the wavelets radiated from the oscillations modify the speed of the wave and cause the reflection. It would be expected that these oscillations would be modified by the rotation of the charges round the earth's field; if there were few rotations during an oscillation the modification would be small, but if there were many it would be important. The waves used in the experiment made about 10^6 oscillations per second, roughly the same as the number of rotations of the electrons, but many more than the number of rotations of ions. If the effective particles were ions, therefore, the rotations would not be important, but if they were electrons the rotations might have some effect which could be detected. What might it be?

Fortunately it was not necessary to investigate this problem theoretically from first principles, because a similar problem had been solved round about 1900 by Lorentz, who had studied the travel of light waves through a transparent crystal placed in the space between the poles of a magnet. He supposed that each molecule of the crystal consisted of an electron attracted by an electric force to a positive charge in such a way that it could oscillate back and forth, under the action of this force, with its own natural frequency. He studied what would happen when a wave forced the electrons into oscillation with another frequency, and how these forced oscillations were modified by the presence of the superimposed magnetic field. He concluded that, at any one frequency, it was possible for two different types of wave to travel through the crystal and that they would have different speeds: they are called *characteristic waves*.

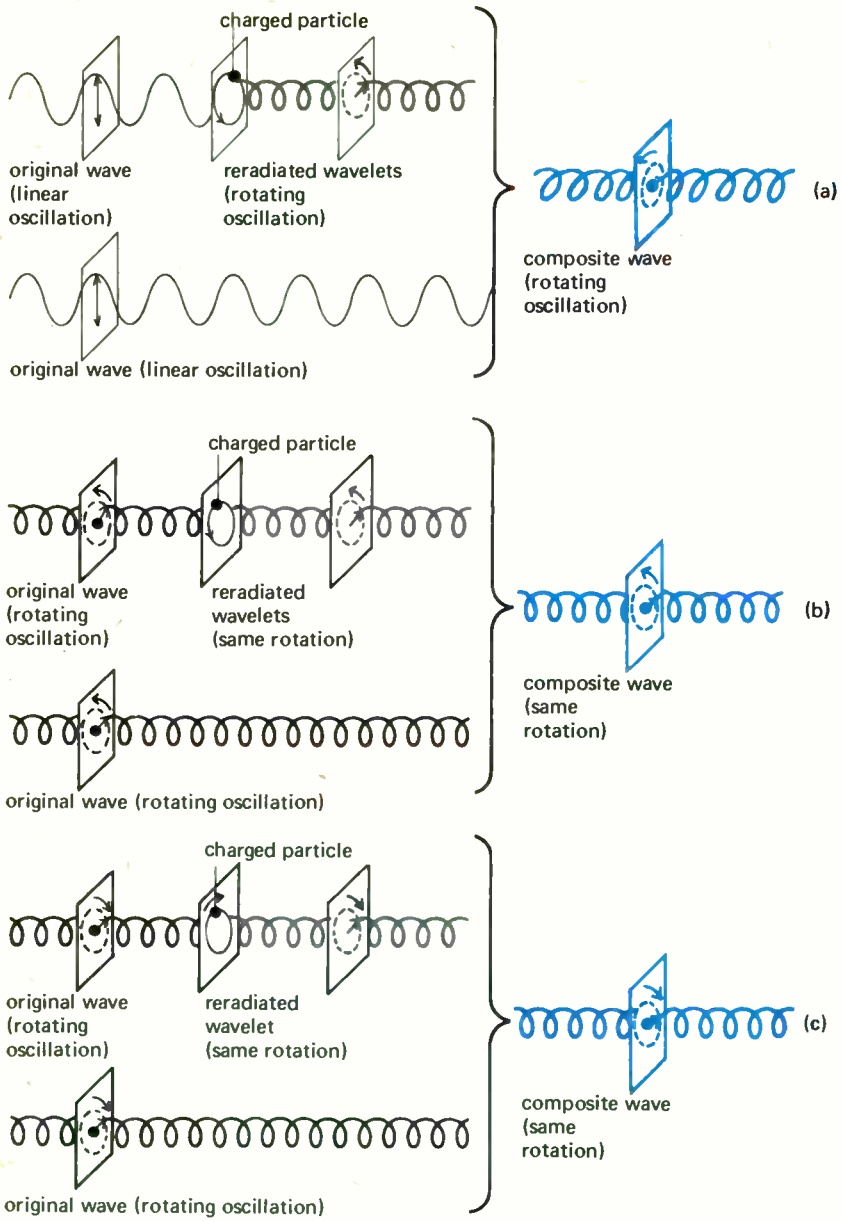
The description of the characteristic waves is given most simply in terms of the small-scale picture of what happens as they travel. If there is no imposed magnetic field the wave forces the electrons into oscillation in the direction of its own electric field. The wavelets

re-radiated from these electrons have electrical oscillations also in the same direction and they add to those in the original wave to produce an oscillating electric field similar to the original one. (The phase, however, is different so that, as already explained, the wave speed is altered.) If, however, a magnetic field is present, the electric field of the oncoming wave sets the electrons into oscillation, not along its own direction, but in a circular motion round the magnetic lines of force, and the electrons then reradiate wavelets with a rotating electric field. These now combine with the original wave to produce a field that oscillates in a more complicated manner.

The word 'polarisation' is used to indicate the nature of the oscillating electric field in a wave: if it oscillates back and forth along a straight line we say the wave has *linear* polarisation; if it rotates we say it has *rotating* polarisation. In the example under discussion the original wave has linear polarisation, but the re-radiated wavelets have rotating polarisations, and when they are added to the original wave they produce a modified wave which also has rotating polarisation. A linearly polarised wave impinging on the crystal thus emerges as one with rotating polarisation; its polarisation has altered.

There are, however, two particular kinds of wave-polarisation that are not altered by the crystal. In each of them the motion of the electrons is similar to the wave-field that drives them: the re-radiated wavelets thus have the same polarisation as the original wave and when they combine with it they do not change its polarisation. These are the characteristic waves; they travel with different speeds and are polarised with electric fields rotating in opposite directions. If a wave with different polarisation impinges on the crystal it is split up into these two waves, which then travel independently.

This theory was developed by Lorentz to explain some of the observations made by Faraday who, in a search for interaction



2·8 The behaviour of characteristic waves.
(a) When a linearly polarised wave passes through an assembly of charged particles in the presence of a magnetic field it causes them to move round in circles. They then reradiate wavelets in which the electric field rotates. The composite wave that results when these wavelets add to the original wave has its electric field rotating so that the polarisation is different from that of the original wave.
(b) If the original wave has its electric field rotating in a certain way it makes the charges rotate in the same way, they reradiate wavelets with the same kind of rotation, and when these are added to the original wave they produce a composite wave whose field also rotates in the original way. In this case the polarisation has not been changed, and the wave is a characteristic wave.
(c) A second characteristic wave is possible in which the rotation is in the opposite sense.

between light and magnetism, had experimented with light beams passing through transparent substances placed between the poles of a magnet. It is interesting to note that the success of the theory was one of the chief reasons in 1900 for the belief that electrons, of the kind which J.J.Thomson had discovered to be moving freely by themselves in gas discharges, also existed, bound closely to positive charges, in the molecules of transparent crystals.

The theory of Lorentz dealt with a problem which was one stage more complicated than that of the Heaviside layer, for Lorentz's electrons were bound by electric forces to positive charges in the atoms and could perform their own natural oscillations, whereas the electrons in the Heaviside layer were free and oscillated only under the action of the driving wave. In another sense, however, Lorentz dealt with a simpler problem by discussing only those situations where light travels either along or at right angles to the imposed magnetic field, whereas in the radio experiments the angle between the directions of the wave's travel and of the earth's magnetic field had an intermediate value. The radio workers had thus to

modify the Lorentz theory in two ways: first, to deal with free and not bound electrons, and second, to deal with situations where the angle between the magnetic field and the direction of travel might take any value. This revised theory, known as the magneto-ionic theory, has played a central part in radio wave investigations of the upper atmosphere. It is still widely used in studies, not only of the upper atmosphere, but of waves in plasmas of the kind that are of interest in cosmology and in the study of thermo-nuclear reactions.

The results of the magneto-ionic theory are contained in an expression usually called the Appleton-Hartree equation. The reason for this name is interesting. In his theory of waves passing through a crystal, Lorentz had taken account of the force that one pair of charges exerts upon another pair; in his mathematics he had expressed it through what is now called the 'Lorentz term'. When Appleton first adapted Lorentz's theory to apply to free electrons (not bound to positive charges) he omitted the Lorentz term; a little later Hartree investigated the matter afresh and concluded that the term should have been included, so Appleton's expressions were suitably modified and the resulting equation was called the Appleton-Hartree equation. Later still, after some quite involved theoretical work, supported by observations of waves reflected from the atmosphere, it was finally concluded that the Lorentz term should, after all, *not* be included in the Magneto-ionic theory for free electrons: Appleton's original equation was thus correct, without Hartree's modification, but in spite of that it is still usual to call it the Appleton-Hartree equation.

Nature of the high-level ionisation

Once the magneto-ionic theory had been developed it was possible to consider in more detail how waves reflected from the Heaviside layer would behave. In the early experiments the wave-frequency

was about 0·75MHz whereas the frequencies of the electron and the ion gyrations about the geomagnetic field (the gyro frequencies) were about 1·5MHz and 100Hz respectively. If heavy ions were the operative charges, the gyro-frequency would be so far removed from the wave frequency that the behaviours of the two characteristic waves would not be very different. Both would be received with similar strengths and the polarisation of the composite wave resulting from their addition would depend on the precise manner in which they were combined, so that its direction of rotation would be continually changing.

If, however, electrons were the operative charges, the wave frequency and the gyro-frequency would be similar and the two characteristic waves would behave differently: theory showed that one would be absorbed much more than the other, and that in the northern hemisphere, where the magnetic field is directed downwards, the remaining one would be polarised with its electric field rotating in a left-handed direction. Where the earth's magnetic field is directed upwards instead of downwards, the sense of rotation would be the opposite.

A crucial experiment was now possible. The polarisation of the down-coming wave was measured in England and it was found that the electric field rotated in a left-handed sense as it would if electrons had caused the reflections. This result seemed to show that electrons, and not ions, were responsible for the reflection, but, to make quite certain, the experiment was repeated in Australia, where the earth's magnetic field is directed upwards, and there the electric field was found to rotate in the opposite sense.

It might seem that these experiments made to test an existing theory and to determine which was the sense of rotation in the reflected wave, would be comparatively simple. So they were, but there were several places in the theory where it was possible to make an error of sign. The present author performed these experiments in

conjunction with Appleton and we jointly came to the conclusion that in England the sense of rotation in the wave was *right-handed* and that this was in accord with theory. When I came to correct the proofs of the written account I realised that there was an error in our analysis of the experiment and I wrote to tell Appleton that the measurements showed *left-handed* rotation. My letter crossed one from him to me saying that he had also found an error, but his concerned the magneto-ionic theory: because he had used a set of units called 'rationalised' he had worked with the incorrect gyro-frequency for the electrons, and the correct theory required *left-handed* rotation. So experiment and theory agreed, after all: it was fortunate that the two mistakes, which cancelled each other, had been noticed independently; we would not have felt the same confidence in our results if we had found a compensating mistake only after an earlier one had led us to search for it. About the same time another worker made similar experiments and published his (wrong) conclusion that the sense of rotation was *right-handed*: he later corrected this result. It is not always easy to be quite certain that all the signs in mathematical expressions are correct.

Once it was known that electrons, and not ions, were the operative charged particles, it could be stated with certainty that their concentration at heights of about 100 kilometres was the smaller of the two numbers mentioned previously, that is, 10^4 per cubic centimetre, so that only about one part in 10^8 of the air was ionised. For comparison it is useful to note that of the particles at ground level one in 10^5 is a helium atom, and this fraction is usually thought to be small and insignificant. The proportion of electrons at 100 kilometres height is one thousand times smaller still and yet, because they are charged and can reflect radio waves, it has been possible to make use of them to provide most of the information discussed in this book.

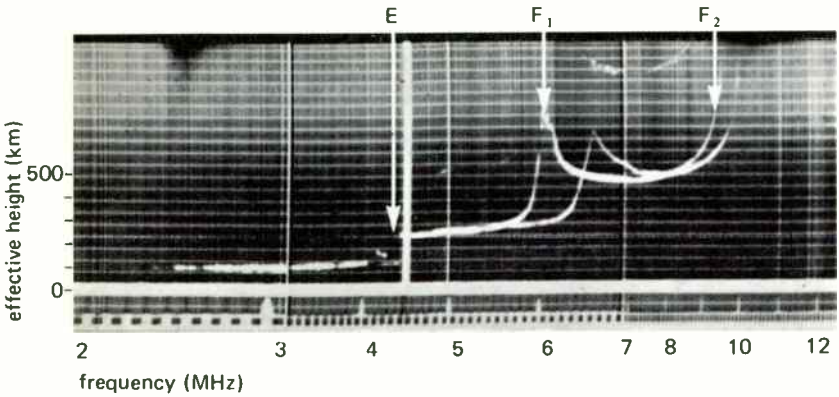
Beyond the Heaviside layer

The early experiments were made with radio waves that penetrated only part way into the layer: the next step was to use waves of greater frequency which would penetrate to greater heights where the electrons were more concentrated. It was expected that, as the frequency was gradually increased, the level of reflection would rise, until it reached the peak of the layer where the electron concentration had its maximum value. Frequencies even greater would nowhere find enough electrons to reflect them vertically: they would escape into space, and the lowest frequency which just penetrated the layer (the *penetration frequency*) would provide a measure of the electron concentration at the peak.

In most scientific investigations the workers have a fairly clear idea of what to expect next. They plan an experiment as an extension of an existing one, usually to measure or observe something inferred from their previous work. In the same way an explorer, having seen a high mountain range in the distance, will make a detailed survey to determine its shape and the height of its peak. But sometimes, in attempting to explain his measurements, the scientist finds something quite unexpected. He can then be said to have made a discovery, rather than to have conducted an investigation. This is just what happened now. The radio-wave frequency was increased until the peak of the layer was reached: waves of greater frequencies penetrated the layer, but they did not go out into space as had been expected; they were reflected by a layer at a greater height. It was as though the explorer, attempting to measure the height of the mountain range, had found another and still higher mountain beyond.

What was this new layer to be called? It had been discovered (and that is the right word) by Appleton and at first it was named the Appleton layer to distinguish it from the lower layer which had been

2·9 An ionogram showing partial split of the F layer into the F1 layer (or ledge) and the F2 layer. The pentration frequencies are marked at E, F1, and F2. The trace is doubled by magneto-ionic splitting, as explained in figure 2·10 on page 75.



named after Heaviside. But what if other layers were discovered later, would it be sensible to go on calling each of them by the name of an investigator? It seemed better to use some more objective nomenclature, and one that could be extended to deal with other layers that might later be discovered. So by general agreement the part of the atmosphere that contains sufficient electrons to affect the travel of radio waves was called the *ionosphere*, and the layers inside it were labelled starting from the bottom by letters of the alphabet in succession. To leave room for some undiscovered layers above and below the known ones, Appleton called the Heaviside layer the E layer, and his new layer the F layer. The height of the E layer was about 100 kilometres, and that of the F layer about 250 kilometres.

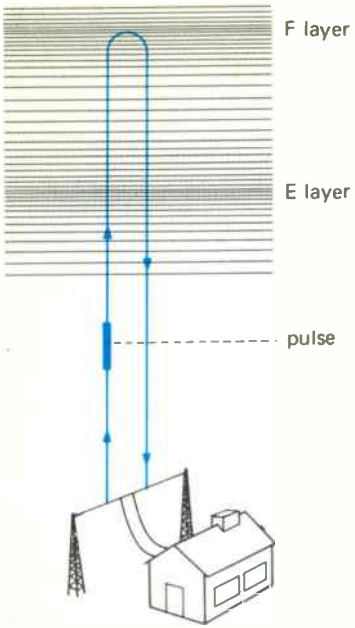
Appleton explained his choice of the names in the following way. In his early theory the electric field of the wave reflected from the lower layer was denoted by the letter E and later, when he discovered the upper layer, the corresponding field was denoted by F. When he came to name the layers he suggested that they be called by the same letters E and F so as to leave room for labelling other layers, above and below, by neighbouring letters of the alphabet.

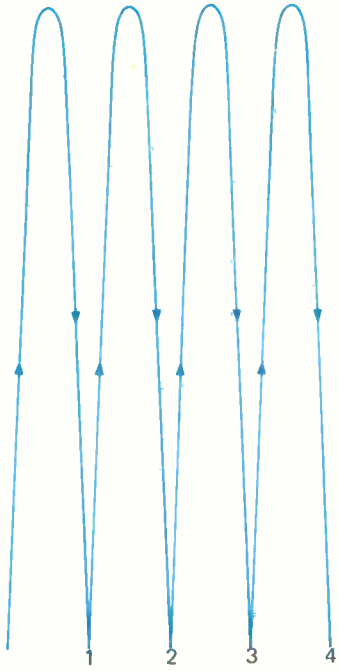
It was convenient to give a special name to the part of the ionosphere immediately below the E layer, since it was there, where collisions between electrons and heavy particles were comparatively frequent, that radio waves were mainly absorbed. It was natural to assign the letter D to that part of the ionosphere, but since there was no evidence that the electrons in it were distributed in a layer, it was called the D region.

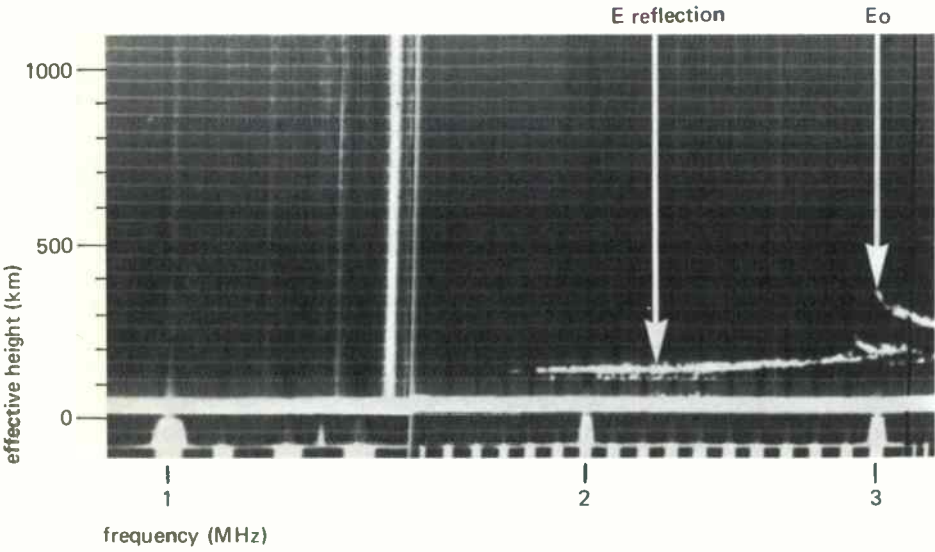
Sounding the ionosphere

After the E layer had been penetrated and the F layer discovered, the next thing to do was to measure the electron concentrations at their peaks. This was simply a matter of finding the frequencies of the waves that would just penetrate them, and inserting these into the appropriate mathematical expressions to derive the peak concentrations. Of all the useful quantities describing radio waves reflected from the ionosphere, these penetration frequencies are the easiest to measure, and they began to be investigated in different parts of the world, especially in England, America and Australia. The apparatus commonly used for sounding the ionosphere is of the form originally used by Breit and Tuve in America: it is called an *ionosonde* and produces a photographic record, called an ionogram, in which the time delay of the echo is automatically plotted against the frequency of the emitted wave. From records of this kind it is easy to read off the penetration frequencies appropriate to the two layers (see figure 2·10).

Sometimes, during the day, and particularly in summer, there appear to be signs of another penetration frequency, intermediate between the clearly marked ones that correspond to the E and F layer peaks. On some ionograms (see figure 2·9) this extra frequency is well marked, on others it is scarcely noticeable: it changes in an orderly way throughout the day and the season, as though it

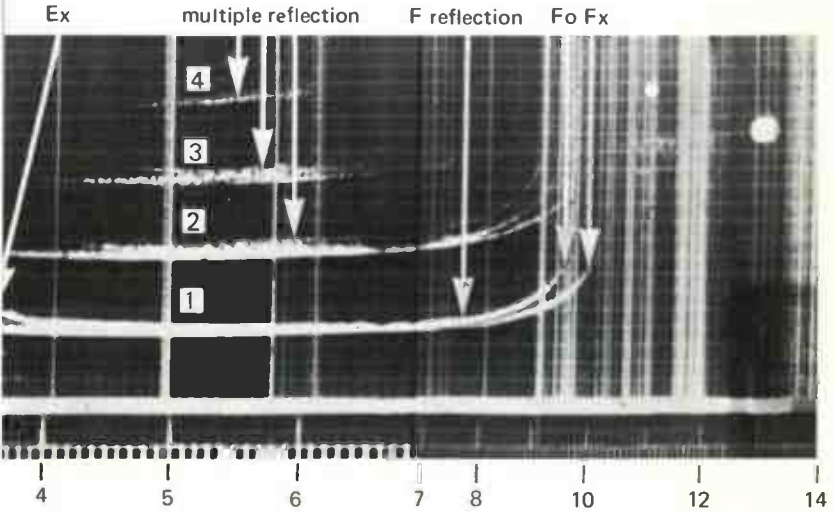


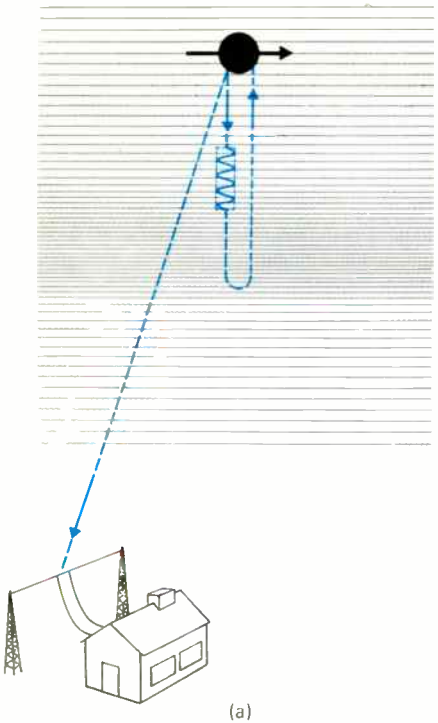




2·10 Sounding the ionosphere. With an *ionosonde*, a short pulse of radio waves is transmitted vertically and received at the same place after reflection from the ionosphere. The radio-wave frequency is changed steadily over a few minutes and the travel time of the pulse is photographically recorded as an ionogram. In an *ionogram* (*below*) the horizontal scale shows the radio frequency of the emitted pulse. The vertical scale shows the effective height at which a sharply reflecting layer would have to be if it were to reflect the pulse with the observed time of travel. Reflection is from the E layer at an effective height of 120 or 150 km for frequencies less than 3 MHz; above that frequency reflection is from the F layer at an effective height of 250 km or more. Between about 5 and 6 MHz, four traces can be seen, corresponding to pulses that have been repeatedly reflected between the F layer and the ground.

The greatest frequency to be reflected from the F layer is the F penetration frequency; the least frequency to be reflected from it is the E penetration frequency. The earth's magnetic field splits the original wave into two *characteristic waves*, so that the penetration frequencies are doubled: they are marked at F_o and F_x, and E_o and E_x on the ionogram. The magneto-ionic theory indicates how these multiple frequencies must be used in calculations.



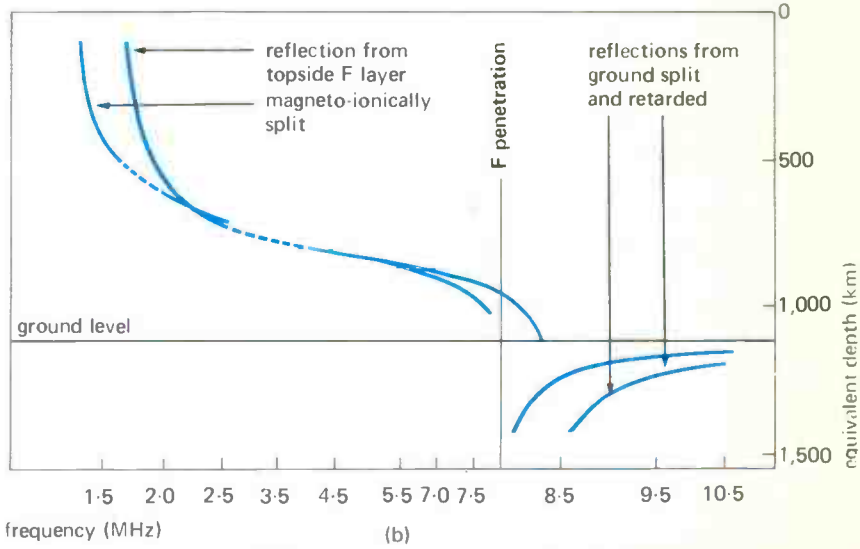


(a)

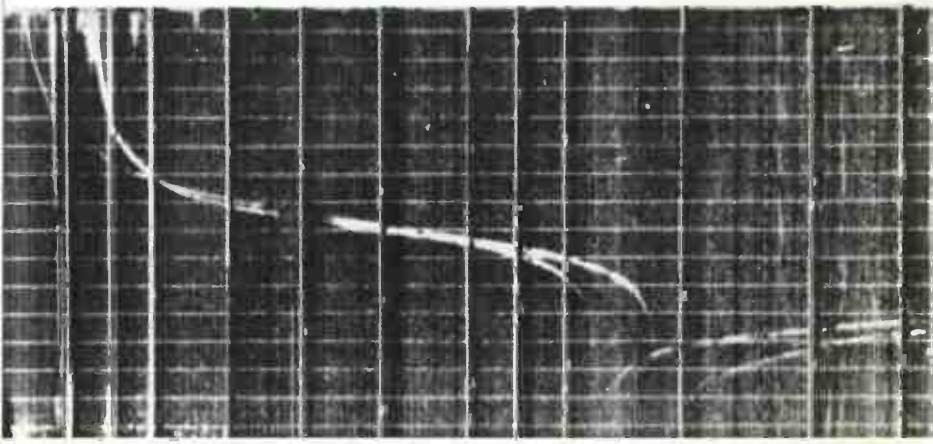
2·11 'Topside' sounding.
(a) An ionosonde in a satellite sends short pulses of radio waves down on to the F layer, from which they are reflected. The time of travel there and back is measured as the radio wave frequency is gradually altered. The results are sent to a ground receiving station by radio telemetry and are there displayed as a topside ionogram.
(b) The explanation of the topside ionogram at (c).

has some fundamental significance. It thus seems that by day the F layer is sometimes split, or partially split, into two sub-layers, which have been named the F1 (below) and the F2 (above), and that the two 'penetration' frequencies can provide a measure of the electron concentrations at their peaks. Since the F1 penetration is not sharp, like that of the F2, it is usually better to speak of the F1 'ledge' of ionisation rather than of the 'peak'.

Ionosondes on the ground can provide information only about the electrons below the peak of the F layer, but if one could be put into a satellite circling the earth at a height of about 1,000 kilometres, so as to send its waves down towards the peak, and if the time taken for these waves to be returned to the satellite after reflection from the upper part of the layer could be measured, it would be possible to investigate the electron distribution above the peak, by a sort of



(b)



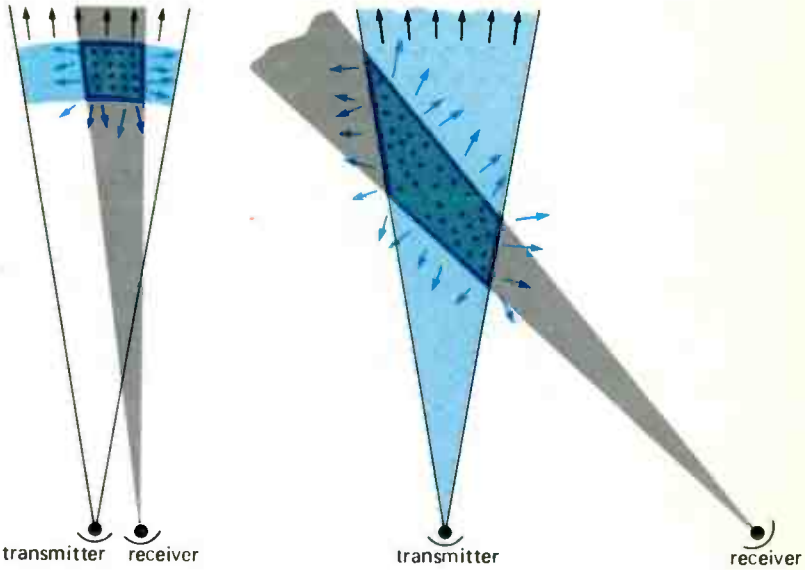
(c)

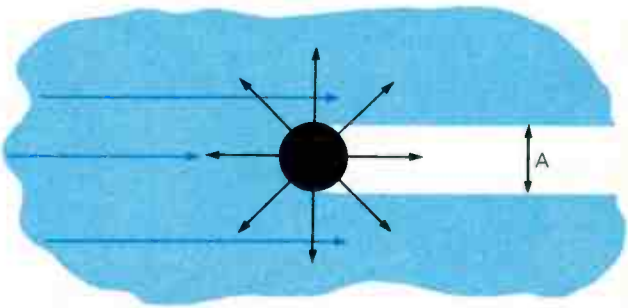
upside down sounding. This is precisely what has been done, first in 1962, in one of the most elegant of the many satellite experiments; it has (most inelegantly) been called a *topside sounder*. The topside ionogram is transmitted by radio to the ground and is recorded as a photographic trace. By combining the information available from ground-based and topside ionosondes it is possible to study the ionosphere right up to the height of the satellite.

But the topside sounder does much more than merely extend upwards what was known from ground based sounders, for it has the advantage of travelling repeatedly right round the earth with a period of about 90 minutes, so that it provides information about the electron distribution in the upper part of the F region over a wide area.

The region above the peak of the F layer has also been explored, since 1958, by means of a technique called *Thomson scatter sounding* which makes use of apparatus on the ground. Although the principles of the method had been in the minds of workers for many years its application had to await the availability of powerful transmitters working in conjunction with large aerials and sensitive receivers. It works as follows. A strong beam of radio waves is sent vertically upwards on a frequency much greater than the penetration frequency of the F layer, so that the waves escape from the earth without being reflected. In their passage through the ionosphere they set the electrons into oscillation so that each reradiates a weak wavelet of its own. These spread radially and when they are received in most directions they have randomly distributed phases since they originate in electrons situated in random positions (see figure 2·3). It is only in the forward direction that they add coherently, in the way explained on page 54, to alter the velocity of the resultant wave. In the Thomson scatter technique, however, use is made of the wavelets that are returned *backwards* towards the ground; here the wavelets add with random phases ('incoherently') to produce only a

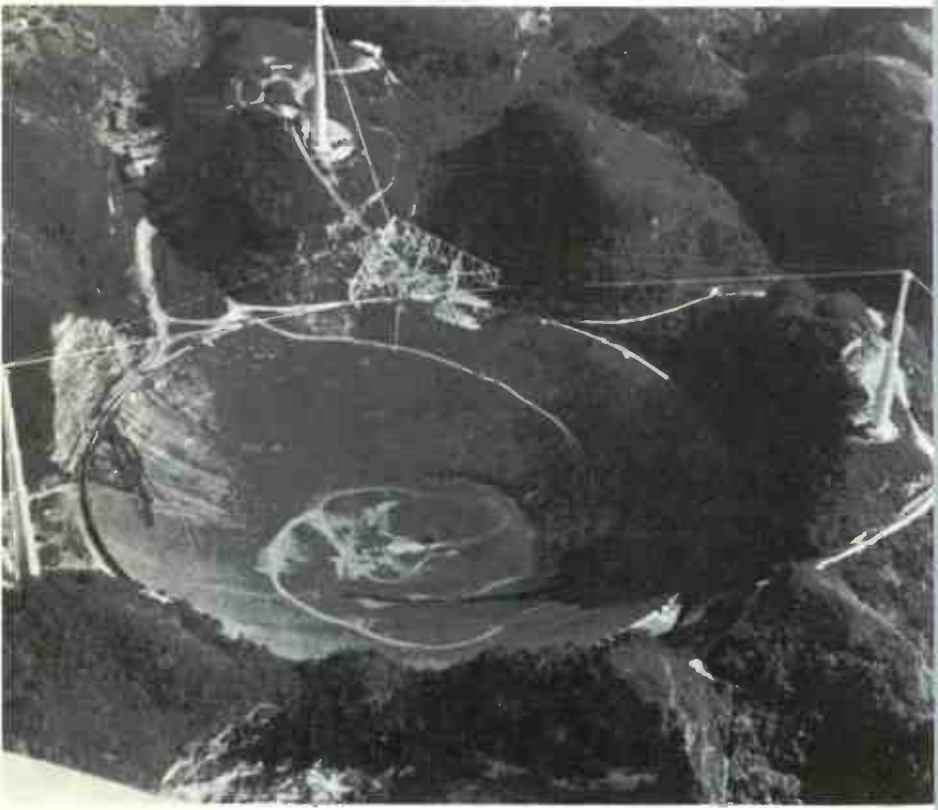
2·12 Thomson scatter sounding. In method (a) a short pulse of radio waves is emitted and the power incoherently scattered backwards from the electrons in the indicated volume is measured at the receiver. In method (b) radio waves are emitted continuously and the power scattered sideways from the electrons in the indicated volume is measured. (c) shows how an electron, on which a radio wave impinges, intercepts and reradiates an amount of power depending on its cross sectional area A.





(c)

2·13 The large aerial used for Thomson scatter sounding in Arecibo, Puerto Rico. It consists of a paraboloidal dish of diameter 300 metres, excavated from the ground together with transmitting and receiving elements supported at its focus by wires attached to three poles.



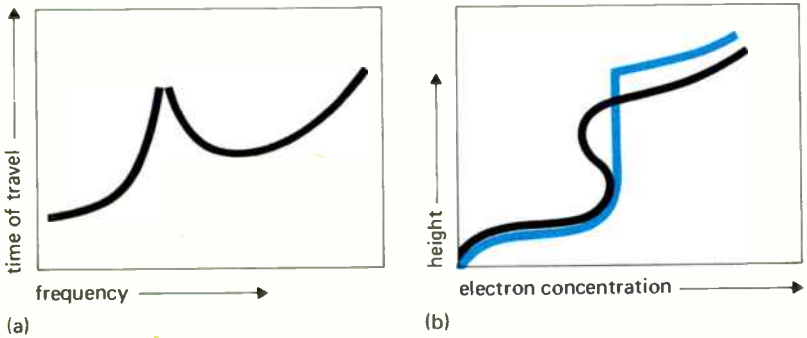
very small power, which nevertheless is sufficient to be collected by a large aerial and measured by a sensitive receiver.

The power returned to the ground can be calculated in the following way. The energy reradiated by each electron is equal to what falls upon its cross-sectional area as the main wave advances. This area was calculated by J. J. Thomson soon after his discovery of the electron. It is extremely small, only about 10^{-25} square centimetres, and a single electron would scatter a power much too small to measure. If, however, it is arranged that all the power from a reasonably sized volume of the ionosphere is collected in the receiver, it can be measured with modern techniques. The measured power is minute; thus, the number of electrons contained in a cube with side 100 kilometres centred at a height of 700 kilometres is about 10^{25} so that when all their cross-sections are added together they amount to only one square centimetre. It is a remarkable fact that the far-reaching results which have been obtained by the method of scatter sounding depend on measurement and analysis of the radio-wave power scattered by these small areas, with size no greater than a postage stamp, at great heights.

By comparing the power scattered from a known volume with the calculated power scattered from one electron it is possible to 'count the electrons' in the volume. In Thomson scatter sounding the operative volume in the ionosphere is determined in one of two ways. If the receiver is located near the transmitter the volume is defined in vertical extent by the length of short pulses of waves emitted periodically, and sideways by the extent of either the transmitting or the receiving radio beam, whichever is the smaller. If, however, the receiver is at some distance from the transmitter it is arranged that the waves are emitted continuously, and the scattering volume is then defined by the volume over which the transmitting and receiving beams overlap.

Results obtained by the methods of the topside sounder and the

2·14 The experimental ionogram (a) would have been recorded if the distribution of electrons in the ionosphere had taken either of the forms shown in (b).



Thomson scatter sounder are complementary: topside sounding provides information about the electron distribution at different times and different places, as the satellite travels over the earth, whereas Thomson scatter sounding provides information continuously at one place.

Measurements of electron concentration have also been made with the help of a radio transmitter taken aloft in a rocket. The Doppler shift in the frequency of the wave emitted from the moving transmitter is measured at the ground. The magnitude of this shift is determined by the velocity of the waves as they leave the transmitter and because this velocity depends on the electron concentration in the surrounding ionosphere it is possible to calculate the concentration at the successive levels reached by the rocket.

Since rocket experiments can be made only occasionally, and at a few places, the method is not so fruitful as those using ionosondes or Thomson scatter, but it has provided valuable additional information, particularly about the region between the E and F layers.

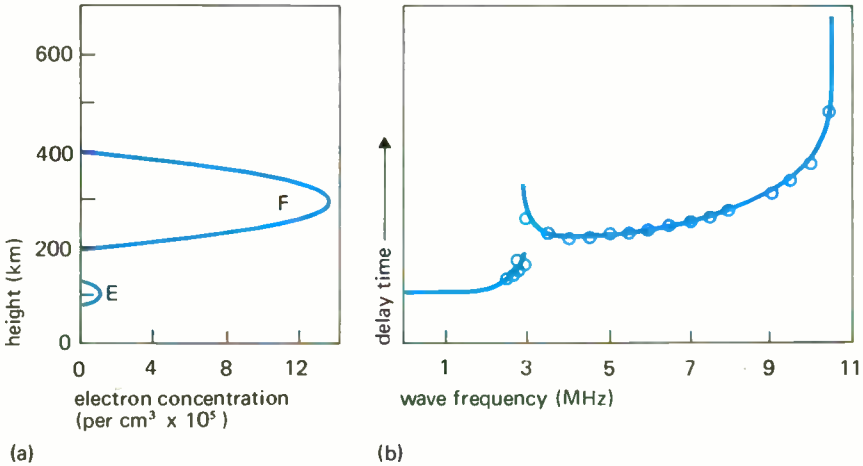
The shapes of the layers

In the early days of ionospheric sounding most attention was paid to the penetration frequency and not much to the shape of the ionogram. But this shape, which relates the travel time of the pulse to the radio-wave frequency, must be related somehow to the height distribution of the electrons in the ionosphere, and in retrospect it seems a little strange that the relationship was not explored more energetically by the early workers: indeed it was not until about 1950 that it was fully investigated. During the 1920s and 1930s, however, some limited use was made, in the following way, of the information contained in the shapes of the ionograms.

The speed of a short pulse of radio waves travelling in the ionosphere depends in a calculable way on the concentration of the electrons: if the concentration is assumed to be known at all heights it is therefore possible to compute the time of travel to the reflection point and back, and if this computation is repeated for several different frequencies the result can be compared with an experimental ionogram. In principle, therefore, it should be possible, by a process of trial and error, to find the electron distribution that provides the best calculated fit to an observed ionogram.

Unfortunately there is a serious difficulty when this procedure is used with ionograms, of the type shown in (a) in figure 2·14, that indicate the presence of two layers. Either of the distributions shown in (b) would correspond to the ionogram shown in (a); each could represent the actual ionosphere, and it is not possible to find a unique solution to the problem. There is always an ambiguity of this kind when two ionospheric layers are superimposed in such a way that the electron concentration decreases to a minimum somewhere between them: it is sometimes referred to as 'the valley ambiguity'. In spite of this difficulty, however, some useful attempts were made in the 1930s to deduce the electron distribution from the ionogram

2·15 Early estimates of electron distributions in the E and F layers. The double-parabolic electron distribution in (a) would give rise to the ionogram (b), which accords reasonably well with the experimental points plotted.



by the method of trial and error.

In days when rapid electronic computers were not available it would have taken too long to postulate several electronic distributions, compute the corresponding ionograms, compare them with the observed ones, and pick out the one that fitted best. Instead, the form of the distribution was described by a mathematical expression and the corresponding ionogram was calculated, not by a process of numerical computation, but by mathematical analysis. For this purpose it was, of course, necessary that the chosen mathematical expression should be amenable to this kind of analysis and it happened that an electron distribution described in terms of two superimposed parabolae was suitable. With this distribution it was

necessary to make the calculation only once; the final result could then be 'scaled' by adjusting six separate quantities corresponding to the heights, the thicknesses, and the peak concentrations of the two layers. These quantities could then be changed to make the calculated result fit the experimental ionogram as closely as possible. Figure 2·15 shows one of the first comparisons of this kind: the difference between the thicknesses of the two layers immediately suggested important consequences, which are discussed on page 121.

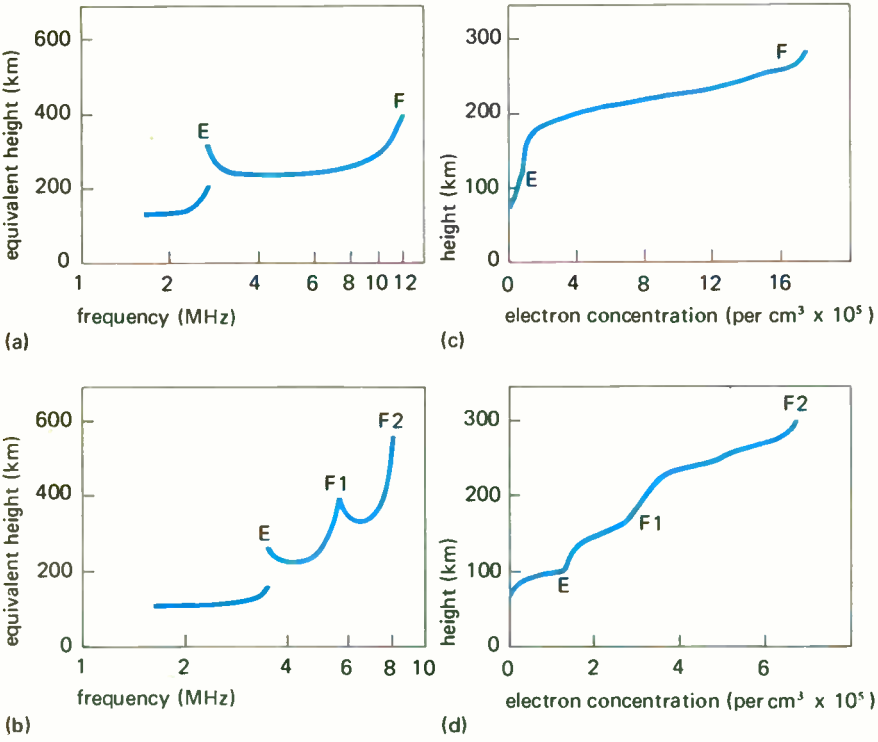
The above-mentioned method of determining the electron distribution is indirect: investigators did not try to deduce the distribution from the ionogram; they proceeded in the opposite direction, and having assumed the distribution of electrons they deduced the ionogram. Later, in the 1950s, when rapid computations could be made with digital computers, the distribution was deduced by a direct process from the ionogram. Even then, however, it was not possible to remove the ambiguity resulting from the possible existence of a 'valley' between the E and F layers. For want of a better assumption it was usually supposed that there was no 'valley' of this kind, and that the electron concentration increased steadily upwards. Figure 2.16 shows the sort of electron distributions that were then computed from experimental ionograms.

The distribution of electrons on the upper side of the F region can similarly be deduced, without any corresponding ambiguity, from the topside ionograms: an example is given in figure 4·2 (page 120).

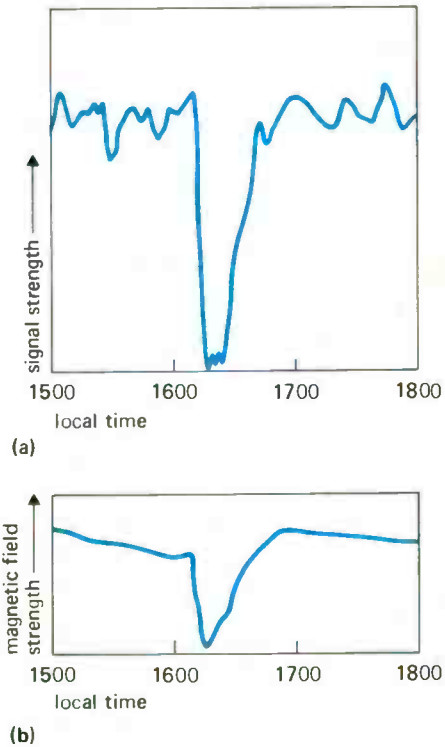
Radio-wave absorption in the D region

In their travel through the ionosphere, radio waves suffer an amount of absorption that depends on the total number of collisions made in unit time by all the electrons in a unit volume, and hence on the product of the electron concentration and the collision frequency.

2·16 The electron distributions at (c) and (d) are calculated respectively from the experimental ionograms at (a) and (b). The relationship between the ionogram (a) and the distribution (c), which correspond to an 'unsplit' F layer, should be compared with the corresponding relationship in figure 2·15 calculated by a different, and less reliable, method.
(b) and (d) correspond to a layer split into an F1 'ledge' and an F2 'peak'.



2·17 A Sudden Ionosphere Disturbance (SID) on 26 November 1936.
(a) shows the strength of the signal received over a distance of 600 km on a frequency 9·6 MHz. During the time between 1615 and 1625 hours the signal was too small to be detected.
(b) shows the strength of the horizontal part of the earth's magnetic field. The sudden decrease in strength is very small, and is only about 1/5,000 of the total.



Although the electron concentration *decreases* downwards from the E layer, the collision frequency (depending on the concentration of neutral air particles) *increases* more rapidly so that absorption is greatest at heights around 80 or 90 kilometres, in the D region.

Observations of the strength of waves reflected from the E and F layers show that this absorption is greater by day than by night, and that it reaches a maximum near midday, as would be expected if the D region ionisation were produced by radiation coming from the sun. The strength of reflected waves is, however, somewhat variable from day to day, and the variation is particularly marked in winter; on some winter days there is unexpectedly strong absorption. We return to these matters on p. 157.

Near the sunspot maximum of 1937 a particularly interesting anomalous feature of radio wave absorption was noticed. It took the form of a very sudden increase of absorption followed by a slower return to normal, the whole change usually lasting between three-quarters and one hour, and it occurred only during daytime. It was given the name *sudden ionosphere disturbance* usually abbreviated to SID. It was not accompanied by any noticeable increase in the penetration frequencies of the E and F layers.*

The anomalous absorption associated with an SID was particularly noticeable in long distance commercial radio communications, and could be so great that received signals would disappear completely and operators would sometimes be misled into thinking their receivers had broken down. It was established in the late 1930s that these SIDs, or fadeouts, occurred simultaneously with the appearance of solar flares. It was also remembered that small irregularities in geomagnetism had previously been found to accompany the observation of some solar flares and the suggestion was therefore made that increases of absorption, and geomagnetic

* More detailed measurements have recently indicated small changes in the penetration frequencies at these times, but the facts are not yet clearly established.

disturbances, were the result of increased ionisation in the D layer produced by intense bursts of ultra-violet or x-radiation, emitted at the time of the visible flare. A burst of radiation of this kind would ionise only on the daylight side of the ionosphere and it was only there that the SIDs were observed.

3 Production and loss of electrons in the ionosphere

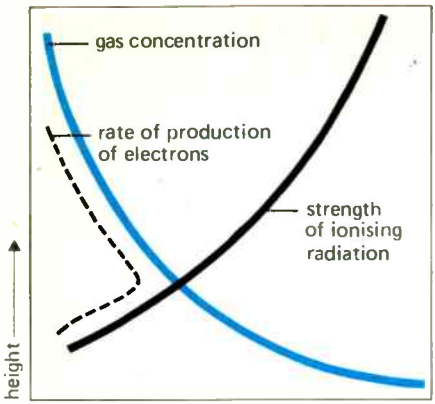
3 Production and loss of electrons in the ionosphere

In this chapter we show how, from observations of the kind mentioned in chapter 2, it has been possible to explain much of the detailed structure of the ionosphere. Until about 1951 the only available information came from ground based radio sounders and the history of the subject starts with the indirect evidence which they provide.

Production of electrons, and Chapman layers

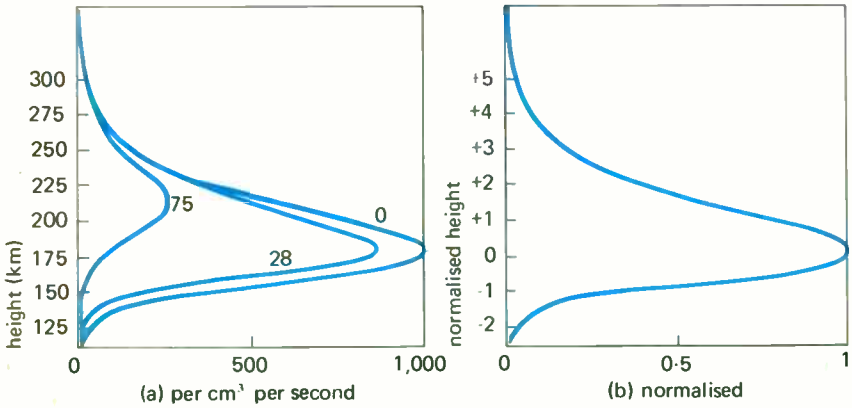
Once the penetration frequencies could be measured it was possible to find how the concentrations of the electrons at the peaks of the layers altered from time to time. To explain these observations it was desirable to have some theory about how the layers were formed. Only the vaguest suggestions had been made previously, mainly because little was known about the composition of the upper atmosphere, or the solar radiation which might ionise it, or what would happen to the ionisation once it was formed. In spite of this lack of detailed knowledge, the first step, as always in the scientific investigation of a complicated problem, was to consider a simpler situation which could be studied theoretically. For this purpose it was first supposed that the concentration of any gas that was to be ionised decreased with height exponentially as described by a scale height H with a magnitude that was left to be determined from the observations. It was next supposed that an ionising radiation fell upon this atmosphere from above, in a direction coming from the sun, so that electrons and positive ions were formed. The electrons were supposed to remain free for some time after they had been produced, but ultimately to recombine with the positive ions to form neutral atoms or molecules.

Under these circumstances it was easy, in the following way, to see that a layer of electrons would be formed. At the top of the atmosphere none would be produced because there would be no



3·1 Left The production of a layer of electrons when ionising radiation falls from above on a gas with an exponential height distribution. Electrons are produced most rapidly at a level where the downward increase of gas concentration is matched by the downward decrease in the strength of the radiation.

3·2 Below A Chapman layer.
(a) The rate of production of electrons calculated for a hypothetical layer in which the scale height is 25 km and the peak rate of production is 1,000 per cm^3 per second at a height of 175 km when the sun's rays arrive vertically. The three curves correspond to solar zenith angles (measured from the vertical) of 0° appropriate to midday at the equator at equinox, 28° (midday London in midsummer) and 75° (midday London in midwinter).
(b) If curves like those at (a) are 'normalised' by measuring heights upwards and downwards from the peak in terms of the scale height as a unit, and rates of production in terms of the peak rate for vertical solar rays, they all take the single form shown at (b).



gas to be ionised: as the ionising radiation penetrated towards the earth it would meet more and more air and would produce more and more electrons. The production of electrons would extract energy from the radiation, which would thereby be weakened, and at a certain height the rate at which the gas concentration *increased* downwards would be matched by the rate at which the strength of the radiation *decreased* downwards. At this level, where the two just balanced, the rate of production of electrons would be greatest; lower down, although the gas concentration would continue to increase, the strength of the radiation would decrease more rapidly, so that the resultant rate of production of electrons would become less. The level at which electrons are produced most rapidly is called the 'peak' of production.

In 1931 Chapman investigated these matters in detail and showed that the rate of production of ionisation would vary with height as shown in the curves of figure 3·2. The corresponding layers of electrons have been called Chapman layers. His paper emphasised the following important points about the curves of figure 3·2.

1 The height of the peak is determined by the concentration and scale height of the atmospheric gas and by the absorbability of the sun's radiation. Less easily absorbed radiation penetrates lower before it forms a layer of electrons. The height does *not* depend on the strength of the radiation.

2 The rate of production of electrons at the peak depends on the strength of the radiation and on its direction of arrival. It is greatest when the radiation arrives vertically and less when it arrives obliquely. If χ measures the angle between the direction of the radiation and the vertical (an angle which the solar astronomers call the Zenith distance of the sun) then the rate of production of electrons at the peak is proportion to $\cos \chi$.

3 If the curves representing layers of all possible shapes are suitably scaled they all look the same. The scaling is done by placing the level

of the peak at the zero on the height scale (so that distances measured upwards are positive and downwards are negative) and measuring all heights in terms of the scale height as a unit; and by calling the peak rate of electron-production unity and measuring all other production rates as fractions of it. The normalised curve is shown in figure 3·2(b); all the other curves shown can be reduced to this form by scaling them as described.

Theoretical results that can be scaled in this way are particularly useful to experimenters, for then they need carry in their minds only one single normalised curve. Several results, deduced from a variety of initial assumptions, can all be converted to the standard one simply by suitably altering the scale and the zero of height. The rates of production of electrons in gases that absorb differently, have different scale heights, and are illuminated by radiations of different strengths, arriving in different directions, can all be studied as one, and, if the assumptions of the theory are fulfilled, they will all be represented by figure 3·2(b) if the scales are suitably adjusted.

The previous discussion is concerned with the rate at which electrons are produced by the incoming radiation, and if there were nothing to counterbalance this production all the available gas would ultimately be ionised. There are, however, processes that result in loss of electrons from the ionosphere, at a speed that is greater when the concentration is greater. If the rate of production remains constant the electron concentration thus builds up until the rate of loss is equal to the rate of production, and after that it does not change. Of course the inclination of the sun's rays actually varies, and with it the rate of production, and at most times the electron concentration is changing; it cannot therefore be said that production and loss are precisely balanced. It happens, however, that it is often sufficiently exact to suppose that the rates of production and loss are equal.* For this reason it has proved useful to discuss the type of electron distribution that would result if the two rates

balanced: a layer of that kind has been called an 'equilibrium layer'.

The rate of loss is determined by a series of complicated reactions between the electrons and the particles of the surrounding air. Because a large part of ionospheric research has been directed to understanding these reactions in detail, it is interesting to enquire why it has proved to be so important.

A low-pressure laboratory without walls

The reactions that can occur between electrons, ions and neutral particles in a gas are similar to, but more complicated than, those that can occur between uncharged particles. Attempts to investigate them in the laboratory have usually been made by first ionising a gas with a flash of ultra-violet light, and then investigating how the concentrations of the resulting charged particles decay when the light is removed. Investigations of this kind are called photochemical. They are difficult to carry out because the charged particles collide frequently with other particles and do not normally last long enough to be investigated. If attempts are made to experiment with a gas at a reduced pressure, so that the collisions are less frequent, a stage is soon reached when the particles last for a time that is limited by their collisions with the walls of the experimental vessel. The ideal arrangement would be to use an ionised gas at a pressure so low that the electrons and ions would last for long enough to be studied, and in a volume so great that each would make several collisions with other particles before it reached the walls. That is just what is available in the ionosphere; thus in the E region the pressure is like that in a good laboratory vacuum, and the electrons travel, on the average, about 100 metres before encountering an air particle.

* This approximation can be made whenever the rate at which the electron concentration changes is much less than the rate at which the electrons are being produced or lost.

If satisfactory measurements were to be made at this pressure in the laboratory an apparatus would be required with dimensions about ten times as large, so that the distance between its walls would be about one kilometre. In the ionosphere, however, there are no walls, and the 'experimental volume' is limited only by the volume within which the concentration of the particles remains effectively unchanged – roughly a cube with a side equal to one scale height. The ionosphere thus provides a 'low-pressure laboratory without walls' and it offers considerable possibilities for the theorist to try out his ideas about what happens to electrons and ions when the pressure is so low that collisions with other particles are important but comparatively rare.

The loss of electrons: experimental results

A free negatively charged electron comes into existence in the ionosphere when the ionising radiation removes it from a neutral particle and leaves that particle as a positively charged ion; it might then disappear in one of two ways (see figure 2·7, p. 62). The first is the direct opposite of the ionisation process: it is one in which an electron and a positive ion collide so that their charges neutralise each other and they recombine to form a single neutral particle. We call this a process of *recombination*. The rate at which electrons are lost in this way depends on the number of encounters between electrons and ions and hence on the product of their two concentrations. If this were the only mechanism removing electrons then, in view of the way in which they were originally produced, the number of electrons would be equal to the number of ions, and the product would be equal to the square of the number of electrons. The important conclusion follows that the *rate of loss of electrons by recombination is proportional to the square of the electron concentration*.

The second process by which electrons can be removed is one in which an electron attaches itself to a neutral heavy particle to produce a heavy negative ion: it is given the name *attachment*. The rate of loss by this process is determined by the frequency with which electrons and neutral particles encounter each other, and this depends on the product of the electron concentration and the (constant) neutral particle concentration. The important conclusion follows that *the rate of loss by attachment is proportional to the first power of the electron concentration* (and not the second power, as for recombination).

One of the first uses of Chapman's theory was to find out whether recombination or attachment is operative in removing electrons from the ionosphere. In an 'equilibrium' situation, in which the rate of production is balanced by the rate of loss, recombination would produce a layer in which the square of the electron concentration was proportional to the rate of production, whereas attachment would produce one in which the concentration was directly proportional to the rate of production. Chapman showed that the rate of production at the peak of a layer is proportional to $\cos \chi$ (χ being the solar zenith angle) so that the electron concentration at the peak should be proportional to $\sqrt{\cos \chi}$ for recombination and to $\cos \chi$ for attachment. Here then was an opportunity to decide which process is operative merely by observing how the peak concentration of electrons varies with the sun's zenith angle.

To make observations at times when the angles are different it is possible to proceed in several different ways. Observations might be made at one place throughout the day; the angle would then be horizontal at sunrise and steepest at noon, and the range would, of course, be greatest in summer when the sun would be highest in the sky at noon. Or observations might be made at noon throughout the year, allowing the sun's elevation to change with the season. Or again observations might be made at many places on the earth at

one and the same universal time, or at local noon at each place, so that the change in the sun's elevation would be the result of geographical separation. All these possibilities were tried, and it was found that, at the peak of the E layer and at the F1 ledge, the electron concentration is roughly proportional to $\sqrt{\cos \chi}$, and not to $\cos \chi$. It was therefore concluded that the electrons are lost by a process of recombination, and not by one of attachment.

Once it had been decided that the electrons in the E and F1 layers are lost by recombination with positive ions, attempts were made to measure the rate of their disappearance. If the ionising radiation could have been suddenly cut off it would have been possible to determine this rate simply by observing the resulting decrease of the penetration frequency. Observations were therefore made at night, when the radiation no longer reached the layer, but they were found to be not very satisfactory because the electron concentration had already decreased a long way before the sun had set, and the remaining decrease during the night was hardly sufficient for any accurate value to be deduced. Observations were also made during eclipses, when the solar radiation is more suddenly cut off, and results like those of figure 5·2 were obtained; the results are, however, difficult to interpret, largely because it is not known whether all the ionising radiation is removed when the visible light is fully eclipsed. Other attempts to measure the rate of loss used the fact that during the day the electron concentration at the peak of a layer does not increase and decrease precisely in step with the intensity of the ionising radiation, but lags behind it, so that the concentration reaches its maximum somewhat after midday, when the strength of the radiation is already declining. The time-lag is, however, comparatively small and is not easy to measure. In spite of much work, in which all these methods were used, it has proved difficult to deduce a reliable value for the 'recombination coefficient' which describes the rate of loss, and only an order of magnitude

can be stated: it is about 10^{-8} units for the E layer and somewhat less for the F1 ledge.* For the F2 layer it has not been possible to suggest any reasonably consistent value.

The experimental results described above led to the belief that the electrons in the E layer and the F1 ledge are formed, in the way envisaged by Chapman's theory, by solar radiations ionising gases that are exponentially distributed in height and that are lost by recombination with positive ions. The presence of two layers (or a layer and a ledge) indicated that two different kinds of radiation are present, perhaps ionising two different atmospheric constituents. As experimental results accumulated it appeared that the expectations of Chapman's theory are not precisely realised, but it was thought that the comparatively small discrepancies could be explained if the temperature is not the same at all heights (so that the gases are not distributed quite exponentially) or if there are electromagnetic forces that move the electrons up and down (see p. 171). It is still considered that the behaviour of the electrons in the E layer can be explained in this way.

The F layer was (and still is) much more difficult to understand. Sometimes it is a single layer and sometimes it is split so that the F1 ledge appears well developed on its lower side. At night it is always single: by day the F1 ledge is most frequently observable in summer and towards the maximum of the solar cycle. When the ledge is observed its penetration frequency behaves as expected for a Chapman recombination layer. The F2 penetration frequency, however, is found not to vary with the inclination of the sun's rays according to theory and it was not possible to decide whether the electrons were lost by recombination (square law) or attachment (linear law).

* The recombination coefficient α is defined by the statement that the number of electrons lost in each second from each cubic centimetre is equal to αn^2 where n is the number of electrons in each cubic centimetre: with this definition α is measured in units $cm^3\ sec^{-1}$. In this book these will be referred to simply as *units*.

The penetration frequency does not decrease after sunset in any clearly defined way; indeed it often increases at times when the layer is no longer exposed to the sun's radiation. During eclipses the ionisation behaves in different ways: sometimes it decreases or remains unchanged, and at other times it even increases. The situation is confused and it was not possible to make any estimate of the rate of loss of the electrons.

Although it was not clear whether the electrons were lost by attachment or recombination, it was usually supposed, without much evidence, that the F2 layer was formed in the way described by Chapman's theory, but that it was subjected to several disturbing influences which masked any simple behaviour. These ideas are no longer accepted and, as we shall see on p. 112, it is believed that the processes involved in the formation of the F2 layer differ from those involved in the formation of the E and F1 layers: the detailed behaviour of the F2 layer, however, still presents unsolved problems, some of which are discussed on page 162 *et seq*.

Ionospheric photochemistry

Once the magnitudes of the recombination coefficients had been estimated, it was possible to see how they compared with the results of laboratory experiments, and of theory. Recombination coefficients had been measured in the laboratory, but they referred to recombination between positive and negative ions of comparable masses, whereas in the ionosphere recombination is between positive ions and negative electrons with very dissimilar masses. The reason for this difference is that in the laboratory discharge tube the pressure is considerably greater than in the ionosphere, so that the free electrons are very rapidly converted into negative ions by attachment before they can recombine with the positive ions: the only recombination that can then be investigated is between the two

kinds of ions. Although the *ion-ion* recombination coefficients measured in the laboratory are of the same order (10^{-7} units) as that estimated for the *electron-ion* recombination in the ionosphere, it was realised that they correspond to a situation that is quite different, and no useful conclusion could be based on their similarity. It was therefore necessary to compare the ionospheric value with those deduced, not from laboratory results, but from theory, and all theory indicated that the coefficient for electron-ion recombination should be much smaller, of the order 10^{-12} units. With this value the E layer would decay only a little throughout the night, and an eclipse would be too short for any change to be noticeable. It was clear that something was wrong, either with the theory or the experiments, and much effort was devoted to finding the error.

In the late 1930s it began to be thought that the discrepancy might be explained in a rather complicated way if the electrons first became attached to neutral particles to form negative ions that afterwards combined with positive ions to form neutral particles. This explanation required that in the E region the number of negative ions should be comparable with the number of electrons. At that time there was no feasible method of measuring the negative ion concentration, but all theoretical reasoning indicated that it must be very small, much too small to account for the large measured recombination coefficient.

Up to 1939, therefore, the state of affairs was that the electron-ion recombination coefficient for the E layer and F1 peak appeared to be about 10^{-8} units, whereas theory suggested it should be about 10^{-12}. The electrons might however disappear by a different process, at a rate that might be compatible with the measured rate, if, in addition to electrons, there were sufficient negative ions in the layers. Although this explanation required a most unlikely concentration of negative ions, there seemed, at that time, to be no other way of explaining the results.

It is useful at this stage to discuss, in more detail, why theory shows that the recombination between electrons and ions is much too slow to account for the observed decays of the E and F1 layers. At first sight it might be expected that if an electron (e^-) and a positive ion (X^+) come near enough to each other they would coalesce to produce a neutral particle (X) as represented by the expression

$$e^- + X^+ \rightarrow X \qquad \textbf{(A)}$$

but, in fact, this expectation is not fulfilled, for the following reason. When particles, such as electrons and ions, collide it is always essential that two important quantities, the *energy* and the *momentum*, should remain unchanged. The requirement that the energy does not change applies to all natural processes, and is the 'law of conservation of energy'; the need for 'conservation of momentum' arises because the mutual forces between the particles are equal and opposite. It turns out that, in a reaction like that labelled **A** which starts with two particles and ends with only one, the two conservation conditions cannot both be satisfied; either the resulting single particle (X) must carry away energy (and that is impossible in the situation envisaged here) or a second particle must exist after the reaction has taken place. In the recombination of an electron with a positive ion an emitted photon (ph) provides the second particle and the reaction can be represented by

$$e^- + X^+ \rightarrow X + ph \qquad \textbf{(B)}$$

Although a reaction of this type permits both conservation conditions to be satisfied simultaneously, the likelihood of it occurring depends on the likelihood of a photon being produced, and that depends on the precise manner in which the electron and ion approach each other. Theory indicates that a suitable approach will occur only seldom, so that recombination by this process is unlikely. When it does occur, accompanied by the radiation of a photon, it is called 'radiative recombination': its speed is represented by a

'radiative recombination coefficient' of the order 10^{-12} units, much too small to account for the decay of the E and F1 layers at night.

Since radiative recombination appeared to be too slow, theorists searched for reactions that would occur more readily. It was realised that electron-ion recombination would be more rapid if two particles were present after the collision, for then both energy and momentum could be more easily conserved. Reactions were therefore investigated in which the electron collided with a positive *molecular* ion (XY^+), consisting of two atoms (X and Y), in such a way that the positive charge was neutralised and the molecule was simultaneously dissociated into its two separate atoms as represented by

$$e^- + XY^+ \rightarrow X + Y \qquad \textbf{(C)}$$

It was concluded that reactions of this type would be much more rapid than those represented by reaction **B**, and that the corresponding recombination coefficients might even be as great as that found for the ionospheric layers. The older idea, which tried to explain the recombination by a mechanism requiring an appreciable concentration of negative ions, was therefore abandoned, and it is now supposed that electrons are lost from the E and F1 layers by reactions of the type **C**. This kind of reaction is called 'dissociative recombination' and at E and F1 heights XY^+ represents a molecular ion either of oxygen (O_2^+) or nitrogen (N_2^+) which is dissociated into its two atoms during the reaction.

The divided F region

Attention was next turned to the loss of electrons from the upper part of the F layer. It has long been realised that, above about 100 or 120 kilometres, some of the oxygen of the atmosphere is dissociated by ultra-violet radiation and is in the form of atoms (O), whereas the nitrogen is not dissociated, and exists only in the form of

molecules (N_2). In the F region the positive ions thus consists of O^+, O_2^+, and N_2^+ and the loss of electrons is by recombination with these. The loss associated with the molecular ions is by dissociative recombination and is rapid, but the loss by recombination to the atomic ions O^+ is radiative and is slow, with a recombination coefficient as small as 10^{-12} units. Now although only rough and unsatisfactory estimates of the rate of loss of electrons from the F2 layer have been made, its appreciable decay throughout the night requires an effective recombination coefficient very much larger than 10^{-12}. Some way must be found to explain how the electrons and the atomic O^+ ions can effectively recombine more rapidly.

The solution was found through one of the reactions involved in the photochemistry of the ionosphere. It is the 'ion-atom interchange' process represented by. .

$$O^+ + N_2 \rightarrow NO^+ + N \qquad \textbf{(P)}$$

in which the atomic ion O^+ changes place with a nitrogen atom in the molecule N_2 to produce a *molecular* ion NO^+: calculations show that this reaction occurs fairly rapidly. Once a molecular ion has been formed the electrons can recombine with it by the dissociative process

$$e^- + NO^+ \rightarrow N + O \qquad \textbf{(Q)}$$

which occurs comparatively easily. Here, then, was a possible mechanism by which electrons might be lost in a roundabout way from the F layer, and when it was followed up in detail it revolutionised the previous ideas about that region, and showed that the layer, previously thought to be formed as a simple layer by the Chapman mechanism was, in fact, formed quite differently.

When the behaviour of the F layer is to be discussed in terms of successive reactions like **P** and **Q**, it is important to know which occurs the more slowly, for that determines the overall rate of loss

of the electrons. Now the dissociative recombination coefficient appropriate to reaction **Q** is constant and is independent of height; but the speed of reaction **P** depends on the number of N_2 molecules available to interact with the O^+ ions, and is therefore greater in the lower parts of the atmosphere. **P** is thus the slower reaction high up and **Q** is the slower low down, so that electrons disappear higher in the atmosphere at a rate determined by **P**, and lower at a rate determined by **Q**.

The important result follows that at low levels the rate of loss of electrons is that corresponding to dissociative recombination **Q**, with a coefficient independent of height; whereas at high levels the loss, controlled by reaction **P**, occurs at a speed which, being dependent on the concentration of N_2, is smaller at the greater heights. This difference explains the splitting of the F region into two parts, F1 and F2, as described in the next section.

Above the F1 ledge

If the rate at which the solar radiation produces electrons were to be plotted against height in the atmosphere it is believed that there would be two peaks in the curve, one at the level of the E layer (about 110 kilometres) and one at the F1 ledge (about 170 kilometres) and that these two peaks would behave (e.g. with respect to changes of the solar zenith distance) like the peak of a Chapman layer (p. 95). It is further thought that, on many occasions, both these levels are low enough in the atmosphere for the recombination reaction **Q** to be operative so that two 'Chapman recombination layers' of electrons are formed with peaks at the E and F1 levels.

At one time it was thought that the existence of the F2 layer, with a peak at a height near 250 kilometres, implied the existence of another peak in the rate of production of electrons, but it now seems that a more acceptable explanation is available in terms of the two

reactions **P** and **Q**. It is supposed that, just above the level of the F1 peak, the concentration of molecular nitrogen (N_2) becomes so small that the loss is determined by reaction **P** rather than by **Q** so that *the rate of loss diminishes upwards as the concentration of* N_2 *diminishes*. Above the F1 peak the rate of production of electrons also diminishes upwards, in the way described by the theory of Chapman, and when equilibrium has been reached the resulting electron distribution is determined by the balance between the production and loss. If the rate of loss decreases upwards more rapidly than the rate of production the *concentration of electrons will increase upwards*. To decide whether this is likely to happen it is necessary to consider how the production and loss vary with height.

The electrons produced from the molecular nitrogen and oxygen are rapidly lost by dissociative recombination. It is the others, produced by the ionisation of atomic oxygen, that are important. Above the F1 peak the rate of production of these electrons is determined by the amount of oxygen available to be ionised, so that the rate decreases upwards in proportion to the concentration of oxygen atoms and at a rate determined by their scale height. The rate of loss controlled by reaction **P** depends, however, on the number of nitrogen molecules, and decreases upwards with the scale height of the nitrogen. Now since the nitrogen molecule (with two atoms) is about twice as heavy as the oxygen atom its scale height is about half as great as that of the oxygen so that *the rate of loss decreases upwards more rapidly than the rate of production*. Although, therefore, the rate of production *decreases* upwards above the F1 peak the electron concentration in an equilibrium layer *increases* upwards and it is these electrons which form the F2 layer. They are produced by the radiation responsible for ionising the F1 layer, and it is no longer considered necessary to explain them by postulating an additional radiation (see figure 3·3, page 110).

Whether or not the F1 peak in the production curve gives rise to

an F1 peak in the electron distribution depends sensitively on the level at which reaction P becomes slower than reaction **Q**, and it happens that sometimes (winter: sunspot minimum) there is no F1 peak in electron concentration and sometimes (summer: sunspot maximum) it is well marked.

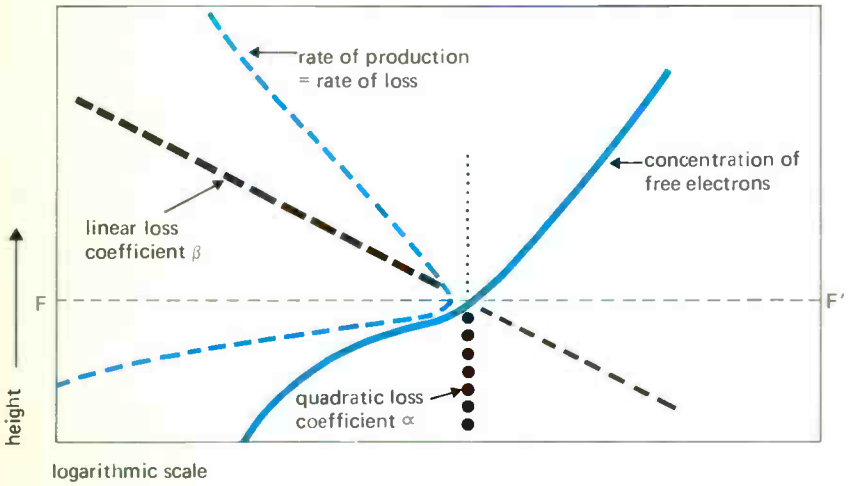
The theory described above is now believed to explain the F region ionisation. The lower part, below the F1 ledge, is fairly exactly a 'Chapman recombination layer' formed by solar radiation that produces electrons most rapidly at the level of the ledge. The ionisation above that level is produced in the upper part of this layer, but since the complicated loss process represented by expressions P and **Q** is operative, the resulting electron distribution is no longer described by the simple Chapman theory but by a theory that takes account of the rates at which atomic oxygen and molecular nitrogen decrease upwards.

At this stage a new problem arises. It has been explained why the electron concentration increases above the level where the electrons are produced most rapidly, but at what height will this increase cease? Of course it will cease at a height where all the oxygen has been converted into ions, but calculations show that height to be not less than a few thousand kilometres. How then is it possible to explain the observed fact that the F2 layer has a peak at about 250 kilometres? There must be some other phenomenon which, becoming important at that level, becomes dominant higher up. This phenomenon is diffusion and is discussed in the next section.

Diffusion and the F2 layer

Let us consider how a layer of electrons will tend to spread out if it is left to itself. It is known that if a thin slab of one gas is somehow inserted in the midst of another uniformly distributed gas, the slab will diffuse outwards into its surroundings until finally it is spread

3·3 A normalised diagram to explain the splitting of the F layer. The horizontal scale is logarithmic, so that straight lines correspond to exponential height variations. The quadratic loss coefficient α is independent of height, the linear loss coefficient β decreases exponentially upwards and their normalised values are equal at the height FF′. The blue curves represent, respectively, the rate of production of electrons (dashed line) and the concentration (N) of the electrons remaining free (continuous line). Under steady conditions the rate of production of electrons is equal to the rate of loss, which is βN above FF′ and αN^2 below. Note that on the logarithmic scale the abscissae for the curve βN are obtained by adding those for β and N, and the abscissae for αN^2 are obtained by adding those for α and twice those for N. FF′ represents the height of the F1 ledge.



uniformly throughout (see figure 3·4). Now the electrons in the ionosphere constitute a somewhat similar layer inserted, by the action of the solar radiation, into the gas of the neutral atmosphere, and they would be expected to diffuse outwards and ultimately to lose their layer-like distribution.

If the surrounding gas, and the layer of gas inserted in it, were both acted upon by gravity the situation would be modified, and when the diffusion process was complete, the gas of the slab would itself be distributed throughout the background gas, not uniformly, but exponentially with its own scale height. Both gases would then be

in diffusive equilibrium as defined on p. 21 and after that the diffusion would cease.

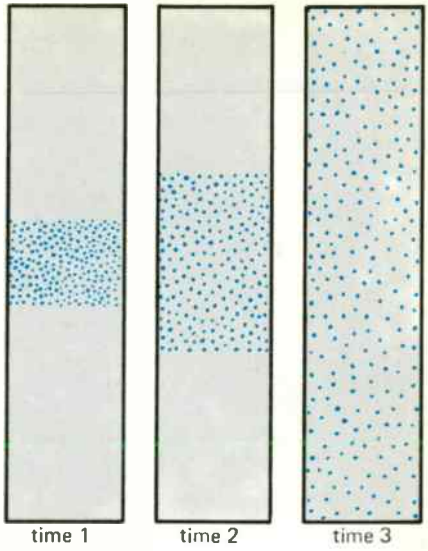
Now the scale height of a gas is determined by the gravitational force acting on it and is inversely proportional to the weight of its atoms or molecules. Since the electrons have weights about 30,000 times smaller than the atoms or molecules of the air, we might expect the corresponding scale height to be 30,000 times greater, about 300,000 kilometres, so that the electrons would extend, with concentration only slightly diminished, up to enormous heights. Meanwhile the heavier positive ions would be left behind, at lower levels, distributed with their own scale height equal to that of the gas from which they were formed. But if this great separation were, in fact, to occur there would be electric forces of attraction between the negative electrons above and the positive ions below, tending equally to pull the electrons down and the ions up. The gravitational forces on the ions are 30,000 times greater than those on the electrons whereas the forces of the electric attraction are equal. Compared with the gravitational forces, the downward electric force on the electrons is therefore much more important than the upward electric force on the ions. When the situation is examined in detail it is found that the electrons and ions finally take up exponential distributions with nearly one and the same scale height and that this height is twice that appropriate to the ions. There is some, very slight, separation between the electron and ion distributions and the resulting electric forces are sufficient to pull the ions up so that their scale height is doubled, and to pull the electrons down so that their scale height is only about 1/15,000th of what it would otherwise be. The combined particles are distributed as though the mass of an ion were shared between it and the electron (of negligible mass) with which it is associated, so that the effective molecular weight is half that of the ions.

It is next necessary to discuss the speed with which the diffusion

occurs. The diffusion of one gas through another is impeded by the collisions between the particles of the diffusing gas and those of the other, background, gas. If the concentration of the background gas is great there will be numerous collisions and the diffusing gas will spread slowly, but if it is small the spread will be rapid. Electrons and ions injected high in the atmosphere, where the concentration of the background gas is small, will thus diffuse rapidly, whereas those injected low, where the concentration is greater, will diffuse more slowly. Above some height where the concentration of the atmosphere becomes small enough, the speed of diffusion will be so great that it will play a major part in determining the final distribution of the electrons and their associated ions.

The present picture of the F2 layer is that its shape is determined by the diffusion of ions and electrons at levels above the peak and by the combined processes of production and loss of electrons at levels below the peak; and that the peak is at a level where the effects of diffusion and of electron loss are equal. Above the peak the layer, consisting of electrons and ions, has a scale height equal to twice that of the ions above. This explanation of the layer is, of course, quite different from the earlier one, which supposed that it was a Chapman recombination layer, with its peak at a height where the electrons were being produced most rapidly.

The new explanation was developed slowly by many workers between 1950 and 1958 but, although it is more plausible than the old, it is not very firmly based on experiment. Indeed it was not until after 1958, when the intensity of the ionising radiation at different heights had been measured by rockets, and the part of the layer above the peak had been explored by satellite, that the theory became more than a reasonable hypothesis. Before that time the most that could be said was that roughly speaking the total decrease of concentration in the average F2 layer during a whole night is of the order expected from the calculated rates of reactions of the types



3·4 A layer of gas molecules (blue) embedded in a uniformly spread gas (grey). The motion of the 'blue' molecules will cause them to spread outwards so that the layer becomes diffused. The diffusion will be impeded by collisions between the 'blue' molecules and those of the 'grey' gas, and will be slower if the 'grey' gas is more concentrated.

P and **Q**, and that the required speed of diffusion is roughly consistent with measurements of diffusion made in the laboratory. The old Chapman theory based on very slender evidence had been overthrown by the new theory when it was realised that the electrons were not lost by a simple recombination process; but the new theory was based on evidence scarcely less slender. It did, however, explain the splitting of the F layer into its two parts, F1 and F2, and it was probably this explanation as much as anything that gained credence for it in the minds of ionospheric physicists. But the firmness with which it was believed, even before measurements had been made in rockets and satellites, provides an interesting example of how readily scientists will accept a neat picture of nature even when the evidence for it is not very convincing.

4 Temperature and composition of the upper atmosphere

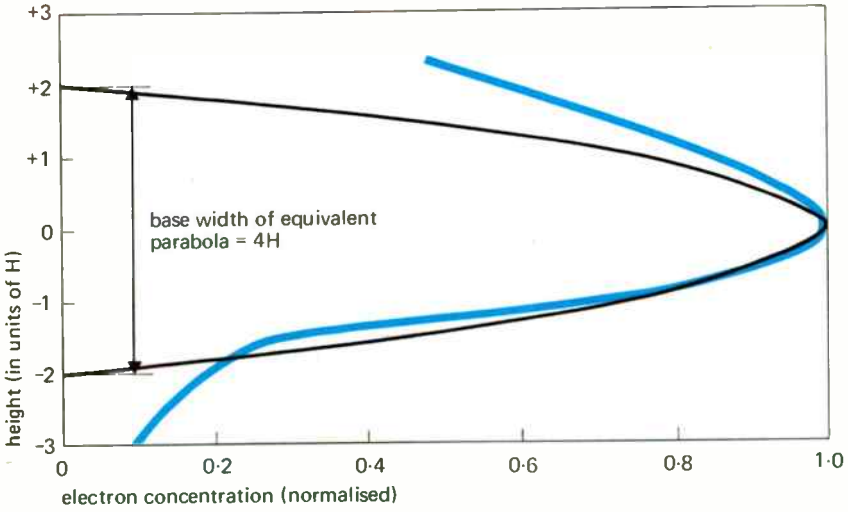
4 Temperature and composition of the upper atmosphere

Difficulties of measuring temperature

Because it is the heat motions of the atmospheric atoms and molecules that carry them to great heights against the attraction of gravity, it is clear that a knowledge of the temperature is one of the key factors in understanding the behaviour of the highest atmosphere. Unfortunately this quantity is one of the most difficult to measure, either with ground-based equipment or with the help of a space vehicle. Much ingenuity has therefore been expended in deducing it indirectly from such quantities as the collision frequency of electrons, the scale heights of electron distributions, or atmospheric densities estimated from measurements of satellite drag. Before describing what has been done let us see why it is impossible to measure the temperature in a simple way with a thermometer carried in a satellite.

When the temperature of an object is measured by putting it in contact with a thermometer, heat energy is transferred between the two until they reach a common temperature: this is then assumed to to be the original temperature of the object. Now the heat energy required to alter the temperature of a body by a given amount depends largely on the mass of the body. If, therefore, heavy and light bodies at two different temperatures are brought into contact, the heat transferred will alter the temperature of the lighter more than that of the heavier. The result is that the final temperature will approximate to the original temperature of the heavy body. If, for example, one body is a light thermometer and the other a heavy pool of liquid, the final temperature of the two will be almost the same as the original temperature of the liquid, and that is what the thermometer will measure. But if, instead, a heavy thermometer and a much lighter drop of liquid are brought into contact, the final temperature will be practically the same as the original temperature of the thermometer: and that is what will be measured, not the

4·1 Down to about $-1\frac{1}{2}$ scale heights the curve (blue) representing the electron concentration below the peak of a Chapman recombination layer is nearly the same as a parabola (black) with a base thickness four times the scale height. Note that the (blue) curve represents an electron concentration that is proportional to the square root of the rate of production shown in figure 3·2.



original temperature of the drop.

Now the smallest mass of any thermometer that could feasibly be used in the upper atmosphere is of the order of 0·1 gram, so that, to measure the temperature satisfactorily, it would have to be put in thermal contact with, say, about 1 gram of air. Suppose the temperature is required at a height of 500 kilometres, where the air is so tenuous that this mass would occupy a sphere of radius about 1 kilometre, then the heat from this large sphere would have to be passed to the thermometer; clearly this is impossible.

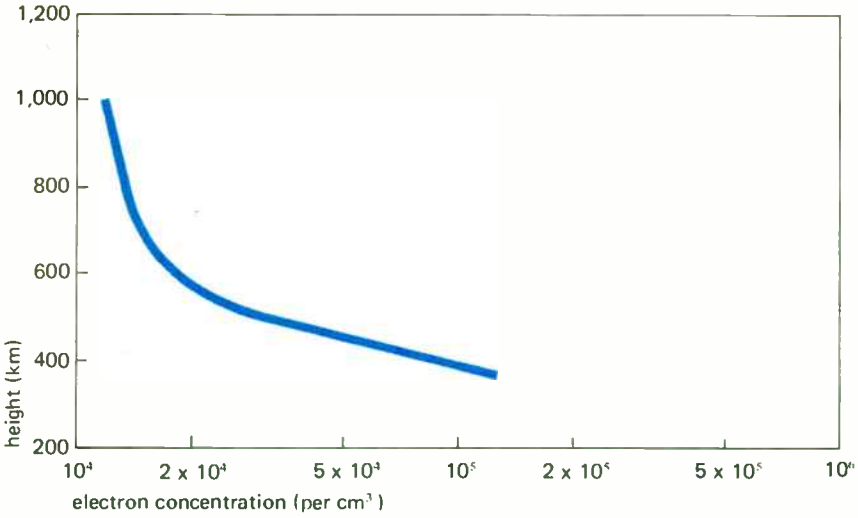
There is another reason why it is difficult to measure the air temperature by means of a thermometer in a satellite. Such a thermometer continually radiates heat into space in the form of infra-red waves and, when in sunlight, it receives radiant heat from the sun. The confusing effects of these radiations are well known when a thermometer is used at ground level in attempts to measure the air temperature, even though the heat exchange with the air is much more rapid there than in a satellite: false results are obtained if the thermometer is heated by sunlight in the day, or if it radiates heat into space at night. In a satellite, the rate of heat exchange with the air is much smaller than at the ground, and the confusing effects of radiation are comparatively much larger; it would not be possible to reduce them sufficiently for reliable measurements to be made.

Deductions from radio soundings

Even if it were feasible to measure the temperature of the air by a direct method it would not have been possible to do so until space vehicles became available in 1950. Earlier, in about 1935, the results of radio-wave sounding were therefore used indirectly to estimate the density and infer the temperature of the upper atmosphere. In the first method the frequency of collision between electrons and heavy particles was deduced from observations of the absorption of radio waves (see p. 61). This collision frequency is determined jointly by the electron concentration (which could be deduced from the appropriate ionogram), by the temperature of the electrons, and by the concentration of the neutral particles. Since the concentration of neutral particles itself depends on the temperature (assumed equal to that of the electrons) it is possible, from a knowledge of the collision frequency, to make reasonable estimates both of the density and the temperature of the atmosphere.

It was found that the collision frequency at a height of about 110

4·2 The electron distribution above the peak of the F layer (the topside ionosphere) plotted logarithmically. The scale heights deduced from the two parts of the curve are approximately in the ratio 1 to 16, as though the ions below 500 km were oxygen atoms and above 600 km were hydrogen atoms, both with temperatures of about 750°. (The curve was deduced from a record made at Singapore just before sunrise on 12 November 1962).



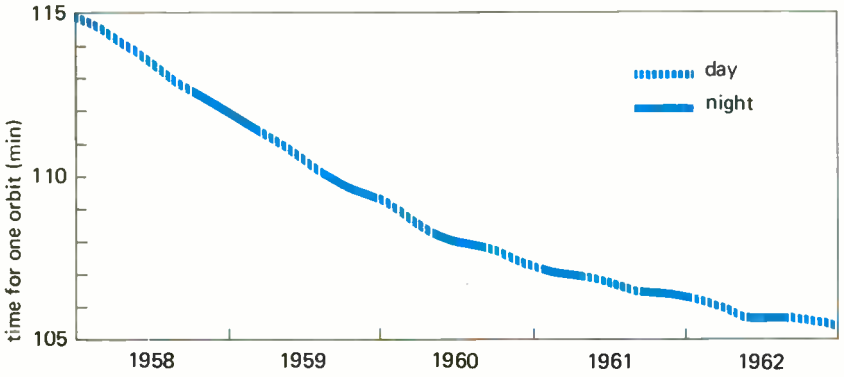
kilometres near the peak of the E layer was 3×10^5 per second, and at about 250 kilometres, near the peak of the F layer, it was $1{\cdot}5 \times 10^3$. The magnitude in the E region corresponded roughly to the temperature (300 degrees) suggested by non-radio observations (see p. 47), but if the temperature had remained the same at greater heights the collision frequency at 250 kilometres would have been less than one per second, whereas the measurements showed it to be at least 1,000

times greater. The facts could be explained only if the temperature above 110 kilometres increased with height until it was about 1,200 degrees at the F layer peak near 250 kilometres.

Soon after this suggestion had been made it was pointed out that a similar conclusion could be drawn, in the following way, from the shapes of the E and F layers. It was realised that if electrons were produced at the rate expected on Chapman's theory, and if they disappeared (as seemed reasonable) by recombination, then the curve relating electron concentration to height below the layer-peak would approximate closely to a parabola with a 'base thickness' equal to four scale heights. Now, as explained on p. 84, it had been possible, by a curve fitting process, to show that some of the measured ionograms were consistent with two superimposed parabolic distributions of electron concentration, and the semi-thicknesses of the parabolae for the E and the (unsplit) F layers corresponded to scale heights of 10 and 40 kilometres. The thickness of the E layer was consistent with a temperature of about 300 degrees, but that of the F layer could be explained only by supposing that the temperature there was greater: if the layer consisted mainly of molecular nitrogen the temperature must be 1,200 degrees but if atomic oxygen dominated it would be 600 degrees. In either case the evidence agreed with that from the collision frequencies that the temperature of the atmosphere at heights above about 150 kilometres was comparatively great.

Although for many years these two arguments provided the main reasons for believing that the temperature at heights around 300 kilometres was much greater than at the ground, it is now known that both were unsound. The conclusions are still thought to be roughly correct but the ways of arriving at them were incorrect, for the following reasons. The argument based on collision frequencies in the F layer depended on measurements made by the method described on p. 61, and it has since been shown that these

4·3 The time taken for a large inflated balloon to orbit the earth as an artificial satellite gradually decreased between 1957 and 1963. The decrease, caused by air drag, was more rapid near 1957 than 1963 and by day than by night. The conclusion is that the air was more dense at sunspot maximum (1957) than at minimum (1962) and by day than by night.



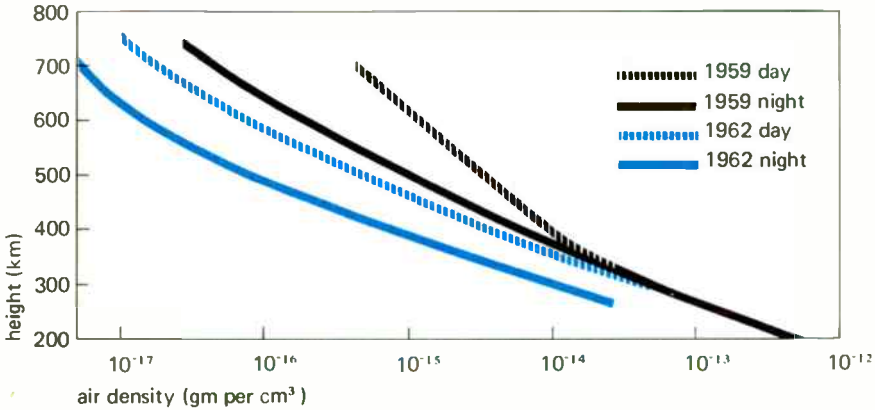
results are not very reliable and that they indicate only an upper limit to the collision frequency. The only sound deduction is thus that the temperature of the F region cannot be greater than about 1,000 degrees, but, so far as these measurements are concerned, it could be less. The argument based on the thickness of the F layer assumed that it was a Chapman recombination layer, whereas it is now known to be formed in quite a different way. It is somewhat surprising that both methods of reasoning are faulty, and yet the conclusion is still believed to be roughly correct.

When the topside sounder (p. 78) was launched in 1962, so that the electron distribution in the top of the ionosphere could be estimated, it was possible to make independent deductions about the

temperature and composition in these higher regions. The theory of the F layer discussed on p. 112 suggests that the distribution of the electrons above the peak has a shape determined by the diffusion of the positive ions which accompany them, in such a way that their scale height is equal to twice that of the gas from which they originated. This scale height can be determined by plotting the topside electron concentration against height, using a logarithmic scale, and measuring the slope of the line. At night, in the region just above the peak of the F2 layer (200–400 kilometres), the scale height usually corresponds to atomic oxygen at a temperature somewhat less than 1,000 degrees: higher up, however (above 500 or 600 kilometres), the slope is about sixteen times as great, and corresponds to hydrogen at the same sort of temperature (figure 4·2). The transition from oxygen to hydrogen occurs at a height that is different in different places and at different times; it is more gradual by day than by night. The detailed implications of these results are discussed, together with results from other experiments, on p. 129. The electron distributions at great heights, deduced in this way with the help of a topside sounder, are confirmed by measurements made from the ground by the Thomson scatter technique.

If deductions about the temperature are to be made from a measurement of scale height the weight of the ions must be known. It can often be estimated from a knowledge of the gases likely to be present but it is more satisfactory to measure it directly. A mass spectrometer in a space vehicle can be used for this purpose, and the results are most reliable when made in satellites some time after launch, so that any gas carried from the ground has had time to diffuse away. The measurements provide information about atomic constitution only at the level of the satellite: when the composition and temperature of the high atmosphere are to be discussed it is desirable that results from mass spectra, topside sounders and Thomson scatter experiments should be considered together.

4·4 Air density at different heights deduced from the timing of satellite orbits. The results are explained by supposing that the temperature of the high atmosphere is greater by day than by night, and at sunspot maximum (1957) than at minimum (1962).



Deductions from satellite drag

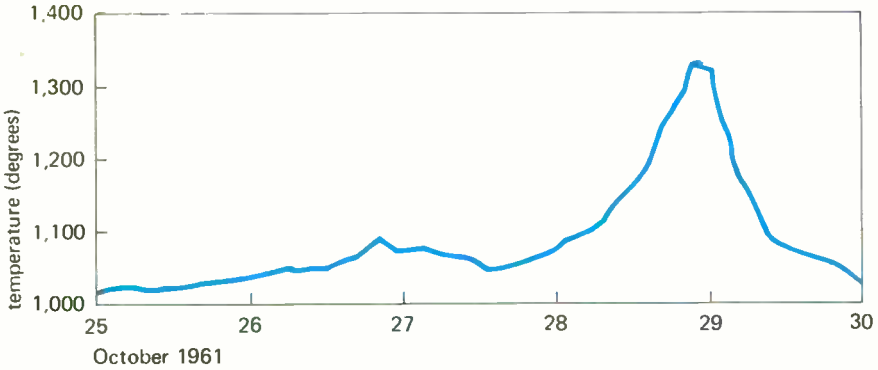
The density of the neutral, non-ionised, atmosphere at great heights has been estimated by observing the drag that it exerts on artificial satellites. The way in which this has been done provides a good illustration of how experiments made for one purpose can sometimes be used for another. It also provides an impressive example of the way in which important scientific results can often be obtained with the simplest equipment, by what has sometimes been called the 'string and sealing wax' type of experiment. Once artificial satellites had been launched the only equipment necessary for the experiments now to be described was a stop-watch and a good pair of binoculars.

The possibility of deducing the atmospheric density at the height of a satellite arises because even at those heights there is enough air to impede the satellite's motion appreciably. For example, if a satellite with a cross-sectional area of ten square metres encircles the earth once at a height of 400 kilometres, it 'runs into' one or two grams of air, which must be pushed out of its way, so that it slows down and moves into a slightly lower orbit. Now the time taken by a satellite to travel round the earth is less when the orbit is lower, and by observing how this time decreases it is possible to deduce the magnitude of the air drag and hence the density of the air.

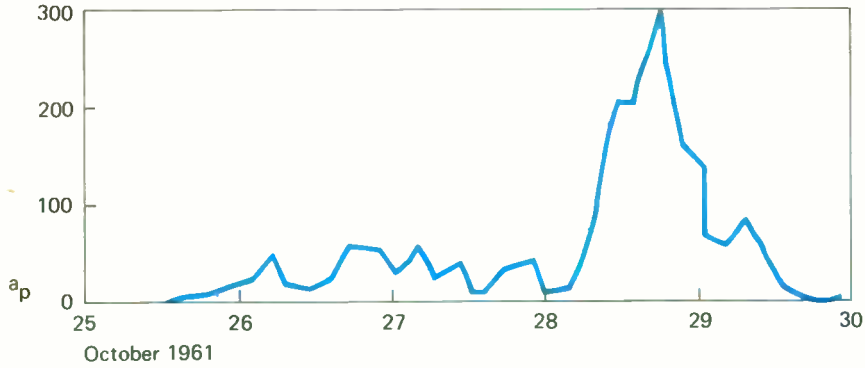
This argument supposes the satellite to be moving in a circular orbit so that the air drag acts equally all the way round, but it can be modified to apply to an elliptical orbit in which the air drag is most important when the satellite is in the densest air, that is, nearest to the earth (at perigee). If observations are made on a satellite in an elliptical orbit it is then possible to deduce the air density near the height and place of closest approach. By making use of different satellites with different perigees, or a satellite whose orbit is continuously changing so that the perigee moves from place to place, it is possible to find the density at different times and places.

To make measurements of this kind, the satellite is observed with a pair of binoculars as it crosses the sky and a note is made of the time at which it passes some point identified on an atlas of the stars. The observations must be made at a time when, although the satellite is in sunlight, the sun has set at the ground, so that the sky is sufficiently dark for the stars and the satellite to be seen. Amateurs have made these measurements in many places, and similar measurements of greater accuracy have also been made at a few places by professional observers using more elaborate equipment. Because the observations of the amateurs are the more numerous and widespread they have proved at least as useful as those of the professionals.

4·5 Increase of temperature at a height of 635 km associated with a magnetic disturbance. The temperature (a) was deduced from observations of the position of a large balloon (3·5 m in diameter) that was orbiting the earth as an artificial satellite. The amount of magnetic disturbance is represented by the 'magnetic character figure' a_p plotted in (b).



(a)



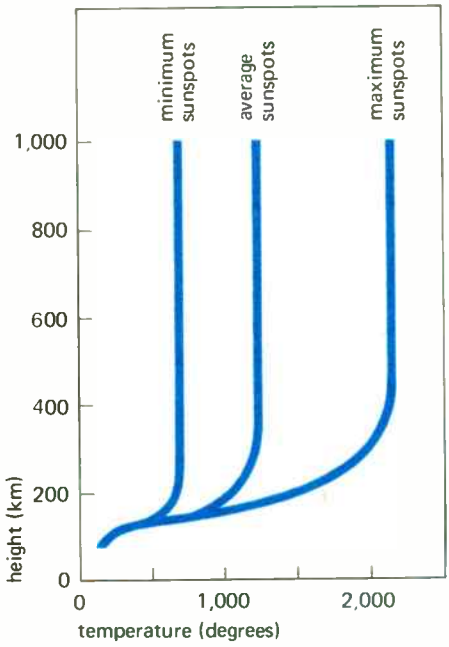
(b)

Figure 4·4 shows how the period of revolution of a large balloon satellite, with perigee height near 350 kilometres, has changed through the years. We first notice the gradual decrease of the revolution time as the air drag takes effect. The position of the perigee gradually drifted round the earth so that it was alternately in light (day) and dark (night) as shown on the graph, and we next notice how the change of the period occurred more rapidly during the day, indicating that the density was then the greater. The rate of change of the period was more rapid in the years 1958 and 1959 near the maximum of the solar cycle than near the minimum (1962) as though the density were greater at maximum. Observations of this kind show that the density at any particular height is not always the same; it is greater by day than by night, and at sunspot maximum than at minimum (figure 4·4). Another striking result is that the density increases during a magnetic storm and the more so the greater the height (figure 4·5). These results showed, for the first time, that the distribution of the high atmosphere, above about 200 kilometres, is very variable and is controlled to a large extent by the sun.

The greater density of the air by day than by night at heights of a few hundred kilometres suggests that there might be a wind in the upper atmosphere, blowing in a direction to pile up the air overhead in the afternoon to remove it in the morning. This varying wind tends to move the electrons of the ionosphere with it, with results which are mentioned later (p. 168).

Gases of the high atmosphere

It has previously been emphasised that lighter gases such as hydrogen and helium will predominate at great enough heights, even if they exist only in relatively small proportions at the ground. Both these gases are so light that some of their most rapidly moving atoms can overcome the earth's gravitational attraction so as to escape from



4·6 Observations of the changes of satellite orbits taken together with theories of heat flow through the atmosphere lead to the belief that at midday the atmospheric temperature varies with height, as shown here for different epochs in the solar cycle. The temperature high in the atmosphere is nearly independent of height and is called the exospheric temperature.

the earth completely. Over sufficiently long times, therefore, they will be present in the upper atmosphere only if they are, somehow, replaced from below. It is supposed that helium is replaced by the radio-active decay of certain minerals, and hydrogen by the dissociation of atmospheric water vapour, and it is possible to compare the rates at which they are produced with the rates at which they escape. It is concluded that the amount of helium entering from the ground and escaping from the top of the atmosphere is so small, compared with the amount present, that it can be neglected. The helium is therefore in diffusive equilibrium, its distribution depends

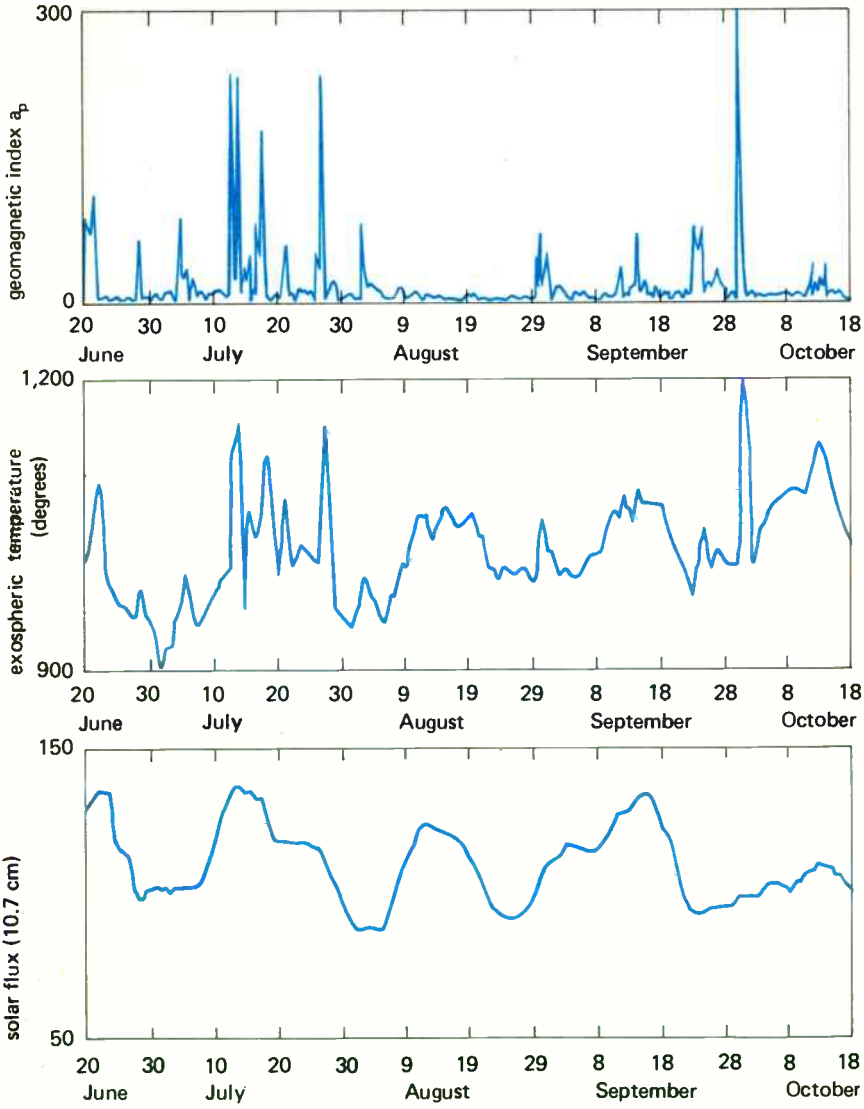
on its scale height, and at a great enough height, depending on the temperature, its density is greater than that of atomic oxygen.

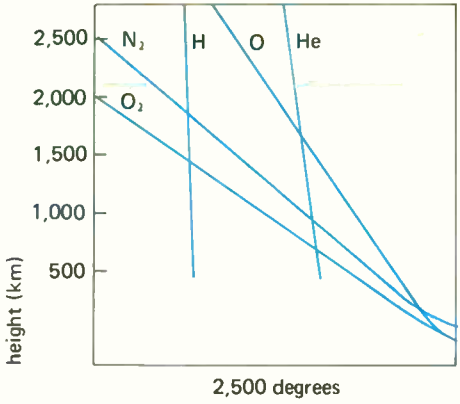
Since the atomic weight of hydrogen is only one-quarter that of helium (at great heights hydrogen is in the atomic form), its atoms are moving more rapidly, and can escape more easily, so the situation is quite different. The amount escaping is comparable with the amount present, and it is a complicated matter to calculate the height distribution. It is probable that on most occasions there is a gradual change from oxygen below to hydrogen above, with helium predominating at an intermediate level and that the heights of transition between the different gases depend on the temperature. . .

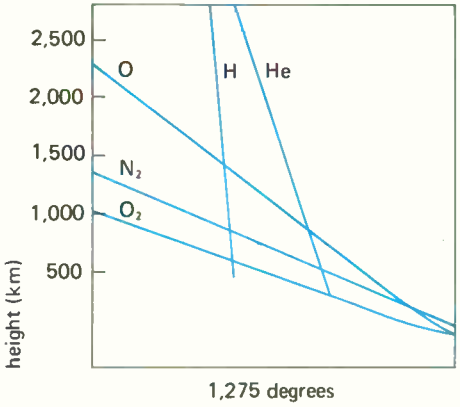
Detailed calculation of the heat flow through the atmosphere suggests that the temperature of the neutral air above about 200 kilometres is very nearly independent of height. This temperature, often called the 'exospheric temperature', determines the density of the highest atmosphere and the relative abundances of the different gases. It can be deduced from measurements of the upper atmospheric density made by observations of satellite orbits. Changes in the exospheric temperature are found to be related to changes in the photon and the particle radiations from the sun.

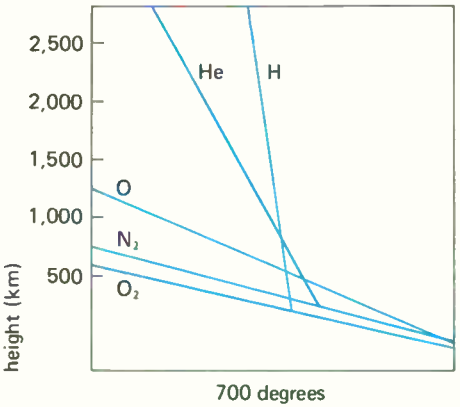
The relation to photon radiation has been established, in a roundabout way, by using radio waves emitted from the sun on a wavelength 10 centimetres. The strength of these solar radio waves has been found to be closely correlated with the measured concentration of electrons in the ionospheric layers, and therefore presumably with the strength of the solar ultra-violet photons which produce them. When the exospheric temperatures are compared with the measured strengths of the solar 10 centimetres waves, they are found to vary in a similar way: there are well marked variations of both quantities with the solar cycle and with a recurrence tendency of about 27 days (the time of rotation of the sun). It thus seems that the temperature is determined to a large extent by the amount of heat

4·7 The height distribution of atmospheric gases depends on the exospheric temperature. Logarithmic plots, in which the slope of the lines (the scale heights) are proportional to temperature, illustrate this behaviour best. The three distributions shown correspond to the three exospheric temperatures of figure 4·6.









4·8 The exospheric temperature fluctuates with a 27-day periodicity in sympathy with the 10 cm solar flux index of activity. It also increases at times when large values of the geomagnetic index indicate that the earth's magnetic field is disturbed.

that the solar photon radiation communicates to the atmosphere.

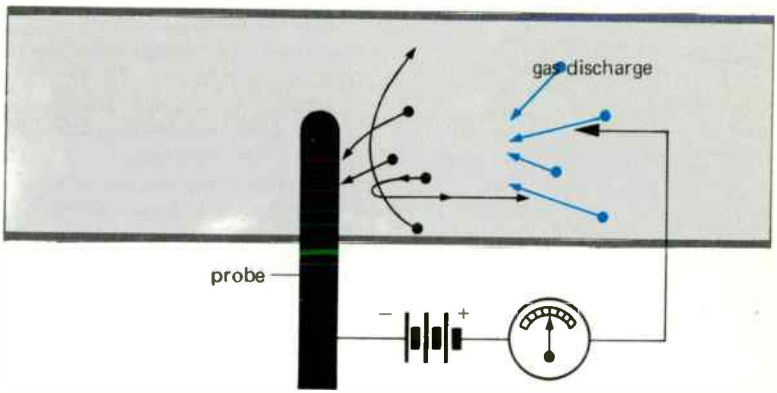
The relation between the exospheric temperature and solar particle radiation is established by comparing it with the amount of magnetic disturbance, which is itself thought to be caused by the arrival of solar particles at the earth. When values of the appropriate index a_p, which represents the extent of the irregular variation of the earth's magnetic field during any one day, are compared with values of the exospheric temperature, it is found that disturbances of geomagnetism coincide with increases of the temperature. These increases are almost certainly not produced directly by the solar particles: it has been suggested that they might be produced indirectly by ionospheric currents or by hydromagnetic waves set up by the arrival of the particles.

Temperatures of electrons and ions

The temperatures discussed in the previous sections are those of the neutral constituents of the atmosphere, even though some of them are deduced from measurements of the scale heights of electrons. In this section methods are described for measuring the temperatures of the electrons and the ions; it is shown that the temperatures of the neutral gas, the ions and the electrons are not always the same, and the reasons for the difference are discussed. First, a method is described by which the temperature of the electrons can be deduced from measurements made in a satellite.

The temperatures of charged particles in laboratory gas discharges are usually measured by a technique, first introduced by Langmuir, that depends on the thermal motions of the particles. A metal wire, called a Langmuir probe, is caused to project into the discharge and the current that flows to it, resulting from the impact of the charged particles upon it, is measured. If the probe is held at a voltage different from that of the discharge, charges of one sign will

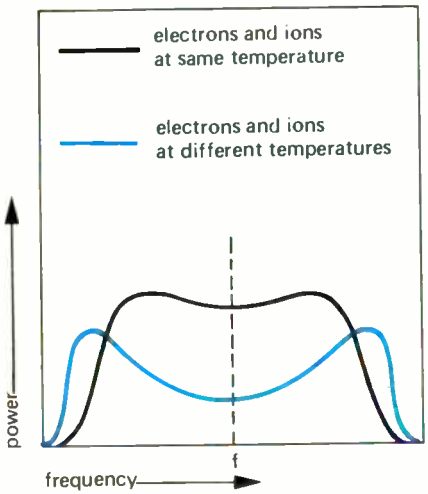
4·9 A Langmuir probe inserted into a gas discharge and held at a voltage more negative than the discharge attracts all the positive ions (blue) and repels the electrons (black). Some of the fastest electrons reach it but many are turned away. As the voltage is changed the number collected, and hence the current to the probe, changes.



be attracted and those of the other sign will be repelled; some of those repelled will, nevertheless, reach it if they are travelling sufficiently rapidly. If the difference of voltage between the probe and the discharge is then increased, some of the slower repelled particles will no longer reach it, and by observing how the total current varies when the voltage of the probe is changed, it is possible to find what proportion of particles have different speeds and thus to deduce their temperature. The details of the technique are complicated but it has been used, admittedly with some difficulty, to estimate the temperatures of the electrons and positive ions in the ionosphere.

When the Langmuir probe method is used in a rocket it is possible to determine the temperatures both of the ions and the electrons. When it is used in a satellite, however, an interesting difficulty arises because the satellite is moving so fast. At any one temperature the speed of the heavy ions is about 170 times smaller than the speed of the light electrons.* It so happens that the speed of a satellite is

* At any fixed temperature the speed of a particle is inversely proportional to the square root of its mass.



4·10 When a wave of frequency f is returned to the earth after being incoherently scattered in the ionosphere, its power is spread over a spectrum of frequencies. The width of the spectrum depends on the temperature of the ions and its shape depends on the ratio of electron to ion temperature.

intermediate between the two. If, then the method of Langmuir is used to measure the speed of the ions, as observed from the moving satellite, it is the satellite's speed that is measured and there is no possibility of deducing the ion temperature. The speed of the satellite is, however, small enough for a reasonable estimate to be made of the speed and the temperature of the electrons.

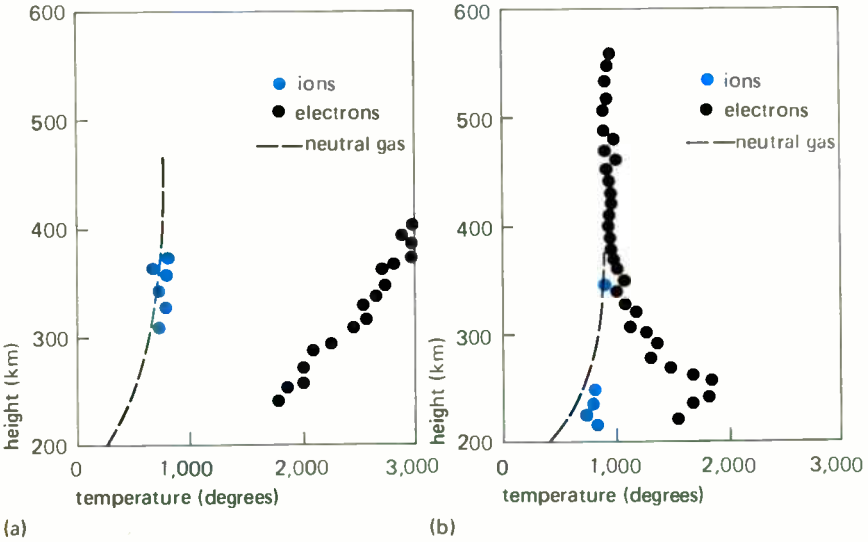
The temperatures of upper atmospheric ions and electrons have also been measured from the ground by an adaptation of the technique of Thomson scatter sounding. It will be recalled that, in this method of sounding, the received wave has been re-radiated from the individual electrons in the ionosphere. These electrons are moving, some in one direction, and some in another, with speeds depending on their temperatures, so that the wavelets which they re-radiate have their frequencies altered by the Doppler effect and when combined at the receiver they form a wave that is spread

in frequency, even though the original up-going wave was radiated only on one frequency. At first sight it would seem that the amount of the spread should be determined by the thermal speeds of the electrons. When measurements were first made, however, the spread was found to be much less than expected, and was appropriate to re-radiation from ions rather than from electrons. This observation stimulated theorists to examine the situation in more detail, and it was then realised that electrical forces of attraction cause the motions of the electrons to be controlled to a large extent by the thermal motions of the ions. It turned out that, although the *electrons* are responsible for the scattering, and their concentration can be correctly deduced from a measurement of the scattered power, the frequency-spreading corresponds to the temperature of the more massive *ions*. An additional valuable result emerged: that if the electrons and the ions were at different temperatures, the shape (not the width) of the spectral distribution of power in the received wave would reveal the fact, and would allow the ratio of the temperatures to be estimated.

Early estimates of the temperatures of the electrons, ions, and neutral particles, based on observed scale heights, direct measurements of temperatures in satellites, and Thomson-scatter experiments, all seemed to be inconsistent and unrepeatable. Soon it was realised that the temperatures of the different particles need not be the same, and theorists got to work to discuss, in more detail than previously, the way in which heat might be shared between electrons, ions and neutral particles. Ideas on the subject are still developing but at present the situation is believed to be roughly as follows.

The temperature of any atmospheric constituent is determined by the balance between the heat put into it and the heat lost from it. Heat is put in mainly at levels where the sun's radiation is most readily absorbed; it is lost either by conduction through neighbouring layers, above or below, or by radiation. Conduction involves the

4·11 Temperatures of electrons, ions and neutral gas near (a) sunrise (0617 hours) and (b) near midday (1331 hours). The temperatures of ions and electrons were deduced from Thomson scatter soundings and those of the neutral particles from observations of satellites.



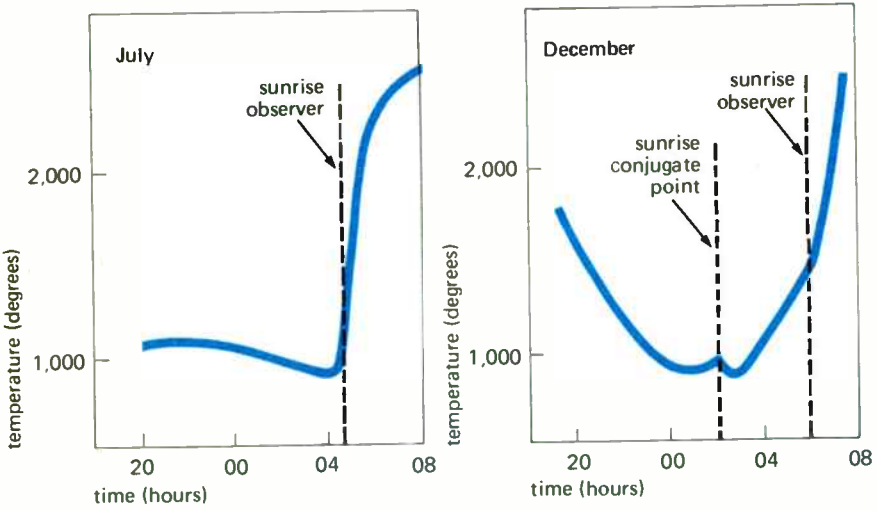
transfer of heat energy from particle to particle by collisions. Radiation occurs if there is some atom or molecule that can absorb energy in thermal collisions with other particles, and can then radiate it as an infra-red photon which, travelling without undue absorption, will carry the energy away from the place of its origin.

In the atmosphere above about 80 kilometres the main radiations that can be absorbed are those that can also ionise, and the most important heat input is in the F1 region at a height of about 200

kilometres. The photons in these radiations have an energy much greater than is needed to ionise the atoms and molecules, and their surplus energy is shared between the electrons and the more massive ions into which the neutral particle is split. Simple mechanics shows that when two bodies of greatly different mass are separated by supplying energy to them the lighter acquires most of the energy of motion. For example, if a bullet is fired from the earth, most of the energy goes into the bullet and practically none into the earth. In the same way when neutral particles are split up by the ionising action of photons most of the energy goes into the light electrons: their initial speeds then correspond to a considerable increase in their temperature. Thus photons of the radiations responsible for ionising the F layer give energy to the electrons equivalent to increasing their temperature to about 10,000 degrees. While the sun is illuminating the atmosphere, new electrons with energies of this magnitude are continuously being produced in a rather thick layer with a peak at a height of around 200 kilometres: they are called photo-electrons. Their motion is controlled by the earth's magnetic field so that they move in a spiral following the geomagnetic lines of force. While they are travelling they part with some of their energy to the surrounding particles. If these are numerous the photo-electrons do not go far before being slowed down; if, however, they are scarce, the photo-electrons travel a long way. Those that travel downwards into the denser atmosphere therefore deposit most of their energy near the place of their origin, whereas some of those that move upwards, into the rarer atmosphere, carry their excess energy to considerable distances. Some indeed may follow the geomagnetic lines of force up to heights where there is not enough atmosphere to absorb their energy; they will then continue down the line and will finally deposit their energy in the denser atmosphere of the opposite hemisphere.

The energy of a photo-electron is passed, by different processes, to the surrounding particles, but since the transfer of motional

4·12 How the temperature of the electrons at a height of 450 km changes near sunrise. In July the temperature of the electrons above the observing point increased when sunlight first reached them at 0400 hours. In December the increase occurred soon after 0300 hours when photo-electrons reached them from the magnetic conjugate point where the sun had already risen. The sun did not rise at the point of observation until 0600 hours.



energy occurs most easily between particles of equal mass it is the surrounding electrons which receive most of it. It is as though a ping-pong ball (the photo-electron) were projected rapidly on a billiard table into an assembly of heavy billiard balls and of ping-pong balls with masses equal to its own, so that it could bounce back and forth amongst them. When it had slowed down it would be found that the other ping-pong balls had gained much more energy than the massive billiard balls. In the same way, when a photo-

electron moves through the background atmosphere consisting of heavier particles and other electrons, it is the electrons, and not the heavier particles, that are heated up.

The temperature to which any group of electrons is raised depends not only on the total amount of heat put into them, but also on how many there are to be heated. When, therefore, the surrounding electrons receive energy from photo-electrons, the temperature to which they are raised depends on how closely they are packed; if their concentration is small their temperature is raised more than if it is great. In the ionosphere the most favourable time for a great increase in electron temperature thus occurs just after sunrise on the F layer, when many new photo-electrons are being produced and when there are comparatively few surrounding electrons to share their energy. Later in the day, although the steeper incidence of the solar radiation produces photo-electrons more rapidly at any one level, the surrounding electrons are more numerous than at sunrise, and their temperature is not increased so much.

After the ionospheric electrons have been heated in this way by the photo-electrons they pass their heat on to the ions and neutral particles, but more easily to the ions in virtue of the electric forces between them. The ions in turn collide with neutral particles, and share their motional energy with them (their masses being similar): finally the neutral particles lose their energy either by radiation or by conduction into the cooler, lower atmosphere.

A characteristic time is associated with the transfer of heat by each of these processes, and if the temperature of the electrons is suddenly increased in a time short compared with one of these, there is not time for the heat to be transferred, and some of the particles become hotter than others. This situation occurs at sunrise when heat energy is fed in too rapidly for the heat transfer processes to keep up with it; the temperature of the electrons then increases considerably, that of the ions by a smaller amount, and of the neutral

particles hardly at all. One or two hours later the processes of heat-transfer catch up with the heat input, and at sufficiently great heights the temperatures of all particles become roughly equal and greater than they were at night.

At some times of the year, and in some places, it happens that sunrise is earlier at one end of a line of force than at the other. Some of the photo-electrons from the sunlit end then travel to the other end, which is still unilluminated, and there they heat up the electrons. Because the sun has not yet risen at that point there is no new ionisation, the background electron concentration is small, and the arrival of the photo-electrons produces a relatively great increase of temperature (figure 4·12).

5 Nature of the ionising radiations

5 Nature of the ionising radiations

Ionisation by photons or particles?

The close fit between experimental results and Chapman's theory of layer formation suggested that the ionosphere was produced by some sort of radiation coming in a straight line from the sun, but its nature was not at first known. It might consist of photons in the ultra-violet and x-ray part of the spectrum, or of particles, either charged or uncharged. Although charged particles would be deflected towards the poles by the earth's magnetic field they could not be dismissed as a possible ionising agency even at low latitudes, for they might arrive at the earth as a neutral stream of mixed positive and negative particles unaffected by the earth's magnetic field.

To decide between particles and photons, advantage was taken of the significant difference between their speeds. Photons travel with the speed of light, but particles move much more slowly; they might be expected, perhaps, to have speeds similar to those that cause aurorae twenty or fifty hours after they have left the sun, or to those that Milne had pictured as ejected from the sun by the pressure of radiation (p. 34).

Observations were therefore made during an eclipse, when the moon intercepts radiation coming from the sun. Because photons travel with the same speed as light, any ionisation produced by them will decrease at the time of the visible eclipse, but the interruption of a stream of particles moving more slowly will occur at a different time and might, at first sight, be expected to be noticeable at the earth somewhat later. It must, however, be remembered that the moon moves round the earth with a speed comparable with that of any particles that might be expected, so that while the particles are traversing the distance between the moon and the earth, the moon moves through a similar distance. When detailed calculations are made it turns out, rather unexpectedly, that the eclipsed particle stream will impinge on the earth *before* the visible eclipse, and that if

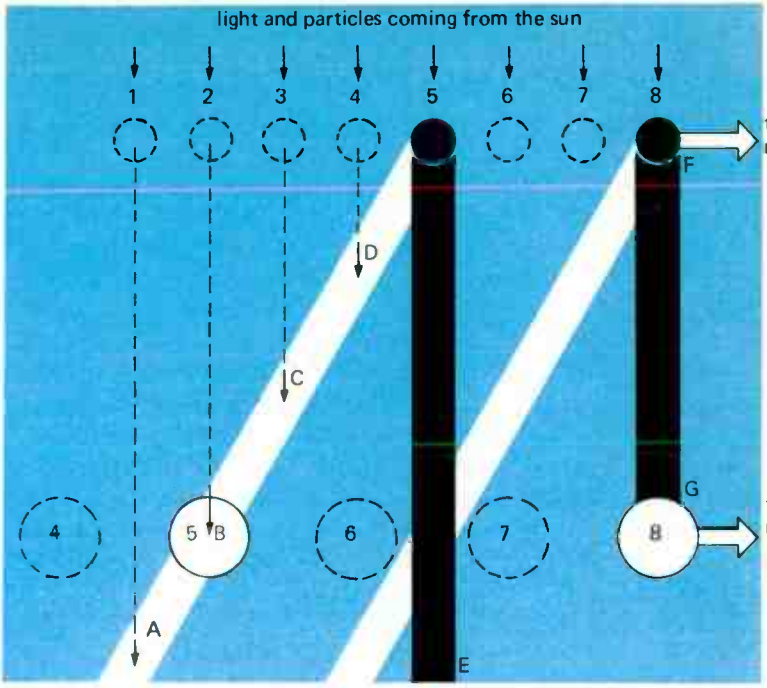
5·1 Top Lunar eclipse of particles and photons. At equally spaced times 1, 2, 3, ... 8 the moon takes up positions 1, 2, 3, ... 8. At time 5 the particles that it intercepted when it was at 1, 2, 3, 4, have reached A, B, C, D, respectively, so that particles are missing from an inclined cylinder. Because light travels so quickly, photons are missing from a cylinder 5E in the direction immediately away from the sun. At time 5, the earth intercepts the cylinder A, B, C, D, and there is a 'particle eclipse'. The moon and earth then move forward, the earth faster than the moon, and at time 8 the earth intercepts the cylinder FG from which visible light, ultra-violet, and X-rays are missing.

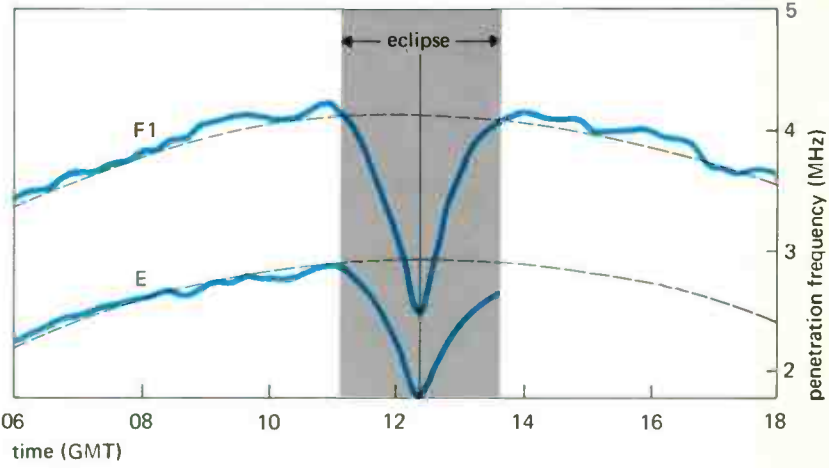
5·2 Bottom Penetration frequencies of the E and F1 layers during the eclipse of 30 June 1954 observed at Inverness (Scotland). The dashed curve shows the average values for the 'control' days before and after the eclipse.

the particles have the speeds expected on Milne's theory the time interval will be about 1½ hours. Here then was the possibility of a simple crucial test. The penetration frequencies of the layers were measured at frequent intervals on a day when there was a total eclipse of the sun and, for comparison purposes, on several days before and after, and the times of the radio eclipse and the optical eclipse were compared.

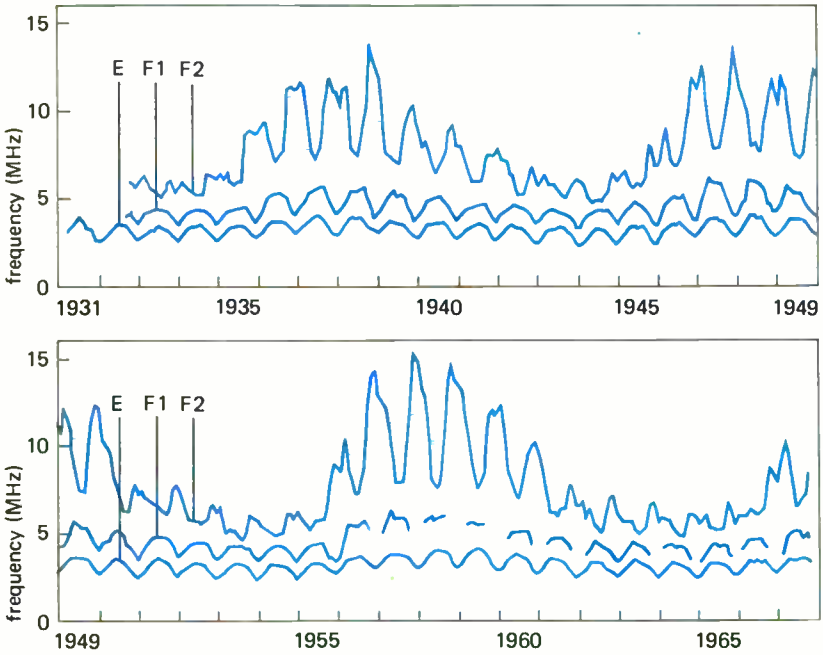
The penetration frequencies of the E and F1 layers were markedly reduced during the visible eclipse, but there were no abnormalities in the preceding few hours. There was then no doubt that the radiation which ionises these layers travels with the speed of light and consists of ultra-violet or x-ray photons. Incidentally, the experiment also confirmed that the radiation does come from the sun.

For the F2 layer the results were, as usual, confusing. The first experiment, made on the occasion of a total eclipse in Canada in 1931, provided a result as convincing as that for the E layer, but on several subsequent occasions no clear decrease was found in the penetration frequency; sometimes, indeed, there has even been an increase at the time of the eclipse. Although experiment cannot indicate clearly whether the radiation which ionises the F2 layer consists of photons or particles it is always supposed that it consists largely of photons. In recent times it has, however, been suggested that ionisation produced by particles may sometimes be added to that produced by photons (see p. 168).





5·3 Noon values of penetration frequencies measured at Slough, England. When the seasonal variations are removed there is seen to be a variation in step with the solar cycle with maxima near 1937, 1948 and 1958. It is noticeable that the seasonal variation for the F2 layer is out of step with that for the others, so that the greatest values of the F2 penetration frequency are found in winter.



Changes in the solar radiation: the solar cycle

Routine measurements of penetration frequencies were started in 1927 and have continued in increasing numbers to the present time, and it has now become clear how the electron concentrations at the peaks of the layers have varied through more than three solar cycles.

The results of observations for the three layers are shown in figure 5·3, from which the eleven-year solar cycle is clearly evident. Even in the early days, when measurements had been made over only part of a cycle, these results were important because they confirmed the previous conclusions, drawn from the early measurements of

geomagnetism and of transatlantic radio signals, that the ionising radiation varies quite appreciably with the solar cycle, whereas it was known that the visible radiation from the sun does not. They showed, moreover, that the variations in the E and F2 layers are noticeably different.

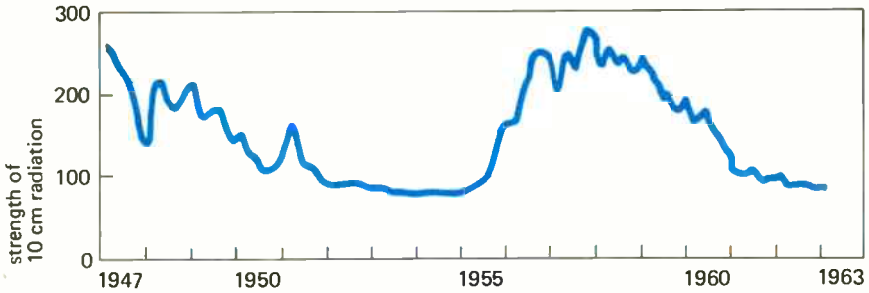
Even during 'quiet' times, when none of the extreme variations accompanying magnetic storms and aurorae are present, the concentration of electrons in the ionosphere is found to vary from day to day, although the variations are smaller than those associated with the solar cycle. Some of these variations exhibit a 27-day recurrence tendency as though they are related to the rotation of the sun; others are more random. It is not easy to decide whether they are the result of variations in the strength of the solar ionising radiation, or are evidence of changes in the earth's atmosphere. That they are related to occurrences on the sun is, however, suggested by observations of radio waves emitted from the sun, in the following way.

Soon after the Second World War the techniques of radar were used to show that the sun emits radio waves of centrimetric wavelengths. Measurements made, since 1946, on a wavelength of 10 centimetres show that the strength varies with the solar cycle, and from day to day, and that it is particularly enhanced during the occurrence of an SID. The day-to-day variations of the electron concentration in the ionosphere are found to be similar to those of the 10 centimetre radiation, and it is concluded that the strength of that radiation can provide a rough measure of the ionising ultraviolet and X-radiation incident from the sun.

Radiations responsible for the E and F layers

As soon as the concentrations of the electrons in the different ionospheric layers were known, attempts were made to explain in detail how they might have been produced by photons from the sun.

5·4 The strength of radio waves emitted from the sun on a wavelength of 10 cm follows the 11-year solar cycle. It provides a useful measure of the sun's ultra-violet and X-radiation.



For each layer there are two measured quantities that might shed light on the nature of these photons: the height of the peak, and the electron concentration at the peak.

Chapman's theory shows that the peak of a layer is formed at a height that depends jointly on the absorbability of the ionising radiation and on the height distribution of the gas to be ionised. The absorbability of any radiation depends both on its wavelength and on the nature of the absorbing gas, so that in seeking to explain the observed height of a layer it is necessary to determine the wavelength of the ionising ultra-violet or X-radiation, its absorbability in the gas to be ionised, and the height distribution of that gas.

The energy of a photon is greater the smaller its wavelength, and only those with wavelengths in the ultra-violet and X-ray part of the spectrum can ionise the gases of the atmosphere. These radiations are absorbed by the air and can be measured only at great heights and with the help of rockets and satellites. Earlier, when observa-

tions had to be made near the ground, it was necessary to infer their strength indirectly by extrapolation from the observable part of the spectrum, which seemed to come from a sun at a temperature of 6,000 degrees. Since observations at ground level extended over a wavelength range only from about 9,000 Å to 3,000 Å, whereas knowledge of the spectrum down to about 10 Å was desirable, the reliability of this extrapolation was doubtful. But it suggested that X-rays with wavelengths less than 200 or 300 Å would be much too weak to have any importance for the ionisation of the atmosphere, and that only longer wavelengths in the ultra-violet part of the spectrum could produce the observed ionisation.

It was next necessary to estimate how far the different gases of the atmosphere would absorb different radiations. Because the appropriate part of the ultra-violet spectrum is absorbed by the air in the laboratory, measurements of absorbabilities were difficult, and the results were often unreliable; they were supplemented by theoretical calculations which, however, were themselves open to doubt, and tended to differ from one worker to another. In spite of the uncertainty of these absorption coefficients, attempts were made to select atmospheric gases that might be distributed suitably in height, and might absorb some parts of the ultra-violet spectrum by the right amount to produce layers with their peaks at the observed heights. In postulating a model of this kind it was, of course, important to ensure that a part of the spectrum that might produce a layer in any one gas was not all used up by producing a higher layer in another gas.

After it had been decided that a particular radiation could produce a peak at a particular height it was next desirable to estimate the rate at which electrons would be produced. This involved an estimate of the intensity of the radiation in the selected wavelength range, and an estimate of the number of electrons produced for each photon absorbed. The concentration of the resulting electrons could then be

calculated by making use of the rate of electron loss deduced from the experiments described on page 100.

There were many uncertainties in making calculations of this kind, particularly those concerning the intensity of the solar radiation, the rate of loss of electrons in the ionosphere, and the absorbability of radiations of different wavelengths, but, in spite of these, several suggestions about the formation of the different layers were made in the years just before the Second World War. It is not surprising that different workers came to different conclusions and that some of them are no longer acceptable.

Soon after the end of the war several new groups of investigators turned their attention to problems of the upper atmosphere and the ionosphere. Some of these brought with them new interests and ideas, particularly to the study of the ionising processes responsible for the production of the different layers. It began to be realised that, although the main part of the visible light from the sun comes from a level, the photosphere, which might be at a temperature of 6,000 degrees, there is evidence, from some of the spectral lines observed visually, that the corona (p. 36) of the sun might be hotter, and might even have a temperature as great as one million degrees. This suggested new possibilities, for at that temperature the intensity of the ultra-violet radiation of shorter wavelength, and of the x-radiation, would be much greater than had previously been supposed. It was known from the visible spectrum that hydrogen was abundant in the sun and there would probably be quite strong radiation of the hydrogen Lyman-alpha line with wavelength 1,216 Å in the ultra-violet part of the spectrum.

During the immediate post-war years there were also advances, both experimental and theoretical, in our knowledge of the absorption of ultra-violet radiation in the different atmospheric gases; and microwave radio techniques, developed during the war for radar, were used to investigate in more detail the various reactions which

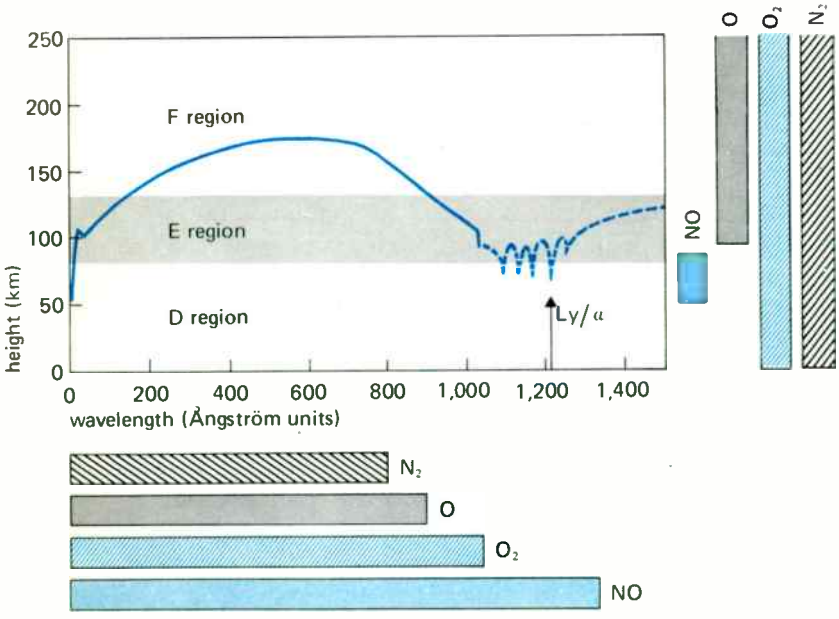
accompanied the loss of electrons. In 1960 space vehicles were first used to investigate the solar ultra-violet and x-radiations at heights where they were no longer absorbed; at first the experiments were made in rockets, and the strength of the radiations were not measured in absolute value, but as techniques developed the absolute strengths were determined from experiments made at greater heights in satellites. As a result it is now possible to understand, in some detail, how the different ionospheric layers are formed.

The new understanding led to some revision of the old ideas about the rate at which electrons are lost from the ionosphere. Whereas earlier it had been necessary to estimate the (unknown) strength of the ionising radiation by combining a knowledge of the electron concentration with an estimate of the rate of loss of electrons, it is now possible to invert the argument and *to deduce* the rate of loss from a knowledge of the electron concentration and the *measured* strength of the ionising radiation.

Present ideas about the factors that influence the production of ionisation at different heights are represented graphically in figure 5·5. Each atmospheric gas can be ionised only by radiations with wavelengths shorter than a certain 'ionisation limit', which is indicated along the horizontal scale: no gas of importance in the ionosphere can be ionised by radiation with wavelength greater than 1,200 Å. The vertical distribution of the gases, determined as described elsewhere in this book, is indicated schematically at the right-hand side of the vertical scale. From the information in this diagram, combined with knowledge of the absorbability of different parts of the spectrum in the different gases, Chapman's theory is used to calculate the peak height of the ionised layer that would be produced by each wavelength: it is shown as a continuous line.

Even those radiations that cannot ionise oxygen or nitrogen are absorbed by these gases, and the broken line represents the heights to which they penetrate before their strength is reduced by a factor

5·5 The formation of the ionospheric layers. The gases available to be ionised are distributed in height as shown on the right-hand side, while the wavelengths of radiations that can ionise them are marked below the horizontal scale. When radiations of different wavelengths are absorbed in the atmosphere they have their strengths reduced by a factor 1/e at the heights shown by the curve. Those that can ionise the major atmospheric gases (the continuous part of the curve) produce electrons most rapidly at these levels. Those that cannot ionise the main gases are indicated by the broken line. The strong Lyman-α radiation at 1216 Å penetrates far enough to ionise the nitric oxide (NO) in the D region.



1/e.* This part of the curve is irregular and there are a number of 'windows' through which radiation of very specific wavelength can penetrate more easily to greater depths. No attempt is made to indicate the precise position of these, with the exception of the important one at 1,216 Å, whose significance is discussed below.

Figure 5·5 shows which radiations are responsible for the ionisation at different heights. The F region is produced by radiations with wavelengths between about 200 and 800 Å, ionising molecular nitrogen and atomic oxygen at heights between about 150 and 170 kilometres. As explained on p. 109, the resulting electron distribution does not always have a peak at this level, when it does it is the F1 peak or ledge. The shape of the F2 layer, at greater heights, is determined by the height variation of the loss process and by diffusion (see p. 112). The E layer near 100 kilometres is produced jointly by x-rays with wavelengths less than about 100 Å, ionising oxygen and nitrogen, and by ultra-violet with wavelengths in the range near 1,000 Å, ionising oxygen.

Ionisation of the D region

In the D region below about 90 kilometres the situation is more complicated. Figure 5·5 indicates that x-rays with wavelengths less than about 20 Å can ionise all the atmospheric gases present at that level, but the other radiations, with wavelengths greater than 1,030 Å which penetrate far enough cannot ionise any of the gases normally present. If, however, some other gas were present, capable of being ionised by the longer wavelengths, the resulting electrons could also be important. Nitric oxide (NO) is a gas of that kind. It is formed, at heights somewhere between 60 and 90 kilometres, when atomic nitrogen, produced by photochemical reactions in the E layer, dif-

* If the radiation were producing ionisation as in a Chapman layer this would be the height at which the rate of production reached its peak value.

fuses downwards and reacts with oxygen. This nitric oxide can be ionised by radiation having wavelengths up to the limit of 1,340 Å, so that if one of the spectral 'windows' (on the dashed part of the curve) transmits solar radiation of sufficient intensity, electrons produced by ultra-violet might be as important as those produced by x-rays. By a remarkable coincidence this is precisely what happens. The Lyman-alpha line is radiated strongly by hydrogen in the sun with a wavelength (1,216 Å) that coincides almost exactly with one of the spectral windows in the atmosphere: it penetrates into the nitric oxide layer and produces a copious supply of electrons at a height around 75 kilometres.

Rocket and satellite measurements of the strengths of the Lyman-alpha and x-radiations, unabsorbed by the atmosphere, show that they are strong enough to contribute roughly equally to the ionisation of the D region. There is, however, the interesting difference that the strength of the x-rays varies by a large factor through the solar cycle, while that of the Lyman-alpha does not change appreciably. The strength of the x-rays also varies from day to day; and with it the concentration of electrons in the D region and the absorption of radio waves.

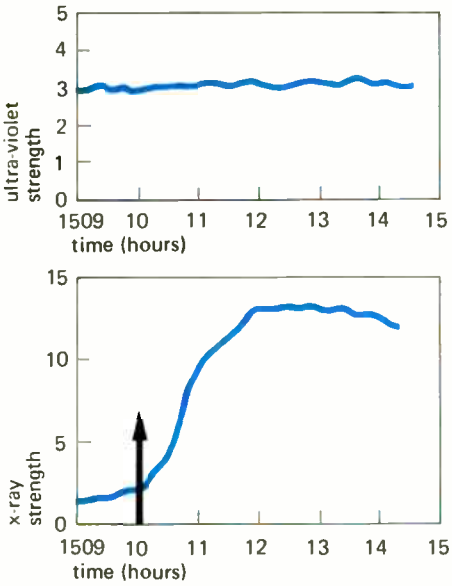
The ionisation in the lowest part of the D region is not, as in the rest of the ionosphere, produced by solar radiation, but by cosmic radiation, consisting of very rapidly moving charged particles coming from distant parts of space and impinging on the earth equally by day and by night. Their great speeds allow them to penetrate to ground level without losing much of their energy, so that the rate at which they produce electrons is proportional to the concentration of the air molecules available to be ionised, and is greater at greater depths. The concentration of electrons at any one height depends not only on the rate of their production, but also on the rate of their loss, and since that increases downwards even more rapidly than the rate of production, the resulting concentration has a peak some-

where in the region of 60–70 kilometres. It is this distribution of electrons, left over from those produced by cosmic rays, which forms the lowest part of the ionosphere, the 'tail' of the D region. Because it is produced by the charged particles of the cosmic rays, which are deviated by the earth's magnetic field, it is more intense near the poles than near the equator.

At first sight it might be expected that, because the cosmic radiation is incident equally strongly by day and by night, the electron concentration at these lower levels would remain constant throughout the 24 hours. For the following reasons, however, this expectation is not realised. At the relevant heights (60 or 70 kilometres) electrons collide so frequently with molecules that most of them become attached to form negative ions; the number of electrons remaining free would therefore be expected to be small as, indeed, it is during the night. By day, however, the sun's radiation falling on the negative ions releases, or detaches, the electrons and allows those that are produced by cosmic rays to play their part in affecting radio waves. The result is that, at levels of about 60 to 70 kilometres, the D region electrons are present by day, but not by night.

The ionisation of the D region varies throughout the day roughly as would be expected if it were produced in the manner described above. At heights around 70 to 80 kilometres it increases gradually from sunrise to midday and decreases towards sunset. At smaller heights, where cosmic rays are the most important ionising agent, the electron concentration increases suddenly near sunrise (when electrons are detached from the negative ions that have been present in the night), remains nearly constant through the day, and decreases suddenly near sunset when the electrons become attached again to form negative ions.

D region ionisation varies with the solar cycle in an interesting way. The part higher up, near 70 or 80 kilometres, which is supposed to be formed by solar ultra-violet and x-radiation, is greater at



5·6 The strengths of the ultra-violet and X-radiations emitted by the sun and recorded in a satellite during an SID which started at the time marked by the arrow. The strength of the X-radiation increased many times while the ultra-violet radiation remained constant.

solar maximum than at solar minimum, as would be expected, since the x-radiation is thought to vary in that same way. The lower part, near 60 kilometres, which is supposed to be formed by cosmic rays, is however greater at minimum than at maximum. This variation is consistant with the observation that the intensity of cosmic rays varies in this same inverse way with the solar cycle.

Although D region ionisation varies in these regular ways through the day and the solar cycle, there are two ways in which it behaves irregularly, and the resulting changes of radio wave absorption are important to users of commercial radio. One of these corresponds to the occurrence of an SID at the time of a solar flare; the other is

characterised by increased ionisation, accompanied by abnormally great radio wave absorption, on some days in winter.

Because SIDs occur at the same time as solar flares (see p. 37), which are observed in the visible H-alpha line of hydrogen, it was at one time thought that the increased ionisation in the D region is produced by a simultaneous 'flare' in the ultra-violet Lyman-alpha line, which is also emitted by hydrogen. When measurements were made in rockets and satellites, however, it was found that the strength of the Lyman-alpha line does not alter appreciably during an SID, but that the strength of short X-rays (with wavelength less than 10 Å) is much increased. It is now believed that it is the increase in X-radiation, and not in Lyman-alpha, that causes the SID.

The winter days of abnormally great absorption (the so-called 'winter anomaly') are probably related to changes in the composition of the atmosphere. Measurements of temperature and pressure made with the help of balloons at heights of about 30 kilometres have shown that sometimes, in regions extending over a few thousand kilometres, there is a sudden heating of the air, a so-called 'stratospheric warming', and evidence is accumulating that warmings of this kind are associated with increases of absorption. If this relation is established with certainty it will be the first clear evidence that changes in the lower atmosphere and in the ionosphere can sometimes have a common cause.

6 The F layer and ionospheric movements

6 The F layer and ionospheric movements

World-wide distribution of electrons

Radio communications over distances so great that the wave has to be guided by the ionosphere were of major importance to both sides during the Second World War. Because the F region seemed to be so variable, from hour to hour, season to season, and through the solar cycle, and so different at different places, it was difficult to choose the best frequency to use on any particular occasion. Nations on the two sides in the conflict therefore established ionosondes in widely different places to make routine measurements of the penetration frequencies of the layers, in the hope that more detailed knowledge would lead to a better choice of frequencies and improve the reliability of communications. When the war was over there was thus extensive knowledge of the ionosphere on a world-wide scale, and it was realised that studies must continue in several places if the behaviour of the layers (particularly the F layer) was to be properly understood.

Many of the wartime observatories were therefore continued in operation so that, from 1945 onwards, the ionosphere was studied, not merely at one or two places where there were active research teams, but as a world-wide investigation which involved close scientific collaboration between widely spread groups of workers. These activities received valuable stimulus from the Union Radio Scientifique Internationale (URSI) – one of the Unions adhering to the International Council of Scientific Unions (ICSU), and its regular meetings, at two-yearly intervals, did much to encourage this kind of international collaboration. During the period of sunspot maximum, 1957–8, the ICSU arranged for a year's study of geophysical phenomena on a world-wide scale. It was called the International Geophysical Year (IGY), and study of the ionosphere formed a large part of it. About 200 ionosondes were widely distributed over the globe, and many of them are still in operation.

Anomalous behaviour of the F layer

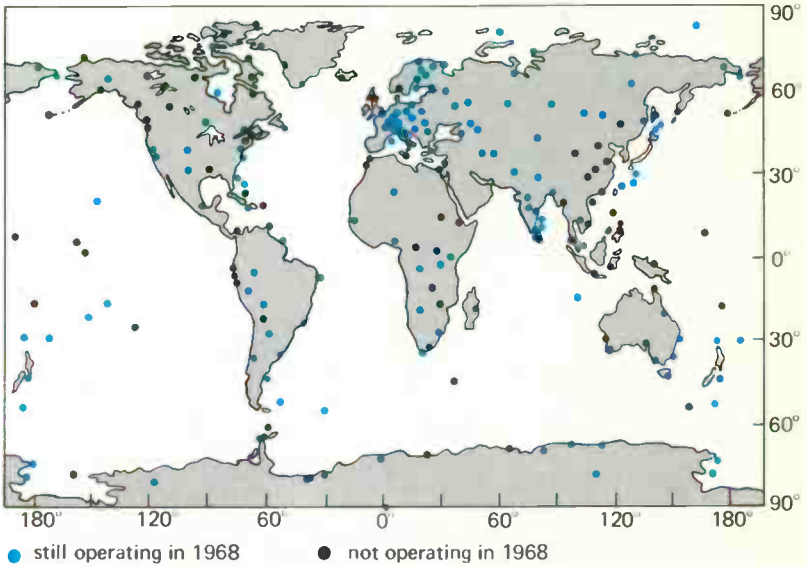
These world-wide observations emphasised what had already been noticed by the first investigators, that the behaviour of the F2 layer is unexpected and is different from that of the E and F1 layers. In the early days, when it was thought that Chapman's theory ought to explain the formation of the F2 layer, it became usual to describe it as a Chapman layer that exhibited a number of anomalies. Later observations showed these anomalies to be so numerous and important that they represent an essential part of the layer's behaviour; moreover present-day theory no longer supposes that the layer, even without the anomalies, is of the 'Chapman' type (see p. 112. In spite of these changes of outlook, however, it is still usual to speak of 'anomalies' in the F2 layer, and the following paragraphs are devoted to describing them.

Most of the anomalies are noticeable in the behaviour of the penetration frequency. If it behaved according to Chapman's theory it would increase towards midday and decrease afterwards, just as it does in the E layer. Often, however, it reaches its maximum value either considerably before, or after, midday, and sometimes there are two maxima, one on each side of noon. Irregularities of this kind are known as diurnal anomalies.

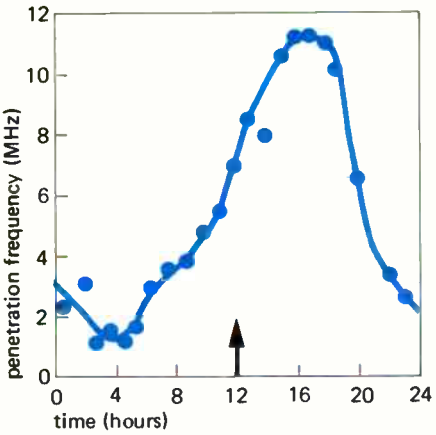
The behaviour of the F layer at night is equally unexpected. After the solar ionising radiation has ceased to reach it* we should expect that the electron concentration would decrease gradually and smoothly throughout the night until the return of the radiation next day. Although some decrease of this kind is observed, it is seldom smooth and often there are increases at times when no radiation is incident at the appropriate heights.

* Because the layer is high in the atmosphere solar radiation may reach it for several hours after sunset at the ground.

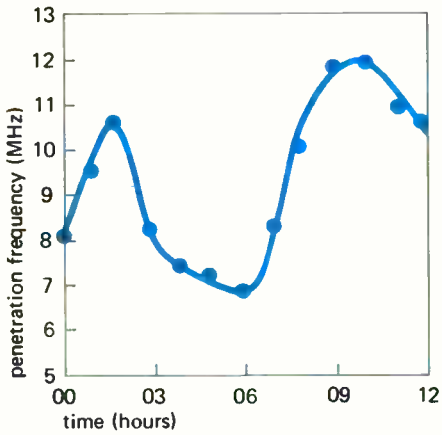
6·1 During the International Geophysical Year (1957–8) ionosondes were operated at the places (including many on small islands in the oceans) shown on this map.



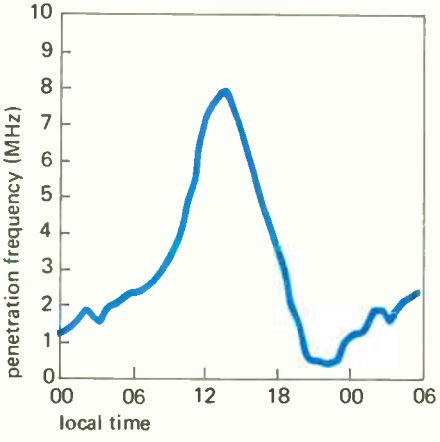
This anomalous behaviour is particularly noticeable in the polar regions during the winter, when no radiation can reach the layer for several weeks on end. The concentration of electrons at the peak of the layer then varies through the 24 hours, much as it would if radiation were arriving; thus at the time when the sun is most nearly rising (what would be midday if it did rise) the concentration is greatest and at times when it is furthest from rising (corresponding



6·2 Left The diurnal anomaly in the F2 layer. The average values for five magnetically quiet days observed at Mauii, Hawaii, in June 1954 reach a maximum in the afternoon.

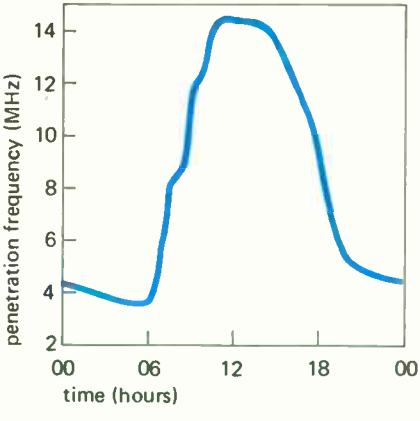


6·3 Left The night anomaly in the F2 layer. The penetration frequency measured at Talara on 28 June 1958 increased during darkness. The measurements also show an example of a diurnal anomaly in which the frequency reached a maximum in the morning hours.

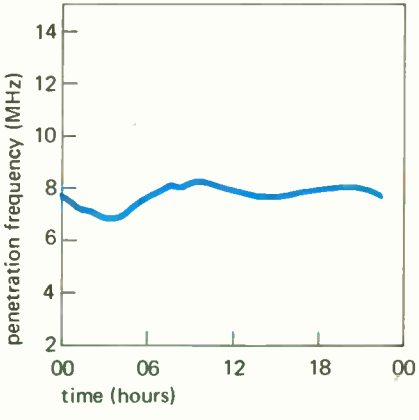


6·4 Left The polar anomaly in the F2 layer. Although the layer is not illuminated during the period of winter darkness, in the Antarctic the penetration frequency reaches a maximum value near 'midday', the time when the sun is most nearly rising.

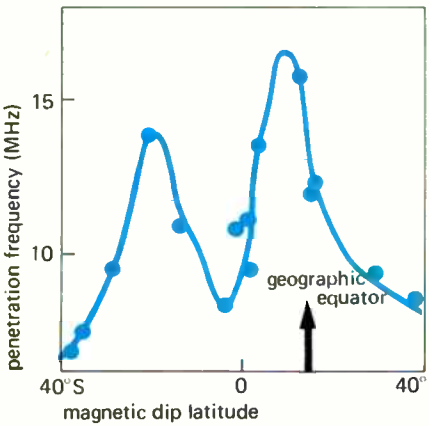
6·5 Below The seasonal anomaly in the F2 layer. Mean values for Slough, England, on magnetically quiet days, show that at midday in June, when the sun was high in the sky, the penetration frequency was less than in December when the sun was low.



December 1956



June 1957



6·6 The geomagnetic anomaly in the F2 layer. At the equinox, when the sun is overhead at midday on the equator, the penetration frequency of the layer is not maximum there, but at places north and south. The minimum between the two maxima is near the *magnetic* dip equator. The curve shows penetration frequencies observed on the American continent at 2100 hours in March 1958 and plotted against dip latitude. The position of the *geographic* equator is shown by the arrow.

to midnight) the concentration is least. There must be some source of electrons that is controlled by the sun, even when its photon radiation cannot reach the layer. In parts of the polar regions during the winter some of the changes in the electron concentration of the F layer occur simultaneously at several places, as though they are related to universal time rather than to the position of the sun.

Studies of the penetration frequency at midday throughout the year reveal another anomaly, known as the 'seasonal anomaly', in which the peak concentration of the layer at mid-summer, when the sun is high in the sky, is *less* than at mid-winter when it is low.

When the extensive wartime measurements came to be examined, they revealed another and most striking anomaly. Chapman's theory suggests that at the equinox the electron concentration at the peak of a layer should be greatest on the equator and should decrease towards the poles in the north and south directions. Measurements show, however, that the concentration at the peak of the F

layer behaves quite differently. As the distance from the equator increases in either direction, the concentration *increases* and reaches a maximum at a latitude of about 20 degrees, and only beyond is there a decrease towards the poles. Instead of an equatorial maximum of concentration there is, in fact, an equatorial minimum, lying between two maxima, one to the north and one to the south.

More detailed examination of the observations revealed the fact that the precise position of this minimum is controlled by the earth's magnetic field. The earth is a huge magnet with lines of force that start in the Southern Hemisphere, end in the Northern Hemisphere and run horizontally (that is, parallel to the earth's surface) at the equator. The magnetic field is, however, not quite symmetrical; it is as though the axis of the terrestrial magnet were inclined to the axis round which the earth spins, so that the magnetic North and South Poles are displaced compared with the geographic ones, and the 'magnetic dip equator', where the field is horizontal, does not coincide with the geographic equator, but is north of it in some places and south of it in others.

When the radio observations were studied in detail it appeared that the equatorial minimum in electron concentration lay, not along the geographic equator, but along the dip equator. This new 'geomagnetic anomaly' demonstrated once more how far the F layer departed from the simple layer of Chapman's theory, and showed that it was somehow controlled by the earth's magnetic field.

It is not only at normal times that the behaviour of the F2 region is difficult to explain: some of the most serious problems arise during magnetic storms. It will be recalled that storms are accompanied by an injection of rapidly moving electrons into the auroral zones, so that luminosity is excited and increased currents flow in those parts of the atmosphere. Both phenomena occur in the E region, at a height near 100 kilometres. At the same time, however, the F region, at a height of about 200 kilometres, is profoundly disturbed, not

only near the auroral zones but over the whole of the earth. The penetration frequency is altered, sometimes it increases and sometimes it decreases, and the layer usually becomes much thicker. It is difficult to account for these phenomena: particularly for the changes remote from the poles and for the reduction in the electron concentration.

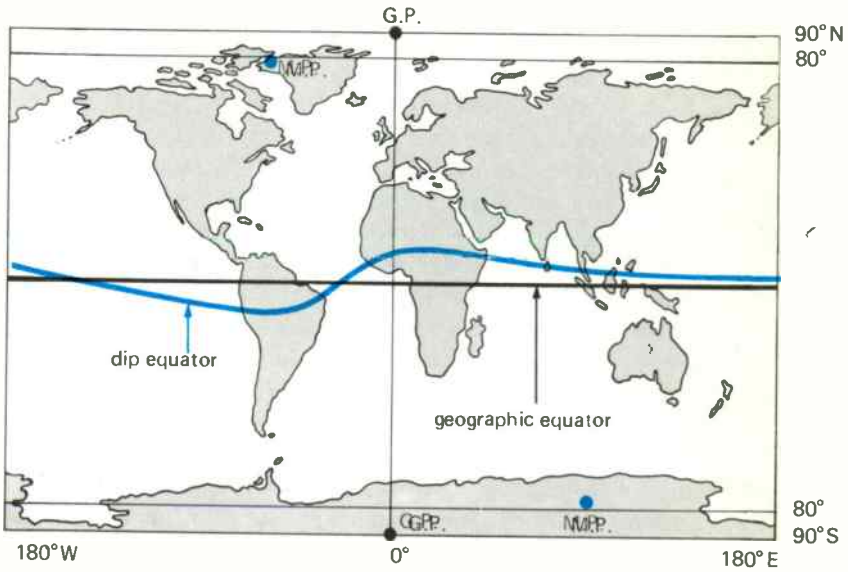
There have been many attempts to explain the anomalies and the storm-time behaviour of the F2 region. Some have ascribed them to changes of temperature of the upper atmosphere, or to ionisation by streams of charged particles superimposed on that produced by ultra-violet, or to mass movements of the air, or to movements of the electrons through the (stationary) air. With one exception, however, there has been no agreement about the reality of the mechanisms suggested. That exception explains the geomagnetic anomaly in terms of movements of electrons in the F layer. Because movements of this kind are closely related to movements in the E layer, and to the dynamo theory of geomagnetism, and have sometimes been suggested as causes for some of the other anomalies and storm-time effects, they are discussed in some detail below.

Theory of the atmospheric dynamo

According to the dynamo theory atmospheric movements drag the conducting ionosphere across the earth's magnetic field so that voltages are induced and currents flow. Calculation of these currents requires a knowledge of the air movements and of the way in which motion of the electrons and ions is controlled by collision with neutral air particles and by the magnetic field of the earth: if these matters are properly taken into account it should be possible to calculate the magnitude of the current and hence of the geomagnetic variations.

The motion of the air at the ground has been investigated by

6·7 The earth's magnetic field is nearly the same as that of a bar magnet at the centre of the earth. The direction of the equivalent magnet is, however, not aligned quite along the direction of the earth's axis of spin, so that the magnetic poles (MP) do not quite coincide with the geographic poles (GP) and the magnetic dip equator, where the earth's field is horizontal, does not coincide with the geographic equator.



analysing measurements of barometric pressure: if the measurements are spread over a long enough time it is possible to sort out the random variations caused by local weather conditions and to isolate regular and persistent oscillations having periods of half a solar and half a lunar day. It is then found that the solar oscillation is about sixteen times as strong as the lunar. At first it was thought that the two oscillations were caused by the gravitational attractions of the sun and the moon, and they were called atmospheric tides, by analogy with the tides in the ocean. In ocean tides, however, it is well known that the lunar one is much the greater, and indeed, since the gravitational attraction of the moon at the surface of the earth is about two and a half times as great as that of the sun, this is just what would be expected. To explain the greater solar 'tide' in the

atmosphere it was at one time thought that the atmospheric gases are so distributed that they tend to oscillate by themselves with a period of half a solar day, so that when the sun tries to drive them with this same period they respond strongly or 'resonate'. Recent measurements of atmospheric composition and temperature at different heights have shown, however, that the gases are, in fact, not distributed in this way and their natural time of oscillation is not sufficiently near to the half solar day for this explanation to be valid: it is now believed that the lunar oscillation is a tidal one, driven by gravity, whereas the solar one is driven by the heat that is put into the atmosphere when the sun's radiation falls upon it. Since the two oscillations are now known to be driven in different ways, it is no longer surprising that the solar is the larger.

It was possible even before the Second World War to use estimates of the magnitudes of the solar oscillation and of the electric conductivity of the ionosphere (based on radio measurements of the electron concentrations at different heights) to calculate the currents that would flow in the atmospheric dynamo, and to see whether they would produce the observed diurnal changes of terrestrial magnetism. Detailed calculation shows that the oscillation at a height of 100 kilometres is about 200 times greater than at the ground, but even with this enhanced movement, the calculated changes of the magnetic field were less than the observed by a factor of at least 400; something was clearly wrong. Several years elapsed before the error was traced, and it was not until about 1950 that a satisfactory explanation was found. It was then realised that although currents could flow in the ionosphere horizontally they could not flow vertically, for if they did they had nowhere to flow to, and when this fact was introduced into the theory it was found to modify the motions of the electrons and ions, and to make an important difference to the magnitude of the calculated conductivity. When the revised conductivity was used in conjunction with the estimated

magnitude of the air speed at 100 kilometres, the calculated geomagnetic variations were found to be similar to the observed ones. One important result was that, near the equator where the earth's magnetic field is horizontal, the conductivity, and hence the currents, would be much enhanced. This conclusion provided a satisfactory explanation of an old observation that the geomagnetic variations near the equator were abnormally large.

The revised theory indicated that the conductivity was greatest at heights near 130 kilometres so that the currents would be expected to flow mainly near that level. When rockets became available for taking equipment into the high atmosphere it was possible to locate the level of the dynamo currents experimentally; indeed, the first experiment to be performed in a rocket was of this kind. It was made in 1949 and employed a magnetometer to measure the strength of the magnetic field at different heights, and to transmit the results to the ground by radio telemetry. As the height increased it was found that the strength of the field at first decreased as would be expected if the earth behaved like a magnetised sphere. (The strength was inversely proportional to the cube of the distance from the centre of the earth.) Above 100 kilometres, however, the field began to vary with height in a different way, and it was concluded that the rocket had then entered the region in which the dynamo currents were flowing. Since that time there have been other experiments of a similar nature: all have shown that the currents flow at heights around 130 kilometres, as theory would suggest.

Movements of the F layer, and the atmospheric motor

The permanent magnetic field of the earth not only influences the direction and magnitude of the current; it also exerts a force on it and causes it to move, just as the current in the armature of an electric motor is caused to move when it is under the influence of the

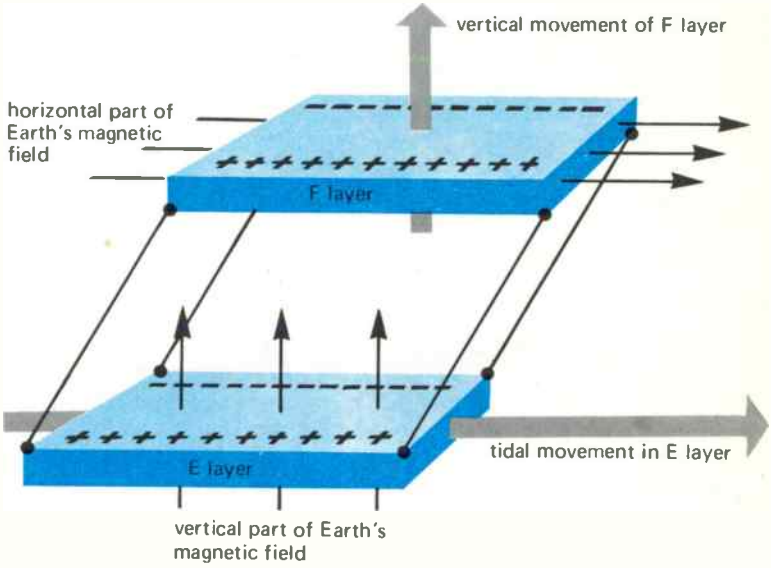
field magnets. This 'motor action' causes a bodily movement of the ionospheric region in which the current flows, and it has been invoked to explain some of the small departures from strict Chapman-like behaviour that have been noticed in the E layer. A similar motor action in the F layer has been suggested to explain some of the anomalies there, which can by no means be described as small departures from a simple model. Let us now see how an atmospheric motor could exist in the F layer driven by the dynamo in the E layer.

If the atmospheric dynamo, situated at a height of about 100 or 130 kilometres, could be connected by wires to the F layer above at heights of about 250 kilometres, it could drive current through it. The current would be much smaller than that in the E layer, and need not be taken into account when considering the geomagnetic variations, but it would be large enough to experience an important force in the presence of the permanent magnetic field. It would constitute an atmospheric motor in the F layer whose movements might explain at least some of the anomalies found there.

Now it so happens that the ionosphere between the two layers can in some respects be thought of as providing 'connecting wires' to join the 'dynamo' in the E layer to the 'motor' in the F layer. This possibility arises because the ionospheric conductivity is greatest along the direction of the magnetic field, and because (over most of the earth) the lines of magnetic force run obliquely, so that they act like wires joining the layers. Once this idea is accepted it is possible to calculate what movements of the F layer might accompany the dynamo current in the E layer.

The calculations have been made in two different ways: one starts with a knowledge of the geomagnetic variations, deduces the currents in the dynamo, inserts what is known of the conductivity in the E layer to derive the voltage which drives these currents, and which is also applied to the F layer, and from this deduces the movements of

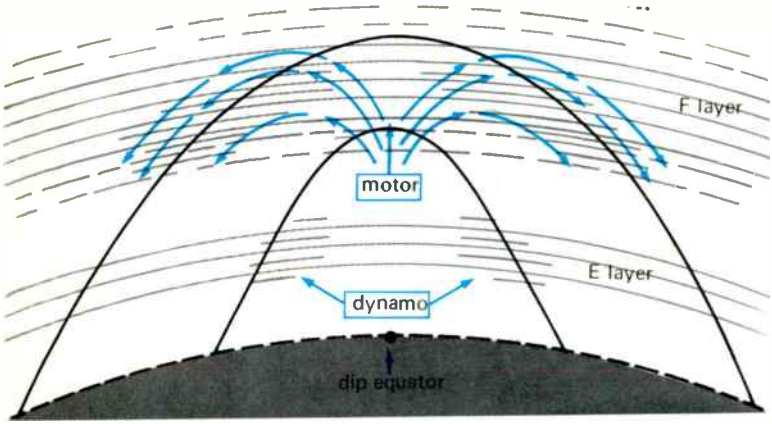
6·8 A model of the atmospheric dynamo in the E layer and the motor in the F layer. Tidal movement of the E layer, combined with the vertical part of the earth's magnetic field, produces a voltage difference which is conveyed to the F layer, as if by wires, along sloping magnetic field lines. The resulting current in the F layer is acted upon by the horizontal part of the earth's field to cause a vertical movement of the layer.



that layer; the other starts with an assumption about the movements in the E layer, deduces from that the voltage applied to the F layer, and then proceeds as before. Both approaches show that movements of considerable magnitude must occur in the F layer but, with one exception, it has not been possible to relate them at all closely to the observed anomalies.

The exception is the geomagnetic anomaly in which, when the sun

6·9 The equatorial anomaly explained as a 'fountain effect'. During the day the action of the atmospheric dynamo in the E layer moves the electrons in the F layer 'motor' upwards over the dip equator and they then diffuse downwards along geomagnetic lines of force so as to accumulate at two places, one on each side of the equator.



is overhead at the equator, the maximum electron concentration in the peak of the F layer is found, not at the equator, but on lines that are approximately parallel to the geomagnetic dip equator and are removed from it by about 20 degrees. To explain this phenomenon it has been pointed out that the distribution of electrons depends, not only on the rates of their production, loss, and diffusion, but also on the movements resulting from the motor action: and that diffusion occurs only with difficulty across a magnetic field so that downwards diffusion will be not along the vertical but along the field lines. Following up these ideas it has been shown that the ionisation produced by day in the F layer at the equator is first

driven upwards by the action of the dynamo beneath and then diffuses obliquely downwards along the lines of force, to build up maxima of electron concentration on two lines, one at each side of the equator. A movement of this kind has been described as a 'fountain effect'. This explanation was suggested soon after the war, but detailed calculations of its magnitude had to await the availability of digital computing methods: it has now been worked out in detail and it seems to account for the observed facts. This agreement between theory and observation is perhaps one of the most cogent reasons for belief in the correctness of the idea of the atmospheric motor and the importance of diffusion near the peak of the F2 layer.

7 The earth's environment

7 The earth's environment

The picture in 1969

Before about 1957, when experiments were first made in artificial satellites, the state of the atmosphere above 200 or 300 kilometres (the F2 layer peak) could only be inferred from ground-based measurements of what lay below. Extrapolation upwards led to the belief that as the atmosphere becomes less dense at great heights, it gradually merges into empty space, or rather into interplanetary space, populated by about one atom in each cubic centimetre. At all times the sun influences the earth's atmosphere through photon radiation, and occasionally also through jets of charged particles that impinge on the earth's atmosphere to produce storms in the ionosphere, accompanied by magnetic irregularities and aurorae. There were, however, no other interactions between earth and sun. In particular the magnetic fields of the earth and the sun each exerted a controlling influence on the ionised gases in its own neighbourhood but neither extended far enough to influence the gas surrounding the other body.

Since the first satellite was launched this picture has greatly changed. The advances in knowledge have been so rapid and fundamental that it would be surprising if it were yet correct in all details. What is described here represents the situation as it is generally accepted in 1969, and although the details may change during the next few years the outline of the picture is not likely to be seriously altered.

The main differences between the old and the new concepts arise because it is no longer thought that the earth and the sun are completely separate bodies with nothing but 'empty space' between them. The earth is now supposed to be immersed in a stream of charged particles flowing steadily outwards from the sun. This *solar wind* distorts the sun's magnetic field and pulls it out along a radial direction so that even when it reaches the earth it is still strong enough to

be important. On arrival at the earth the wind wraps round the geomagnetic field, and envelopes it, as a stream of air wraps round and envelopes a solid obstacle in its path. The geomagnetic field, however, is more like a deformable object than a solid one, and it is distorted by the solar wind passing over it. The region of this distorted geomagnetic field is called the *magnetosphere*. In parts of it charged particles can oscillate back and forth trapped by geomagnetic lines of force; and particles from the solar wind can enter these trapping regions in ways not yet fully understood.

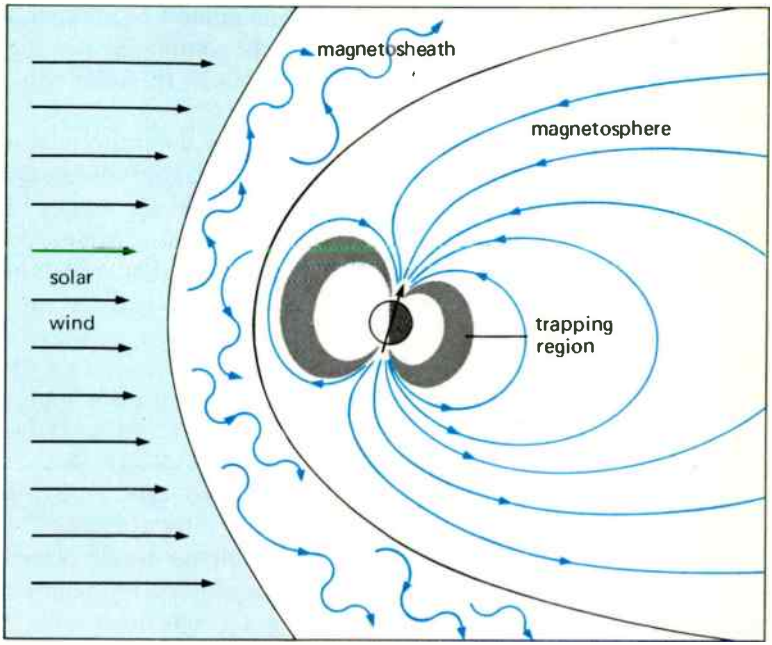
During a solar disturbance bursts of charged particles emitted from active areas on the sun are superimposed on the normal solar wind. In their approach to the earth they modify the solar magnetic field, already stretched out towards the earth by the normal wind, and when they impinge on the magnetosphere they are responsible for magnetic storms and aurorae. One result of their arrival is to speed up some of the particles that are oscillating in the trapping regions, so that they leave those regions and enter the ionosphere with sufficient energy to produce the aurora and storm effects.

In this chapter we elaborate some details of this modern picture. The reasoning which has led to it is naturally detailed and complicated; although mention is made of the kind of experimental observation on which it is based, no attempt is made to present it in detail. Attention is directed first to the environment of the sun, second to that of the earth, and finally to the intermediate region where the two environments meet.

The sun and the solar wind

The first suggestion that a 'wind' might be continually blowing outwards from the sun came from observations of the tails of comets. It had been known for a long time that these pointed in a direction away from the sun, a fact which had been attributed to the pressure

7·1 When the solar wind encounters the magnetic field of the earth, a shock wave is set up (the magnetosheath) and the wind flows round it. The region inside the magnetosheath (into which the wind does not pass) is called the magnetosphere. In it the earth's magnetic field is distorted and there are regions in which rapidly moving charged particles can be confined.



that the photons of sunlight exerted on the gases of a comet's tail. Some comets, however, have two tails, only one of which can be ascribed to the pressure of the photons, and in 1951 it was suggested that the second was caused by interaction between charged particles in the comet and a stream of charged particles moving outwards from the sun. From a measure of the direction of the tail the speed of these particles was calculated to be about 300 or 500 kilometres per second. The next step in the argument came from theoretical investigations of the conditions in the corona, the sun's ionised atmosphere. Spectroscopic and other evidence had suggested

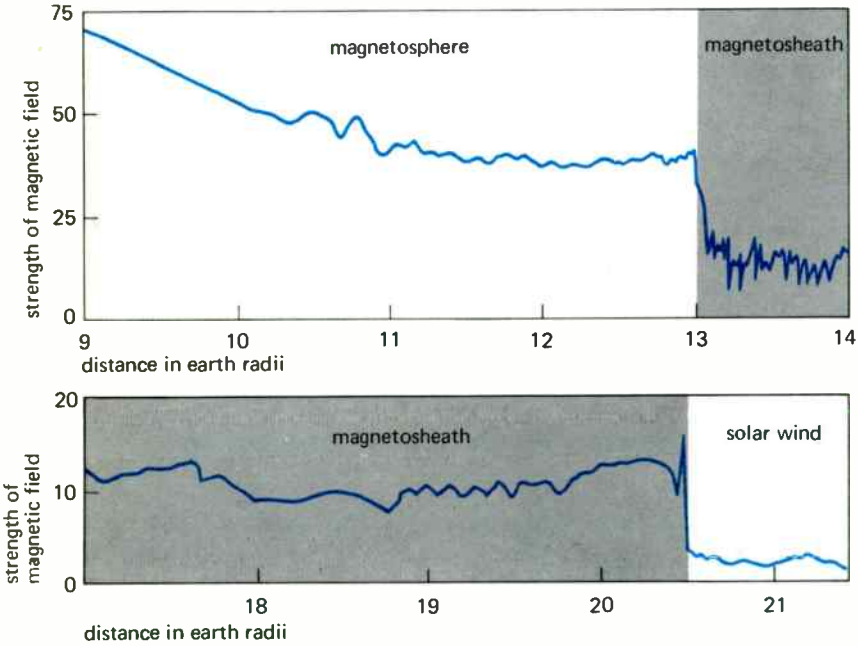
7·2 The Mrkos comet in August 1957. The straight tail consists of ionised gases blown away from the comet by the solar wind. From the fact that the tail does not lie precisely in a radial direction from the sun the speed of the wind can be estimated. The more diffuse, curved, tail results from the pressure of sunlight on the uncharged gases in the comet.

that its temperature might be as great as one million degrees, and theory showed that the ionised gas must then be continually moving outwards with a speed roughly the same as that of the solar wind calculated from observations of the comets.

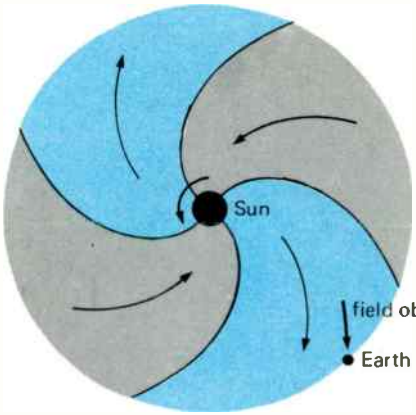
The wind consists of charged particles, electrons and protons, and is therefore a good conductor of electricity, and it was next necessary to consider how it might affect the magnetic fields of the earth and the sun. When a conductor moves in a magnetic field, currents are induced in it, just as they are in the moving wires of a dynamo, or in the atmospheric dynamo low in the ionosphere. These currents produce their own magnetic field and if, for any reason, they are moved, the field moves with them. This phenomenon is illustrated by a well-known demonstration experiment which involves the use of a superconducting ring of lead. The lead is placed over a bar magnet and is then cooled to a very low temperature so that it loses all its resistance (it becomes superconducting). A current is then induced in the ring by withdrawing the magnet, and the presence of this current can be detected through the deflection of a small compass needle placed nearby. If the ring is then moved it carries its current and its magnetic field with it, so that wherever it is it will deflect a compass needle. It can be said that the field has been 'frozen in' to the ring through the agency of the circulating current that has been induced in it. In just the same way when currents are induced in the highly conducting solar wind they move with it and carry a magnetic field along with them. The result is that the solar magnetic field 'frozen into' the wind becomes distorted by the wind's movement so that the lines of force are stretched radially from the sun.

In the solar wind the distribution of the magnetic field is determined by the distribution of the moving particles. There can be an inverse situation in which the distribution of moving particles is determined by a magnetic field. Near the earth, for example, the charged particles of the high atmosphere spiral along the earth's

7·3 Measurements made in space probes show that the magnetic field strength decreases as the distance from the earth increases. There are two sharp changes in this decrease, one at a distance of about 13 earth radii (the boundary of the magnetosphere) and one at about 20·5 earth radii (the outer boundary of the magnetosheath).



magnetic lines of force so that ultimately their distribution is controlled by the shape of the field, instead of conversely as in the solar wind. What is the reason for this important difference? It is concerned with the relative amounts of energy stored in the moving particles and in the magnetic field. A stream of particles possesses energy by virtue of its motion, and if it is brought to rest the energy appears in another form. For example, if a stream of bullets from a machine-gun is stopped by a sheet of metal the energy of motion is converted into heat and the temperature of the metal increases. Energy is also stored in any place where a magnetic field exists, even



7·4 The solar wind carries the sun's magnetic field outwards in a path which takes a spiral shape because the sun is rotating. During 1963–4 the field was directed alternately towards and away from the sun in neighbouring sectors. As these were carried round by the solar rotation, they caused the field observed near the earth to vary in a way which repeated after 27 days.

if it is empty space. This fact is not so obvious, but is demonstrated in a striking manner if the current flowing in the coils of a large electromagnet is suddenly interrupted by the opening of a switch; a violent spark and an electric discharge, passing between the points of the switch, are then evidence of the energy that has to be dissipated as the magnetic field is removed from the magnet.

If in any volume the energy residing in the moving particles greatly exceeds that of the magnetic field, then the particles will determine the shape of the field; whereas if the magnetic field has the greater energy it will determine the distribution of the particles. In the solar wind the particle energy is the greater and the wind drags the field lines with it; near the earth, in the magnetosphere, the energy of the field is the greater, and the distribution of the ionisation is determined by it.

The intimate relationship between the particles and the magnetic field suggests that, if the field can be investigated, for example by

means of a space probe, it should be possible to make deductions about the distribution of the associated charged particles. Many space probes have carried magnetometers to measure the field for this purpose. It is found that, out to about 13 times the earth's radius, the strength of the field decreases in inverse proportion to the cube of the distance (figure 7·3), as would be expected on the well established supposition that it is similar to that of a small bar magnet placed at the centre of the earth.* At greater distances, however, there are two rather sudden changes. First, near about 13 earth radii, it becomes irregular and smaller than would be expected from a bar-magnet field, and second, further out, at about 20 radii, it decreases again and becomes smooth and more or less independent of distance. The innermost region is in the magnetosphere, the most remote is in the solar wind, and the region between, where the field is irregular, is called the *magnetosheath*.

In the solar wind, the magnetic field is directed roughly, but not exactly, radially towards or away from the sun (figure 7·4). The precise direction is explicable in terms of the fact that the sun rotates on its axis once in about 27 days; the shapes of the magnetic lines of force therefore have a shape similar to that of a jet of water projected radially from a rotating garden hose. The angle between the wind and the sun-earth line is then determined by the speed of rotation of the sun and by the speed of the solar wind. An interesting observation is that the magnetic field is sometimes directed towards, and sometimes away from, the sun, any one direction being maintained for about five or seven days, so that in a single solar rotation there might be four or five reversals. The implications of this observation are not yet clear, but there is little doubt about its importance.

* Although the currents in the atmospheric dynamo produce a measurable deviation from the inverse cube law near the current sheet (see p. 171) they have a negligible effect at greater heights.

The solar wind and the earth

To understand what happens when the solar wind impinges on the earth it is useful to consider how a stream of air is deviated so as to flow round a solid obstacle in its path. The deviation is caused by a wave that travels back from the obstacle into the oncoming stream, as though it carried a message that the obstacle is in the way and must be avoided. In air this wave travels with the speed of sound. If, however, the air stream has a speed greater than the speed of sound the air particles arrive at the obstacle before they have received the message that it is there: the flow lines continue straight, right up to the surface, and when they reach it they have to make an abrupt change of direction. The sudden change of direction, accompanied by a sudden change of pressure and of the speeds of the particles, occur in a thin layer, called a shock wave. A shock wave is always produced if the relative speed of the air and the obstacle is supersonic; the air might, for example, be at rest, and the object might be an aircraft moving through it with supersonic speed: the arrival of the shock wave at the ground then produces the well-known supersonic bang.

The situation with the solar wind is similar, except that ordinary sound waves cannot travel in it; the appropriate type of wave, called a hydromagnetic wave, is one that can travel in a medium, like the solar wind, containing charged particles and a magnetic field. The speed of the hydromagnetic wave in the wind is considerably smaller than the speed of the wind itself so that when the wind encounters the earth a shock wave is set up. Outside this wave the wind is unaltered; inside its direction changes suddenly.

There is another difference between the impact of the solar wind on the earth and the impact of a neutral wind on a solid sphere. It is the magnetic field and not the solid earth itself that acts as the obstacle in the path of the solar wind. It is to this obstacle, round which

the solar wind flows, that the name magnetosphere is given.

The shock wave in the solar wind is not formed precisely at the place where the magnetosphere starts, but is some distance in front, and between it and the magnetosphere there is an intermediate region where the streaming motion of the particles becomes converted into a random motion. This region is the magnetosheath, and in it the randomly moving particles carry with them an irregular magnetic field that can be detected when a space probe passes through, as shown in figure 7·3, at distances between 13 and 20·5 earth radii.

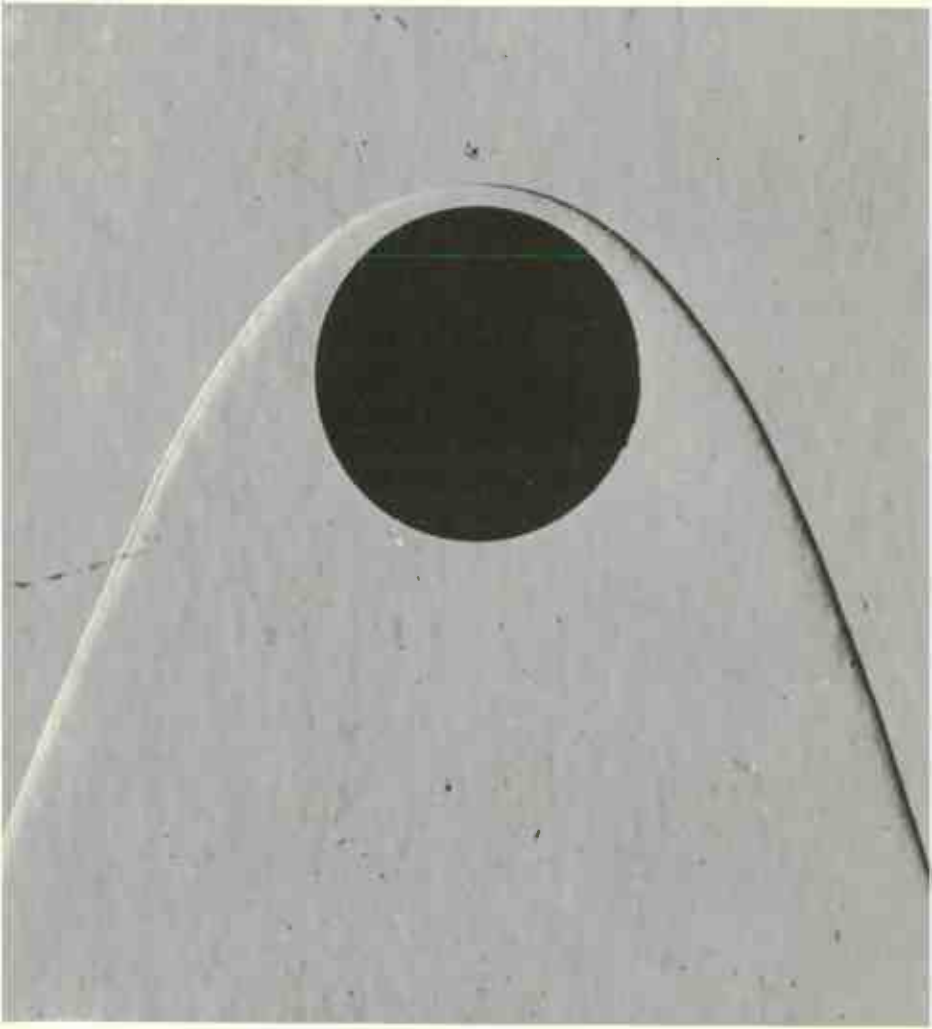
The magnetosphere

On the earthward side of the magnetosheath the particles of the solar wind do not penetrate the geomagnetic field. This region is the magnetosphere, and in it the field energy exceeds the particle energy, so that motions of particles are controlled by the field. It can be thought of as an isolated region, containing particles and magnetic field, that acts as an obstacle round which the solar wind flows. It has a shape determined by the earth's magnetic field but is distorted by the solar wind blowing past it and is more like a deformable balloon than a rigid obstacle: it is pulled into the shape of a pear, with the stalk pointing away from the sun. Because the axis of the earth's magnetism makes an angle with the direction of the solar wind the final form of the distorted magnetosphere is as shown in figure 7·1.

Inside the magnetosphere the motion of charged particles is controlled by the geomagnetic field, and to a first approximation it can be said that particles neither enter nor leave it from the solar wind: this statement is refined later.

Below the magnetosphere there is the ionosphere, where a moving electron makes so many collisions that guidance by the geomagnetic field is comparatively unimportant. The ionosphere merges into the

7·5 A shock wave in air: the photograph of a shock wave in a supersonic wind tunnel when a stream of air moving with a speed 8·6 times the speed of sound impinges on a sphere.



magnetosphere at heights where the collisions become sufficiently infrequent to allow the magnetic field to play an important role.

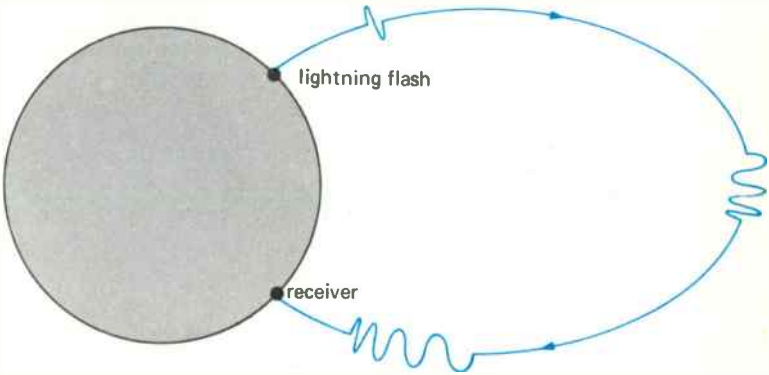
Many of the atoms in the magnetosphere exist in the form of ions, positively charged. When an electron, also charged, approaches one of these its path is bent: the bending is less if it is travelling more rapidly. Bending of the path, of this kind, constitutes a 'collision' and interferes with the geomagnetic control. The more rapidly moving electrons, which suffer least deviation, are thus more closely tied to the geomagnetic field lines: some of them, leaving the ionosphere at one end of a line, travel along it with a spiral motion, and reach the ionosphere at the magnetic 'conjugate point' in the opposite hemisphere. They are discussed on p. 137.

Under some circumstances (p. 193) the rapidly moving particles travel back and forth along field lines in the magnetosphere without penetrating the ionosphere: they are said to be 'trapped' and unless something happens to release them from the trap they will remain in the magnetosphere indefinitely. The regions in which the trapping occurs are often called Van Allen belts, after the investigator who first discovered them.

Whistlers and magnetospheric electrons

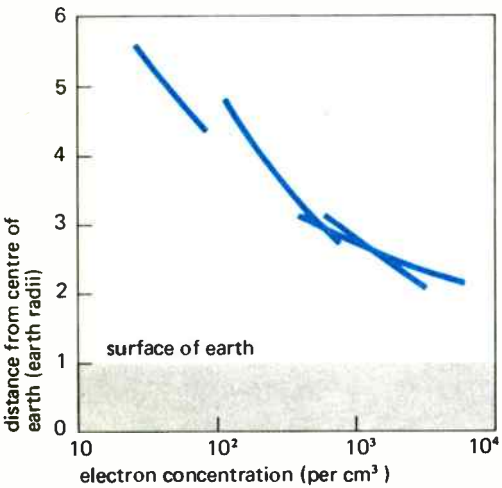
The concentration of electrons in the magnetosphere can be measured from the ground, in a simple way, by observing those naturally occurring electromagnetic radiations (known as 'whistlers') that can be heard if a long wire is attached to a telephone receiver, preferably through an audio frequency amplifier of the kind used in gramophones. Each is a whistle of descending pitch that lasts about one second: it can be roughly represented by the word 'pee-oo'. In the past they have frequently been observed by chance. A particularly early and extensive series of observations was recorded in 1886 by those who used the telephone to the observatory on the top of the

7·6 The production of a whistler. A lightning flash radiates an impulsive electromagnetic wave, which travels along a geomagnetic field line through the electrons of the ionosphere and magnetosphere. As the pulse travels, it develops oscillations, more rapid in front and slower at the rear, and is received at the other end of the field line as a whistle whose pitch is first high and then low. The behaviour of the pulse is similar to that of the pulse produced in the surface of water when a stone is dropped in.



Zugspitze mountain in Austria; they have also been heard on submarine telegraph cables. More recently they have been studied systematically and the changes in pitch have been recorded.

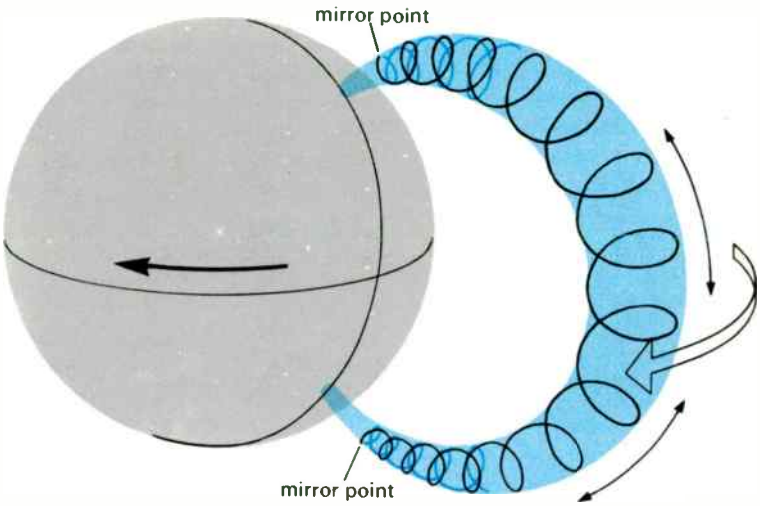
In 1953 it was shown that a whistler is produced when the wave radiated from the electric discharge in a lightning flash in one hemisphere travels, along a geomagnetic line of force, to the other hemisphere, and that the nature of the whistle depends on the electron concentration in the ionosphere and magnetosphere through which it passes. A lightning flash radiates electromagnetic waves with frequencies covering a very wide range; some, for example, are received on an ordinary radio receiver, where they are heard as 'atmospherics'. Others have frequencies so small that they correspond to audible sound: there is no need to use a radio receiver to detect them, a wire connected to a telephone suffices. When the flash occurs all frequencies are radiated simultaneously and the noise in a nearby receiver sounds like a bang and not like a musical note. When the waves reach the ionosphere those of audio frequency are compelled to travel along the lines of force of the geomagnetic field,



7·7 The electron concentration in the magnetosphere deduced, by four different investigators, from observations of whistlers.

and the higher frequencies travel more rapidly than the lower. They follow the field lines out to a great distance and down into the other hemisphere, and it is there that they can be received as a whistler, with the higher frequencies preceding the lower. The time lag between the reception of the high and the low frequencies depends on the total distance travelled, and on the way in which the electrons are distributed along the line of force that is followed, particularly on the electron concentration at the topmost part of the line, where it is farthest from the earth. By making observations at different places, where the lines of force reach out to different distances, it is thus possible to investigate the electron concentrations at places remote from the earth. Some results are shown in figure 7·7. The measured concentrations refer to the slower electrons, having speed like those in the ionosphere; they far outnumber the high-speed electrons trapped in the Van Allen belts.

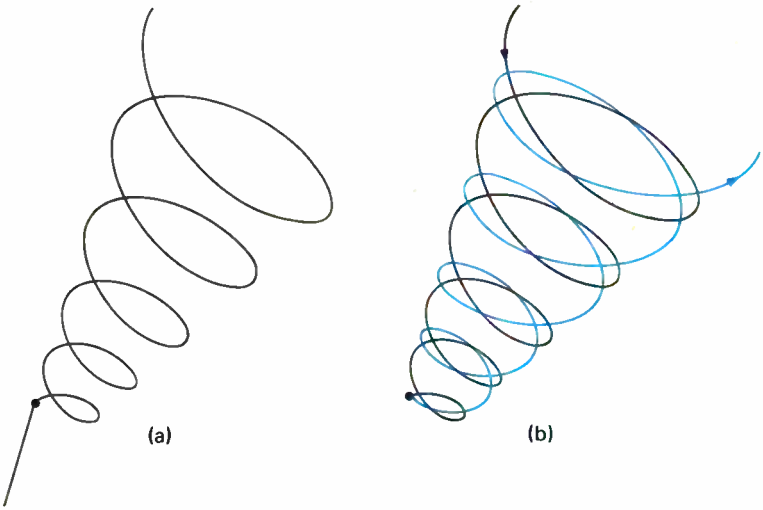
7·8 Charged particles moving with the right speed and direction can be trapped by the geomagnetic field so that they move in a spiral along a field line, bouncing back and forth between two mirror points. The spiral is more closely wound at the ends than at the centre, and the whole spiralling motion drifts round the earth from west to east for electrons and from east to west for protons.



Trapped particles

Those charged particles that are moving sufficiently rapidly make few collisions in the magnetosphere and travel along geomagnetic field lines in spirals, which can be closely or loosely wound, according to the circumstances. If the field lines run parallel to each other, any one particle continues to spiral along one line, always in the same direction and in a spiral of the same closeness. But the earth's field lines are not parallel; they crowd together at the ends, near the poles, so that a particle's motion is more complicated. As it moves in to a region where the lines are closer together its spiral path becomes tighter and, if conditions are right, it may finally become so tight that the forward motion stops completely, and only the circular motion around the line remains. At that point the particle starts to progress along the line in the backward direction, and the spiral opens out again as it travels. At the other end of the line the process is repeated,

7·9 If a trapped particle has its mirror point sufficiently deep in the atmosphere it may collide with an atmospheric particle, so as to be removed from its trapped spiral motion, and be 'dumped' into the lower atmosphere, as at (a). If the mirror point is high the chance of collision is small and the particle will probably retrace its path, as at (b).



again the spiral closes up until the forward motion stops, and the particle travels back along the line of force once more. In this way the spiralling particle is kept continually bouncing back and forth from one 'mirror point' to another, moving along a line of force in a spiral that is loosely wound at the centre and closely wound at the ends.

If a mirror point is so near the earth that it is inside the ionosphere, where there is a high probability that a particle will collide with an ionospheric constituent, the motion is seriously disturbed, the particle ceases to be trapped, and, yielding up its energy in ionisation or illumination, it is lost to the magnetosphere; it is said

to be 'dumped' into the ionosphere. But if the mirror point is at a sufficiently great height the particle continues to bounce back and forth without interruption; it is then said to be trapped.

Trapped particles of this kind can be observed by means of apparatus carried in satellites, and it is important to know which, if any, are likely to be deposited into the ionosphere. The desirable information can be derived from observations of the tightness of the spiral at any point along its length, for if this is known the position of the mirror points can be calculated. The particle will penetrate closer to the earth if the spiral is more loosely wound.

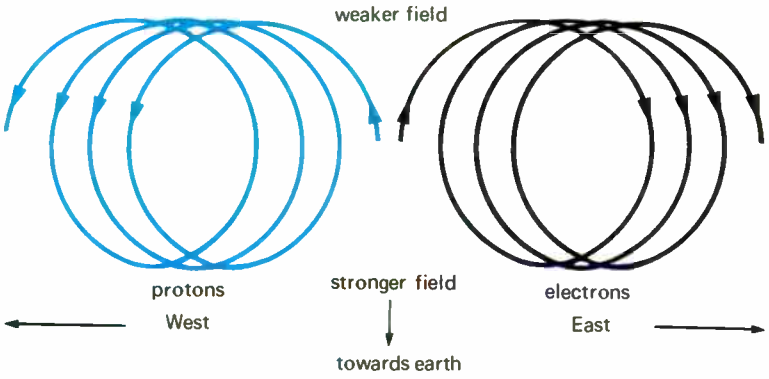
We have not yet finished with the motion of the trapped particle, for it has another component of importance for our story. The radius of the circular motion in the spiral depends on the strength of the magnetic field; it is less when the field is greater. To understand the importance of this let us look at the spiralling particle along the direction of a line of force (see figure 7·10). We expect to see the projection of its motion as a circle whose radius depends on the strength of the field, but we must remember that the field is slightly weaker at places more distant from the earth, so that the radius of the circular motion is slightly larger on the outside than on the inside. The motion is, in fact, not strictly circular at all: it takes a looped form that corresponds to a gradual movement of the circle to one side. While this is happening the bouncing back and forth along the field line continues: the result is that the bouncing spiral motion gradually drifts sideways and ultimately moves right round the earth. There is an important difference between the behaviour of electrons and protons in this context. Because they have different signs of electric charge, they execute their circular motions in opposite senses, and since the direction of the sideways drift is related to the direction of the motion in the circle, the two kinds of spiralling particles drift in opposite directions. The paths of the electrons drift round the world to the east and those of the protons towards the

west, and the different signs of the two charges makes this double drift equivalent to a current flowing from east to west in a ring round the earth.

The different components of these combined motions occur at different rates. The time of rotation round a line of force does not depend on the speed of the particle, but is determined only by its mass and the strength of the magnetic field. The strength of the field at a distance of about four times the earth's radius is such that the gyration time of electrons is about 10^{-6} seconds and of protons about 10^{-3} seconds. The time of oscillation back and forth between mirror points depends on the positions of the points; if they are located about 500 or 1,000 kilometres above the earth, the time taken for both electrons and protons to bounce back and forth is about one minute. The time taken for the spiral path to drift sideways right round the earth is about 30 minutes.

The region inside which electrons and protons can be trapped is shown schematically in figure 7·1. It does not occupy the whole of the magnetosphere; its precise position is determined by the depth at which the mirror points can be situated without the particles being dumped by 'scattering'. This depth depends, of course, on the geomagnetic field, and if this were symmetrically distributed over the earth, the trapping region would be symmetrical. But the earth's field is, in fact, unsymmetrical, and this is for two reasons. First it is distorted by the drag that the solar wind exerts on the outside of the magnetosphere, so that the trapping region on the night side is drawn out, and that on the day side is compressed. The result is that the lowest part of the trapping region lies in rings, that, instead of being circles round the two magnetic poles, have the form of ovals stretched out towards the night side of the earth. If any leakage were to take place by 'dumping' of particles out of the trap, it would be expected along these ovals, since it is there that the region penetrates most deeply into the ionosphere.

7·10 The motion of a charged particle as seen by an observer situated on the geomagnetic equator and looking towards the north. The magnetic field is directed into the paper and the particles rotate nearly in circles. At the greater distance from the earth the field is a little weaker and the path is less curved, so that the circular motion gradually moves sideways. Electrons and protons which rotate in opposite directions move sideways in opposite senses.



But that is not the end of the story. Quite apart from any distortion resulting from the solar wind, the magnetic field of the solid earth itself is not strictly symmetrical. It is only to a first approximation that it can be likened to the field of a bar magnet at the centre of the earth: in particular there is a region of abnormally weak field in the South Atlantic ocean. The reason for this anomalous feature is not known, but the occurrence of the anomaly is important because over it the trapping region comes much lower than anywhere else on the earth, and if particles were to be dumped into the ionosphere it is there that they might be expected.

Solar disturbances and ionosphere storms

The preceding paragraphs have described the earth's environment at times when the sun is quiet. When it is disturbed the environment is modified so that magnetic storms, ionosphere storms and aurorae occur: they are discussed in turn below.

A disturbance on the sun is accompanied by the expulsion of an energetic stream of charged particles, mainly electrons and protons, that travel through the solar wind towards the earth. The particles in the stream are much more numerous and are moving more rapidly than those in the quiescent solar wind, and when they impinge on the magnetosphere they compress it on the daylight side. The compression is accompanied by a sudden increase in the geomagnetic field at the ground: this is the first indication of a magnetic storm, it is called a 'sudden commencement', and is experienced nearly simultaneously all over the earth. It does not occur quite simultaneously because the compression of the earth's field has to travel, with a finite speed, as a hydromagnetic wave motion through the magnetosphere and the ionosphere.

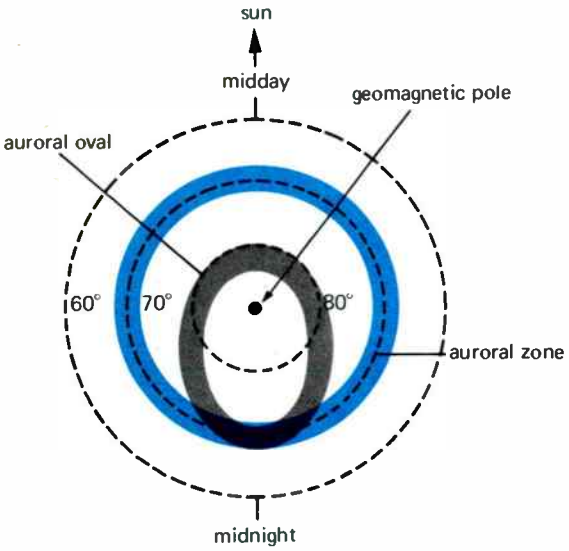
The next thing that happens is that somehow, by processes not yet properly understood, some of the particles from the solar stream enter the trapping regions of the magnetosphere, where they begin to oscillate and drift in the manner previously described. The trapped electrons drift to the east and the protons to the west, so that between them they correspond to a ring of current surrounding the earth roughly centred over the equator, at a distance of about three or four earth radii. This 'ring current' produces a magnetic field at the ground, observable as part of the storm: it is called the main phase of the storm.

At intervals during the main phase, energetic particles leave the magnetosphere and enter the ionosphere along the night side of the auroral oval, where the trapping region comes nearest to the earth.

These excite the light of the aurora; they also produce a local increase in the conductivity of the dynamo region, so that a strong 'polar electro-jet' current flows, and makes itself evident at the ground through small changes of magnetic field, called polar sub-storms. As a sub-storm progresses the auroral luminosities move to fill the oval more completely.

It is not at all clear where the energetic particles that enter the ionosphere come from. They are not simply the storm particles emitted by the sun for, as was realised long ago, their energies are much greater than would correspond to the time taken by those particles to reach the earth. They must be particles that have somehow been stored in the magnetosphere and have then been released and subjected to some accelerating mechanism, triggered by the arrival of the storm particles. The fact that the aurora appears first at the night-time edge of the polar oval has suggested to some investigators that the energetic particles originate and are accelerated in the long tail of the magnetosphere on the side remote from the sun.

According to these views, the aurora on any one occasion is distributed roughly along the auroral oval, where the distorted trapping region comes nearest to the ground, and is strongest on the night side of that oval. It had previously been observed that aurorae occurred most frequently along the circular 'auroral zones' removed in latitude from the magnetic pole by about 22 degrees. These two apparently contradictory statements have been reconciled by noting that the ovals extend on the night side to the auroral zones and that the earth rotates under them so that, as seen from the earth, these parts of the oval coincide, in succession, with all parts of the circular auroral zones. Moreover, on any one occasion, aurorae are strongest and more numerous on these night portions of the ovals, so that when all aurorae are studied together they are found, statistically, to cluster round the auroral zones.



7·11 During any one auroral display the aurorae are observed to be located near an auroral oval, shown grey in the diagram. They are most likely to occur at night on the side of the oval that is farthest removed from the geomagnetic pole. When many displays of this kind are taken together the most common occurrence is in a zone that is occupied in succession by the night portions of these ovals. This auroral zone is roughly circular round the geomagnetic pole.

Polar cap absorption (PCA) events

During the solar-cycle maximum that coincided with the IGY it began to be realised that, in addition to ionosphere storms of the kind just mentioned, that follow the outbursts of solar flares by twenty or thirty hours, there are sometimes disturbances of a different type that follow more rapidly. These are marked, not by magnetic or auroral disturbances, but by intense absorption of radio waves noticeable throughout the region inside the auroral zones. These occurrences, known as polar cap absorption, or PCA, events, have two important features. First they are accompanied by a considerable increase in the electron content of the ionospheric D-region which absorbs radio waves most easily, and second, they

do not occur only along the auroral zones, or on the auroral ovals, but over the whole of the regions inside the zones. The speed of the responsible particles can be found from the time interval between the solar flare and the increase of ionosphere absorption: then it can be asked whether any known particles of that speed could penetrate to D region heights (about 80 kilometres) to produce the ionisation required. When the calculations are made it appears that protons would roughly satisfy the required conditions. It seems, therefore, that PCA events are caused by streams of protons that are ejected from the sun, are guided by the magnetic field to the regions inside the auroral zones, and there penetrate to the D region to produce ionisation by collisions with the air particles.

Two points about this explanation require especial mention. The first is that these solar particles, unlike the slower ones that arrive twenty to thirty hours after the flare, have sufficient speed to penetrate to the D region and to give up their enegy there. There is no need in their case to look for some local source of acceleration to give them the speed necessary to penetrate the ionosphere.

The second point is that sometimes the particles do not appear to have come from the sun by the most direct route. In one experiment, for example, in which equipment in a satellite was used to record the times at which particles of different speeds arrived, the results were consistent with the idea that all the particles had been emitted simultaneously from the sun and had travelled the same distance, the fastest arriving first; but the distance they had travelled was considerably greater than the sun-earth distance. This experiment supported other indirect evidence, based on a comparison between travel times and penetration depths into the ionosphere, that showed that the route from the sun to the earth was often a round-about one.

One suggested explanation is that the protons emerge from the sun in a direction that would not lead them to impinge on the earth, but

8 The ionosphere and radio communications

8 The ionosphere and radio communications

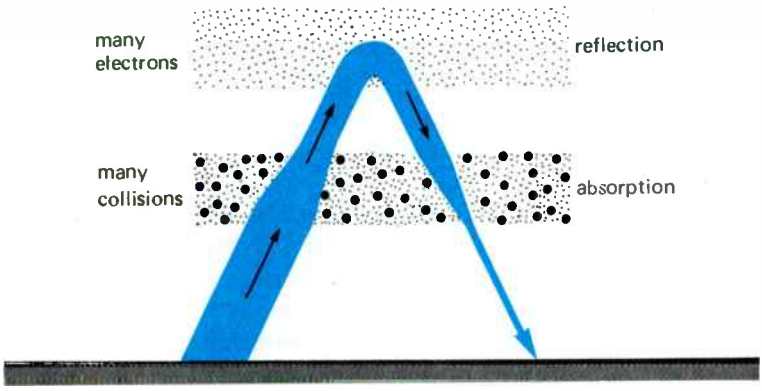
Frequencies useful for communications

This chapter discusses the part played by the ionosphere in helping (and sometimes in hindering) radio communications. In commerical radio telegraphy the engineer is usually concerned with distances so great that the ground wave is weak and he must rely on the reflected wave to produce a signal. With sound and television broadcasting, however, where he is concerned with producing an undistorted signal at a comparatively short distance, he prefers to rely on the direct wave travelling over the ground, any reflection from the ionosphere is a nuisance, and can cause distortion: moreover if two transmitters are working on the same frequency a wave reflected from one can interfere with receivers near the other.

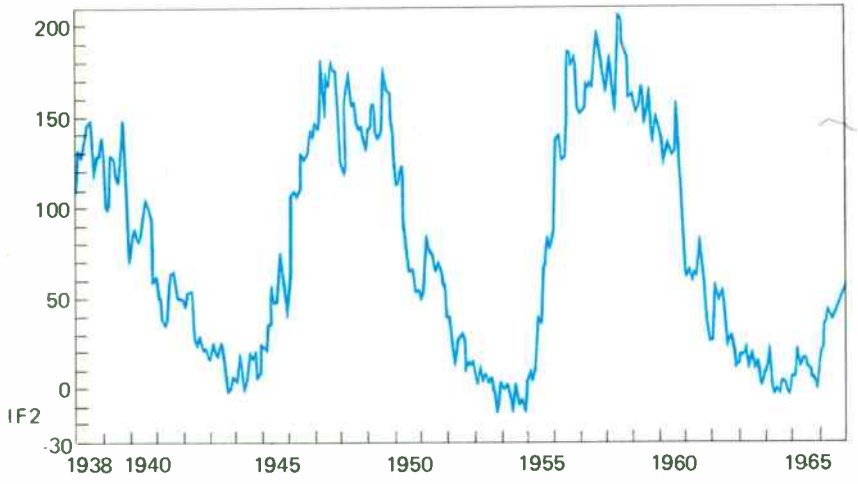
First let us consider the situation in which the ionosphere is necessary for the communication system to work, the case of long distance transmission, usually by means of dot-dash signals. In previous chapters it has been shown that the ionosphere can do two different things to a wave: it can reflect it and it can absorb it; the electrons responsible for the reflection are at the greater heights (usually in the F layer) and those responsible for the absorption are lower down, in the D region, where the collision frequency is greater. Moreover as the frequency of a wave becomes greater the concentration of electrons required to reflect it increases, and the amount of absorption decreases. Thus, if the frequency is too great there will not be enough electrons high up to reflect it; if it is too small the electrons low down will absorb it. There is a relatively narrow band of useful frequencies lying between those that are 'reflection limited' and those that are 'absorption limited', and the communication engineer must choose his frequencies within this band if he is to provide an efficient service.

The changes that are always taking place in the ionosphere make the choice difficult; knowledge based on the world-wide

8·1 If there are sufficient electrons high up, radio waves are reflected, and if there are too many electrons low down, they are absorbed. If the frequency is too great there are not enough electrons to reflect (reflection limitation); if it is too small the absorption is too great (absorption limitation).



8·2 The behaviour of the F layer at any place can be forecast roughly if the magnitude of an index called IF2 is known. The index is derived from measurements made at a few key observatories and varies systematically through the solar cycle.

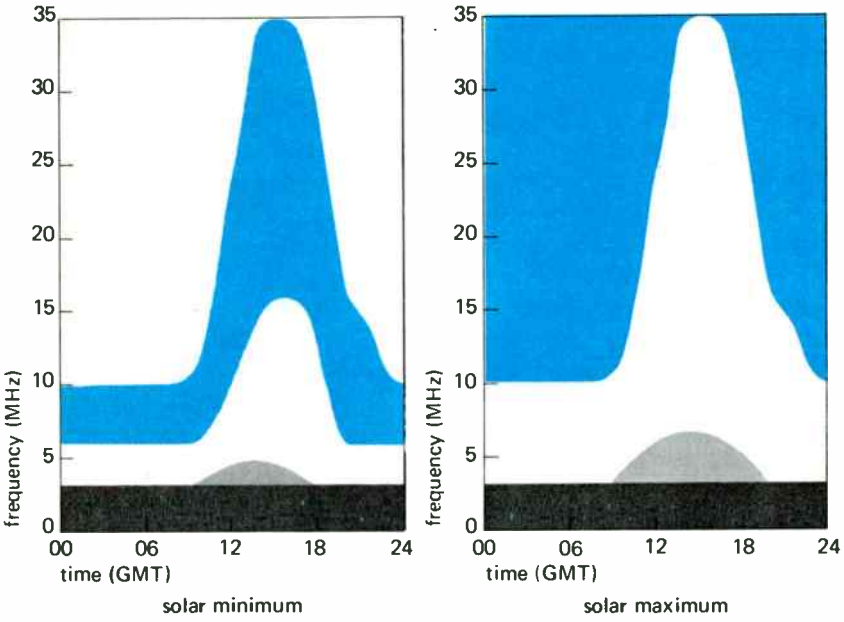


distribution of ionospheric observatories is therefore used in attempts to forecast, for some months ahead, what these changes will be. For this purpose the behaviour of the F2 layer is described in terms of an index, IF2, based on observations of the penetration frequency made at about ten key observatories over many years. The index is found to vary in step with the solar cycle so that by examining at any one time the way in which it changed at similar parts of previous cycles it is possible to forecast, by extrapolation, how it will behave in the succeeding few months. If the index is known the daily and seasonal behaviour of the F2 layer at any place in the world can be estimated. The behaviour of the D region is easier to forecast, since it depends in a more regular way on the inclination of the sun's rays: the days of 'abnormal winter absorption' and the occurrence of sudden ionosphere disturbances present the greatest difficulties.

An engineer who is concerned with providing communication over some particular path first selects the places where he thinks the waves will be reflected, possibly near the middle for a once reflected wave, or at other places for waves reflected between earth and ionosphere more than once. He then uses the forecast knowledge of the F layer and the D regions to construct a diagram, like that of figure 8·3, to show how waves reaching the ionosphere at these points will behave. The curves show how the reflection-limited frequencies and the absorption-limited frequencies will change throughout the day, and for satisfactory working a frequency that lies in the uncoloured region, between the two curves, must be selected. Note that the usable range of frequencies is smaller at the minimum of the solar cycle than at the maximum.

This method of frequency selection, based on a knowledge of the ionosphere as studied at a limited number of observatories, is good enough for overall frequency planning, but because the ionisation varies to some extent unpredictably from day to day, it does not

8·3 The uncoloured regions show the frequencies which could be used for communications between London and Halifax (Nova Scotia) at different times of the day in December 1954 (solar minimum) and December 1958 (solar maximum). In the blue region the frequencies were reflection limited, in the grey they were absorption limited. Frequencies less than 3 MHz were reflected from the E layer and were not used.



necessarily select the best frequency for use at any one time.

Ideally the choice of working frequency should be based on ionospheric information at the actual time of the transmission and over the same path. A method of providing this information that has the advantage of being operated from the transmitter, without the need for co-operation from the receiver, works as follows. The transmitter is caused to emit a series of short radio-wave pulses that are reflected by the ionosphere to distant points: on arriving there some of the energy is returned as a scattered wave towards the transmitter, where it is recorded as an echo. The time delays of these echoes thus correspond to all the distances reached by the transmitted waves:

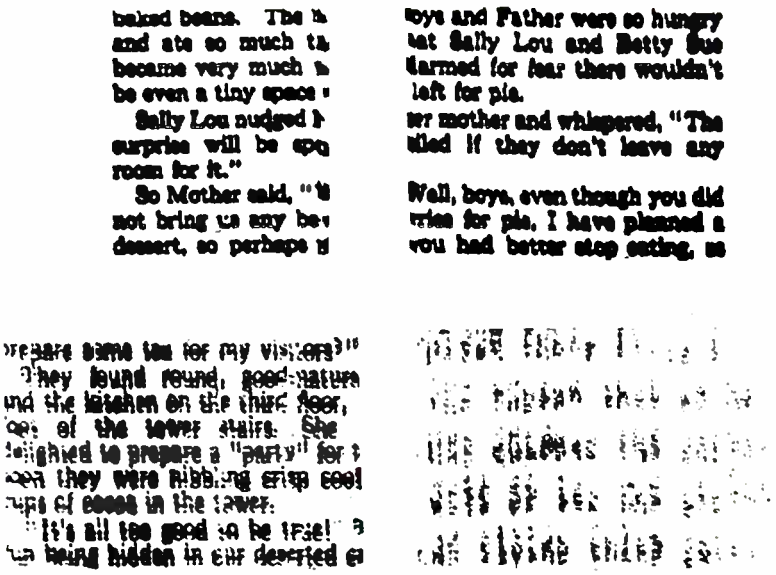
the frequency of the transmitter is then adjusted until one of the echoes corresponds to the distance of the desired receiving point. It is best to choose the highest frequency that reaches the required point, for then the absorption is least. The probing waves are best sent out in a narrow rotating beam so that the regions from which waves are scattered back are displayed on a two-dimensional map. An arrangement of this kind has proved useful in selecting the best frequency for communicating with aircraft crossing the Atlantic.

Imperfections of ionospheric propagation

For long-distance transmission the engineer could not do without the help of the ionsophere, but it is not always as perfect for the purpose as he would like. He may be concerned with several different kinds of transmission, for example, teletype in which the signal consists of a rapid succession of 'dots' and 'dashes' operating automatic printing devices at the receiver; or facsimile picture transmission in which the picture is scanned by a writing-pen moving rapidly back and forth in one direction while a sheet of paper is moved more slowly in a perpendicular direction; or television in which the scanning is performed electronically; or sound broadcasting in which the strength (or the frequency) of the radio wave is altered in sympathy with the sound vibrations. Whatever the type of transmission it is desirable that the signal should arrive at the receiver not only with sufficient strength but also without distortion.

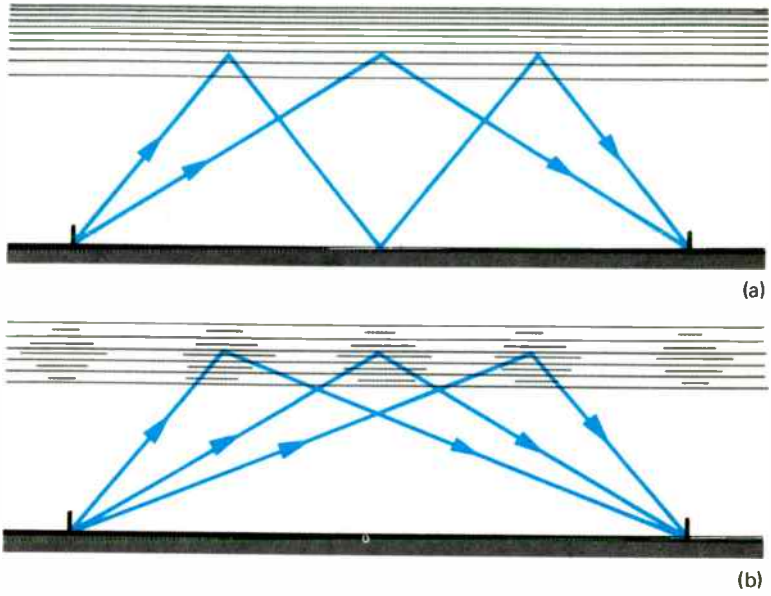
Sometimes the waves travel by more than one path: then the strength of the received signal may vary, as changes in the lengths of the paths cause the relative phases of the waves to alter. The variations, known as fading, can be quite rapid if the lengths of the transmission paths are changing rapidly: several essential parts of a message transmitted, for example, by means of the dots used in teletype communications, can then fail to be recorded.

8·4 When a facsimile picture is transmitted random distortion is liable to occur if speeds are used which require the paper to move faster than ten inches per second.



When more transmission paths than one are present, a short dot may arrive at the receiver not just once, but twice, or even several times, so that a teletype message is seriously confused. It is found in practice that reliable communication can usually be achieved only if there are less than about 200 dots per second. The transmission of a facsimile picture must be made more slowly by radio than by land-line; the kind of distortion which can occur if the speed is too great is shown in figure 8·4. Broadcast music is distorted if the time delay between the different paths is comparable with the time period of the musical note to be transmitted. Distortions like these, produced by

8·5 Multipath transmission. A signal may travel by more than one path either because (a) it is reflected repeatedly between ground and ionosphere or (b) the ionosphere is irregular.



multiple transmission paths, are often more disturbing than the fading that accompanies them.

Multiple transmission paths can be caused either by many reflections between the ionosphere and the ground, or by reflections from a rough and irregular ionosphere. They can never be avoided, but can often be minimised by working close to the frequency that is reflection-limited, for then the steeper rays, which would form the multiple paths, cannot be reflected. For very special purposes it has sometimes been possible to send messages at great speeds by using a frequency that is reflected only once.

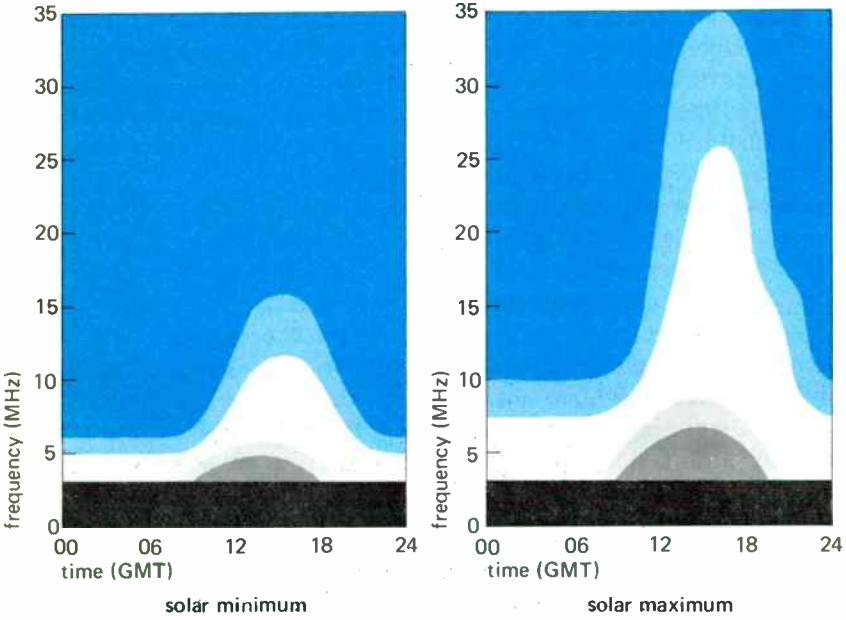
In high-quality audio broadcasting the ionosphere is nearly always a nuisance rather than a help. Because there must be no distortion there must be only one transmission path, and for that purpose the ground wave alone is suitable, and any wave reflected from the ionosphere must be weak in comparison. Ideally, the frequency should be so great that no reflection is possible from the ionosphere: that is achieved with u.h.f. (ultra-high frequency) sound broadcasting and television. But when smaller frequencies are used, as in medium-frequency broadcasting, a reflected wave is received unless it is much absorbed. By day there are enough electrons low in the ionosphere to absorb it strongly, but at night the electrons low down are fewer, the absorption is weaker, and at sufficiently great distances the reflected wave becomes noticeable compared with the ground wave. The useful range of the station is then restricted to distances of about 100 kilometres. In order to extend it, special aerials are used, designed to radiate as little energy as possible upwards, in those directions where it might reach the receiver after reflection from the ionosphere. It is not possible, however, to exclude radiation at lower angles in directions that will be reflected to listeners at still greater distances; the reception of these ionospherically reflected signals from distant high-frequency broadcasting transmitters can cause very bothersome interference, particularly in the winter when it is dark during the main hours for broadcast listening.

When it is necessary to broadcast speech or music to distances so great that the wave must travel by reflection from the ionosphere, distortion cannot be avoided. In overseas broadcasts of this kind from Great Britain some of the worst conditions are encountered when the reflection occurs over the tropics in the evening, for at that time and place the upper part of the ionosphere is often very irregular and rough, a phenomenon known as 'equatorial spread F'. Broadcasts can be violently distorted in this way just at the popular time for evening listening.

At this same time, high-speed teletype signals are also often found to be useless. Before the reason for this was known, operators at Singapore transmitting to London, where it was midday, often attributed the breakdown in communication to the inefficiency of the staff at the receiving end, and they referred to the corresponding lull in traffic as the 'London lunchtime effect'.

One type of interference is of interest, although it proves operationally troublesome only with powerful transmissions. It arises when two waves interact with each other as they pass through the ionosphere, so that the message being transmitted by one is, to some extent, transferred to the other. The process has been called 'ionospheric cross-modulation', or the 'Luxemburg effect' from the sending station that was first found to produce this kind of interference.

What happens is this. We saw on p. 61 that, when a wave is partially absorbed in the ionosphere, the electrons are heated up so that they move more rapidly and make more collisions with particles of the surrounding air. Suppose then that a wave which we want to receive is being transmitted through the ionosphere and that a second wave is suddenly switched on. This second wave heats up the electrons, they make more collisions, and absorb the wanted wave a little more strongly. If the second, or disturbing, wave is repeatedly switched on and off the wanted wave is received alternately weak and strong, and if the switching of the disturbing wave is at a speed that corresponds to a musical note, the note is also heard on the wanted wave. For the effect to be strong the electrons must have time to heat up and cool down during the switching period: this requires the frequency to be less than about 125 oscillations per second, and it is the bass notes in music or speech that are most easily transferred from the disturbing to the wanted wave. Cross-modulation of this kind has sometimes been found troublesome to broadcast services; it is more intense when the disturbing wave is more powerful, for then the electrons are heated more strongly.



The effects of ionospheric disturbances

Sudden ionosphere disturbances (SIDS) (p. 87) and ionosphere storms (p. 198) are responsible for many serious disruptions of radio communications. The effect of an SID is mainly to increase the electron content of the D region and to increase the absorption, so that often the signal can no longer be received. During a sudden fadeout of this kind the signal becomes weak in a time interval of a few seconds and then returns to normal during the next three quarters of an hour or so. Fadeouts can interrupt long-distance radio communications so suddenly and so completely that engineers have sometimes thought that the power supply to their

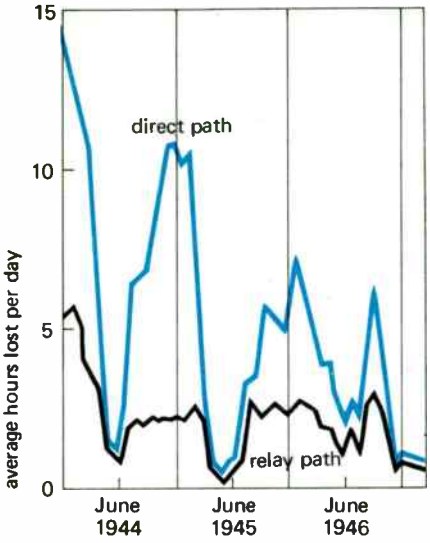
8·6 During an ionosphere storm the reflection limited frequency is decreased and the absorption limited frequency is increased. The diagrams show how these changes reduce the range of frequencies available for transmission between London and Halifax (Nova Scotia). The heavily coloured regions show the frequencies which are reflection limited (blue) and absorption limited (grey) at normal times. The lightly coloured regions mark the frequencies which are rendered useless by the storm. It is clear that the effect of a storm is more drastic at solar minimum than at solar maximum.

receivers had failed. Because SIDs, and with them fadeouts, are caused by intense bursts of X-rays coming from the sun they occur only on the daylight side of the earth, and equally at all latitudes.

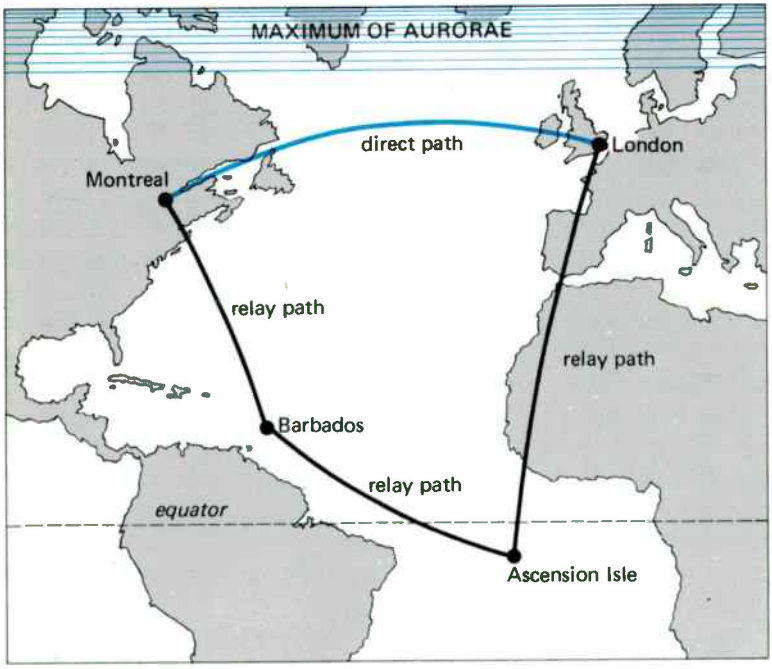
In contrast to SIDs, ionosphere storms are caused by streams of charged particles deflected by the earth's magnetic field towards the auroral zones, where they impinge equally on the day and night sides of the earth. They are accompanied by an *increase* of electron concentration in the D region, particularly near the auroral zones, and by changes in the F layer that, over much of the earth, take the form of a *decrease* of electron content. The D region change causes increased absorption and the wave frequency that is absorption-limited is increased; a wave passing near the auroral zone may become too weak to be useful, and the occasion is then called an 'auroral blackout'. The decrease of the F layer ionisation causes a reduction of the wave frequency that is reflection-limited.

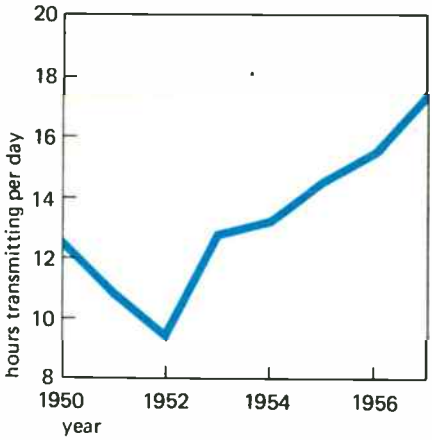
The two effects of an ionsophere storm are illustrated in figure 8·6, which shows how the changes in the 'absorption-limited' and the 'reflection-limited' frequencies combine to reduce the range of usable frequencies. The reduction is most significant at the minimum of the solar cycle, when the usable range is small even in the absence of a storm, and on transmissions near the auroral zone, where the change in absorption is most pronounced.

Ionosphere storms very seriously interfere with communication.



8·7 In radio communications between London and Montreal the number of hours lost because of unsuitable ionospheric conditions is much greater if the waves travel by the direct path than if they are kept away from the auroral zone by using relay stations at Ascension Island and Barbados.





8·8 The number of hours suitable for communications on the Britain–Australia circuit was less in years near the minimum of the solar cycle (1952) than in other years, because at these times ionosphere storms restricted the choice of operating frequency (see figure 8·6).

Because they usually start suddenly, the operator is taken by surprise, and he does not know how much the frequency should be changed to restore working conditions. A series of very damaging storms occurred about 1952 near the minimum of the solar cycle. They showed a tendency to repeat after 27 days as the solar rotation returned long-lived storm regions to the centre of the sun's disc. Figure 8·7 shows how these storms reduced the serviceability of one particular communication link.

If the path of the wave passes near, but not necessarily over, the auroral zone, the effect of a storm can be severe, and it is quite surprising what important improvement can sometimes be achieved by using a path in only a slightly different direction. Thus, for example, the radio link between London and Montreal is very susceptible to interruption in this way (indeed it was once used as a reliable guide to the occurrence of ionosphere storms), whereas the link between London and New York, which is a little further removed from the auroral zone, is much less affected. Relay stations can often be used with advantage to send a signal by a route that does not pass near the auroral zone; the kind of improvement then obtained is illustrated in figure 8·7.

It would be valuable if operators could be warned in advance when an ionosphere storm was expected and when it had ended. During the International Geophysical Year special attempts were made to

forecast the occurrence of storms by making use of all relevant information, including observations of the light emitted during the solar flare that preceded them, the location of the flare on the sun's disc, the solar magnetic field associated with it and evidenced by its effect on the spectral lines in the sun's light, and the radio waves emitted from the sun during the disturbance; and in recent years it has been possible to use space probes to detect the presence of streams of charged particles during their travel from the sun to the earth. But unfortunately no reliable method for warning radio operators of the onset of a storm has yet been found.

Although storms are most frequent near the maximum of the solar cycle, they are often more disturbing to communications near the minimum, for reasons previously mentioned. At these times their disruptive effect can be minimised by recalling their tendency to repeat after an interval of 27 days.

The effects of atomic explosions

When a storm occurs the ionosphere is upset by those unpredictable streams of charged particles that are injected from the sun: it can also be upset in a more predictable way by the injection of charged particles from a nuclear explosion. Explosions of this kind, made from rockets at heights of about 100 kilometres, produce copious energetic electrons that travel back and forth trapped on geomagnetic lines of force. Some of these are 'dumped' into the D region and cause strong absorption of radio waves, first noticeable at two places; one near the explosion and one at the other end of the line of force that passes through it. With time the trapped electrons move in an easterly direction and after about one hour they surround the whole earth, so that the increased absorption is more widely spread; it continues as long as there are sufficient trapped electrons leaking into the D region: sometimes it lasts for several days.

Because the natural disruption of communications that occurs during an ionosphere storm, or the artificial disruption produced by an atomic explosion, could be exploited in an embarrassing way by an enemy in time of war, the military are interested in world-wide communication systems immune from these disabilities. For many purposes very long waves are more satisfactory, because they are reflected, rather than absorbed, in the D region. They continue to be useful during a naturally occurring storm and are less seriously interferred with by an atomic explosion. But the ionosphere is on the whole an unsatisfactory mirror, and those concerned with military communications, which must be usable literally at all times, think in terms of providing additional communication facilities by placing satellites in orbit to provide reflectors which (they hope) will always be present.

Sporadic-E ionisation

Another type of unusual ionospheric occurrence, called sporadic-E ionisation, can be troublesome to radio communications. It consists of an abnormal increase of ionisation at a level of about 100 kilometres over limited areas of radius 1,000 or 2,000 kilometres. At middle latitudes sporadic-E occurs most frequently near midday and in summer: there is no generally accepted theory to account for it, and it has not proved possible to forecast in detail when it will be present.

When it occurs in an intense form it can reflect waves in a single hop to a distance of about 2,000 kilometres on a frequency that would normally be reflection-limited. It can thus happen that a transmission intended only for comparatively local reception by the ground wave may sometimes be reflected by sporadic-E to much greater distances where it may cause interference to someone whom it was never intended to reach. In summer, waves reflected in this

8·9 Television is usually transmitted on frequencies so great that they are not normally reflected by the ionosphere, but sometimes very intense sporadic E-ionisation may reflect waves from a distant station sufficiently strongly to produce serious interference.

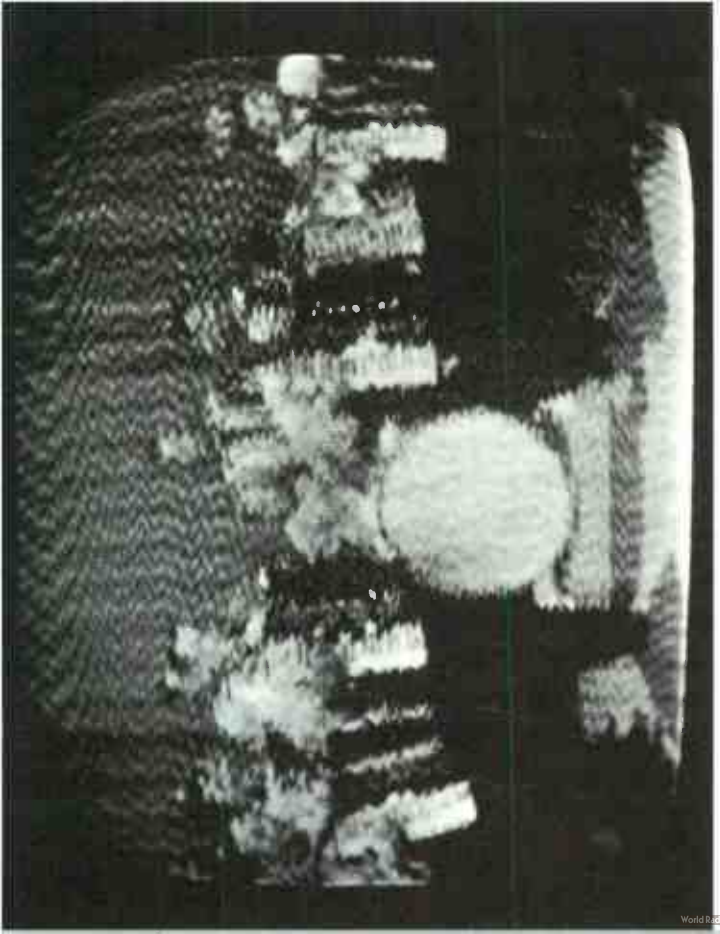
way from the European continent have frequently interferred with reception of local television transmissions in England.

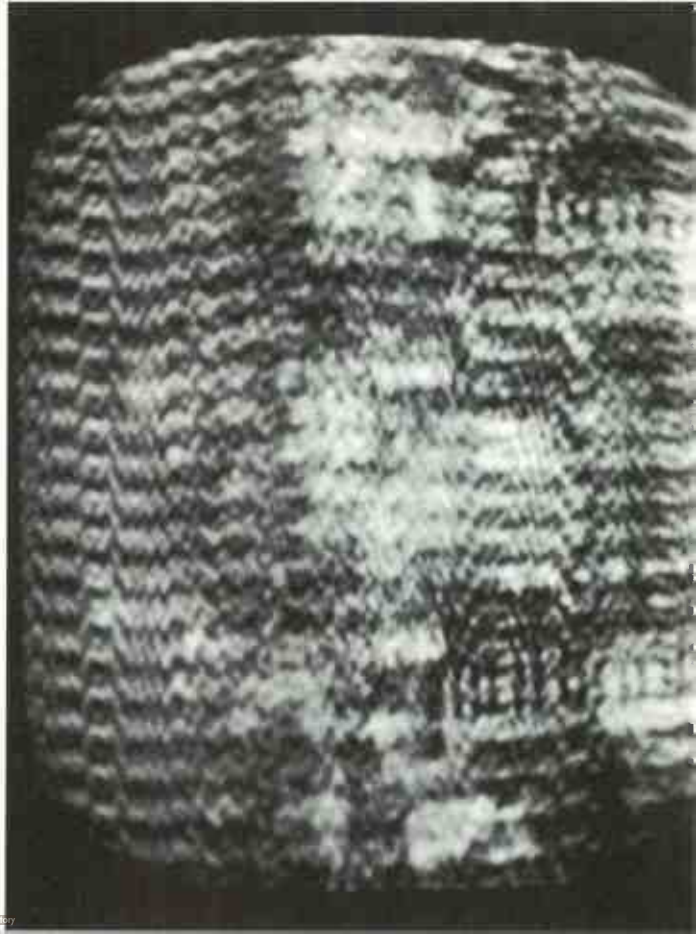
Some special transmission systems

The communication systems described above make use of waves reflected in the usual way by the ionosphere. There are, in addition, some more specialised systems that make use of irregularities in the ionosphere, and that involve the return of waves by methods different from those of ordinary reflection. Although these have been used only to a limited extent, they are of especial interest for readers of this book because it was studies of ionospheric physics that first suggested their feasibility. Whereas the possibility of long-distance radio communication was discovered empirically by Marconi, the methods now to be described were first suggested on theoretical grounds.

One system uses waves that have been scattered from randomly occurring irregularities in the ionosphere. A strong beam of waves is sent nearly horizontally to meet the ionosphere, at a height of about 100 kilometres, at a distance of about 1,000 kilometres. A frequency in the neighbourhood of 35 MHz is used, so great that it escapes into space instead of being reflected. There are, however, comparatively small irregularities in the ionosphere, of size comparable with the wavelength, which scatter a small part of the wave in all directions, but with a preference for the forward direction. Some of the forward scattered waves reach the ground at distances of about 2,000 kilometres, and can be received there by sufficiently sensitive equipment.* Because the received signal comes from several different irregularities, all moving in a random manner, it fades rapidly and is

* This forward scattered wave arises from irregularities in the ionosphere and is not to be confused with the backward scattered wave (used in Thomson-scatter sounding) that is present even in a completely smooth or regular ionosphere.





much distorted. We might ask why a system of this kind is used at all: why use a frequency on which most of the energy is lost into outer space, only a small proportion being scattered back to earth, and one on which the received signal is fading violently? There are two reasons. One is that, because the frequency used is greater than those normally reflected from the ionosphere, it is easier to find a part of the frequency spectrum that is not occupied by other users; but a more important reason is that this system of communication is only slightly disrupted by fadeouts or by ionosphere storms, or even by polar blackouts. The waves are not unduly absorbed by the extra ionisation that is present in the lower ionosphere on those occasions simply because the frequency is so great; and storm effects in the upper ionosphere are of no importance because the waves are scattered lower down. Indeed, the scattered signal is usually *stronger* during a fadeout or storm because an increase in the general background ionisation assists the scattering process. This method of transmission has been used chiefly in the polar regions, where blackouts are very troublesome.

The second special type of communication system makes use of meteors of the kind that are continually entering our atmosphere from outside with speeds of about 30 kilometres per second: they produce straight trails of ions and electrons at a height of about 90 kilometres. If a trail happens to lie in the proper direction it can reflect waves from a given transmitter to a given receiver, and since the electron concentration is great a high radio frequency can be used, higher than that normally employed. But the trails rapidly decay, they remain useful as reflectors only for about half a second, and can be used only for rather special types of transmission: it is necessary to wait until a trail is formed in the proper position to reflect, and then to send the message very rapidly before the trail decays. The message is therefore stored electronically, and a continuous probing signal is emitted: when a meteor trail is formed

in a convenient place it reflects the probing signal strongly to the receiver, which is then caused to send a signal back to the transmitter where it initiates a rapid transmission of the stored message. If there is still more message to be sent after the trail has died away, it is transmitted when the next appropriate trail is formed. Because reflection occurs only when a meteor trail is in the proper position a system of this kind can be used with considerable secrecy. A great advantage of the system is that the ionised trails, having been produced by particles travelling in a straight line, are themselves very straight, so that waves are reflected from them as if from a smooth mirror. There are none of the difficulties associated with rough reflectors or multipath transmission, and very great signalling speeds can be achieved with little liability to error.

Appendices

Appendix 1
Some fundamental physical ideas

Some of the fundamental physical ideas referred to in this book may not be familiar to all readers. It is the purpose of this appendix to explain them in simple terms.

Waves and wave motion

Wave motions of many different kinds occur in nature. Some, like waves on the surface of water, can be seen as they travel, but others, like radio waves or sound waves, are noticeable only when they produce an effect in a receiver. This book is largely concerned with a type of wave motion called *electromagnetic*; it includes *radio, infra-red, visible light, ultra-violet*, and *x-rays*; there is also passing reference to a closely allied kind called *hydromagnetic waves*, and to *sound waves*. Because all of these are invisible it is convenient first to describe some of the properties of waves in terms of those seen on water if a bar dipping into the surface is caused to oscillate up and down. If the oscillation is produced by a crank on a shaft rotating at constant speed, as shown in figure A1, it is called *simple harmonic*. At one instant the surface of the water takes up a shape like that at (a) in figure A2; at successive times, separated by equal short intervals, the shapes are like those at (b) and (c) etc. The sinuous shape seems to travel along the surface of the water away from the oscillating bar, and keeps the same shape as it travels and has a definite and measurable speed. We say that the displacement of the surface, which determines its shape, travels through the water as a *wave motion*; it is the *shape* that travels, not the water, which merely provides something that supports the wave as it travels; this is called the *medium* through which the wave travels. In general terms a *wave motion* is a *disturbance that travels through a medium without the medium itself travelling*.

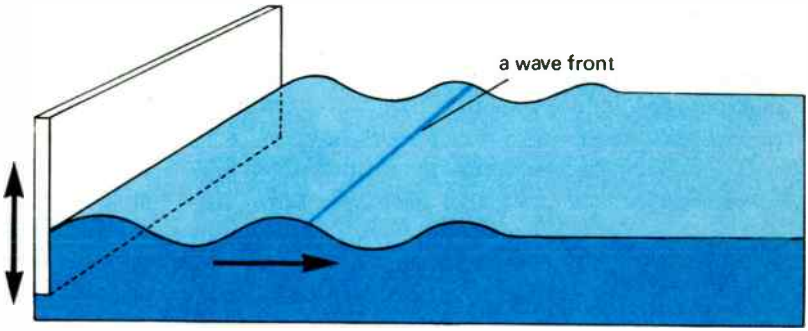
The length of one of the oscillations, from one crest of the wave to the next, is called the *wave length*; the number of oscillations made

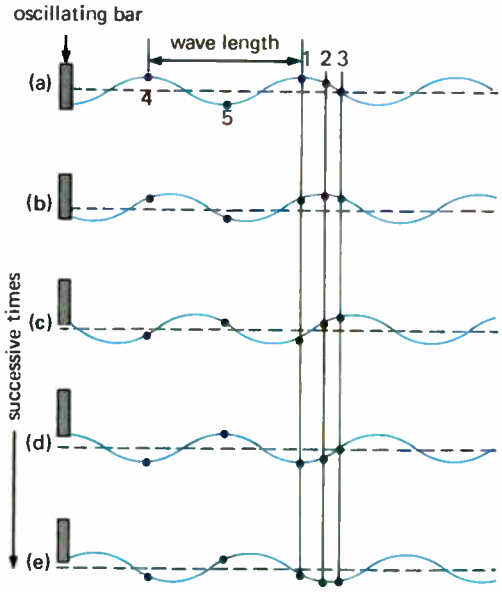
A1 Top If a bar dipped into water is made to oscillate vertically in the way shown it will give rise to waves which travel over the surface while the water itself does not travel. A line that connects all points at the same height on the surface is called a wave front.

A2 Bottom If waves are produced as in figure A1, the surface of the water assumes the shapes shown at (a), (b), (c), etc. after successive equal intervals of time. Individual points on the surface, such as 1, 2, 3, 4, 5, will each perform oscillations whose timings, or 'phases', are different. Points 1 and 4 will oscillate in step, or 'in phase'; 1 and 5 are out of step or 'in anti-phase'; the oscillation of 2 lags behind that of 1, or 'lags in phase'. The distance between an oscillation and the nearest one that is in phase with it is called a 'wave length'.

by the driven bar in one second is called the *frequency* and it is stated in units called the *Hertz*, after the name of the investigator who first demonstrated the existence of radio waves. Thus a wave in which there are 10 oscillations per second is said to have a frequency of 10 Hertz, written 10 Hz. A wave with a frequency of one million oscillations per second is said to have a frequency of one *Mega Hertz* (*MHz*), and in this book frequent mention is made of radio waves with frequencies such as 3 MHz, that is, 3 million oscillations per second.

As the wave passes, each point on the surface of the water oscillates in just the same way as the end point; it has a simple harmonic oscillation with the same frequency, but points at successive distances perform oscillations each a little later, and there is an increasing time lag between them. Thus, for example, at time (a) in figure A2 point 1 is at the top of its oscillation, point 2 is approaching the top and will reach it at time (b); point 3 is passing through the midpoint of its oscillation and will not reach the extreme until time (c). Two oscillations which are not in step are said to have a *phase difference*; the one that occurs earlier is *'advanced in phase'* and the one that occurs later is '*retarded in phase*'. Oscillations, like those at





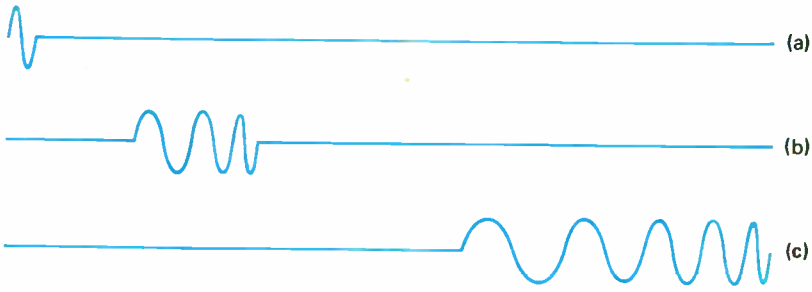
1 and 4, separated by one wavelength, are always in step, and are said to be *in phase*; oscillations like those at 1 and 5, separated by one-half wavelength, are always just out of step, so that one is at one end of its oscillation when the other is at the other end; they are said to be *in anti-phase*.

If the driving rod is long enough the waves travel so that each wave crest or wave trough is a straight line perpendicular to the direction of travel, as shown in figure A1. Each of these lines is called a *wave front*, and along it all the oscillations are in phase. If the water is shallower at one end of a wavefront than at the other the speed is different at the two ends and the wave bends round; analogous situations with sound waves are described on page 24 and with radio waves on page 56.

If the frequency is greater (that is, if there are more waves emitted or received per second) there are more waves within any given distance, so that the wavelength is smaller. If waves of all frequencies travel with the same speed, as with radio waves in empty space, then *the wavelength is inversely proportional to the frequency*. For example, it is common knowledge that the shorter wave broadcasting stations radiate on the greater frequencies.

The motion of the bar (figure A1) that launches a wave on the surface of the water need not be simple harmonic, but might consist of a single back and forth stroke; the wave would then start out with a shape like that at (a) in figure A3. A wave of this kind becomes modified as it travels, and at different distances it takes different forms, as in (b) and (c); oscillations appear, they are closer together at the front (smaller wavelength and greater frequency) and further apart at the back (larger wavelength and smaller frequency). Oscillations develop in this way from an initial impulsive wave only if the wave speed depends on the frequency, as it does for water waves; an analogous example when waves travel through the ionosphere is mentioned on page 191.

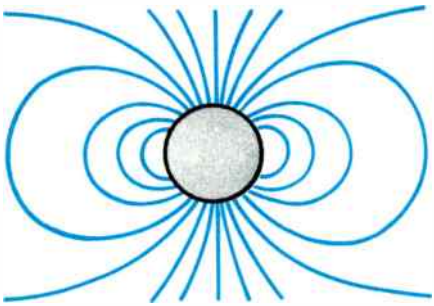
A3 If an impulsive wave like that at (a) is launched on the surface of water it develops oscillations as it travels, and at successive times assumes shapes like those at (b) and (c), with the more rapid oscillations in front.



Electromagnetic waves

Before describing electromagnetic waves it is convenient to explain what is meant by an electric field. If a charged particle experiences a force when it is placed at any particular point, there is said to be an *electric field* at that point, and *the strength of the field* is measured in terms of the magnitude of the force. We can, in imagination, suppose a particular region to be explored by measuring the forces on several small test charges of this kind so that the strength of the field is known everywhere. Lines can then be drawn to indicate the direction of the field at each point; these are called *lines of electric force* and provide a useful picture of the field over the region explored. Since an electron would repel (or attract) one of these small test charges in a direction radially away from itself, the lines of force surrounding an electron are straight lines running radially.

The *magnetic field* at a point is described in a similar way by imagining it to be explored by means of a very small compass needle placed at the point, and the direction of the field is defined to be the direction in which the needle sets. *Magnetic lines of force* can then be drawn to show the direction of the field at each point. The lines of force surrounding a magnetised sphere run as indicated in figure A4, and their form can be mapped out by sprinkling iron filings on a piece of card placed symmetrically round the sphere. The earth, which is a magnetised sphere, is surrounded by lines of force of this kind.

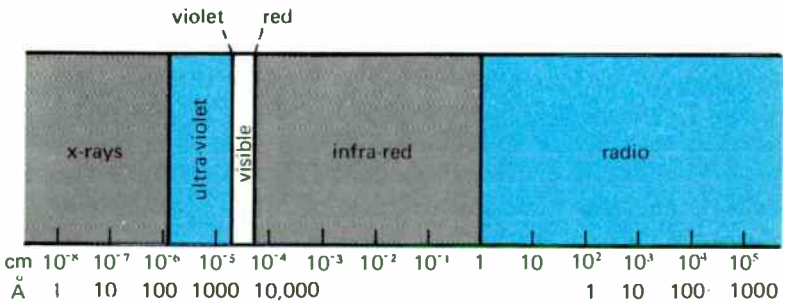


A4 The lines of magnetic force surrounding a magnetised sphere. A small compass needle placed at any point will set along the direction of the line passing through that point. The earth is a magnetised sphere of this kind.

In 1865 Maxwell showed theoretically that if the strength of the electric field at any point could be made to oscillate simple harmonically, then a 'wave of electric field' would travel out, so that at distant points the field would also oscillate with the same frequency but with a phase lag that increased with the distance. These oscillating fields could, ideally, be explored with the aid of small test charges placed at a succession of increasing distances. The forces on the charges, and the positions which they would take up, would then follow a moving pattern with a shape similar to that of a wave travelling over water. Maxwell showed that in these waves the electric oscillations are accompanied by magnetic oscillations and he therefore called them *electromagnetic waves:* here we are concerned only with their electric fields.

When Maxwell realised that electromagnetic wave motion was possible he suggested that light and the invisible radiations in the ultra-violet and infra-red were wave motions of this kind, and that it should be possible to produce electromagnetic waves with electrical equipment, and that they would be found to have the same speed as light, and have many properties in common with it. In 1887 Hertz made the first transmitter of radio waves and it soon became accepted that they are of the same nature as light, visible and invisible; all

A5 The spectrum of electromagnetic waves. The wavelengths cover an enormous range and are conveniently measured, in different parts of the spectrum, in metres, centimetres, or Angstrom Units (A). One Angstrom Unit equals 10^{-8} cm.



are electromagnetic, and all travel through empty space with the same speed, the only difference being that they have very different frequencies, or, to state the same fact in another way, all have different wavelengths.

When the properties of electromagnetic waves are described in terms of their wavelength, it is usual to speak of *a spectrum*, and to name the different parts of it as shown in figure A5. In order of decreasing wavelength they are *radio, infra-red, visible, ultra-violet*, and *X-rays*. The wavelengths cover a very wide range, those of radio are conveniently measured in metres or centimetres, those of light are of the order of 10^{-5} centimetres and of x-rays 10^{-8} centimetres. It is often convenient to measure a wavelength of light, ultra-violet or x-rays in terms of a unit equal to 10^{-8} centimetres, it is called the *Ångstrom unit*, after a leading spectroscopist, and is denoted by Å. Visible light has wavelengths which range from 6,000 Å in the red to 3,900 Å in the violet; ultra-violet radiations has wavelengths less than 3,900 Å. The division between short ultra-violet and long x-rays is arbitrary, and in this book it is considered to be in the neighbourhood of 100 Å.

Photons

Although the description of electromagnetic waves as given by Maxwell, and outlined above, is sufficient for understanding their travel, it was realised round about 1900 that when they encounter matter they behave in many ways like particles. The name *photon* is given to these equivalent particles, and their behaviour is described by quantum theory. For the purpose of this book the important point is that *the energy of a photon is inversely proportional to its wavelength*, and is therefore greatest at the X-ray, and least at the radio, end of the spectrum. This energy is of importance in deciding whether a particular electromagnetic radiation can produce changes in atoms or molecules on which it falls. Such changes can be produced only if the energy of the photon is great enough; thus X-rays, with short wavelength, can easily remove electrons from all atoms so as to ionise them, but radio waves, with much greater wavelength, cannot ionise, while the intermediate ultra-violet can ionise some atoms and not others.

Photons must be described as particles only in a very special sense, since they have no mass and no charge, nor are they deflected by electric or magnetic fields. Their travel is guided by the electromagnetic waves with which they are associated; if these arrive at a point simultaneously by two or more paths they may sometimes add up (as described in the next section) to cancel each other with the result that no photons arrive there.

Superposition of waves

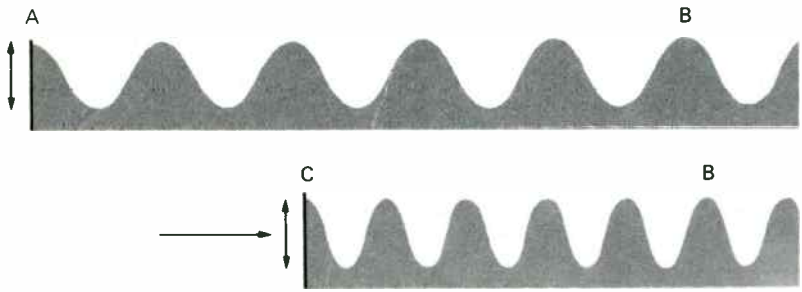
If two trains of water waves, of the same frequency, reach a point by paths of different length the movements that they produce in the water are added together, and the resulting oscillation depends on their phase difference. If the oscillations are in phase they both tend

to move the surface in the same direction at all times and a large oscillation results. If they are in anti-phase, one tends to move it up and the other down, and the two cancel out, and there is no motion. If the two waves are produced by two 'transmitters' oscillating in step, the question of whether or not they arrive at a point in phase or in anti-phase depends on the difference in length of the two paths they have traversed. If the *path difference* contains a whole number of wavelengths the waves add, if it contains a whole number plus a half they arrive in anti-phase and cancel each other. Corresponding phenomena occur with other types of wave motion: an example with radio waves is given on p. 51.

The Doppler effect

When a source of waves moves towards or away from an observer he finds that the frequency is increased or decreased: the phenomenon is called, after its discoverer, the *Doppler effect* (figure A6), and the change of frequency is called the *Doppler shift*. It is very noticeable when a train or an automobile sounding a warning note passes near an observer, when the effect is heard as a change in the pitch of the note. Consider the water waves of figure A6 and suppose the source remains stationary at A and emits five waves per second. The wave crest emitted at the start of an interval of one second reaches B (where AB is the distance travelled by the waves in one second) at the end of the interval and there are a number of waves equal to the frequency (5) in the length AB. Suppose now the oscillating bar moves with its driving mechanism towards B: during the one second it emits the same number of waves (5) and the first of these reaches B, as before, but the bar itself is now at C, nearer to B, so that the five waves are contained in the shorter distance CB. The wavelength has therefore been decreased. But once they have been launched the waves travel without regard to the movements of the source,

A6 The Doppler effect. The stationary source at A emits five waves into the distance AB in one second. If the source had been moving towards B, so that after one second it was at C, the five waves would occupy the distance CB; their wavelength would be shorter and therefore the frequency as observed at B would be greater, although the frequency of the source itself has not altered.



and because the wavelength is decreased the frequency has been increased. If the source had been moving away from B the wavelength would have been increased and the frequency would have been decreased. These changes of frequency constitute the *Doppler shift*.

Atoms, molecules, and ions

A gas consists of small particles, atoms or molecules. An *atom* consists of a *nucleus* which is positively charged, surrounded by negatively charged *electrons* in number sufficient to balance the positive charge on the nucleus so that the atom as a whole is uncharged. Most of the mass of the atom is in the nucleus which (for oxygen and nitrogen) is about 30,000 times heavier than an electron. If sufficient energy is supplied to an atom, for example by bombarding it with rapidly moving electrons or by irradiating it with photons of sufficient energy, it is possible to remove one of the light electrons so that it becomes free and leaves behind a comparatively heavy particle carrying a positive charge equal to that on the negative electron which has left it. This (heavy) particle is called a *positive ion* and the process is called *ionisation*.

A *molecule* consists of two or more atoms. This book is mainly

concerned with molecules of oxygen or nitrogen, each of which contains two atoms.

Temperature

All the particles in the atmosphere, atoms, molecules, ions, or electrons are continually moving about and colliding with each other. At any one temperature the particles of one type (e.g. the neutral oxygen molecules) have speeds spread over a wide range, and their average speed is determined by the temperature. When the temperature is greater the speed of the particles is greater and at a sufficiently low temperature they all come to rest completely. Although this limiting temperature cannot be attained in practice its magnitude can be calculated, and it is convenient to measure all temperatures from it as an *absolute zero of temperature*. A scale in which temperatures are measured from this zero and in which degrees have the same size as in the centigrade scale, is called the *absolute scale*. The absolute zero is 273 degrees below the melting point of ice: a temperature on the absolute scale is therefore equal to the temperature on the Centigrade scale plus 273 degrees. Most of the temperatures mentioned in this book are measured on this scale and are quoted simply as *degrees*: thus the boiling point of water (100°C) would be given as 373 degrees.

Particles of different gases at the same temperature have average speeds inversely proportional to the square root of their masses; thus the speed of an electron with a mass of only about 1/30,000 that of an oxygen molecule is about 170 times greater than that of the molecule.

Appendix 2
Recent advances

This appendix discusses recent advances of two kinds. One concerns the relations between the ionosphere in the polar regions and the long 'tail' of the magnetosphere on the night side of the earth. The other concerns a varied assortment of naturally occurring radio and hydromagnetic waves that can be used in various ways to study the structure of the upper ionosphere.

The magneto-tail and the polar ionosphere

Satellites penetrating to great distances from the earth have usually moved in very elongated elliptical orbits so that they repeatedly traverse the distance between apogee (the most distant point) and perigee (the nearest point). The longest axis of the ellipse maintains roughly the same direction in space while the earth moves round the sun, so that if an orbit starts with its longest axis in the sun-earth direction it is perpendicular to that direction after three months, back in it after six months and perpendicular again after nine months. If, therefore, the apogee is viewed from the earth throughout a year, it seems to swing round from the daylight side to the night side and back again, and the satellite explores all the region inside a circle traced out by its moving apogee. This statement is not in fact quite correct: the position of the long axis of the ellipse does not remain quite fixed in space because the asymmetry of the earth's gravitational field causes it to move round, or precess, slowly. For our present purpose, however, this precession is unimportant.

The apogees of the satellites important for the present discussion were at distances of about 15, 24, 32, 40 and 68 earth radii, and as they moved round they fell on the corresponding circles. The situation can be visualised if circles with these radii are drawn on figure 7·1 (page 181) with the earth as centre; it will then be seen that all are outside the magnetosphere on the sunward side, cross the magnetospheric boundary somewhere, and are inside the magnetosphere on

the anti-sunward side. It is by the use of these satellites that the magnetosphere has been explored on the side away from the sun.

The exploration has been assisted by two long-distance space-probes each of which crossed the sun-earth line once on the side remote from the sun, one at a distance of about 500 earth radii and the other at 1,000 earth radii. Useful information has also come from a Russian satellite rotating in a closed orbit round the moon: as the moon drifts round the earth this satellite moves with it and returns information from the distance of the moon, about 60 earth radii.

The most informative experiments carried by these satellites and probes have been designed to measure magnetic fields and particle concentrations. They have shown that the magnetosphere extends for at least 80 earth radii in the anti-solar direction, so that the stalk of the 'pear' is, in fact, much more elongated than would appear from figure 7·1. This elongation is called the magneto-tail. It is stretched out in the direction away from the sun and carries the magnetic lines of force embedded in it. These lines of force end, at the earth, in the polar regions, whereas lines ending at lower latitudes are not stretched out into the magneto-tail but keep comparatively near to the earth and return to it in the opposite hemisphere.

Measurements of particle concentrations, made on the two very distant probes, appear to show that at 500 and 1,000 earth radii the magneto-tail no longer exists as an ordered structure, but there are signs of irregularities near the sun-earth line, as though the earth leaves a kind of 'wake' in the solar wind as it moves round the sun.

The difference between the closed field lines, in the inner part of the magnetosphere, and the open ones in the tail, is accompanied by important differences in the electron concentration in the two regions. As the magnetospheric plasma moves by diffusion along the lines of force one part can diffuse out to great distances in the tail whereas the other part is held comparatively near to the earth, moving only up to the topmost point of the earthbound field lines. There is a kind of

plasma wind blowing out from the polar regions along the open lines of force: it has been called the *polar wind*. It is continually abstracting plasma from the outer magnetosphere and from the polar regions so that in these places the electron concentration is smaller than in those where the field lines return to earth. There is thus an 'outer magnetosphere' where the concentration is small (about one electron per cubic centimetre) and an inner part, called the *plasmasphere*, where it is greater (about 10^3 per cubic centimetre). The comparatively sharp boundary between the two is called the *plasmapause*.

The boundary is at a distance of about 4 earth radii over the equator and, following a field line to higher latitudes, it approaches nearer to the earth, until it falls within the topside ionosphere (1,000 km) at a (geomagnetic) latitude of about 60°. When the electron concentration is measured at this height, with the help of a topside sounder moving towards the pole, it is found to decrease abruptly as the sounder crosses the boundary.

The existence of the two regions of different electron concentration was discovered by two different experimental methods before its explanation was given in terms of the polar wind. In one the concentration of electrons was measured directly in satellites and rockets and was shown to decrease comparatively sharply at heights now recognised as being near the plasmapause.

The two regions were recognised in another way from observations of whistlers (see page 190). It will be recalled that the note of a whistler changes with time in a way determined by the distribution of electrons along a geomagnetic line of force. In earlier experiments whistlers were recorded at comparatively low latitudes where the field lines lay entirely inside the plasmasphere, and they led to results like those of figure 7·7. Later, however, records were made in the polar regions, where the field lines reach out to much greater distances, and it was these records that first showed the existence of the plasmapause. Extended observations from these regions have led

to a fairly detailed knowledge of the position of the plasmapause: and have shown that it moves inwards when there is a storm in the magnetosphere and ionosphere.

Recent work on whistlers has emphasised a tendency for workers in a specialised subject to invent their own terminology. In this particular subject it is curiously anthropomorphic; they liken the shape of the record made by some whistlers to the shape of a human nose and call them 'nose whistlers', while the shape of a curve depicting the height variation of electron concentration is called the 'profile' of the concentration, and a sharp kink in it, corresponding to the plasmapause, is called a 'knee' or sometimes even 'Carpenter's knee' after its discoverer.

The structure of the magnetic field in the magneto-tail has one particular feature that is probably of considerable significance. The field lines terminating in the south polar region are directed outwards from the earth, whereas those in the north polar region are directed inwards towards the earth. When, therefore, they are stretched out parallel to each other in the magneto-tail, those on the south side point outwards while those on the north side point inwards. Between the two there is a neutral sheet where there is no magnetic field. This neutral sheet is occupied by a current, made up of ions and electrons, and the magnetic fields on the two sides of this current sheet have opposite directions. A situation of this kind can be unstable, and under some circumstances the energy stored in the current sheet can be passed to charged particles to accelerate them; they can then travel along lines of force towards the polar regions of the earth where they appear as energetic particles. There is increasing evidence to show that an explanation along these lines probably accounts for some, at least, of the energetic particles found from time to time in the polar ionosphere.

If the energy of the current sheet is passed, in the way suggested, to energetic particles, the shape of the magnetic field is altered. The field

lines that were originally running parallel to each other in opposite directions become connected across the plane where the neutral sheet was, and a kind of bridge is formed so that a line runs out from the earth on one side, crosses the bridge, and returns to the earth on the other side. The phenomenon is therefore often described as 'field line reconnection'. Theorists are at present busy examining its consequence.

These ideas about the relation between the polar ionosphere and the magneto-tail are new, and not all are necessarily correct. It seems likely, however, that in the near future they will be found to account for some of the unusual features of the polar ionosphere. For example, it now seems probable that energetic particles are entering the polar regions at all times along field lines that come from the magneto-tail. When they impinge on the ionosphere they produce electrons simultaneously over most of the polar regions and probably account for that part of the ionisation that follows 'universal time' rather than 'local time' (see page 166).

Natural audio-frequency radiations

The whistlers discussed in the previous section, and on page 190, are audio-frequency radiations originating in lightning flashes. An intensive study of their nature has revealed that the magnetosphere is permeated by a miscellaneous collection of audio-frequency radiations of different kinds, many of them having other origins, some capable of penetrating the ionosphere so as to be observable on the earth. Observation of some of these radiations have led to deductions about the structure of the magnetosphere and upper ionosphere.

The radiation coming from a lightning flash and travelling through the ionosphere as a whistler has the polarisation appropriate to one of the possible characteristic waves (see page 64). This polarisation and that of the other characteristic wave depend in a sensitive way on

the relative abundance of the different kinds of positive ion that accompany the electrons in the ionosphere. If there are two or more kinds of ion, for example oxygen, helium or hydrogen, and if their relative concentration changes with height, then at certain levels, depending on the ratios of the ion concentrations, the two characteristic waves have identical polarisations. When the wave that forms the whistler reaches these heights, some of its energy is transferred to the other characteristic wave and a new whistler with a different frequency variation is radiated; it appears differently on the experimental record and is called an *ion whistler*. This phenomenon has been used to determine the relative concentrations of the positive ions of atomic oxygen, helium and hydrogen in the upper atmosphere.

When the first records of whistlers were made it was often found that, amongst the typical discrete whistles of descending tone, there was also a warbling sound, containing superimposed notes of ascending and descending pitch. In England, where the first records were made, the sound was reminiscent of that made at dawn by a colony of birds, a sound known as 'the birds' dawn chorus'. The radio-wave phenomenon was therefore called 'the dawn chorus', a title that seemed particularly appropriate because these radiations were often most intense near dawn. Later experiments, made in other parts of the world, have shown that this kind of natural radiation does not always occur near dawn; moreover it appears that England is one of the few places where dawn is greeted by the birds in this way. The radio phenomenon has therefore been re-named and is now called merely *chorus*. Although there is no universally agreed explanation, it is generally thought that the radiations are produced by electrons moving along magnetic lines of force with a speed equal to that of one of the characteristic waves. They can then pass their energy to the wave, and a sufficiently large bunch of electrons moving in this way could excite the kind of radiation that is observed.

When nature's audio-frequency waves are observed in satellites

other kinds of noise are also detected. They have not the tuneful nature of whistlers or 'chorus', but, when listened to with the help of a pair of telephones, they produce an irregular rushing or swishing noise. Although no certain origin has yet been suggested for this sort of noise, it has been shown to have one particular characteristic that has proved useful in the study of the ionosphere. In the 'swishing' noise the energy is spread, roughly equally, over a wide band of frequencies, but it is found that the audio spectrum is terminated sharply at the low-frequency end, and that the cut-off frequency is different at different times and places. It has, moreover, been established that it depends on the nature of the positive ions near the satellite and that it can be used to determine the relative concentrations of the ions of oxygen, helium and hydrogen at that level. When deductions made from observations of this kind are added to deductions made from 'ion whistlers' the world-wide distribution of the three types of ions can be examined at different times and compared with results like those of figure 4·7 (page 130) derived theoretically.

Magnetic micropulsations

The earliest investigations of the earth's magnetic field were made by studying the position taken up by lightly suspended compass needles which could move quickly enough to indicate the normal daily variations of the field and the main features of its change during a storm. There were signs of comparatively rapid variations of the field, called magnetic *micropulsations*, but the magnets found them difficult to follow. In recent years they have been intensively studied by methods that do not use suspended compass needles and that can follow changes taking place over a wide range of intervals from about 100 seconds to 0·1 second.

Micropulsations represent the extremely-low-frequency end of a spectrum of natural oscillations with frequencies that extend right up

to thousands of Hertz (i.e. time of oscillation 1/1000 sec). All these oscillations are evidence of waves travelling through the magnetosphere and ionosphere. The nature of those with the highest frequency is determined predominantly by the electrons: they are electromagnetic waves (whistlers and 'chorus' fall into this class). The nature of those, including micropulsations, that have the lowest frequency is determined by ions: they are hydromagnetic waves.

The origins of micropulsations are not clearly established. Some may be generated by the interaction of fast charged particles with hydromagnetic waves much as the 'chorus' is supposed to have been produced by the interaction of charged particles with electromagnetic waves. Others may arise from instabilities that occur on the surface of the magnetosphere as the solar wind sweeps over it, much in the same way as waves are produced on the surface of water when a wind blows across it. The waves associated with micropulsations are very long, an oscillation period of one second corresponding to a wavelength of about 1,000 km in the magnetosphere, and those with longer periods to even longer wavelengths. Since any particular kind of wave can be made to reveal structure greater but not less than its own wavelength, micropulsations have been used to investigate the structure of the magnetosphere on a scale of 1,000 km upwards. They have, for example, been used to deduce the distance of the magnetopause (the boundary of the magnetosphere) and to find how it moves during a magnetic storm; also to deduce the concentration of electrons and ions far out in the magnetosphere.

Selective bibliography

The publications listed below provide an introduction to further reading and themselves contain further references. The subject discussed in this book is advancing so rapidly that it is important to note the date of any publication.

1 An account at an elementary level that assumes little specialised knowledge

H.S.W.Massey and R.L.F.Boyd, *The Upper Atmosphere*, Hutchinson, London/Philosophical Library, New York, 1958.

2 Specialised books on the ionosphere and magnetosphere (1960–9)

J.A.Ratcliffe (ed.), *Physics of the Upper Atmosphere*, Academic Press, London/New York, 1960.

R.C.Whitten and I.G.Poppoff, *Physics of the Lower Ionosphere*, Prentice Hall, London/New York, 1965.

K.Davies, *Ionospheric Radio Propagation*, Constable, London/Dover, New York, 1965.

C.O.Hines *et al.* (eds.), *Physics of the Earth's Upper Atmosphere*, Prentice Hall, London/New York, 1965.

W.N.Hess, *The Radiation Belt and Magnetosphere*, Blaisdell (Book Centre), London/Waltham (Mass.), 1968.

H.Rishbeth and O.K.Garrioth, *Introduction to Ionospheric Physics*, Academic Press, London/New York, 1969.

3 Publications which illustrate the state of knowledge at different times in the past

P.O.Pedersen, *The Propagation of Radio Waves*, Danmarks Naturvidenskabelige Samfund, 1927.

S.Chapman, 'Some Phenomena of the Upper Atmosphere', *Proc. Roy. Soc.*, **132A**, 353, 1931.

'Meeting for Discussion of the Ionosphere', *Proc. Roy. Soc.*, **141A**, 697, 1933.

D.F.Martyn and O.O.Pulley, 'The Temperature and Constituents of the Upper Atmosphere', *Proc. Roy. Soc.*, **154A**, 455, 1936.
H.R.Mimno, 'The Physics of the Ionosphere', *Rev. Modern Physics.*, **9**, 1, 1937.
E.V.Appleton, 'Regularities and Irregularities in the Ionosphere', *Proc. Roy. Soc.*, **162A**, 451, 1937.
E.V.Appleton, 'The Ionosphere', *Occ. Notes, Roy. Astronomical Soc.*, **3**, 33, 1939.
J.A.Fleming (ed.), *Terrestrial Magnetism and Electricity*, McGraw-Hill, London/New York, 1939.
S.Chapman and J.Bartels, *Geomagnetism* (2 vols.), Oxford U.P., London/New York, 1940.
A.L.Green, 'Early History of the Ionosphere', *Associated Wireless Australia Tech. Rev.*, **7**, 177, 1946.
A.P.Mitra, *The Upper Atmosphere*, Asiatic Society, Calcutta, 1952.

4 Literature for detailed study

Results of investigations of the ionosphere and magnetosphere are regularly reported in the following publications: *The Journal of Atmospheric and Terrestrial Physics*; *Planetary and Space Science*; *The Journal of Geophysical Research*; *Space Science* (published annually and containing papers read at Symposia of COSPAR Council on Space Research); *Space Science Review*; *Geophysical Review*.

Acknowledgments

Acknowledgment is due to the following for the photographs (the numbers refer to the pages on which the photographs appear).

28–9 courtesy Geophysical Observatory, Alaska; 33, 35 and 183 © Mount Wilson and Palomar Observatories; 36 courtesy Professor C.W.Allen; 72, 74–5 (bottom) and 77 (bottom) courtesy Director of the Radio and Space Research Station, Slough; 80 courtesy Cornell University, which is supported by the Advance Research Projects Agency under a contract with the Air Force Office of Scientific Research; 210 courtesy Cables and Wireless Ltd; 221 courtesy British Broadcasting Corporation.

The diagrams were made by John Messenger and Design Practitioners Limited from the author's own sketches.

Index

References in *italics* refer to figures in the text.

J.A.Ratcliffe

16s
80p

Sun, Earth and Radio

An introduction to the ionosphere and magnetosphere

The upper atmosphere consists of two layers of charged particles – the magnetosphere and the ionosphere – which determine the behaviour of radio waves. This book is an introduction to both wave and medium. It shows how this branch of science has developed since radio waves were first used to study the upper atmosphere in 1924.

Introductory chapters describe how radio waves have been used, first from the ground and later from space vehicles, to investigate the earth's upper atmosphere and the effect upon it of solar radiations. Dr Ratcliffe then gives an account of present-day knowledge of the layers of the upper atmosphere, how they affect the transmission of radio waves and how this new knowledge has led to greatly improved communications.

J.A. Ratcliffe FRS has worked for most of his life at the Cavendish Laboratory at Cambridge and was head of the Radio Research Section set up to study the ionosphere. From 1960 to 1966 he was director of the Radio and Space Research Station at Slough in England.

With 15 black and white photographs and 83 two-colour diagrams

Cover design by Design Practitioners Limited

Price (in UK only) 16s net
80p net

SBN 303 17895 7