

Atmospheric Diffusion Modeling

Karl B. Schnelle, Jr.

Vanderbilt University

- I. Use of Models and Relationship to Atmospheric Pollution
- II. Atmospheric Turbulence
- III. Gaussian Plume Model for Continuous Sources
- IV. Diffusion Coefficients
- V. Plume Rise
- VI. Accidental Releases of Material to the Atmosphere
- VII. Calculated Concentration and Averaging Time
- VIII. Atmospheric Concentrations with Diffusion Ceilings
- IX. Multiple Sources and Receptors
- X. Designing a Chimney to Limit Ground-Level Concentrations
- XI. The New Models

GLOSSARY

Atmospheric turbulence Random velocity and pressure fluctuations occurring in the atmosphere.

Deterministic model Mathematical model in which each variable and parameter can be assigned a definite fixed number.

Diffusion ceiling Stable layer of air above the mixed layer, which restricts the vertical motion of a diffusing substance.

Diffusion coefficient Parameter of the Gaussian mathematical model which must be determined by measurement. Also called the sigma value.

Dosage Concentration–time relationship that is related to the effects that a polluting substance has on a receptor.

Evolutionary process Stochastic process in which the probability distribution is a function of time.

Gaussian model Mathematical model that is characteristically bell-shaped. The Gaussian model is the common

representation of the crosswind and vertical concentration distribution of a diffusing substance.

Inversion Positive temperature gradient or increase in temperature with elevation, resulting in adverse conditions for dispersion of pollutants.

Isopleth Lines of constant concentration in a diagram or plot of a mathematical model of concentration versus distance.

Lapse rate Decrease in temperature with height; that is, a negative temperature gradient.

Mixing depth Thickness of the turbulent region next to the ground, in which atmospheric properties are well mixed.

Plume rise Distance a smoke plume rises above its emission point due to the inherent momentum and buoyancy of the plume.

Richardson number Dimensionless ratio describing the relative importance of convection and the turbulence generated by mechanical shear.

Stability of the atmosphere Term describing the

dynamic characteristic of turbulence in the atmosphere. **Stationary process** Stochastic process in which the probability distribution function is not a function of time.

Stochastic model Model in which the principle of uncertainty is introduced. Variables and parameters can only be assigned a probability of lying between a range of values.

Temperature gradient Change of temperature with height. It is normally negative in the lower atmosphere; that is, the temperature decreases with height under normal atmospheric conditions.

AN ATMOSPHERIC DIFFUSION MODEL is a mathematical expression relating the emission of material into the atmosphere to the downwind ambient concentration of the material. The heart of the matter is to estimate the concentration of a pollutant at a particular receptor point by calculation from some basic information about the source of the pollutant and the meteorological conditions.

Deterministic, statistically regressive, stochastic models, and physical representations in water tanks and wind tunnels have been developed. Solutions to the deterministic models have been analytical and numerical, but the complexities of analytical solution are so great that only a few relatively simple cases have been solved. Numerical solutions of the more complex situations have been carried out but require a great amount of computer time. Progress appears to be most likely for the deterministic models. However, for the present the stochastically based Gaussian type model is the most useful in modeling for regulatory control of pollutants.

Algorithms based on the Gaussian model form the basis of models developed for short averaging times of 24 hr or less and for long-time averages up to a year. The short-term algorithms require hourly meteorological data, while the long-term algorithms require meteorological data in a frequency distribution form. Algorithms are available for single and multiple sources as well as single and multiple receptor situations. On a geographical scale, effective algorithms have been devised for distances up to 10–20 km for both urban and rural situations. Long-range algorithms are available but are not as effective as those for the shorter distance. Based on a combination of these conditions, the Gaussian plume model can provide at a receptor either

1. the concentration of an air pollutant averaged over time and/or space, or
2. a cumulative frequency distribution of concentration exceeded during a selected time period.

The U.S. Environmental Protection Agency (EPA) has developed a set of computer codes based on the Gaussian model which carry out the calculations needed for regulatory purposes. These models are available from the Applied Modeling Research Branch, Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711.

I. USE OF MODELS AND RELATIONSHIP TO ATMOSPHERIC POLLUTION

To make a calculation, the source of pollution must be defined and the geographic relation between the source and the receptor of pollution must be known. Furthermore, the means of transport between the source and the receptor must be understood. Thus the *source-transport-receptor* trilogy must be quantitatively defined to make the desired computation.

A. The Source

Weather affects many kinds of pollution sources. For example, on a cold day more fuel will be used for space heating. Hot weather brings on a greater use of space cooling, which produces a greater demand on electricity production.

Defining the source is difficult in most cases. One must consider first whether it is mobile or stationary and whether the method of emission is from a point or a line, or more generally from an area. Then its chemical and physical properties must be determined, most appropriately by sampling and analysis, but this is not always possible. One must turn to estimation or perhaps to a mass balance over the whole emission process to determine the amount of material lost as pollutant. The major factors required to describe the source are

1. Composition, concentration, and density;
2. Velocity of emission;
3. Temperature of emission;
4. Diameter of emitting stack or pipe; and
5. Effective height of emission.

From these data the flow rate of the total emission stream and of the pollutant in question can be calculated.

The source problem is further complicated when the emission is instantaneous instead of a continuous stream. Mass balances become impractical in such a case. Measuring the factors mentioned previously may become impractical as well. Grab samples with subsequent laboratory analysis might be used to determine concentration. If flow

rate can be measured or estimated, as well, then the total emission can be calculated.

B. The Receptor

In most cases legislation and subsequent regulations will determine the ambient concentrations of pollutant to which the receptor is limited. Air quality criteria delineate the effects of air pollution and are scientifically determined dosage-response relationships. The relationships specify the reaction of the receptor or the effects when the receptor is exposed to a particular level of concentration for varying periods of time. Air quality standards are based on air quality criteria and set forth the concentration for a given averaging time. Thus, the objective for a calculation is to determine if an emission will result in ambient concentrations that meet air quality standards.

Usually, in addition to the receptor, the locus of the point of maximum concentration, or the contour enclosing an area of maximum concentration, and the value of the concentration associated with the locus or contour should be determined. The short time averages that are considered in regulations are usually 3 min, 15 min, 1 hr, 3 hr, or 24 hr. Longer time averages are one week, one month, a season, or a year.

C. Transport

Understanding transport begins with the three primary factors that affect the mixing action of the atmosphere: radiation from the sun and its effect at the surface of the earth, rotation of the earth, and the terrain or topography and the nature of the surface itself. These factors are the subjects of basic meteorology.

The way in which atmospheric characteristics affect the concentration of air pollutants after they leave the source can be viewed in three stages:

- Effective emission height
- Bulk transport of the pollutants
- Dispersion of the pollutants

1. The Effective Emission Height

After a hot or buoyant effluent leaves a properly designed source, such as a chimney, it keeps on rising. The higher the plume goes, the lower will be the resultant ground level concentration. The momentum of the gases rising up the chimney initially forces these gases into the atmosphere. This momentum is proportional to the stack gas velocity. However, velocity cannot sustain the rise of the gases after they leave the chimney and encounter the wind, which will cause the plume to bend. Thus mean wind speed is a

critical factor in determining plume rise. As the upward plume momentum is spent, further plume rise is dependent upon the plume density. Plumes that are heavier than air will tend to sink, while those with a density less than that of air will continue to rise until the buoyancy effect is spent. When the atmospheric temperature increases with altitude, an inversion is said to exist. Loss of plume buoyancy tends to occur more quickly in an inversion. Thus, the plume may cease to rise at a lower altitude, and be trapped by the inversion.

Many formulas have been devised to relate the chimney and the meteorological parameters to the plume rise. The most commonly used model, due to Briggs, will be discussed in a later section. The plume rise that is calculated from the model is added to the actual height of the chimney and is termed the *effective source height*. It is this height that is used in the concentration prediction model.

2. Bulk Transport of the Pollutant

Pollutants travel downwind at the mean wind speed. Specification of the wind speed must be based on data usually taken at weather stations separated by large distances. Since wind velocity and direction are strongly affected by the surface conditions, the nature of the surface, predominant topologic features such as hills and valleys, and the presence of lakes, rivers, and buildings, the exact path of pollutant flow is difficult to determine. Furthermore, wind patterns vary in time, for example, from day to night. The Gaussian concentration model does not take into account wind speed variation with altitude, and only in a few cases are there algorithms to account for the variation in topography. For the future, progress in this area will come through numerical solutions of the deterministic models.

3. Dispersion of the Pollutants

Dispersion of the pollutant depends on the mean wind speed and atmospheric turbulence, described in more detail in the next section. Dispersion of a plume from a continuous elevated source increases with increasing surface roughness and with increasing upward convective air currents. Thus a clear summer day produces the best meteorological conditions for dispersion, and a cold winter morning with a strong inversion results in the worst conditions for dispersion.

D. Uses of the Models

Atmospheric diffusion models have been put to a variety of scientific and regulatory uses. Primarily the models are used to estimate the atmospheric concentration field in the

absence of monitored data. In this case, the model can be a part of an alert system serving to signal when air pollution potential is high and requiring interaction between control agencies and emitters. The models can serve to locate areas of expected high concentration for correlation with health effects. Real-time models can serve to guide officials in cases of nuclear or industrial accidents or chemical spills. Here the direction of the spreading cloud and areas of critical concentration can be calculated. After an accident, models can be used in *a posteriori* analysis to initiate control improvements.

A current popular use for atmospheric diffusion models is in air quality impact analysis. The models serve as the heart of the plan for new source reviews and the prevention of significant deterioration of air quality. Here the models are used to calculate the amount of emission control required to meet ambient air quality standards. The models can be employed in preconstruction evaluation of sites for the location of new industries. Models have also been used in monitoring network design and control technology evaluation.

II. ATMOSPHERIC TURBULENCE

Diffusion in the atmosphere is dependent upon the mean wind speed and the characteristics of atmospheric turbulence. These factors determine the stability characteristics of the atmosphere. Turbulence consists of horizontal and vertical eddies that mix the pollutant with the air surrounding it. Thus, in a smoke plume, the turbulence decreases the concentration of the pollutants in the plume and increases the concentration in the surrounding air. When turbulence is strong, the pollutants are dispersed more rapidly. Strong turbulence exists in an unstable atmosphere in which vertical motion is enhanced. Maximum instability occurs in the summer on a clear sunny day in the early afternoon. Conversely, when turbulence is weak, a very stable atmosphere is present. This condition would be most pronounced on a clear winter day, early in the morning just as dawn breaks. The diffusion parameters in the models describing turbulent mixing are quantitatively specified in relation to the stability condition; therefore, stability, surface roughness, and wind conditions must be explicitly defined.

Turbulent eddies are formed in the atmosphere by convection and by geologic and manmade structures. Convection occurs when the air is heated from below by the warm surface of the earth and the buildings and pavement covering it. Whenever the temperature decreases rapidly with height, convection is most pronounced and may persist up to many hundreds of meters above the surface on

clear days. It is these convective eddies or currents that result in the vertical air motions that birds and glider pilots use to climb into the upper air without the expenditure of other energy.

Mechanical eddies result from the shearing forces produced when the wind blows over the surface of the earth. At ground level the wind speed is zero, and it reaches a maximum usually at many thousands of meters above the surface. Mechanical turbulence increases with increasing wind speed and is greater over rough surfaces than over smooth surfaces. Terrain roughness can be characterized by a roughness length z_0 which is proportional to the size of the eddies that can exist among the roughness elements. The roughness length is relatively small if the roughness elements are close together. For smooth sand, z_0 is about 0.10 cm. Over cities, z_0 can increase to several meters. (Typical values of z_0 are listed later in Table V.)

A. Adiabatic Lapse Rate and Potential Temperature

Lapse rate is the rate of temperature decrease with height in the atmosphere. If we consider the hypothetical case of a bubble of air rising through a dry atmosphere with no mixing or heat exchange between the bubble and its environment, the bubble will expand and cool. The cooling is said to be at the dry adiabatic lapse rate (DALR). Thus the DALR can be defined as the negative of the temperature gradient that is established as dry air expands adiabatically while ascending. It should be noted that the term *temperature gradient* has the opposite algebraic sign from the lapse rate. Both temperature gradient and lapse rate are used when discussing the temperature structure of the atmosphere.

To calculate the temperature gradient for a reversible adiabatic expansion of dry air, the ideal gas equation of state is combined with the hydrostatic equation of motion and integrated. The following equation results,

$$\text{DALR} = \lambda = -dT/dz = (g/g_c)(\gamma - 1)/R\gamma \quad (1)$$

where g/g_c is the ratio of acceleration due to gravity to a conversion constant, λ , the ratio of specific heats, R , the ideal gas constant, T , the absolute temperature, and z , the altitude.

For dry air with a molecular weight of 29, $\gamma = 1.41$, $g/g_c = 1.0 \text{ gmf/gm}$, and $R = 84.78 \text{ (}\frac{\text{gmf}/\text{cm}^2 \text{ liter}}{\text{g mol } ^\circ\text{C}}\text{)}$

$$\lambda = 0.995^\circ\text{C}/100 \text{ m} \quad (2)$$

Air saturated with water vapor releases heat as it cools, and liquid water condenses. Thus the adiabatic temperature decrease is greater than that given by Eq. (1). The wet adiabatic temperature gradient ranges from about -0.9°C per

100 m in the polar region to about -0.4°C per 100 m in the tropics.

The potential temperature is defined as the temperature resulting when dry air is brought adiabatically from its initial state to a standard pressure of 1000 mb (mb = millibar). For an adiabatic expansion of an ideal gas:

$$T_0 = T(P_0/P)^{\gamma-1/\gamma} \quad (3)$$

where T_0 is the temperature at the standard pressure P_0 . If $P_0 = 1000$ mb, then $T_0 = \theta$, the potential temperature. Substituting for T_0 , taking the logarithm of both sides of the equation, and differentiating with respect to z , one can derive the following equation:

$$\frac{1}{\theta} \frac{d\theta}{dz} = \frac{1}{T} \left[\frac{dT}{dz} + \left(\frac{\gamma-1}{\gamma} \right) \left(\frac{g/g_c}{R} \right) \right] \quad (4)$$

The last term can be identified as DALR or λ . For changes in pressure of 200 mb or less, θ is approximately equal to T within 5% error. Thus the potential temperature gradient can be approximated by

$$\frac{\Delta\theta}{\Delta z} = \left[\left(\frac{dT}{dz} \right)_{\text{actual}} + \lambda \right] \quad (5)$$

B. The Richardson Number and Stability

The relative importance of convection and the turbulence generated by mechanical shear can be judged by the Richardson number. In defining the Richardson number, it is convenient first to define a stability parameter s , which will be used later in modeling smoke plume releases,

$$s = \frac{g}{T} \left(\frac{\Delta\theta}{\Delta z} \right) \quad (6)$$

The term s can be thought of as being proportional to the rate at which stability suppresses the generation of turbulence. Turbulence is also being generated by mechanical shear forces at a rate proportional to $(\partial\bar{u}/\partial z)^2$. The Richardson number is the ratio of these two processes.

$$\text{Ri} = \frac{s}{(\partial\bar{u}/\partial z)^2} = \frac{g}{T} \frac{(\partial\theta/\partial z)}{(\partial\bar{u}/\partial z)^2} \quad (7)$$

The Richardson number is a turbulence indicator and also an index of stability. Meteorologists classify atmospheric stability in the surface layer as unstable, neutral, and stable. Strongly negative Richardson numbers indicate that convection predominates, winds are weak, and there is a strong vertical motion characteristic of an unstable atmosphere. Smoke leaving a source spreads rapidly vertically and horizontally. As mechanical turbulence increases, the Richardson number approaches zero, and the

TABLE I Richardson Number and Stability

Classification	Richardson number	Comment
Stable	$\text{Ri} > 0.25$	No vertical mixing, winds weak, strong inversion, mechanical turbulence damped, negligible spreading of smoke plume
Stable	$0 < \text{Ri} < 0.25$	Mechanical turbulence weakened by stable stratification
Neutral	$\text{Ri} = 0$	Mechanical turbulence only
Unstable	$-0.03 < \text{Ri} < 0$	Mechanical turbulence and convection
Unstable	$\text{Ri} < -0.04$	Convection predominant, winds weak, strong vertical motion, smoke rapidly spreading vertically and horizontally

dispersion of a smoke plume decreases, approaching neutral stability where $(\partial\theta/\partial z) = 0$. Finally, as the Richardson number becomes positive, vertical mixing ceases, and mechanical turbulence is damped. The atmosphere becomes stably stratified, and very little vertical dispersion of a smoke plume occurs. Table I summarizes these conditions.

C. Stability Classification Schemes

1. Pasquill–Gifford Stability Classification

As a simplified measure of stability, Gifford modified a system of stability classification based upon suggestions by Pasquill at the British Meteorological Office. In this classification it is assumed that stability in the layers near the ground is dependent on net radiation as an indication of convective eddies and on wind speed as an indication of mechanical eddies. Insolation (incoming radiation) without clouds during the day is dependent on solar altitude, which is a function of time of day and year. When clouds are present, the extent of their coverage and thickness decreases the incoming and outgoing radiation. Daytime insolation must be modified according to the existing cloud cover and ceiling height. At night, the radiation condition is judged solely on cloud cover. Six stability categories are defined in Table II.

2. Turner's Stability Classification

Turner has taken the Pasquill–Gifford work and produced a stability classification that is based on hourly meteorological observations taken at weather bureau stations. This stability classification has been made completely objective so that an electronic computer can be used to compute stability. There are seven stability classes:

TABLE II Pasquill–Gifford Stability Categories

Surface wind (measured at 10 m)		Strong	Day insolation moderate	Slight	Night	
(m/sec)	(mph)				Thinly overcast or $\geq 4/8$ cloudiness ^a	$\leq 3/8$ cloudiness ^a
<2	4.5	A	A–B	B	—	—
2–3	4.5–6.7	A–B	B	C	E	—
3–5	6.7–11.2	B	B–C	C	D	E
5–6	11.2–13.4	C	C–D	D	D	D
6	13.4	C	D	D	D	D

^a The degree of cloudiness is defined as that fraction of sky above the local apparent horizon that is covered by clouds.

NOTES:

1. Insolation is the rate of radiation from the sun received per unit of earth's surface.
2. Strong insolation corresponds to sunny mid-day in summer. Slight insolation corresponds to similar conditions in mid-winter.
3. For A–B, B–C, and so forth, values take the average of A and B values.
4. Night refers to the period from one hour before sunset to one hour after dawn.
5. Regardless of wind speed, the neutral category D should be assumed for overcast conditions during day or night and for any sky conditions during the hour preceding or following night.

1 mph = 0.4470 m/sec

1 m/sec = 2.237 mph

A—extremely unstable	D—neutral
B—moderately unstable	E—slightly stable
C—slightly unstable	F—moderately stable

1. Extremely unstable
2. Unstable
3. Slightly unstable
4. Neutral
5. Slightly stable
6. Stable
7. Extremely stable

Stability class is determined as a function of wind speed and the net radiation index and is listed in [Table III](#). The net radiation index ranges from 4 to –2. An index of 4 is given to the highest positive net radiation, which is radia-

tion directed toward the ground; an index of –2 is given to the highest net negative radiation, which is radiation directed away from the ground. An unstable condition occurs with high positive net radiation and low wind speed; neutral condition with cloudy skies or high wind speeds; stable conditions with high negative net radiation and light winds. The net radiation index is determined in reference to [Table IV](#) and the following procedure.

1. If the total cloud cover is 10/10 and the ceiling is less than 7000 ft, use net radiation index equal to 0 (whether day or night).
2. For nighttime (between sunset and sunrise):
 - a. If total cloud cover $\leq 4/10$, use net radiation index equal to –2.
 - b. If total cloud cover $> 4/10$, use net radiation index equal to –1.

TABLE III Stability Class As a Function of Net Radiation and Wind Speed

Wind speed (knots)	Net radiation index						
	4	3	2	1	0	-1	-2
0,1	1	1	2	3	4	6	7
2,3	1	2	2	3	4	6	7
4,5	1	2	3	4	4	5	6
6	2	2	3	4	4	5	6
7	2	2	3	4	4	4	5
8,9	2	3	3	4	4	4	5
10	3	3	4	4	4	4	5
11	3	3	4	4	4	4	4
≥ 12	3	4	4	4	4	4	4

TABLE IV Insolation As a Function of Solar Altitude

Solar altitude (a)	Insolation	Insolation class number
$60^\circ < a$	Strong	4
$35^\circ < a < 60^\circ$	Moderate	3
$15^\circ < a \leq 35^\circ$	Slight	2
$a \leq 15^\circ$	Weak	1

3. For daytime:

- Determine the insolation class number as a function of solar altitude from [Table IV](#).
- If total cloud cover $\leq 5/10$, use the net radiation index in [Table III](#) corresponding to the insolation class number.
- If cloud cover $> 5/10$, modify the insolation class number by following these six steps.
 - Ceiling < 7000 ft, subtract 2.
 - Ceiling ≤ 7000 ft but $< 16,000$ ft, subtract 1.
 - If total cloud cover equals $10/10$, subtract 1. (This applies only to ceilings ≥ 7000 ft since cases with $10/10$ coverage below 7000 ft are considered in item 1 above.)
 - If insolation class number has not been modified by steps (1), (2), or (3), assume modified class number equal to insolation class number.
 - If modified insolation class number is less than 1, let it equal 1.
 - Use the net radiation index in [Table III](#) corresponding to the modified insolation class number.

In urban areas a large amount of heat is retained in the buildings and pavement after the sun goes down. This heat is reradiated at night. During the day the surfaces of an urban area are more reflective and become hotter, thus producing more convective eddies. For these reasons, convective turbulence in an urban area is not as insignificant as it is in the rural areas, and urban areas are rarely as stable. Thus, it is possible to combine stability categories (6) and (7)—or (5), (6), and (7)—into one category when using Turner's classification for urban modeling.

D. The Planetary Boundary Layer and Its Thickness

The earth's surface exerts a drag on the atmosphere that results in an airflow that is similar to that around a sphere in a wind tunnel. This drag influences wind speed and the mixing of atmospheric-borne substances up to a height of as much as 2 km. In the daytime this region is typically turbulent. At night, with weak winds, the thickness of the turbulent layer can be as low as a few tens of meters. The region is called the planetary boundary layer and is a region in which the atmosphere experiences surface effects through vertical exchange of momentum, heat, and mass in the form of moisture.

The thickness of the planetary boundary layer is characterized by the thickness of the turbulent region next to the ground. Since atmospheric properties are well mixed by the turbulence within the layer, the layer is sometimes called the mixing depth.

The height of the lowest inversion can also be used to describe the thickness of the planetary boundary layer. In the daytime, temperature typically decreases with height within the planetary boundary layer and increases with height above it. Any region in which the temperature increases with height is called an inversion. In an inversion, turbulence is suppressed. Thus in the daytime the mixed layer can be approximated by the height of the lowest inversion.

On clear nights with weak winds, infrared radiation establishes a ground-based inversion. Only the bottom of the boundary layer becomes turbulent under these conditions. Thus the heights of the inversion and the mixed layer are considerably different.

1. Neutral Conditions

Within the mixed layer, the Earth's rotation influences the vertical wind shear and therefore the intensity of turbulence. Thus, under neutral conditions, the depth of the mixed layer should be proportional to the surface friction velocity u_* and the Coriolis parameter f ,

$$u_* = \sqrt{\tau/\rho} \quad (8)$$

$$f = 2\Omega \sin \phi \quad (9)$$

where τ is surface stress, ρ air density, ϕ latitude, and $\Omega = 7.29 \times 10^{-5}$ sec $^{-1}$, the rate of the earth's rotation. The thickness can then be estimated as

$$h \simeq 0.2 \frac{u_*}{f} \quad (10)$$

Since u_* increases with wind speed and surface roughness, h increases with these two parameters as well.

2. Unstable Conditions

During the day the depth of the mixed layer grows as a function of the surface heat flux H . If the initial daytime surface temperature is T_0 and the temperature at time t is T after an adiabatic temperature profile has been established, then conservation of heat energy requires:

$$\int_{t_0}^t H dt = \frac{C_p \rho h (T - T_0)}{2} \quad (11)$$

where C_p is the specific heat of air.

From this we can calculate the depth of the mixed layer noting that the ratio $(T - T_0)/h$ is equivalent to $\lambda - dT/dz$:

$$h = \left[\frac{2 \int_{t_0}^t H dt}{C_p \rho (\lambda - dT/dz)} \right]^{1/2} \quad (12)$$

Equation (12) is a simplified expression which can be affected by atmospheric phenomena that cause the inversion layer to change.

3. Stable Conditions

The thickness of the lowest continuously mixed turbulent layer can be estimated from

$$h \simeq 0.4 \sqrt{\frac{u_*}{f} L} \quad (13)$$

where L is the Monin–Obukhov length to be defined in the next section. This equation has had relatively little success in correlating measured values.

4. Holzworth's Method

Mixing depth has been determined by Holzworth of the U.S. Environmental Protection Agency in reference to the morning and afternoon. The morning mixing height is calculated as the height above ground at which the dry adiabatic extension of the morning minimum surface temperature plus 5°C intersects the vertical temperature profile observed at 12:00 Greenwich Mean Time. The afternoon mixing height is calculated in the same way except that the maximum surface temperature observed between 12:00 and 16:00 local standard time is substituted for the minimum temperature plus 5°C. A compilation of the mixing depth data has been prepared by the U.S. Environmental Protection Agency.

E. Surface Layer Wind Structure

There is no precise definition of surface layer. Typically the vertical fluxes of momentum, heat, and moisture are large at the surface and decrease to zero at the top of the planetary boundary layer. It would be well to make a definition of surface layer such that the fluxes remain constant. However, a change of 10% in flux is tolerable. Assuming that decrease in flux is linear with height, a feasible definition of surface layer is therefore 10% of the mixing depth.

The near constancy of fluxes in the surface layer implies that wind direction does not change with height. Thus the mean wind can be described by \bar{u} only, the mean wind speed. Mean variables in the surface layer are also functions only of height, and the variation with height is controlled by the surface stress, the vertical heat flux at the surface, and the terrain roughness.

Because both mechanical and convective forces determine turbulence, Monin and Obukhov proposed a similarity theory that introduced two scaling parameters, the friction velocity u_* and the length L , where

$$L = -\frac{\rho C_p T u_*^3}{kgH} \quad (14)$$

with upward heat flux H taken as positive. Note that L is computed from measurements as close to the surface as

possible and taken as independent of height. The value of L depends only on u_* and H and is about –10 m on strongly convective days, –100 m on windy days with some solar heating, and approaches infinity in purely mechanical turbulence. Since heat flux H is downwind at night, L is positive and small in light-wind, stable conditions.

According to the Monin–Obukhov hypothesis, various statistics of atmospheric parameters, when normalized by proper powers of u_* and L , are universal functions of z/L . They define a nondimensional wind shear S , which is a function of z/L ,

$$S = \frac{k_a z}{u_*} \frac{\partial \bar{u}}{\partial z} = \phi_m \left(\frac{z}{L} \right) \quad (15)$$

1. Neutral Air

For neutral air, that is, in the case of purely mechanical turbulence, $S = 1.0$ and

$$\bar{u} = \frac{u_*}{k_a} \ln \left(\frac{z}{z_0} \right) \quad z \geq z_0 \quad (16)$$

the classical logarithmic wind profile. Here k_a is the von Karman constant. The value of this constant varies from 0.35–0.43; usually a value of 0.40 is used for the case in which measurements have not been made.

The term z_0 is the value of z at which \bar{u} vanishes. It represents turbulent eddy size at the surface and is a measure of roughness. Thus z_0 is termed the *roughness length*. Table V lists typical values of z_0 and u_* .

2. Unstable and Stable Air

The relations below have been established for unstable and stable air:

TABLE V Values of z_0 and u_* for Use in Vertical Wind Speed Profiles

Type of surface	z_0 in cm	u_* in m/sec ^a
Smooth mud flats, ice	0.001	0.16
Smooth snow	0.005	0.19
Smooth sea	0.02	0.22
Level desert	0.03	0.23
Snow surface, lawn grass to 1.0 cm high	0.1	0.26
Lawn, grass to 5 cm	1–2	0.38–0.43
Lawn, grass to 60 cm	4–9	0.51–0.65
Fully grown root crops	14	0.75
Pasture land	20	0.87
Suburban housing	60	1.66
Forests, cities	100	2.89

^a For \bar{u} at 2.0 m = 5.0 m/sec, $k_a = 0.40$.

$$\phi_m\left(\frac{z}{L}\right) = \left(1 - 15\frac{z}{L}\right)^{1/4} \quad (\text{unstable}) \quad (17)$$

$$\phi_m\left(\frac{z}{L}\right) = \left(1 + 5\frac{z}{L}\right) \quad (\text{stable}) \quad (18)$$

Equation (15) can then be integrated to give

$$\bar{u} = \frac{u_*}{k_a} \left[\ln\left(\frac{z}{z_0}\right) - \psi_m\left(\frac{z}{L}\right) \right] \quad (19)$$

where

$$\psi_m\left(\frac{z}{L}\right) = \int_0^{z/L} [1 - \phi_m(\xi)] \frac{d\xi}{\xi} \quad (20)$$

Applying Eqs. (17) and (18) to this integral results in the following equations for the velocity distribution in unstable and stable flow:

$$\begin{aligned} \bar{u} &= \frac{u_*}{k_a} \left\{ \ln\left(\frac{z}{z_0}\right) - 2 \ln\left[\frac{1}{2} \left(1 + \frac{1}{\phi_m}\right)\right] \right. \\ &\quad \left. - \ln\frac{1}{2} \left[\left(1 + \frac{1}{\phi_m^2}\right)\right] \right. \\ &\quad \left. + 2 \tan^{-1}\left(\frac{1}{\phi_m} - \frac{\pi}{2}\right) \right\} \quad (\text{unstable}) \end{aligned} \quad (21)$$

$$\bar{u} = \frac{u_*}{k_a} \left[\ln\left(\frac{z}{z_0}\right) + 5\left(\frac{z}{L}\right) \right] \quad (\text{stable}) \quad (22)$$

It should be noted that for neutral conditions, the integral of Eq. (20) is zero, and Eq. (19) reduces to Eq. (16). Furthermore, Eq. (19) can be solved for the friction velocity u_* ,

$$u_* = \frac{k_a \bar{u}}{\ln(z/z_0) - \psi_m(z/L)} \quad (23)$$

Thus Eq. (23), with appropriate measurements, can be used to determine u_* . Equation (23) shows that u_* increases with wind speed \bar{u} and roughness z_0 at a given height z .

3. The Power Law for the Wind Profile

It is common engineering practice to describe the wind profile with a power law

$$\frac{\bar{u}}{u_m} = \left(\frac{z}{z_m}\right)^p \quad (24)$$

where \bar{u} is the average wind speed at height z , u_m average wind speed measured at height z_m , and p power whose value is dependent upon stability conditions and surface roughness. It can be shown that

$$p = \frac{\phi_m(z/L)}{\ln(z/z_0) - \psi_m(z/L)} \quad (25)$$

TABLE VI Estimates of Power p in Velocity Profile Equation

	Stability class					
	A	B	C	D	E	F
Urban	0.15	0.15	0.20	0.25	0.30	0.30
Rural	0.07	0.07	0.10	0.15	0.35	0.55

where z should be interpreted as the average height of the layer involved. For the case of neutral conditions with strong winds this equation reduces to

$$p = 1/\ln(z/z_0) \quad (26)$$

The geometric mean value of z over the layer to be dealt with is then more appropriate. For a windy day with some sunshine $L = -100$ m, with $z_0 = 20$ cm, a value appropriate for grasslands with some trees, $z = 100$ m and $z_m = 10$ m, and $p = 0.20$ with mechanical turbulence and 0.145 when convection is added. It is apparent that convection can have a significant effect on the value of p .

Values of p have been found that range from 0.02–0.87. Table VI lists values of p that can be used successfully up to a height of 200 m. This table reflects the effect of surface conditions by presenting values of p for urban and rural areas. Stability effects are given according to the Pasquill–Gifford classes.

III. GAUSSIAN PLUME MODEL FOR CONTINUOUS SOURCES

The physical picture of a smoke plume can be developed by considering first a puff of smoke emitted as a point source. The puff is made up of a gas or small particles which follow the direction of the wind with the speed of the wind. Small eddies cause dilution and expansion of the puff about the centerline of its path by pumping fresh air into the puff. Large eddies buffet the puff about and transport it downwind. Linking together an infinity of puffs results in the formation of continuous release from a point.

An instantaneous photograph of a continuous plume would show violent fluctuations and a meandering flow. Figure 1 shows an instantaneous plume compared with time-averaged plumes. The 10-min average shows a smoothed concentration profile; the 2-hr average is the same but has a reduced maximum and is more spread out. The easiest to describe mathematically is the time-averaged case, which appears as a Gaussian curve.

The Gaussian plume model for atmospheric diffusion has emerged as the most commonly used mathematical technique for dispersion calculations from continuous sources. There are a number of factors in its favor.

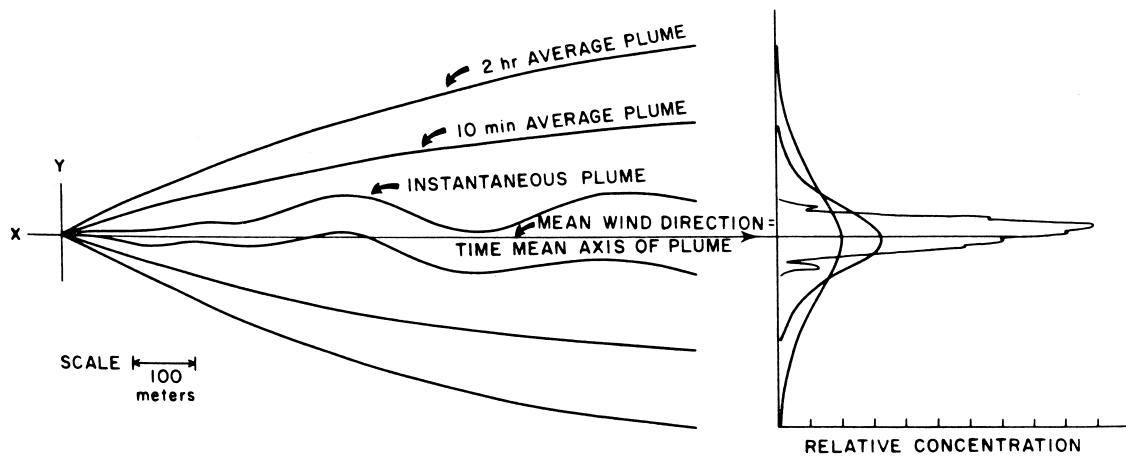


FIGURE 1 Comparison of an instantaneous plume with a time-averaged plume.

1. It produces results that agree with experimental data as well as any model.
2. It can be obtained as an analytical solution to the deterministic model for constant wind speed and diffusion coefficient.
3. Since the turbulent motion of the atmosphere is essentially a random process, a Gaussian format is suggested to describe the nature of the resultant action of the atmosphere.
4. Other concentration distributions would fit the data equally well, but the Gaussian is not mathematically complex, and it is conceptually satisfying.

There are also a number of limitations that reduce the effectiveness of the Gaussian model for general use. Wind shear is not considered. All meteorological factors are assumed to remain constant over space and over the averaging time used for the calculation. Calm winds are excluded, and bulk transport of the pollutant along the mean wind direction is assumed to be much greater than diffusion in the mean wind direction. No transient-response case can be handled by the Gaussian plume model, and chemical and physical transformations of the plume can be handled only in a very simple manner. It is assumed that the plume travels on flat terrain, or at best, several simple terrain corrections are made, and the plume is reflected off the ground and any diffusion ceiling imposed, such as an inversion.

Some of these limitations can be overcome by the use of other model types. For example, predictions of downwind concentrations from instantaneous releases or puff diffusion can be treated by reference to similarity theory. Shear fields can be introduced in the numerical solution of the deterministic model, and the statistical theory can assist in developing the Gaussian model itself.

A. The Statistical Approach to the Turbulent Diffusion Process

The dispersion of an ensemble of small particles or cluster of molecules in space must take place through the random motions of the ensemble. In this random or stochastic process, the variables are usually functions of time, and the value of the variables can be specified only in terms of a probability distribution.

There are two ways to look at a cloud of diffusing molecules: the Eulerian viewpoint or the Lagrangian viewpoint. In the Eulerian viewpoint, the diffusion equations would be derived from a consideration of concentrations and flux at a fixed point in space. Such quantities are easily observable. It is possible to arrive at the same results from the Lagrangian viewpoint, which focuses on the history of the random movements of the diffusing material. Statistical properties of the random motions would have to be mathematically described to produce useful results.

The Eulerian approach to diffusion can be regarded as relating to the ensemble average concentration field of the diffusing molecules. In turbulent diffusion, one must distinguish between the ensemble average concentration field and an instantaneously observable one. The kinematics of diffusing particle movements leads to elucidation of certain properties of the turbulent diffusion process. These considerations are of prime importance in arriving at a satisfactory understanding of the complex turbulent phenomenon. The results of the Lagrangian treatment of the random motion complements the information derived from the Eulerian treatment.

1. Taylor's Theorem

To relate the ensemble average field to the kinematics of the diffusing particle movements, we turn to probabilistic

considerations of dispersion through random movements. When the probability distribution is not a function of time, in a random or stochastic process, the process is said to be *stationary*. Other stochastic processes are known as *evolutionary*. In turbulent diffusion, when the turbulent field is homogeneous in temperature, mean velocity, and turbulent intensity, the velocity of a diffusing particle becomes a stationary process. In a stationary process many properties become independent of time, for example:

- Ensemble mean value
- Square of the mean value
- Any chosen product of the mean value

We may then choose the y -component of the velocity vector to be zero by measuring velocities relative to a frame of reference moving with the mean value of the ensemble. Thus, $v(t) = 0$ and v^2 is a constant independent of time.

A stationary stochastic process can be characterized by the autocorrelation function of velocity $R(\tau)$, where τ is a time delay. The autocorrelation function measures the persistence of a given value of the random variable concerned. For particle diffusing velocities, when the particle possesses a given y -directed velocity component $v(t)$, a short time later it is likely to have a velocity of similar magnitude and sign, $v(t + \tau)$. The velocity covariance can be formed for the particle as the mean of the product of the two velocities.

$$\text{covariance} = \overline{v(t) \cdot v(t + \tau)} \quad (27)$$

where

$$\lim_{\tau \rightarrow 0} \overline{v(t) \cdot v(t + \tau)} = \overline{v(t)^2} \quad (28)$$

If there are no organized flow structures such as standing waves present, after a long time the two velocities become independent of each other, and essentially any resemblance to the original velocity $v(t)$ is a pure coincidence. Under these conditions the covariance becomes the product of the mean of the two velocities.

$$\lim_{\tau \rightarrow 0} \overline{v(t) \cdot v(t + \tau)} = \overline{v(t)} \cdot \overline{v(t + \tau)} \quad (29)$$

where $\overline{v(t)} = 0$ and $\overline{v(t + \tau)} = 0$ by definition. Note also, since the process is stationary, the covariance is independent of whatever time t is chosen for the basis of calculations, which implies that the velocity covariance is an even function.

The autocorrelation function is chosen to be a nondimensional function of the covariance. Thus,

$$R(\tau) = \frac{\overline{v(t) \cdot v(t + \tau)}}{\overline{v^2}} \quad (30)$$

Note that $\lim_{\tau \rightarrow \infty} R(\tau) = 0.0$ and for all τ

$$-1.0 \leq R(\tau) \leq 1.0 \quad (31)$$

It has become customary to identify $(\overline{y^2})^{1/2}$ with σ_y , the width of a plume at a fixed value of downwind distance x . Since $\overline{y^2}$ refers to particles diffusing after a fixed diffusion time, this is only an approximation. However, it is a useful approximation, and we can write

$$\frac{d\sigma_y^2}{dt} = \frac{d\overline{y^2}}{dt} = 2y \frac{dy}{dt} = \overline{2yv} \quad (32)$$

and then

$$\frac{d\sigma_y^2}{dt} = 2 \int_0^t \overline{v(t)v(t + \tau)} dt \quad (33)$$

Replacing $\overline{v^2}$ with σ_v^2 , Eq. (30) can be substituted into Eq. (33), producing Taylor's equation:

$$\sigma_y^2 = 2\sigma_v^2 \int_0^t \int_0^{t'} R(t') dt' dt \quad (34)$$

2. Some Approximations

For small diffusion times at $t \rightarrow 0$, $R(t) \rightarrow 1.0$ and Eq. (34) becomes

$$\sigma_y^2 \simeq \sigma_v^2 t^2 \quad (35)$$

or

$$\sigma_y \propto t$$

For large diffusion times $t \rightarrow \infty$, and

$$\int_0^t R(t') dt' \rightarrow T$$

where T is a constant known as the time scale.

$$\sigma_y^2 \simeq 2\sigma_v^2 T t \quad (36)$$

or

$$\sigma_y \propto t^{1/2}$$

If we now assume that $R(t)$ can take a simple exponential form

$$R(t) = \exp\left(-\frac{t}{T}\right) \quad (37)$$

then from Eq. (34),

$$\sigma_y^2(t) = 2\sigma_v^2 T^2 \left[\frac{t}{T} - 1 + \exp\left(-\frac{t}{T}\right) \right] \quad (38)$$

Now define a function of (t/T) such that

$$f_y\left(\frac{t}{T}\right) = \frac{\sigma_y}{\sigma_v t} = 2^{1/2} \frac{T}{t} \left[\frac{t}{T} - 1 + \exp\left(-\frac{t}{T}\right) \right]^{1/2} \quad (39)$$

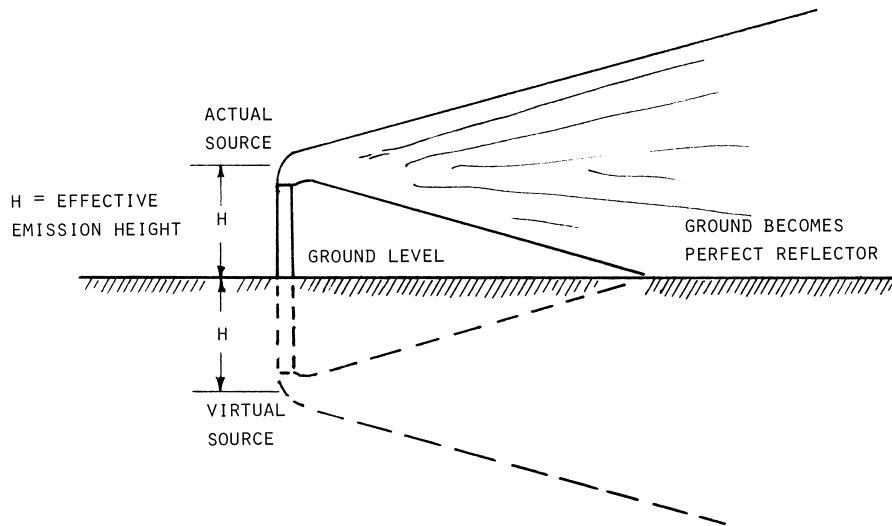


FIGURE 2 Earth-boundary problem.

B. Gaussian Diffusion Formula—Continuous Release

1. Surface Releases

The spatial distribution of concentration is expressible in terms of the shape of the crosswind and vertical concentration profiles. These profiles are expressed in terms of their dimension σ_y , defined in the last section, and σ_z , which can be similarly defined. The parameters σ_y and σ_z in this case are standard deviations of the concentration distribution.

The continuity equation can then be written as a function of wind speed $u(z)$ and the source strength or emission rate Q .

$$Q = \int_0^{+\infty} \int_{-\infty}^{+\infty} u(z) c(x, y, z) dy dz \quad (40)$$

where $c(x, y, z)$ is the concentration distribution. For a ground level release, where the variation of wind speed with height is neglected, the Gaussian formulation that follows satisfies Eq. (40).

$$C(x, y, z) = \frac{Q}{\pi \bar{u} \sigma_y \sigma_z} \exp \left[-\frac{1}{2} \left(\frac{y^2}{\sigma_y^2} + \frac{z^2}{\sigma_z^2} \right) \right] \quad (41)$$

2. Elevated Releases

Most continuous sources that we deal with are located near the earth's surface. Thus eventually the plume will intersect the earth, and we are then forced to account for this physical barrier to the flux. The problem has usually been handled by borrowing the virtual image source technique from heat conduction studies. It is assumed that an iden-

tical virtual source is located symmetrically to the actual source with respect to the ground, as shown in Fig 2. The resultant equation is

$$\begin{aligned} \bar{C}(x, y, z) = & \frac{Q}{2\pi \bar{u} \sigma_y \sigma_z} \exp \left[-\frac{1}{2} \left(\frac{y}{\sigma_y} \right)^2 \right] \\ & \times \left\{ \exp \left[-\frac{1}{2} \left(\frac{z-H}{\sigma_z} \right)^2 \right] \right. \\ & \left. + \exp \left[-\frac{1}{2} \left(\frac{z+H}{\sigma_z} \right)^2 \right] \right\} \end{aligned} \quad (42)$$

In this formula, H represents the height of the source above the surface. The formula predicts that the concentration is zero at the source and remains effectively zero for some distance from the source. It then rises rapidly to a maximum and thereafter falls off to approach the curve for surface release asymptotically.

If the wind speed \bar{u} should go to zero, the concentration predicted by formula (42) would approach infinity. In practice, calculations are limited in calm winds conditions to $\bar{u} = 0.5$ m/sec.

C. The Accuracy of the Gaussian Plume Model

The accuracy of the Gaussian diffusion model has been reviewed in a note prepared by the [American Meteorological Society 1977](#) Committee on Atmospheric Turbulence and Diffusion. The Committee estimates can be summarized as follows:

1. For models discussed above, employing sigma values, estimating concentrations in ideal circumstances where there is uniform terrain, steady meteorology,

- with source and ambient parameters measured by research-grade instruments, the observed maximum downwind ground-level concentration value should be within 10–20% of that calculated for a ground level source and within 20–40% for an elevated source.
2. In most real-world applications for which the controlling meteorological parameters are measured from a tower, conditions are reasonably steady and horizontally homogeneous (less than ~50%) variation from the spatial and temporal average during the experiment) and no exceptional circumstances exist that could affect the atmospheric dispersive capacity in ways not accounted for by the model, accuracy within a factor of 2 can be expected.

Some important meteorological circumstances that can be classified exceptional are as follows:

1. Aerodynamic wake flows of all kinds, including stack downwash, building wakes, highway vehicle wakes, and wakes generated by terrain obstacles
2. Buoyant fluid flows, including power plant stack plumes and accidental releases of heavy, toxic gases
3. Flows over surfaces markedly different from those represented in the basic experiments, including dispersion over forests, cities, water and rough terrain
4. Dispersion in extremely stable and unstable conditions
5. Dispersion at great downwind distances (greater than 10–20 km)

In addition, any physical or chemical process, such as chemical reaction, dry deposition, resuspension, or precipitation scavenging, produces additional uncertainties in model predictions.

D. Summary of the Method of Application of the Equations

1. Determine stability conditions from Pasquill–Gifford table or by Turner's method.
 2. Estimate values of σ_y and σ_z as a function of downwind distance and stability class for stated averaging time. (A set of measured values must be available. See the next section.)
 3. For short stacks and for ground level sources, measure \bar{u} at an elevation of 10 m. This \bar{u} may be used for downwind dispersion distances up to 1.0 km.
- For tall stacks or for downwind dispersion distances greater than 1.0 km, \bar{u} should be determined at the mean height of the plume or averaged through the vertical extent

of the plume. Equation (24) could be used to calculate the effect of altitude on wind speed.

It has become the practice to set \bar{u} to the value it would have at stack height, using Eq. (24) to adjust the wind speed from the measured height.

Set $\bar{u} = 0.5$ m/sec for calm winds.

4. Determine source strength Q and effective stack height H .

5. Calculate time-averaged concentration. The averaging time is the same used to determine the σ values.

IV. DIFFUSION COEFFICIENTS

Diffusion coefficients or sigma values must be determined by measurement. Several significant studies have been carried out in which values are reported in the literature. Time permitting, it is reasonable and most accurate to measure, at the site of interest, the meteorological parameters required to determine the sigma values. Research-grade instrumentation is required, but it is not complex for level terrain with grass or agricultural ground cover and very few obstacles in the path of the diffusing materials. The more complex the surrounding terrain and buildings become, the more complex the instrumentation must be to carry out the necessary measurements.

To make an estimate of σ_y , or σ_z , the stability class must first be determined. The two typing techniques of Pasquill–Gifford and Turner discussed previously can be used. Then a series of curves or formulas are referenced to find values for σ_y and σ_z as a function of stability class, downwind distance, and averaging time. For the values of T_y and T_z that follow, averaging time should be considered to be 1 hr.

[Figure 3](#) presents a series of curves adapted by Gifford from a report of an experimental program to determine σ_y and σ_z . Work was not carried out beyond about 1.0 km, and the curves are very tentative beyond that distance.

Because calculators and computers are in widespread use, an analytical formula is possible and most desirable. Briggs has made use of several sets of experimental curves and theoretical concepts regarding asymptotic limits to produce the formulas in [Table VII](#).

Previously, Eq. (39), a theoretical formula for σ_y , was developed from a statistical model. Taylor's work suggests that formulas for σ_y and σ_z similar to Eq. (39) can be written as follows:

$$\sigma_y = \sigma_{\theta} x f_y \left(\frac{x}{\bar{u} T_y} \right) \quad (43)$$

$$\sigma_z = \sigma_e x f_z \left(\frac{x}{\bar{u} T_z} \right) \quad (44)$$

where

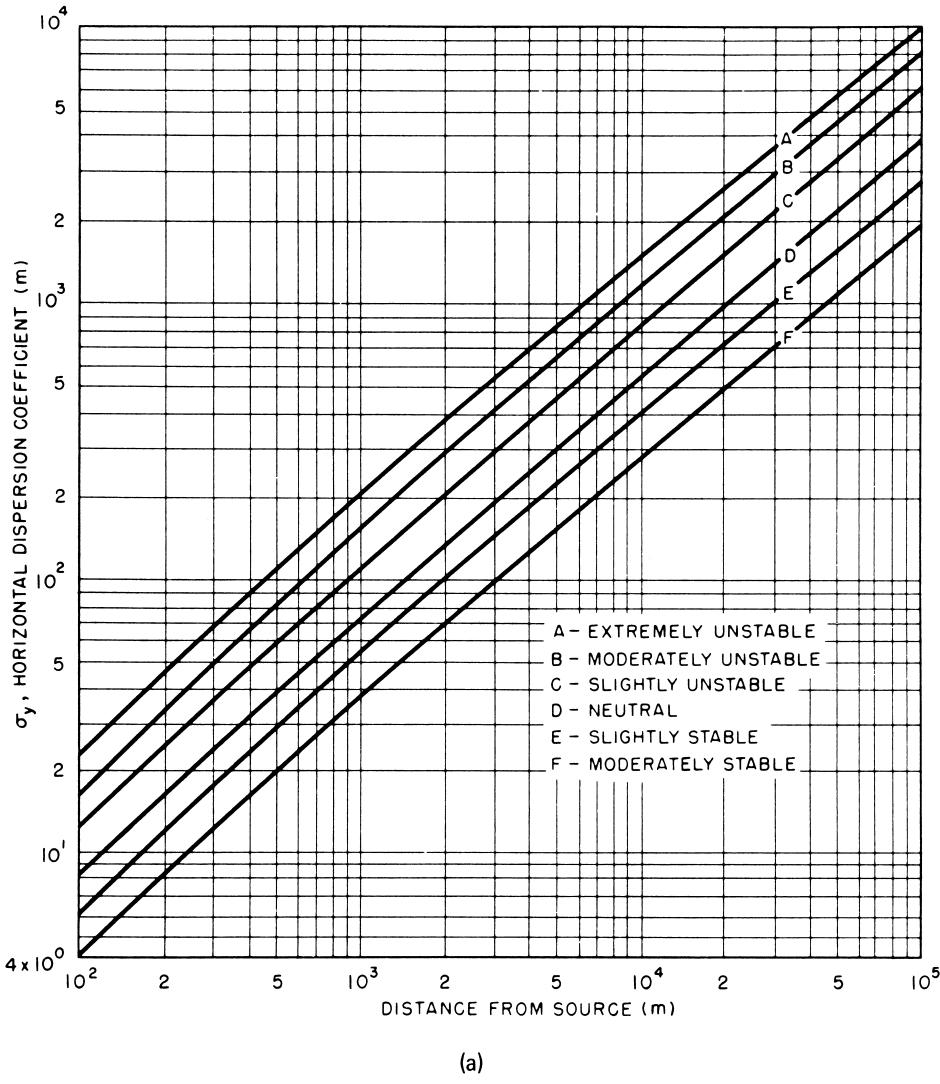


FIGURE 3 (a) Horizontal dispersion coefficient as a function of downwind distance. (b) Vertical dispersion coefficient as a function of downwind distance.

$$\sigma_\theta = \frac{\sigma_v}{\bar{u}} \quad (45)$$

$$\sigma_e = \frac{\sigma_z}{\bar{u}} \quad (46)$$

and the diffusion time $t = x/\bar{u}$.

Then σ_y and σ_z are completely determined by observations of σ_θ and σ_e , the standard deviations of the wind direction fluctuations in the horizontal and vertical directions, respectively. These quantities can be determined from wind bivane measuring instruments, which respond to both vertical and horizontal fluctuations.

An empirical relationship for f_y has been developed.

$$f_y = (1 + 0.031x^{0.46})^{-1} \quad \text{for } x \leq 10^4 \text{ m} \quad (47)$$

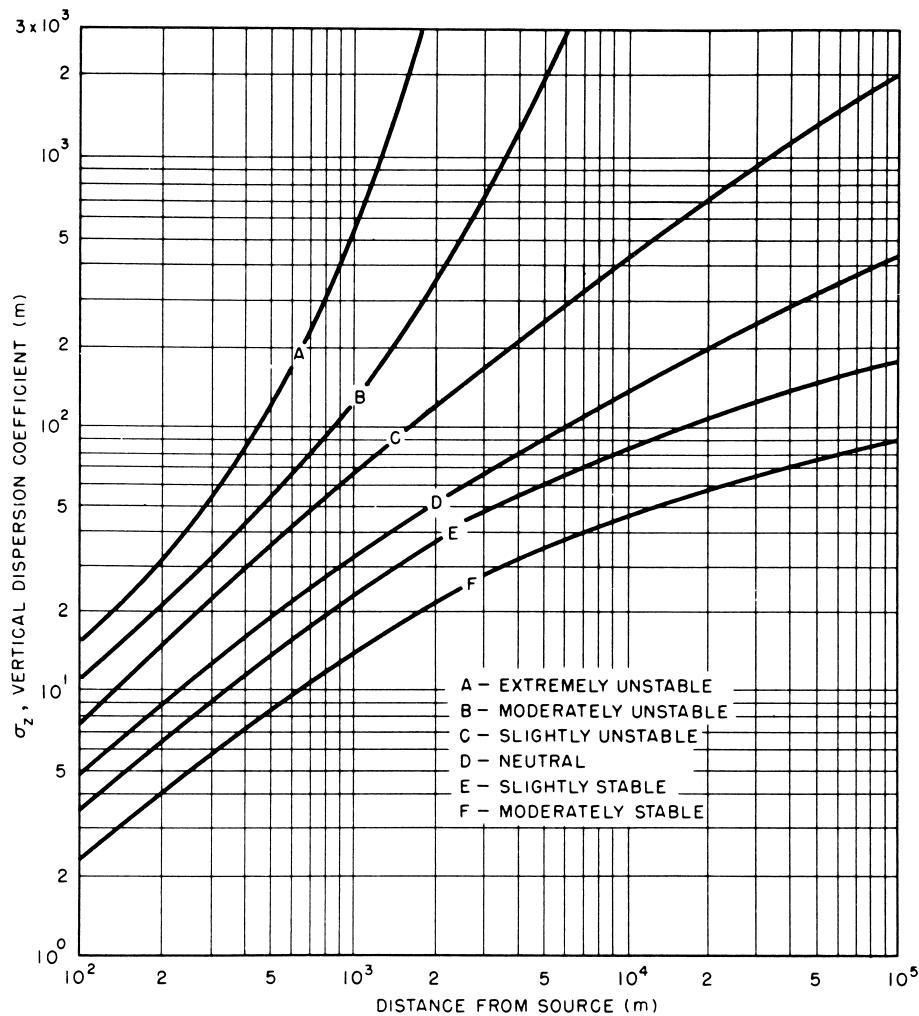
$$f_y = 33x^{-1/2} \quad \text{for } x > 10^4 \text{ m} \quad (48)$$

A universal function for f_z is difficult to determine since few data are available on vertical concentration distribution.

V. PLUME RISE

A. The Effects of Density Differences on Atmospheric Diffusion

Up to this point we have been dealing with diffusion as if the plume did not exhibit a density different from that of the surrounding air. Most chimney plumes, especially those resulting from combustion processes, are made up of gases whose composition is very similar to air. The density difference comes about due to the higher than ambient temperature in the plume. This higher temperature causes



(b)
FIGURE 3 (Continued)

the plumes to be buoyant and induces the phenomenon called *plume rise*.

Some plumes are emitted at about ambient temperature but are made up of gases whose composition and density are much different from the surrounding air. These gases possess a positive buoyancy if they are lighter than air and a negative buoyancy if they are heavier than air. In either case, the plume has a momentum due to the velocity of emission as well as buoyancy. The momentum contributes to plume rise in the initial stages close to the chimney.

The effect of buoyant or negatively buoyant plume gases on diffusion is to

1. Introduce a systematic vertical motion that affects the randomness of the turbulence,
2. Influence the energy chain of the turbulent motions, and

3. Influence the bodily motion of the whole plume, thus altering the effective emission height to be used in the diffusion models.

It is usual to write the effective emission height as the sum of the actual height of the chimney h_s and the plume rise Δh as

$$H = h_s + \Delta h \quad (48)$$

It is assumed that the plume is emitted with a velocity sufficient to escape the downwash effects of the wake behind the chimney. It is also assumed that the chimney is of sufficient height, and the terrain is sufficiently level, so that the effects of buildings and terrain can be neglected. Upon leaving the chimney, the plume usually encounters a crosswind which causes the plume to bend. The effect on the plume can be described in the following stages:

TABLE VII Formulas Recommended by Briggs ($10^2 < x < 10^4$ m)

Pasquill type	σ_y, m	σ_z, m
Open-country conditions		
A	$0.22x(1 + 0.0001x)^{-1/2}$	$0.20x$
B	$0.16x(1 + 0.0001x)^{-1/2}$	$0.12x$
C	$0.11x(1 + 0.0001x)^{-1/2}$	$0.08x(1 + 0.0002x)^{-1/2}$
D	$0.08x(1 + 0.0001x)^{-1/2}$	$0.06x(1 + 0.00015x)^{-1/2}$
E	$0.06x(1 + 0.0001x)^{-1/2}$	$0.03x(1 + 0.0003x)^{-1}$
F	$0.04x(1 + 0.0001x)^{-1/2}$	$0.016x(1 + 0.0003x)^{-1}$
Urban conditions		
A-B	$0.32x(1 + 0.0004x)^{-1/2}$	$0.24x(1 + 0.001x)^{-1/2}$
C	$0.22x(1 + 0.0004x)^{-1/2}$	$0.20x$
D	$0.16x(1 + 0.0004x)^{-1/2}$	$0.14x(1 + 0.0003x)^{-1/2}$
E-F	$0.11x(1 + 0.0004x)^{-1/2}$	$0.08x(1 + 0.00015x)^{-1/2}$

Initial or Jet Phase. As the plume begins to bend in the crosswind, the vertical momentum is nearly all converted to horizontal momentum. This action takes place within a distance of about three to five stack diameters above the stack. During the process the mass increases about 30 times due to the vigorous mixing. The upward velocity of the plume persists due to buoyancy.

Thermal Phase. In the thermal phase, mixing is due to self-generated turbulence. The plume retains a smooth shape and continues to show a moderate rise. The most dominant effect determining the path of the plume centerline is the total excess heat of the plume.

Breakup Phase. It has been observed that plume rise continues after the thermal phase, progressing to a situation in which atmospheric turbulence begins to dominate the mixing. At this point there is a breakup of the plume into distinct parcels, with a nearly stepwise increase in plume diameter. The effect is more pronounced in strong turbulence.

Diffuse Phase. Finally, far enough downwind, the plume appears again to attain some cohesiveness, though it is now large and more diffuse. Atmospheric turbulence continues to dominate, but growth of the plume is slow.

The continuing rise of a positively buoyant plume effectively increases emission height and reduces subsequent ground level concentrations. Ground level concentration is roughly proportional to the inverse square of the effective emission height. Thus a plume rise that doubles the effective emission height reduces ground level concentration by about a factor of four. It is apparent that the accurate prediction of plume rise is most desirable.

B. Plume Rise Formulas

There have been many plume rise observations, with the purpose of developing a correlation between chimney and meteorological variables. At best, observation of plume rise is difficult due to turbulence in the atmosphere, which causes unpredictable fluctuations of the plume. Therefore, much of the data are erratic, and most of the empirical relations are valid only for the particular chimneys for which they were developed.

It is generally agreed that the equations prepared by Briggs are most representative of the greatest number of situations. Briggs was careful to study each set of data and to employ the widest possible conditions in developing his correlations. These equations form the basis of all of the U.S. Environmental Protection Agency computer algorithms. It is also assumed that the plumes are unaffected by building wakes or stack-tip downwash.

The equations under subsections 1 and 2 following are recommended for plumes dominated by buoyancy. The buoyancy flux parameter is defined as

$$F = g V_s \frac{d_s^2}{4} \left(1 - \frac{T_a}{T_s} \right) \quad (49)$$

where g is the acceleration due to gravity, 9.8 m/sec^2 , V_s efflux velocity of the plume from the top of the stack in m/sec , d_s inside diameter of stack at the point of plume efflux in meters, T_s absolute temperature of stack gases as emitted in K , and T_a absolute ambient air temperature in K . It is assumed that the plume rises through the thermal phase and then ceases to rise. The final plume rise and the distance to the final plume rise in meters are calculated as follows:

- For buoyancy dominated plumes, unstable or neutral conditions A, B, C, and D,

$$x^* \equiv 34F^{2/5} \quad \text{for } F \geq 55 \text{ m}^4/\text{sec}^3 \quad (50)$$

$$x^* \equiv 14F^{5/8} \quad \text{for } F < 55 \text{ m}^4/\text{sec}^3 \quad (51)$$

Distance to the final rise = $3.5x^*$ in meters. The final plume rise is

$$\Delta h = \frac{1.6F^{1/3}(3.5x^*)^{2/3}}{\bar{u}} \quad (52)$$

For distances less than $3.5x^*$ m,

$$\Delta h = \frac{1.6F^{1/3}(x)^{2/3}}{\bar{u}} \quad (53)$$

- For buoyancy dominated plumes, stable conditions E and F, the stability parameter is used.

$$s \equiv \frac{g}{T_a} \left(\frac{\Delta\theta}{\Delta z} \right) \quad (54)$$

If $(\Delta\theta/\Delta z)$ is not given, use

0.02 K/m for E stability

0.035 K/m for F stability

The plume rise is given by,

$$\Delta h = 2.6(F/\bar{u}s)^{1/3} \quad (55)$$

and the distance to final rise is:

$$x_f = 2.07 \frac{\bar{u}}{s^{1/2}} \quad (56)$$

When the stack gas temperature is greater than the ambient air temperature, it must be determined whether the plume rise is buoyancy dominated or momentum dominated. For this purpose a crossover temperature $(\Delta T)_c$ is calculated.

3. For momentum dominated plumes, unstable or neutral conditions A, B, C, and D, the form of the equation for $(\Delta T)_c$ is determined by the value of the buoyancy flux parameter F , Eq. (49).

For F less than 55,

$$(\Delta T)_c = 0.0297 T_S V_S^{1/3} / d_S^{2/3} \quad (57)$$

For F equal to or greater than 55,

$$(\Delta T)_c = 0.00575 T_S V_S^{2/3} / d_S^{1/3} \quad (58)$$

When $\Delta T = T_S - T_a$ is less than $(\Delta T)_c$, the assumption is made that plume rise is dominated by momentum. For plumes dominated by momentum in unstable and neutral conditions, the plume rise is calculated from,

$$\Delta h = 3d_S V_S / \bar{u} \quad (59)$$

4. For momentum dominated plumes, stable conditions E and F, the stability parameter s , Eq. (54) is used.

For the case where the atmosphere is stable, the crossover temperature is calculated from:

$$(\Delta T)_c = 0.01958 T_a V_{SS}^{1/2} \quad (60)$$

Once again if ΔT is less than ΔT_c , the plume rise is assumed to be momentum dominated. In this case of a stable atmosphere, the plume rise is calculated from:

$$\Delta h = 1.5 [V_S^2 d_S^2 T_a / (4T_S \bar{u})]^{1/3} s^{-1/6} \quad (61)$$

Equation (59), for unstable-neutral momentum rise is also evaluated. The lower result of these two equations is used as the resulting plume rise. Since momentum rise occurs quite close to the point of release, the distance to final rise is set equal to zero.

VI. ACCIDENTAL RELEASES OF MATERIAL TO THE ATMOSPHERE

A class of problems that has drawn recent interest deals with accidental releases of material to the atmosphere. Two types of accidental releases are of primary concern. A

catastrophic rupture of a pipeline or a vessel, or a spill from a vessel, can produce a release that lasts from a few seconds to a few minutes. This results in a burst of material or a puff-type release. Gases or liquids may leak from around seals, pipe joints, and valves or from cracks or holes in vessels. This type release may start slowly and increase in size. High pressure releases of both gasses and liquids from pressure relief valves or pressure seal ruptures could occur. These high pressure releases may be accompanied by flashing of the fluid to a vapor-liquid mixture. If this kind of a release lasts from 10 min to half an hour or more, it could be described as a small continuous release.

To model accidental releases of the kind described above requires both a source emission model and a transport and dispersion model. When an accidental source emission occurs, the initial emission and acceleration phase gives way to a regime in which the internal buoyancy of the puff or plume dominates the dispersion. This regime is followed by transition to a regime in which the internal turbulence dominates the dispersion. There is then a transition from dominance of internal buoyancy to dominance of ambient turbulence. Models describing source emissions are beyond the scope of this article. This article will describe only the transport and dispersion models where ambient turbulence dominates.

Two types of dispersion models are used to describe these releases when the puff or plumes are neutrally or positively buoyant. When there is an instantaneous release or a burst of material, we make use of a *puff model*. In this model the puff disperses in the downwind, crosswind, and vertical directions simultaneously. Computer codes written for puff models usually have the capability of tracking multiple puff releases over a period of time. When the release rate is constant with time, the puff model can be mathematically integrated into a *continuous model*. In this case, dispersion takes place in the cross wind and vertical directions only. The mathematical expressions are those discussed in Section III. The dispersion coefficients used, however, may differ from those described in Section IV. Plume rise equations from Section V for positively buoyant plumes may be used in conjunction with these dispersion models. The equations of current models indicate that they are well formulated, but the application of the models suffers from poor meteorological information and from poorly defined source conditions that accompany accidental releases. Thus, performance of these models is not adequate to justify their use as the sole basis for emergency response planning, for example.

A. Instantaneous Puff or Dispersion Model

To determine concentrations at any position downwind of an instantaneous source, the time interval of travel after the

time of release and the diffusion in the downwind, lateral, and vertical directions must be considered. A suggested equation to estimate downwind puff concentration from an elevated release at height H is:

$$C = \frac{2Q_T}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_z} \exp\left[-\frac{1}{2}\left(\frac{x - \bar{u}t}{\sigma_x}\right)^2\right] \\ \times \exp\left[-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right] \exp\left[-\frac{1}{2}\left(\frac{H}{\sigma_z}\right)^2\right] \quad (62)$$

Here Q_T represents the total mass of the release, the σ 's refer to dispersion coefficients following the motion of the expanding puff, and the 2 in the numerator accounts for assumed ground reflection, which is consistent with the continuous plume models. Note that there is no dilution in the downwind direction by the wind. The wind serves to move the centerline of the puff in the downwind direction. This motion is accounted for by the product $\bar{u} \cdot t$ term in the exponential involving σ_x . In the term $\bar{u} \cdot t$, t is the time after release and \bar{u} is the average wind speed in the x direction. Thus, $\bar{u} \cdot t$ is the distance down wind the puff has traveled after release.

The total exposure that would be experienced by a ground level receptor at a point (x, y, o) is given by

$$C_{TOT} = \int_{x/\bar{u}}^{\infty} C(x - \bar{u}t, y, o) dt \quad (63)$$

Assuming that the puff passes rapidly overhead, σ_y and σ_z will be constant during the integration. Furthermore, diffusion along the x -axis is neglected by comparison with the mean wind resulting in

$$C_{TOT} = \frac{Q_T}{\pi\bar{u}\sigma_y\sigma_z} \exp\left[-\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \frac{z^2}{\sigma^2}\right)\right] \quad (64)$$

This equation has the same form as Eq. (41). However, the σ 's now should be values representative of puff dispersion.

B. Dispersion Parameters for Puff Models

The Pasquill–Gifford dispersion parameters as represented by Fig. 3 have been used in some algorithms to represent puff dispersion coefficients. There is a set of coefficients due to Slade which result in significantly reduced spread of the plume and calculated centerline concentrations which are greater than those found from the Pasquill–Gifford parameters of Fig. 3. The Slade coefficients are presented in Table VIII. Much less is known about diffusion in the downwind direction than in the lateral and vertical directions. Thus, there is no good estimate of σ_x , and it is customary to assume that it is equal to σ_y .

TABLE VIII Estimation of Dispersion Parameters

Parameter	Condition	Downwind distance		Approximate power function (x in meters)
		100 m	4000 m	
σ_y in m	Unstable	10.0	300.0	$0.14(x)^{0.92}$
	Neutral	4.0	120.0	$0.06(x)^{0.92}$
	Stable	1.3	35.0	$0.02(x)^{0.89}$
σ_z in m	Unstable	15.0	220.0	$0.53(x)^{0.73}$
	Neutral	3.8	50.0	$0.15(x)^{0.70}$
	Stable	0.75	7.0	$0.05(x)^{0.61}$

C. Momentum Dominated Jets

A model has been developed for momentum dominated releases. Here it is assumed that the release is a jet pointed either upward or downward. The jet must be perpendicular to the wind flow and there must be no obstructions to the wind flow. It is assumed that choked flow exists and that the release reaches sonic flow at the point of emission. The rise of the jet can then be calculated from the following equation.

$$\Delta h = 2.4(V_S d_S / \bar{u}) \quad (65)$$

It was found that the jet will expand at such a rate that its radius equals about 0.4 times its distance from its source. Therefore, a horizontal jet inhibited by the ground would not be well represented by this model. Also, this equation cannot be used if the jet directly impinges upon the ground.

D. Elevated Dense Gas Releases

After its initial rise due to momentum, a dense gas puff or plume will begin to sink. Eventually the plume centerline strikes the ground surface.

The available models for dense gas release trajectories are much more complex than the models previously discussed in this article. The following model may be applied to dense gas releases for downwind distances of no more than a few hundred meters. In this region, internal turbulence will dominate over ambient turbulence.

First calculate the initial jet Richardson number Ri_o from the following equation to determine if the gas is dense enough to have a significant effect on the plume trajectory.

$$Ri_o = g V_S d_S ((\rho_S - \rho_a)/\rho_a) / u_*^3 \quad (66)$$

In this equation, u_* is the ambient friction velocity roughly equivalent to $\bar{u}/15$ for wind measured at 10 m and a roughness factor of 1.0 cm, ρ_S is the initial density of the gases released, ρ_a is the density of the ambient air and d_S and V_S are the initial release diameter and velocity. If $R_o \gg 1.0$, the density effects will be important.

The model predicts the following maximum initial rise:

$$\Delta h/d_S = 1.32(V_S/\bar{u})^{1/3}(\rho_S/\rho_a)^{1/3} \times (v_S^2 \rho_S / (d_S g (\rho_S - \rho_a)))^{1/3} \quad (67)$$

The downwind distance x_g to touchdown is given by the relation:

$$x_g/d_S = (V_S \bar{u} \rho_S / (d_S g (\rho_S - \rho_a))) + A \quad (68)$$

where

$$A = 0.56((\Delta h/d_S)^3((2 + h_S/\Delta h)^3 - 1.0)^{1/2} \times (\bar{u}^3 \rho_S / (d_S g V_S (\rho_S - \rho_a)))^{1/2}) \quad (69)$$

and $h_S/\Delta h$ is the ratio of stack height to maximum rise.

The ratio of maximum concentration C_{\max} at a given downward position to the initial concentration C_o at the point of maximum rise is given by:

$$C_{\max}/C_o = 1.688(V_S/\bar{u})(\Delta h/d_S)^{-1.85} \quad (70)$$

The same ratio at the point x_g is given by:

$$C_{\max}/C_o = 9.434(V_S/\bar{u})((h_S + \Delta h)/d_S)^{-1.95} \quad (71)$$

E. Dense Gas Releases at Ground Level

Perhaps the majority of accidental releases are at ground level or near to the ground surface. In this case, the models reviewed previously are not applicable, and a new set of models involving dense gas slumping are being developed. The tendency of a dense gas to slump when spilled on a ground surface is similar to the spread of thick molasses on a table top. It should be noted that gases like ammonia with molecular weights near that of air and which are thus usually considered buoyant, will slump if emitted as a cold cloud. The major factor here is density and not molecular weight.

Slumping is not the only problem that must be considered when dealing with surface spills. Most dense clouds contain aerosols which require the following phenomena to be modeled:

- Evaporation and condensation of liquid drops in the cloud.
- Heat and mass transfer with the underlying surface.
- Radiation flux divergence.
- Chemical reactions.

Furthermore, materials with lower volatility may diffuse or seep into the soil before evaporating.

There are models available to account for these effects. However, at this writing they are not sufficiently developed to warrant inclusion in this discussion. Some of these available models are extremely complex, requiring large

computer codes and involving long run times. Discussion of these models is beyond the scope of this article.

VII. CALCULATED CONCENTRATION AND AVERAGING TIME

For the purposes of calculation, averaging time is usually divided into two major periods, a short term of one day or less and long time periods of one month or more. The most common short-term averaging periods used are 24 hr, 8 hr, 3 hr, and 1 hr. For long-term averaging, a period of a month, a season (3 months), or a year is most frequently employed.

A. Short-Term Averaging Time Corrections

The Gaussian estimation formulas suggested for determining concentration are time-averaging equations. The concentration calculated depends upon the sampling or averaging time associated with the values of the dispersion coefficients used. Frequently it is desired to estimate concentrations for other sampling times. A power law is suggested as a possible conversion law for use with single sources and averaging times of 24 hr or less. Thus,

$$\bar{C}_s = \bar{C}_k \left(\frac{t_k}{t_s} \right)^P \quad (72)$$

where C_s is the concentration for time t_s , C_k concentration for time t_k , t_s longer averaging time, t_k shorter averaging time, and P power. Values of P have ranged from 0.17–0.75. The suggested value is 0.17.

B. Long Time Period Averaged Concentration

The time-averaged ground level concentration, as given by Eq. (42), is integrated in the crosswind.

$$\bar{C}_{CW} = \int_{-\infty}^{\infty} \frac{Q}{\pi \sigma_y \sigma_z \bar{u}} \exp \left\{ -\frac{1}{2} \left[\left(\frac{y}{\sigma_z} \right)^2 + \left(\frac{H}{\sigma_z} \right)^2 \right] \right\} dy \quad (73)$$

$$\bar{C}_{CW} = \frac{2^{1/2} Q}{\pi^{1/2} \sigma_z \bar{u}} \exp \left[-\frac{1}{2} \left(\frac{H}{\sigma_z} \right)^2 \right] \quad (74)$$

To determine the concentration due to a wind blowing from a particular sector over a long period of time, Eq. (74) is divided by the width of the sector at the distance x . Within a sector it is assumed that the wind directions are distributed randomly over the averaging period and that the effluent is distributed uniformly in the horizontal plane. For a 16-point wind rose, the sector angle is $360/16 = 22.5^\circ$, and the sector width is thus $(2\pi/16)x$. Thus the long-term average in any sector becomes

$$\bar{C}_{LT}(\theta) = \left(\frac{2}{\pi}\right)^{1/2} \frac{Q}{\sigma_z \bar{u} (2\pi x/16)} \exp\left[-\frac{1}{2} \left(\frac{H}{\sigma_z}\right)^2\right] \quad (75)$$

For sources that emit continuously over a long period of time—monthly, seasonal, or annual—average concentration isopleths can be calculated. In addition to dispersion coefficient data, stability frequency of occurrence information is required. For a given stability class S , wind speed class u , and sector θ , multiply \bar{C}_{CW} by the frequency f_n , the fraction of time that the wind blows toward the sector, thus,

$$\bar{C}_{LT}(\theta, u, S) = f_n(\theta, u, S) \left(\frac{2}{\pi}\right)^{1/2} \frac{Q}{\sigma_z \bar{u} (2\pi x/16)} \times \exp\left[-\frac{1}{2} \left(\frac{H}{\sigma_z}\right)^2\right] \quad (76)$$

The net average concentration is then obtained by summing the long-term averages for each sector from each stability class and wind speed.

$$\bar{C}_{NET} = \sum_{\theta} \sum_u \sum_S \bar{C}_{LT}(\theta, u, S) \quad (77)$$

This model can be employed for single or multiple sources or for large-area sources.

VIII. ATMOSPHERIC CONCENTRATIONS WITH DIFFUSION CEILINGS

A. Diffusion Ceilings Defined

When a stable layer of air exists near the surface of the earth, the effect is to produce a diffusion lid and thus to restrict the vertical speed of pollutants. A well-known example of the phenomenon is the morning haze layer over cities. This haze layer consists of particulate and gaseous emissions that react to form small particulates. On winter mornings, the haze layer is particularly noticeable due to the strong radiation inversion and to the emission into the inversion layer of products of combustion from space heating. In the early morning just at dawn, the layer of smoke can be noticed around rooftops, close to the altitude where it was first emitted. This is due to the very poor vertical mixing in the stable air of the inversion. The top of the haze layer may even be below the altitude of the top of the inversion layer. As the sun comes up, mixing of the smoke layer near the ground is noticed because the inversion layer breaks up beginning near the surface. Later the vertical mixing of the pollutant diminishes the haze layer, and it may vanish.

Many times diffusion lids result from a stable layer aloft which effectively traps plumes below in a relatively un-

stable region of air near the surface of the earth. Mixing by diffusion proceeds in this unstable air. However, the stable layer aloft limits the diffusion upward much as the ground limits the diffusion downward.

The model developed to account for the mixing under either type of stable layer employs the idea of complete reflection from the diffusion lid, as in the case for the plume contacting the ground. Now that the plume is trapped between lid and bottom, multiple reflections are allowed for in the model. Thus the continuous point source model becomes

$$\bar{C} = \frac{Q}{2\pi\sigma_y\sigma_z\pi} \exp\left[-\frac{1}{2} \left(\frac{y}{\sigma_y}\right)^2\right] \left\{ \exp\left[-\frac{1}{2} \left(\frac{z-H}{\sigma_z}\right)^2\right] + \exp\left[-\frac{1}{2} \left(\frac{z+H}{\sigma_z}\right)^2\right] + E_T \right\} \quad (78)$$

where

$$E_T = \sum_{n=1}^{n=i} \left\{ \exp\left[-\frac{1}{2} \left(\frac{z-H-2nL}{\sigma_z}\right)^2\right] + \exp\left[-\frac{1}{2} \left(\frac{z+H-2nL}{\sigma_z}\right)^2\right] + \exp\left[-\frac{1}{2} \left(\frac{z-H+2nL}{\sigma_z}\right)^2\right] + \exp\left[-\frac{1}{2} \left(\frac{z+H+2nL}{\sigma_z}\right)^2\right] \right\} \quad (79)$$

and L is the height of the stable layer. The number of reflections is given by n , where $n=4$ is sufficient to include all the important contributions. Lacking a measured value of L , the mixing depth previously defined is a good value to employ in the equations. It should be noted that in the limit as n becomes infinitely large, E_T can be replaced by an integral from $n=0$ to $n=\infty$. The result of the integration is such that at ground level

$$\lim_{n \rightarrow \infty} \bar{C} = \frac{Q}{(2\pi)^{1/2} \sigma_y L \bar{u}} \exp\left[-\frac{1}{2} \left(\frac{y}{\sigma_y}\right)^2\right] \quad (80)$$

B. Plume Trapping

In the case of plume trapping by a very deep inversion or by an upper air subsidence inversion when L is very large, it has been assumed that the number of reflections is reduced to one. Thus, at ground level on the centerline, the equation describing plume trapping would be

$$\begin{aligned}\bar{C}_T = \frac{Q}{\pi \sigma_y \sigma_z \bar{u}} & \left\{ \exp \left[-\frac{1}{2} \left(\frac{H}{2\sigma_z} \right)^2 \right] \right. \\ & + \exp \left[-\frac{1}{2} \left(\frac{H+2L}{\sigma_z} \right)^2 \right] \\ & \left. + \exp \left[-\frac{1}{2} \left(\frac{H-2L}{\sigma_z} \right)^2 \right] \right\} \quad (81)\end{aligned}$$

C. Fumigation

Fumigation describes the rapid downward mixing to the ground of material aloft that has accumulated during a period of strong atmospheric stability. Under a strong inversion the plume was described as fanning. Thus fumigation usually describes the condition that results during the breakup of a fanning plume formed during a nocturnal inversion. A model for ground level concentrations can be determined by integrating the equation for average concentrations including the ground reflection term from $z=0$ to $z=\infty$ and then considering the material in the cloud to be distributed uniformly through a layer of height L . The equation is thus the same as developed previously for the diffusion ceiling problem:

$$\bar{C}_F = \frac{Q}{(2\pi)^{1/2} \sigma_y L \bar{u}} \exp \left[-\frac{1}{2} \left(\frac{y}{\sigma_y} \right)^2 \right] \quad (82)$$

D. Possible Algorithms

Faced with the necessity of computing ground level concentrations when a diffusion lid exists, there are several possible algorithms. One of the earliest algorithms employed was based on the assumption that a plume edge could be defined. The usual assumption is that the plume extends vertically to the point at which the concentration is 1/10 that found at the maximum on the plume centerline. At this point, the plume has a halfwidth of $2.15\sigma_z$. The distance to this point defined as x_L occurs when

$$H + 2.15\sigma_z = L \quad (83)$$

or

$$\sigma_z = 0.47(L - H) \quad (84)$$

Up to the point x_L , Eq. (42) is used to calculate concentrations. After a distance equal to $2x_L$, it is assumed that the plume is completely mixed in the vertical, and Eq. (82) is used. Between x_L and $2x_L$, an arithmetic mean of the two equations proportional to the fraction of the distance between $2x_L$ and x_L is used; or on a concentration distance plot, the two end points are simply joined by a smooth curve.

Another similar algorithm has been proposed with Eq. (42) being used in stable conditions or for unstable conditions with unlimited mixing. When mixing is limited by a layer L units in depth with $\sigma_z > 1.6L$, Eq. (78) is used, and for $\sigma_z < 1.6L$ Eq. (82) is used. In case H and Z are both greater than L , then $C = 0$.

IX. MULTIPLE SOURCES AND RECEPTORS

A. Multiple Stacks in Line

When N equivalent stacks are located close to each other, for instance, in line, it has been found that the concentration effect of the multiple stacks is less than that of the sum of the individual stacks; that is, the resultant concentration is less than N times that of a single stack concentration. The centerline concentration calculated on the basis of a single stack is multiplied by the factor $N^{4/5}$.

B. Multiple Sources and Receptors

The effect at a single receptor from multiple sources is accounted for by adding up the concentrations as calculated for each individual source. Thus it is assumed that the superposition principle applies. When there are also multiple receptors, the effect of multiple sources is accounted for similarly at each receptor. While there are more than two to three sources and two or three receptors, the use of a digital computer to carry out the calculation is recommended.

C. Concentration Isopleths

To calculate concentration isopleths, it is recommended that a coordinate grid be laid out to cover all sources and receptors. The origin of the grid should be taken as the southwestern most source or receptor. Since a digital computer program is virtually required for such a calculation if more than one source is present, the grid spacing can be relatively small. One hundred grid points would probably be a maximum. Calculation of the concentration at each grid point is carried out, and the effects of each source are added at each grid point. With the concentration known at each grid point, the isopleths can be drawn in by eye or by a graphic plotter. The results are particularly useful when siting ambient monitoring instruments.

X. DESIGNING A CHIMNEY TO LIMIT GROUND-LEVEL CONCENTRATIONS

The models discussed up to this section are an idealized picture where the wind-speed profile is uniform, the

temperature gradient is constant, and, thus, turbulence is constant over the whole of the plume. Application is to relatively flat or gently sloping terrain with no obstructions and in which aerodynamic wakes are not influential. The stability imposed remains constant in both the vertical and horizontal directions. Moreover, other than the earth-boundary problem, no effects of chimney downwash, building, or terrain have been discussed. Real plumes encounter varying layers of stability, downwash behind the chimney, and effects of buildings and terrain. However, many applications require treatment of situations where effective emission heights are low enough that the plumes may be influenced by aerodynamic effects of building and terrain located very close to the source. Also, the wind passing over a chimney will cause wake effects behind the chimney that may result in lowering the effective height of the emission.

The problems mentioned above are due to the *separation* of flow around a building or behind a chimney. In the flow of fluids past objects and through orifices or similar restrictions as between buildings, fluid vortices are shed periodically downstream from the object. Objects subject to this separation phenomena include chimneys, tall columns in chemical processing plants, pipe galleries suspended above ground, electrical transmissions lines, etc. These vortices could cause vibrations in the structures if the frequencies are near the natural frequency of the structure. The vibrations could result in sound waves being generated or, in the worst case, damage to the structure. Atmospheric dispersion effects can result in the cavities that form behind buildings and hills. In this section we will discuss these effects with an eye toward designing an actual chimney.

Chimney design is set through the Good Engineering Practice (GEP) requirements of EPA. The document entitled "Guidance for the Determination of Good Engineering Practice Stack Height (Technical Support Document for the Stack Height Regulation)," June 1985, EPA, Office of Air Quality Planning and Standards, EPA-450/4-80-023R, Research Triangle Park, North Carolina, provides the technical background for the specification of a chimney. Of course, the higher the chimney, the lower the downwind ground-level concentration will be. Therefore, a maximum height to be used in dispersion calculations from chimneys has been set. This height is 65 m measured from the ground-level elevation from the base of the stack. This height is then modified according to the height of nearby structures and significant terrain features. The GEP rules do not limit the height to which a stack may actually be constructed. Since chimneys do not require Federal Aviation Administration lighting or marking below 200 ft (61 m), modeling should first be done at a maximum height of 200 ft to determine if ground-level

concentrations permitted will be satisfactory. At this point, even if the required modeling for a permit is satisfied, an additional look at the chimney is required to determine if all possible meteorological, building, and terrain effects have been examined. These effects, which are usually not required to be examined could result in serious air quality standard violations even though the original chimney design met the required GEP and produced acceptable dispersion modeling results.

A. Principal Plume Models

Meteorologists have identified four principal models of plumes. These plumes are designated as follows:

1. Unstable—Looping Plume

Unstable conditions are associated with highly convective conditions and a high degree of turbulence. These plumes tend to produce plumes which loop up and down very vigorously. The plumes may touch the ground at irregular intervals causing high ground-level concentrations at the point of touchdown. However, these high concentrations will be of short duration. The Gaussian Model, Eq. (42), is applied in this condition, but it is not as successful as in the following neutral coning condition.

2. Neutral—Coning Plume

This plume occurs for neutral or nearly neutral atmospheres. The cone point is at the top of the chimney, and the plume then spreads out moving downwind with a fairly regular shape. Equation (42), the Gaussian Model, with the Briggs' Plume Rise Model is best described by this plume dispersing under neutral D conditions. However, in addition to neutral category D, Pasquill-Gifford stability categories B, C, E, and F also produce plumes which are approximately coning and can be described by Eq. (42).

Coning plumes which are buoyancy dominated in light winds develop a relatively high plume rise. This results in low ground-level concentrations. High winds cause a greater spread of these plumes with greater dilution and lower ground-level concentrations. Somewhere in between a critical wind speed should exist which produces the highest ground-level concentration.

3. Stable—Fanning Plume

Under stable conditions which occur most likely in the early morning the plume is transported downwind as a flat ribbon with some horizontal dispersion that gives the plumes a fanning look. Vertical diffusion is so slow that the ground-level concentration is negligible. The inversion

will begin to break up as the morning progresses. Finally the plume will be well mixed to the ground in high concentrations as described by Eqs. (78), (79), (81), and (82). This condition could produce some high ground-level concentrations within a narrow band width. These fumigations do not persist more than 30–45 min, and thus this condition may not produce the most critical ground-level conditions. In the case of high-level inversions the plume can be trapped for a significant part of the day. High concentrations on the ground could result in this case, and from tall chimneys trapping may result in the critical design conditions. Eq. (81) applies in this case.

4. Lofting

In the case when an inversion is established from the ground up as in the late afternoon, the plume may be able to penetrate the inversion. The plume can now diffuse upward, but it is prevented from diffusing into the stable air below. This could be a favorable condition since the plume would tend to diffuse away above the stable layer resulting in zero ground-level concentration. However, eventually the material in the plume above the inversion comes to the ground and could cause a problem further downwind.

B. Locating the Maximum Concentration

As suggested in the case of coning above, there is a critical wind speed where the maximum concentration can be found. For both looping and coning plumes where Eq. (42) can be used to describe the plumes, a maximum concentration can be found for each stability category by differentiating the equation and setting the resultant derivative equal to zero. The approximation formulas found for maximum concentration and its location are useful in guiding a more accurate determination by a point-by-point or a grid search.

The maximum concentration must occur on the centerline, and the ground-level concentration is desired. Therefore, both y and z are set equal to zero in Eq. (42) before differentiation.

Because the maximization takes place in the downwind direction x , exponential equations as a function of x , the downwind distance, are used to represent σ_y and σ_z , respectively.

$$\sigma_y = ax^p \quad (85)$$

$$\sigma_z = bx^q \quad (86)$$

The effective emission height is also a function of x through its relation to plume rise Δh . The plume rise $\Delta h \propto x^{2/3}$ until the final plume rise is reached. Thus in maximizing Eq. (42), it is assumed that the final plume rise has been reached.

Equations (85) and (86) are substituted into Eq. (42), then Eq. (42) is differentiated with respect to x and set equal to zero. The result of this operation is to determine the location of the maximum as

$$x_{\max} = \left[\frac{H}{b\sqrt{(p+q)/q}} \right]^{1/q} \quad (87)$$

The maximum concentration is then given by

$$\left[\frac{C_{\max}\bar{u}}{Q} \right] = \frac{b^{(p/q)} \left(\frac{p+q}{q} \right)^{(p+q)/2q} \exp \left[-\left(\frac{p+q}{2q} \right) \right]}{\pi a H^{(p+q)/2q}} \quad (88)$$

The right-hand sides of Eqs. (87) and (88) are functions only of the effective emission height H and stability that determines a , b , p , and q .

The simplest assumption about σ_y and σ_z is that the ratio σ_y/σ_z is a constant independent of x . This leads to the more familiar equations for maximum concentration and distance to that maximum

$$x_{\max} = x \quad \text{where} \quad \sigma_z = \frac{H}{\sqrt{2}} \quad (89)$$

$$\frac{C_{\max}\bar{u}}{Q} = \frac{2}{e\pi H^2} \cdot \frac{\sigma_z}{\sigma_y} \quad (90)$$

C. Chimney, Building, and Terrain Effects

Before much research had been done, simple rules of thumb were used to avoid the problems of stack and building downwash. These models are reviewed here.

1. To Prevent Stack Downwash

As the wind encounters any bluff object, a wake will form on the lee side of the object. Due to the boundary layer effects and the increase in local velocity around the object, the pressure in the wake of the object must be less than that in the surrounding outer atmosphere. If the momentum of an emitted smoke plume is not large, the low pressure area can cause the plume to be sucked down behind the stack in the aerodynamic downwash. Up to one third of the stack height may be lost in this manner. To be assured that these troubles are avoided, the velocity of emission, V_s , of the plume should be greater than 1.5 times the maximum expected wind speed.

$$V_s > 1.5 \bar{u}$$

2. To Prevent Building Downwash

The effect that aerodynamic downwash can have on a building is to induce reverse flow that is shown by the

velocity profile on the lee side of the building. Improperly located stacks which are too short may provide an effluent that will actually be sucked back into the buildings. To prevent such building downwash make all stacks 2.5 times the building height h_b . Material emitted at this height is ejected into more or less unperturbed flow. As a result significant dilution can occur before the plume reaches the ground so that ground level concentrations are reasonable.

$$h_s > 2.5 h_b$$

3. Flow Patterns Near Buildings

The separation of flows in boundary layers along exposed surfaces of bluff objects is well known in the study of fluid mechanics. The separation is controlled either by geometry or by aerodynamic effects. If the separation lines are not controlled by geometry, their locations tend to fluctuate with variations in the overall flow pattern.

Before the 1970s the accepted conceptual model of flow patterns near buildings was relatively simple. Recently these models have become more complicated and accurate. They show that the flow pattern is a strong function of obstacle geometry. The flow characteristics include a displacement zone which exists upwind of the obstacle where the incident fluid first becomes affected by the presence of the obstacle. Within this zone the fluid attempts to travel over and around obstacles, producing changes in wind speed and wind direction. A pressure higher than ambient will be measured at the front surface of the interfering object as the approaching wind decelerates. Since the wind speed is also a function of vertical distance above ground level, a downward directed pressure gradient will be established near the upwind face to the obstacle. This pressure gradient is responsible for the downward directed velocity that occurs near the front of the obstacle. At ground level this downward flow must move out from the building causing the approach flow to separate from the ground upwind of the building. This results in an eddy in front of the lower portion of the windward side of the obstacle. Above this windward eddy, the incident flow of the fluid strikes the windward face of the obstacle. The fluid then moves upward and/or to the obstacle's side depending on the proximity of the roof or the obstacle's side edges. The resulting viscosity-induced boundary layer separates from the exposed surface at sharp edges of the sides and roof where the flow cannot follow the abrupt change in direction of the obstacle's construction. The separated boundary layers move out into the surrounding fluid as shear layers. If the obstacle is sufficiently long, reattachment of the flow to the obstacle's outer surfaces will occur at some location downstream, and separation will eventually occur again at the termination of the obstacle. If the obstacle is not sufficiently long, reattachment does not occur, but the separated layers curve inward toward the obstacle's surface feeding into a zone of recirculation that exists immediately downwind of the body.

4. Dispersion Over Complex Terrain

Models for the assessment of environmental impacts of air pollutants released in the vicinity of hills or mountains differ somewhat from the normal bi-Gaussian plume formula. Furthermore, there are two specific concerns associated with the release and transport of atmospheric pollutants in an area with complex terrain. First, the impingement of plumes on elevated terrain surfaces may result in dangerously high ground-level concentrations. Second, there is the possibility of unusually long periods of stagnation within the atmospheric boundary layer in lower, valley-like regions. Several disastrous incidents of this nature have occurred in the past.

The flow patterns and velocity distribution around elevated complex terrain (hills, mountains, or ridges) need to be understood before quantitative predictions of pollutant dispersion can be made. The broad aspects of the structure of these stratified flows include:

1. determining if the streamlines from upwind impinge on the hill, go around the hill, or go over the summit of the hill,
2. describing the region of the separated or recirculating flow in the lee of the hill,
3. finding the effects of heating and cooling of the surface.

Observations and model experimentation indicate that the air flow over elevated terrain occurs in approximately horizontal planes around the topography. This observation has been used to estimate the surface concentrations caused by upwind sources of pollution. However, there has been little experimental work or atmospheric data describing how strong the stratification strength is required for any given streamline starting below the top of the hill to be deflected around, instead of over, the hill.

XI. THE NEW MODELS

The models presented have their greatest use in regulatory applications. These models do not require on-site monitoring instrumentation. Rather, the meteorological measurements used in the U.S. are the routinely collected data from the Weather Bureau Stations of the National Oceanic and Atmospheric Administration frequently

located at the major airports in the various states. The United States Environmental Protection Agency (USEPA) and environmental agencies charged with regulation of air pollution in Western Europe have prepared a series of computer algorithms based on these Gaussian Models. The computer programs resulting from these algorithms rely on classifying the stability of the atmosphere, usually with the Pasquill-Gifford method or a method similar to it. Then, depending upon whether the situation to be examined is a rural or urban setting, dispersion parameters are selected which are functions only of the stability and downwind distance. The process is simple, easily applied, and the resultant concentration calculated is usually a reasonable estimate for most regulatory purposes. Furthermore, since all emission sources controlled by the USEPA or other governmental agencies use the same programs applied in the same way, everyone arrives at the same answers. Also, avoiding the use of on-site instrumentation reduces the problems that would arise due to the representativeness of the data and the maintenance, sensitivity, and accuracy of the instrumentation.

A. Using Turbulent Statistics

These procedures for estimating dispersion have had a long and useful life. It now appears that in the future the use of improved procedures for making estimates of dispersion from measured turbulent statistics will be employed. Crosswind dispersion for a time period τ can be estimated by averaging the wind direction θ over moving time-intervals of duration $x/\bar{u}\beta$ and using the following relation:

$$\frac{\sigma_y}{x} = [\sigma_\theta]_{\tau,x/\bar{u}\beta} \quad (91)$$

where σ_θ is in radians. From a limited number of field measurements β has been estimated to be about 4.0. Making these calculations will, of course, be dependent upon making accurate measurements of the wind fluctuations.

Vertical dispersion from an elevated release can be estimated from the elevation angle ϕ of the wind using

$$\frac{\sigma_z}{x} = [\sigma_\phi]_{\tau,x/\bar{u}\beta} \quad (92)$$

Values of the vertical dispersion over the time period τ can be estimated by averaging values of ϕ in radians over moving time intervals of duration $x/\bar{u}\beta$ prior to the calculation of σ_ϕ .

B. Boundary Layer Characterization

The initial step taken by modelers in the U.S. and Western Europe in moving away from the Gaussian Modeling procedure mentioned above has come in the attempt to more precisely define the planetary boundary layer. Then, the

necessary values for the determination of the dispersion parameters come from the characterization of the boundary layer.

The atmospheric boundary layer has been visualized to consist of three major divisions: the forced convection part of the surface layer; then above this layer, the local free convection part of the surface layer; and finally at the top, the mixed layer. (See Section II, D and E.)

A dimensional analysis predicts that the flow properties are functions of z/L , where z is the height above the surface, and L is Monin-Obukhov length, L [Eq. (14)]. Field data support this theory. The mechanical production of turbulent energy comes from the shearing action of the mean flow, whereas the heating of the atmosphere adjacent to the ground produces the buoyant turbulent energy. Therefore, the ratio z/L is a measure of the relative importance of buoyancy versus shear effects. The stability of the atmosphere can be judged from the ratio of z/L as follows:

- if $z/L > 0$ the atmosphere is stable,
- if $z/L = 0$ the atmosphere is neutral, and
- if $z/L < 0$ the atmosphere is unstable.

When $z/L < -1$, buoyant production of turbulence is greater than mechanical production of turbulence; when $z/L > -1$ mechanical production of turbulence exceeds buoyant production of turbulence; and when buoyant production is negligible, and all the turbulence is created by wind sheer, i.e., mechanical turbulence; $-0.1 < z/L < 0.1$.

Furthermore, the quantities in the Monin-Obukhov length scale are approximately constant throughout the surface layer. This length scale is the dominant parameter governing diffusion and the flux-gradient relationships in the surface layer.

1. Forced Convection Part of the Surface Layer

In the forced convection layer, wind shear plays a dominant role, and the Monin-Obukhov similarity hypothesis applies. To develop their similarity theory Monin and Obukhov idealized the field of motion that is frequently used in the atmosphere near the ground. It was assumed that all statistical properties of the temperature and velocity fields are homogenous and do not vary with time. The steady mean motion was considered to be unidirectional at all heights in the x -direction. Second-order terms in the equations of the field were considered to be negligible. The scale of motion was considered to be small enough to omit the Coriolis Force, and the radiative heat inflow was neglected. In the surface layer the turbulent fluxes are approximately constant at their surface values.

2. Local Free Convection Part of the Surface Layer

If the Monin-Obukhov length, L , has a small negative value owing to the increase of surface heat flux, or to a decrease of u_* , the friction velocity, or if z increases with L remaining nearly constant, z/L can become quite large negatively. Then, a free-convection-type layer exists. The condition is termed local free convection where a windless condition is indicated in which all turbulent motion is buoyant and created by strong surface heating.

With z growing large while L remains constant indicates that local free convection is attained by going far enough above the surface but remaining in the atmospheric boundary layer. Consequently in the local free convection layer, z continues to be a significant scaling parameter.

3. Top or Mixed Layer—Convective Velocity Scale

In this mixed layer the boundary layer height z_i is the controlling parameter. At night the air near the ground is usually stably stratified with little or no turbulence above the mechanically produced layer. This mixed layer is usually 10–100 m deep. On cloudy, windy nights the layer can be deeper. During the daytime the air is normally stratified above the convective boundary layer, which is 500–2000 m deep in the afternoon. Above the mixed layer the air is free from friction effects, and other surface effects are negligible. The geostrophic approximation can then be used.

C. The New Models

The new models are written to accept wind, turbulence, and temperature profiles from detailed on-site observations, if available. The profiles depend on the surface heat and momentum fluxes and the mixing depth which can be measured. If only a minimum of information is available from routine observations of wind speed near the surface, insulation, cloudiness, and temperature, then these vertical profiles are generated through surface and mixed layer scaling models. The meteorological variables that affect the plume dispersion are generated as averages over the appropriate depth of the boundary layer. Mixed layer heights are calculated from boundary layer models that use available observations and a surface energy balance. The newer plume rise formulations have explicit treatments of the effects of convective turbulence and partial plume penetration. Terrain effects are treated by dividing the streamline height to result in a consistent treatment of receptors above and below stack height.

Two other significant effects are accounted for in the new models. Surface releases and convective effects both

cause the vertical distribution to deform from the Gaussian shape. In the case of surface releases, the rapidly changing wind-speed profile causes a changing of the turbulence eddy structure to include larger eddies near the surface which affects the vertical concentration distribution. A separate treatment of this phenomenon in the model is then required. Under extreme convective conditions, up to two thirds of the air may be descending. This results in a skewed distribution of vertical velocities. A probability density function differing from the Gaussian is constructed to account for the difference.

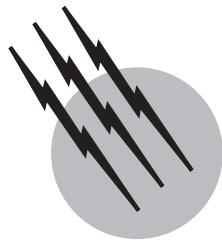
SEE ALSO THE FOLLOWING ARTICLES

AEROSOLS • ATMOSPHERIC TURBULENCE • CLIMATOLOGY • METEOROLOGY, DYNAMIC • POLLUTION, AIR • POLLUTION CONTROL

BIBLIOGRAPHY

- American Meteorological Society (1977). "Committee on Atmospheric Turbulence and Diffusion (1978)." *Bull. Am. Meteorol. Soc.* **59**(8), 1025.
- Briggs, G. A. (1969). "Plume Rise," U.S. Atomic Energy Commission, TID-25075, Clearinghouse for Federal Scientific and Technical Information, National Bureau of Standards, U.S. Department of Commerce, Springfield, Virginia.
- Csanady, G. T. (1973). "Turbulent Diffusion in the Environment," D. Reidel, Boston.
- Gifford, F. A. (Jan–Feb 1976). "Turbulent Diffusion-Typing Schemes: A Review," *Nuclear Safety* **17**(1), 68–86.
- Hanna, S. R., Briggs, G. A., Deardorff, J., Eagan, B. A., Gifford, F. A., and Pasquill, F. (1977). "AMS Workshop on Stability Classification Schemes and Sigma Curves—Summary of Recommendations," *Bull. Am. Meteor. Soc.* **58**, 1305–1309.
- Hanna, S. R., Briggs, G. A., and Hosker, R. P., Jr. (1982). "Handbook on Atmospheric Diffusion," Technical Information Center, U.S. Dept. of Energy, DOE/TIC-12223 (DE82002045).
- Hanna, S. R., and Paine, R. J. (1989). "Hybrid Plume Dispersion Model (HPDM) Development and Evaluation," *J. Appl. Meteor.* **21**, 206–224.
- Haugen, D. A., ed. (1975). "Lectures on Air Pollution and Environmental Impact Analysis," American Meteorological Society, Boston.
- Holzworth, G. C. (1972). "Mixing Heights, Wind Speeds, and Potential for Urban Air Pollution Throughout the Contiguous United States," Environmental Protection Agency, Office of Air Programs, Research Triangle Park, North Carolina.
- Panofsky, H. A. (1969), "Air Pollution Meteorology," *Am. Scientist* **5**(2), 269–285.
- Panofsky, H. A., and Dutton, J. A. (1984). "Atmospheric Turbulence," Wiley, New York.
- Pasquill, F. (1976). "Atmospheric Dispersion Parameters in Gaussian Plume Modeling. Part II. Possible Requirements for Change in the Turner Workbook Values," EPA-600/4-76-030b, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.
- Pasquill, F., and Smith, F. B. (1983). "Atmospheric Diffusion," Wiley, New York.
- Randerson, D., ed. (1984). "Atmospheric Science and Power Production," Technical Information, Office of Scientific and Technical Information,

- United States Department of Energy, DE 84005177 (DOE/TIC-27601), National Technical Information Service, U.S. Department of Commerce, Springfield, Virginia.
- Schnelle, K. B., Jr., and Dey, P. R. (2000). "Atmospheric Dispersion Modeling Guide," McGraw-Hill, New York.
- Schnelle, K. B., Jr., and Schnelle, K. D. (1983). "A New Look at Chimney Design," *Environ. Prog.* **2**(2), 91–103.
- Scorer, R. (1958). "Air Pollution," Pergamon Press, London.
- Seinfeld, J. H., and Pandis, S. N. (1998). "Atmospheric Chemistry and Physics," Wiley, New York.
- Slade, D. H., ed. (1968). "Meteorology and Atomic Energy 1968," U.S. Atomic Energy Commission, TID-24190, Clearinghouse for Federal Scientific and Technical Information, National Bureau of Standards, U.S. Department of Commerce, Springfield, Virginia.
- Smith, M., ed. (1968). "Recommended Guide for the Prediction of the Dispersion of Airborne Effluents," The American Society of Mechanical Engineers, New York.
- Taylor, G. I. (1921). "Diffusion by Continuous Movements," *Proc. London Math. Soc., Ser. 2*, **20**, 196.
- Turner, D. B. (1994). "Workbook of Atmospheric Dispersion Estimates," 2nd ed., Lewis Publishers, Boca Raton, Florida.
- Turner, D. B. (August 1997). "The Long Life of the Dispersion Methods of Pasquill in U.S. Regulatory Air Modeling," *J. Appl. Meteor.* **36**, 1016–1020.



Atmospheric Turbulence

Albert A. M. Holtslag

*Department of Meteorology and Air Quality
Wageningen University, The Netherlands*

- I. Introduction
- II. Atmospheric Turbulence and Mean Motions
- III. Turbulent Kinetic Energy
- IV. Turbulence in the Atmospheric Surface Layer
- V. Turbulence in the Atmospheric Boundary Layer
- VI. Spectrum of Atmospheric Turbulence
- VII. Atmospheric Turbulence and Modeling
- VIII. Turbulent Dispersion of Air Pollutants

GLOSSARY

Atmospheric boundary layer (ABL) Lower part of the atmosphere that is directly influenced by the presence of the Earth's surface and its typical characteristics.

Convection Fluid flow driven by vertical temperature differences.

Dispersion Spreading of a gas or particles through the air.

Flux (or more precisely flux density) Transport of a quantity per unit area in unit time.

Mechanical turbulence Turbulence driven by (vertical) wind shear.

Spectrum Distribution of (co-)variance over frequencies or wave numbers.

Surface layer Approximately the lowest 10% of the atmospheric boundary layer.

Wind shear Rate of change of wind component with (vertical) distance.

TURBULENCE is chaotic fluid flow characterized by the appearance of three-dimensional, irregular swirls. These swirls are called eddies, and usually turbulence consists of many different sizes of eddies superimposed on each other. In the presence of turbulence, fluid masses with different properties are mixed rapidly. Atmospheric turbulence usually refers to the small-scale chaotic flow of air in the Earth's atmosphere. This type of turbulence results from vertical wind shear and convection and usually occurs in the atmospheric boundary layer and in clouds. On a horizontal scale of order 1000 km, the disturbances by synoptic weather systems are sometimes referred to as two-dimensional turbulence. Deterministic description of turbulence is difficult because of the chaotic character of turbulence and the large range of scales involved. Consequently, turbulence is treated in terms of statistical quantities. Insight in the physics of atmospheric turbulence is important, for instance, for the construction of buildings and structures, the mixing of air properties, and the dispersion of air pollution. Turbulence also plays an

important role in weather forecasting and climate modeling. Here, we review the basic characteristics of atmospheric turbulence with an emphasis on turbulence in the clear atmospheric boundary layer.

I. INTRODUCTION

Atmospheric turbulence owes its importance in meteorology and air quality to its characteristic of mixing air with different properties efficiently. In practice, turbulence may cause engineering problems, because it shakes structures such as bridges, towers, and airplanes, causing failure of such systems in extreme cases. In the case of airplanes, most interest is in relatively small-scale vertical motion that produces discomfort and occasionally severe damage and injuries. Information on both the intensity and the scale of the turbulence is needed to estimate the consequences of such turbulence. Similarly, turbulent fluctuations in the horizontal motions during severe storms can be fatal to tall buildings or bridges, particularly if resonance (e.g., forcing of a system at its natural frequency) occurs.

Where does turbulence occur and how does it relate to the scale of other processes in the atmosphere and climate system? [Figure 1](#) gives an overview of important atmospheric and climate processes as a function of horizontal scale (extent) and time scale (duration). At the lower left side, the figure starts with a length scale of 100 m and increases up to the circumference of the Earth. The corresponding time scales vary between seconds and centuries. From small to large scales, we note the typical sizes and energy ranges of atmospheric phenomena as (mechanical) turbulence, convection, a cumulus (fair weather) cloud, a thunderstorm, the sea breeze, and synoptic weather systems which appear as low- or high-pressure systems on weather maps. [Figure 1](#) also shows some of the relevant processes in the climate system which affect the atmosphere on larger temporal and spatial scales, such as atmosphere–land interactions and the ocean circulation.

Atmospheric turbulence usually refers to the three-dimensional, chaotic flow of air in the Earth's atmosphere with a time scale of less than 1 sec to typically 1 h. The corresponding length scales are from 1 mm (thus, five orders of magnitude smaller than the axis in [Fig. 1](#) indicates) up to the order of 1 km. This type of turbulence includes the mechanical turbulence by vertical wind shear and turbulence by convection in the lower part of the troposphere, known as the atmospheric boundary layer (ABL). The depth of the ABL can vary over land between tens of meters during night up to kilometers during daytime. Over sea, the depth is typically a few hundred meters and rather con-

stant on the time scale of a day. Turbulence in the ABL transfers momentum, heat, moisture, carbon dioxide, and other trace constituents between the Earth's surface and the atmosphere. It also impacts on air pollution and the erosion of soil.

Most of the atmosphere above the ABL is not turbulent, although turbulence can occur throughout the atmosphere. For instance, cumulus-type clouds, which may grow into thunderstorms, are always turbulent through convection produced due to the heat released by the condensation of water vapor. Turbulence can also occur in clear air above the ABL; most of this is produced in layers of strong vertical wind shear at the boundary between air masses. Such layers tend to slant with slopes of order 1–100; their vertical extent is typically several hundred meters to a few kilometers. This so-called “Clear-Air-Turbulence” has surprised many airplane passengers. On the larger horizontal scales, the disturbances by synoptic weather systems are sometimes referred to as two-dimensional turbulence on top of a large-scale mean flow.

Its capability to mix air with different properties efficiently introduces atmospheric turbulence into different types of application. For instance, the correct formulation of atmospheric turbulence and the transfer of quantities between the Earth's surface and the atmosphere are essential parts of atmospheric models for the prediction and research of weather, climate, and air quality. Because of the mixing capacity of turbulence, chimney plumes are diluted and spread over larger volumes than they would be without turbulence. Strong local peaks of pollution are prevented, and otherwise clean air is polluted. Thus, the air pollution meteorologist must distinguish between conditions of strong and weak turbulent mixing.

Turbulence also sets up irregular blobs of temperature and moisture in the atmosphere. Refractive indices for sound waves and electromagnetic waves depend primarily on temperature and moisture. Therefore, there are irregular volumes of low and high refractive index that act as imperfect diffraction gratings and partially reflect these types of waves. There are two types of applications of this phenomenon. In the first application, pulsed waves can be emitted, typically from the ground (radar, sodar, lidar), and their reflections can be recorded at the ground. The characteristics of the reflected waves provide information about the turbulence and sometimes the mean motion at the point of reflection (by measuring Doppler shifts). The second application involves the interference of turbulence with waves. In these cases the waves themselves are used for other purposes such as communication. The properties of turbulence must be known to estimate the distortion of the waves or the reduction of their energy.

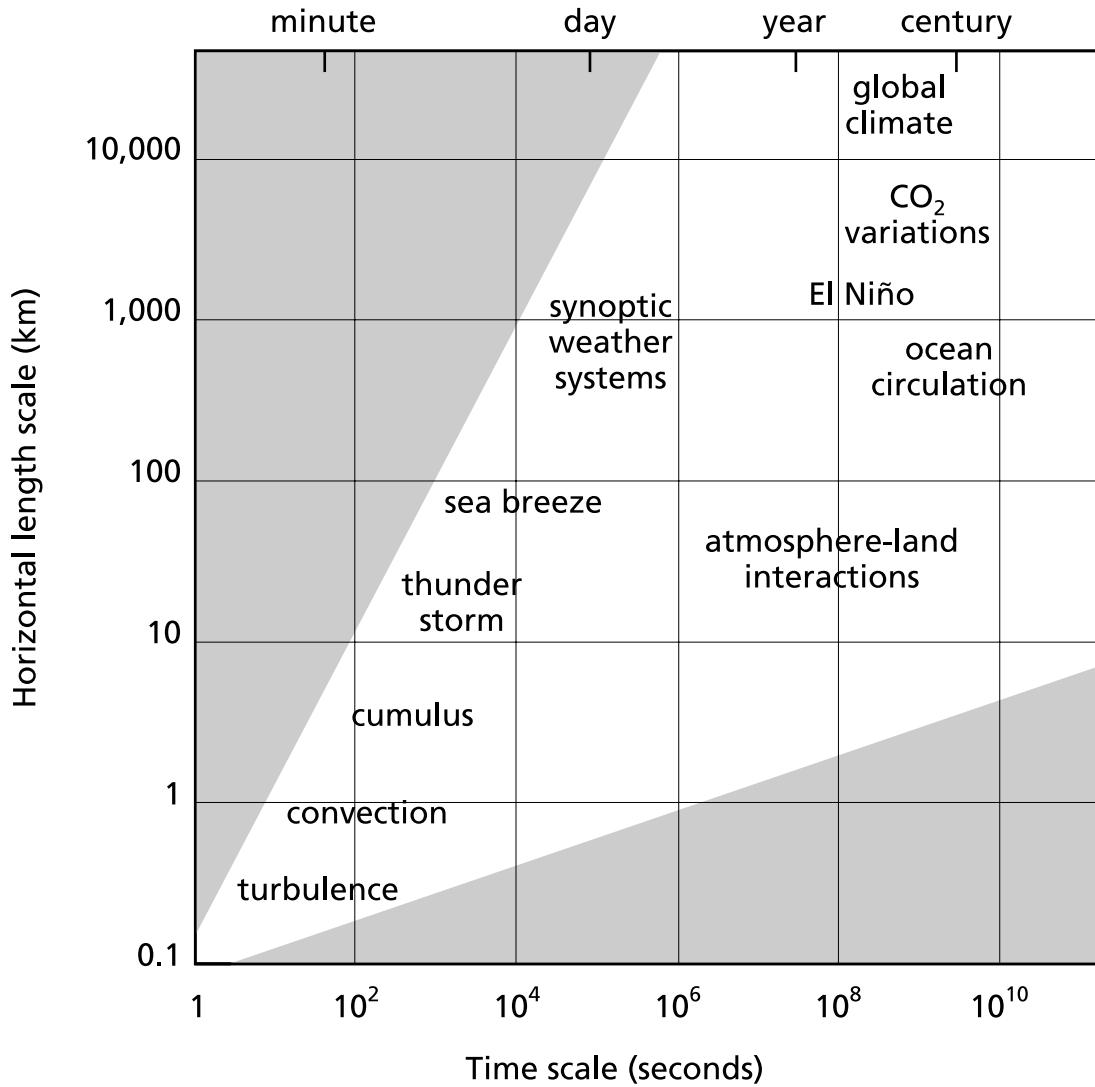


FIGURE 1 Characteristic spatial and temporal scales for atmospheric processes related to weather and climate.

II. ATMOSPHERIC TURBULENCE AND MEAN MOTIONS

There is an enormous range of scales in atmospheric motion, as indicated in Fig. 1. For many subjects the detailed description of the small-scale turbulent motions is not required. In addition, the randomness of atmospheric turbulence makes a deterministic description difficult. As such, there is a need to separate the small scales of atmospheric turbulence from “mean” motions on the larger scales. Let C denote an atmospheric variable, such as specific humidity. Then \bar{C} represents a mean or “smoothed” value of C , typically taken on a horizontal scale of order 10 km or a time scale in the order of 30 min to 1 h. A local or instantaneous value of C would differ from \bar{C} . Thus, we have

$$C = \bar{C} + c, \quad (1)$$

where c represents the smaller scale fluctuations. Note that we use lower case for the latter (often primes are used as well to indicate fluctuations). In principle, the fluctuations around the mean motion may reflect gravity waves as well as turbulence. When convection is present, gravity waves typically do not exist. In other cases, gravity waves may coexist with turbulence, and if the wind at the same time is weak, there may be no turbulence. If turbulence exists, it is usually more important for most atmospheric applications because it mixes more efficiently than gravity waves.

In order to make the mathematical handling of c tractable, it must satisfy the so-called Reynolds postulates.

These require, for example, that $\bar{c} = 0$. After a quantity has been averaged to create a large-scale quantity, further averaging should produce no further changes. The mean of the summation of two variables A and C should produce $\overline{A \pm C} = \overline{A} \pm \overline{C}$. A further condition is that a mean variable \overline{C} must be differentiable, since terms like $d\overline{C}/dt$ occur in the atmospheric equations. In practice, not all these conditions are rigorously satisfied.

If the Reynolds postulates are fulfilled, then the averaging for the product of two variables provides

$$\overline{AC} = \overline{A}\overline{C} + \overline{ac}. \quad (2)$$

The second term at the right-hand side of Eq. (2) is known as the turbulent covariance. Similarly, the turbulence variance of a quantity is given by $\overline{C^2} - \overline{C}^2$ (which is the square of the standard deviation). If in Eq. (2) A represents one of the velocity components, then \overline{AC} is the total flux of C , and the second term at the right-hand side of Eq. (2) represents a turbulent flux of C . For instance, \overline{uc} and \overline{wc} are the horizontal and vertical turbulent fluxes of the variable C , respectively. Here, u and w are the turbulent fluctuations of the horizontal and vertical velocities. Near the surface, the mean vertical wind \overline{W} is usually small, and, thus, the total vertical fluxes are normally dominated by the turbulent contributions.

Turbulent fluctuations, variances, and fluxes of variables are normally influenced by the vertical variation of temperature in the atmospheric layer of interest. Since pressure decreases with altitude, air parcels, which are forced to rise (sink), do expand (compress). According to the first law of thermodynamics, a rising (sinking) parcel will cool (warm) if there is no additional energy source such as condensation of water vapor. This is called a dry adiabatic process. It can be shown that in the ABL, the temperature variation with height for a dry adiabatic process is $dT/dz = -g/C_p$ (here g is gravity constant and C_p is specific heat at constant pressure). The value for g/C_p is approximately 1 K per 100 m. An atmospheric layer which has such a temperature variation with height is called neutral (at least when there is no convection arising from other levels). In that case, $\Theta = T + (g/C_p)z$ is constant, where Θ is called the potential temperature. (Note that the previous definition for potential temperature is not accurate above the boundary layer.)

In a neutral layer, vertical fluctuations in atmospheric flow can maintain themselves. If the potential temperature of the atmospheric layer increases with height, vertical displacements are suppressed. This is called a stable condition. On the other hand, when the potential temperature decreases with height, vertical fluctuations may be accelerated. Consequently, this is called an unstable condition. Thus, in considerations with vertical fluctuations and

atmospheric stability, we have to deal with potential temperature and not with the actual temperature. Similarly, the vertical flux of sensible heat is connected to turbulent fluctuations of potential temperature; e.g., it reads as $\overline{w\theta}$ (in mK/s). The latter relates directly to the energy per time and unit area H by $H = \rho C_p \overline{w\theta}$ (in W/m²), where ρ is density of the air (in kg/m³).

III. TURBULENT KINETIC ENERGY

The kinetic energy of atmospheric motion per unit of mass E is given by half of the sum of the velocities squared in the three directions (as in classic mechanics), e.g., $E = (U^2 + V^2 + W^2)/2$. Similar to Eq. (2), we can separate between the Mean Kinetic Energy \overline{E} of the mean atmospheric motions and the Turbulent Kinetic Energy (TKE or e) of the smaller scale fluctuating motions by turbulence. Thus, e is given by $e = (\overline{u^2} + \overline{v^2} + \overline{w^2})/2$. It is possible to derive a budget equation for e , which in its basic form reads as:

$$de/dt = S + B + D - \varepsilon \quad (3)$$

Here, de/dt is the total variation of e with time (the sum of local variations and those transported with the mean air motion). The other terms represent shear production of turbulence S , buoyancy production or destruction effects B , transport and pressure redistribution effects D , and finally the dissipation term ε .

Let us consider the physics behind the terms in (3) briefly. The quantity S in (3) measures the rate of production of mechanical turbulence, which arises when the fluid is stirred. For example, in the horizontal x -direction the term is given by $-\overline{uw} d\overline{U}/dz$. It depends primarily on vertical variations of wind or, near the ground, on wind speed and surface roughness. It is almost always positive. The quantity B is the rate of production or breakdown of turbulence by buoyancy effects (such as heat convection). It is given by $gw\overline{\theta_v}/\overline{\Theta_v}$, where the subscript “ v ” indicates a virtual temperature. The virtual temperature accounts for the influence of water vapor in the air on the buoyancy, because moist air is lighter than dry air. Thus, we note that B depends on the vertical variation of virtual potential temperature and its flux. The term D in Eq. (3) represents divergence and pressure redistribution terms. These have a tendency to cancel near the surface. Finally, the term ε reflects the molecular dissipation of turbulence into heat, and this term is always positive. In fact, ε is typically proportional to e/τ , where τ is the characteristic time scale for the turbulent mixing process.

In daytime conditions over land, the buoyancy term B is typically positive and creates turbulence. This is the case because the atmosphere is almost transparent to

visible light on clear days. Even on cloudy days, significant amounts of solar radiation can reach the surface. As a result, solar radiation warms the land surface, creating moderate to strong temperature gradients near the surface. This produces convection throughout the ABL, and the whole ABL is heated by this convection. Thus, the largest turbulent eddies may have the size of the thickness of the ABL due to convection. When there is wind, the term S is also positive and mechanical turbulence coexists with heat convection. The thickness of the daytime ABL over land is typically of order of 1 km, but it may be three times as large with strong heating. It increases rapidly in the morning and very slowly after the time of maximum heating.

The buoyancy term B is usually negative over land at night and therefore destroys turbulence in such cases. This is due to the cooling of the surface by the loss of infrared energy, which is strongest in clear nights. If there is some wind, mechanical turbulence occurs; S is positive and B is negative. But the ratio of $-B/S$ increases upward and becomes nearly critical at a height of typically 100 m. Above this level, turbulence is weak and intermittent, but still is able to extend the cooling to several hundred meters. Farther up, cooling proceeds by infrared radiation only, and the flow is not turbulent. As a result, we must distinguish between the depth of the turbulent layer h and the depth in which nighttime cooling is felt. For problems of pollution prediction, h is more important. If nighttime ABL moves over a city, heat convection is enhanced by the warm city surface, and a thicker and more turbulent mixed layer develops. In this case, h is usually between 100 and 300 m.

The ratio $-B/S$ is known as the flux Richardson number Ri_f . It tends to be positive on clear nights, negative on clear days (with slow or moderate winds), and near zero on cloudy, windy days or nights. Qualitatively, it describes the relative importance of heat convection and mechanical turbulence. It is used in modeling atmosphere motions, in describing the character of turbulent dispersion of air pollutants, and in atmospheric stability problems. When the value of the flux Richardson number exceeds 0.25, B and ε oppose S sufficiently to destroy mechanical turbulence. The value of 0.25, which separates turbulence from no such turbulence, is called the “critical” Richardson number. Actually, the flux Richardson number is often replaced by the so-called gradient Richardson number Ri_g , which is usually of the same order of magnitude but is measured more easily.

The discussion so far has concentrated mostly on turbulence over land, where its characteristics may change radically from night to day (depending on the heating and cooling of the surface). Over water, this diurnal variation is not very strong. In any case, on windy days mechanical turbulence normally dominates the ABL ($S > B$). Then the

thickness of the ABL is mostly proportional to wind speed, but it also depends on surface roughness and other factors.

IV. TURBULENCE IN THE ATMOSPHERIC SURFACE LAYER

The lowest 10% of the ABL is called the atmospheric surface layer. In this layer, the behavior and the characteristics of all variables are relatively simple. It is also an important layer because most human activity takes place in the surface layer. In the surface layer, changes of mean wind direction with height are small and usually negligible; also, vertical fluxes of momentum, heat, and moisture vary little relative to their surface values, so that the vertical variation can usually be neglected. These simplifications are normally valid over homogeneous terrain. Regions of hilly terrain or rapidly changing ground cover or topography do present special problems.

Let us consider a neutral surface layer in stationary and horizontal homogeneous conditions. In that case, $de/dt = 0$, $B = 0$ (because of the adiabatic temperature profile and a zero heat flux at the surface), and usually also $D = 0$. Then, Eq. (3) provides a balance between shear production S and dissipation of kinetic energy ε . In the neutral surface layer with mean wind speed \bar{U} in the x -direction, we have $S = -\bar{uw}_0 d\bar{U}/dz$, where uw_0 represents the vertical flux of horizontal momentum near the surface. This momentum flux is directly related to a friction force on the atmospheric motion by the drag of the surface. Since the momentum flux plays a dominant role in the surface-layer turbulence, it is used to define a characteristic turbulent velocity scale u_{*0} by

$$u_{*0}^2 \equiv -\bar{uw}_0. \quad (4)$$

Here, u_{*0} is known as the (surface) friction velocity. It is known from observations that in the neutral surface layer the standard deviation of the velocity components are all proportional to u_{*0} , and, therefore, ε is proportional with u_{*0}^2 . As an example, the standard deviation of vertical velocity fluctuations is given by $\sigma_w \approx 1.3u_{*0}$.

In the neutral surface layer the relevant length scale of the turbulence is given by the height z above the surface, since this height directly impacts on the size of the turbulent eddies transporting momentum and scalars near the surface. Thus, the time scale of turbulence $\tau \propto z/u_{*0}$. Then with $\varepsilon \propto u_{*0}^2/\tau$, we have $\varepsilon \propto u_{*0}^3/z$. Equating shear production and dissipation provides for neutral conditions:

$$\frac{d\bar{U}}{dz} = \frac{u_{*0}}{\kappa z}, \quad (5)$$

where κ is a constant that absorbs all proportionality factors. This constant is known as the “Von Karman”

constant. The value for the latter is usually taken as 0.4, although its exact value is controversial. The observed variation is about 10% around 0.4, and there have been experiments indicating that κ may vary with flow parameters.

According to Eq. (5), the vertical gradient of mean wind speed (known as wind shear) is directly related to friction velocity and the actual height above the surface. In fact, the latter two quantities could also have been used *a priori* to estimate the wind shear in the surface layer by noting their physical dimensions. In the surface layer, the actual wind shear arises due to the presence of the surface. At the surface the mean wind speed should be zero. Consequently, integration of (5) provides

$$\overline{U}_z = \frac{u_{*0}}{\kappa} \ln\left(\frac{z}{z_0}\right). \quad (6)$$

Here, \overline{U}_z is the mean wind speed at height z , and z_0 enters as an integration parameter for the surface such that $\overline{U}_0 = 0$. The parameter z_0 is known as the surface roughness length, since it reflects the aerodynamic roughness of the surface. Its value varies from (much) less than 1 mm over smooth ice and water up to several meters over cities or forests. It may change with height and wind direction, because turbulence near the surface is sensitive only to the roughness of the ground cover; higher up, the turbulence senses the larger scale topography and upwind roughness.

In principle, Eq. (6) is valid for the height range $z_0 \ll z \ll h$ in neutral conditions. Indeed, it has been shown that Eq. (6) is in good agreement with observations in the lower atmosphere up to 100 m or higher in windy conditions. This is typically the case when over flat terrain the wind speed at a height of 10 m $\overline{U}_{10} > 6$ m/s or so. Therefore, the neutral wind profile can be used to estimate u_* in such conditions, when observations of wind speed at one level are available and if proper estimates for the roughness length can be made. Consequently, also the standard deviations of wind speed fluctuations and turbulent kinetic energy can be estimated from single wind speed in such conditions. Note that over water surfaces, the standard deviations of wind speed and turbulent kinetic energy are less for the same wind speed because of the lower surface roughness.

So far, we have considered neutral conditions in the surface layer. In cases with light to moderate wind speeds, the effects of surface heating and cooling also become relevant. Consequently, the virtual sensible heat flux $\overline{w\theta_{v0}}$ at the surface needs to be considered as an important variable as well [because it impacts on the buoyancy term in Eq. (3)]. The combination of surface heat flux with friction velocity defines a characteristic length scale L , given by

$$L \equiv -\frac{u_{*0}^3}{\kappa g w \overline{\theta_{v0}} / \Theta_v}. \quad (7)$$

This length scale is known as the Monin-Obukhov length. It is defined to have the opposite signs as B in Eq. (3). It appears that the ratio $-z/L$ or behaves as a Richardson number from large negative values in strong convection through near zero in pure mechanical turbulence and to positive values in stable air (e.g., at night over land). Unlike the Richardson numbers, however, there is no critical value for z/L .

In the surface layer over homogeneous terrain, so-called surface-layer (or Monin-Obukhov) similarity theory can be used to study the combined effects of convective and mechanical turbulence on the profiles and distributions of mean and turbulence characteristics. For instance, if any variable with the dimensions of velocity is normalized by the friction velocity u_{*0} , the resultant dimensionless quantity then only depends on z/L . As an example of the application of this type of similarity or “scaling,” consider the standard deviation of vertical velocity σ_w . Surface-layer similarity theory requires

$$\sigma_w/u_{*0} = f(z/L), \quad (8)$$

where $f(z/L)$ should be a “universal” function. The latter can be inferred from measurements or sometimes deduced from more sophisticated theories. The same type of procedure applies to other variables, such as moisture and potential temperature. For these variables, their surface fluxes are needed as well to define characteristic scaling quantities. Surface-layer similarity is confirmed by many observations over different types of surfaces. Over the ocean too, surface-layer scaling gives good results in the lowest 10% of the ABL. However, the horizontal velocity scales do follow another type of scaling as discussed in Section V.

Finally, applying surface-layer similarity, the wind speed gradient of (5) can be generalized into

$$\phi_m \equiv \frac{\kappa z}{u_{*0}} \frac{d\overline{U}}{dz} = g(z/L), \quad (9)$$

where $g(z/L)$ is another function. The latter can also be integrated to obtain a more general wind profile than (6) indicates. Equation (9) applies well for a large range of stability conditions over generally homogeneous terrain up to 100 m or so, as long as proper estimates for the Monin-Obukhov length are utilized. Procedures are available in the literature to estimate the Monin-Obukhov length from routinely available observations.

V. TURBULENCE IN THE ATMOSPHERIC BOUNDARY LAYER

As mentioned above, in the surface layer the vertical variation of the turbulent fluxes can be neglected. As a result,

the scaling parameters are usefully defined in terms of surface fluxes. Above the surface layer, the influence of the turbulent boundary layer depth (denoted by h) becomes relevant. Consequently, dimensionless groups of variables will typically depend on z , h , and L . From these length scales we can form three dimensionless groups, namely, relative height z/h , the surface-layer stability ratio z/L (as above), and the stability parameter h/L . The latter indicates the (in)stability of the whole ABL.

In Fig. 2, we used the relative height z/h vs the overall stability h/L for unstable cases in the ABL. Note that logarithmic scales are used for the axis to stress the importance of the surface layer in these cases. The diagram applies for horizontally homogeneous and stationary cases without clouds in the ABL. Several regions are indicated

with their characteristic scaling variables (as explained shortly). For $z/h < 0.1$, we note the surface layer as discussed above. It appears that for $-z/L > 0.5$ (dashed line) to 1 (solid line), the convection is so strong that the influence of mechanical turbulence can be neglected. In that case, friction velocity drops from the list of relevant variables, and the turbulent condition is solely determined by the surface virtual heat flux and the height z . These variables define a characteristic velocity scale w_f by

$$w_f^3 \equiv \frac{g}{\Theta_v} \overline{w\theta_{v0}} z. \quad (10)$$

Consequently, during free convection in the surface layer $\sigma_w \propto w_f$.

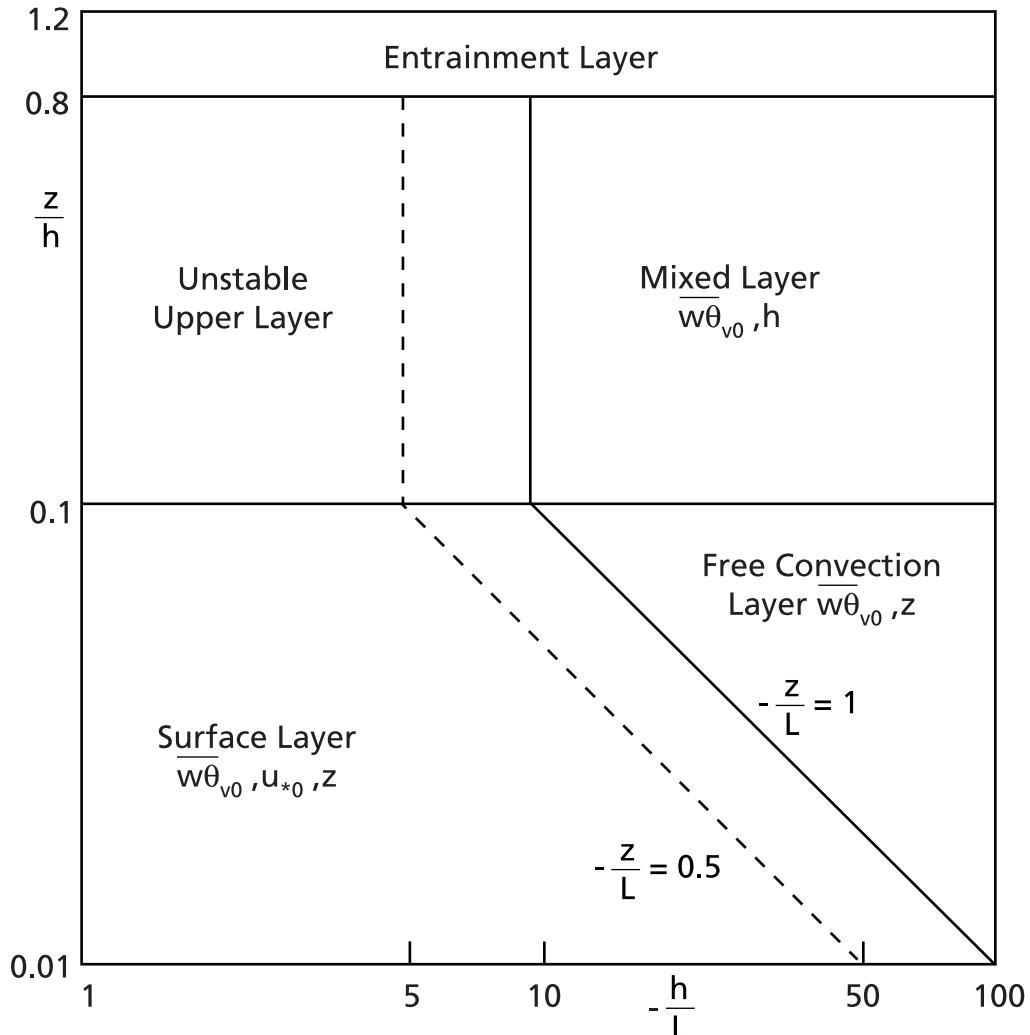


FIGURE 2 A diagram indicating the various similarity regions in the unstable atmospheric boundary layer. The indicated scaling parameters are defined in the text. [After Fig. 1 in Holtslag, A. A. M., and Nieuwstadt, F. T. M. (1986). *Boundary-Layer Meteorol.* **36**, 201–209.]

Above the surface layer ($z/h > 0.1$), one can distinguish between unstable conditions and mixed-layer conditions. This depends on the value of h/L . In fact, for $-h/L > 5$ (dashed line) to 10 (solid line), so-called mixed-layer scaling applies. In such cases the buoyancy term B of Eq. (3) dominates the budget for turbulent kinetic energy across the convective ABL. In the mixed layer, the scaling length is the boundary layer depth h itself because the largest eddies are of the size of h and these are very effective in mixing across the whole boundary layer. Consequently, a scaling velocity w_* is defined by

$$w_*^3 \equiv \frac{g}{\Theta_v} \overline{w\theta_{v0}} h. \quad (11)$$

Note that by definition we have $(w_*/u_{*0})^3 = -h/(\kappa L)$ and also $(w_f/u_{*0})^3 = -z/(\kappa L)$, meaning that the ratios of the velocity scales are alternative measures for stability in the ABL and in the surface layer, respectively.

In convective conditions it appears that the standard deviations of vertical and horizontal wind are all proportional to w_* , and the variation with height is relatively small. In fact, the precise value of $-h/L$ is unimportant above the surface layer once the ABL is in the convective state. In unstable conditions with $-h/L < 5$, it appears that all scaling variables may influence turbulence. Thus, in that case we expect

$$\frac{\sigma_w}{u_{*0}} = f\left(\frac{z}{h}, \frac{h}{L}\right), \quad (12)$$

where $f(z/h, h/L)$ is another universal function. It is noted that the shape of the latter function should allow for the correct limits of σ_w when the surface layer is approached and also when the stability ratio $-h/L$ is varied. Note that horizontal wind fluctuations, in contrast to vertical wind fluctuations, are not so strongly affected by distance from the ground in the surface layer; instead, horizontal fluctuations depend mostly on the size of the large eddies, which are limited only by the ABL depth h .

Standard deviations of turbulent fluctuations for potential temperature, and passive scalars such as specific humidity, also follow surface-layer and mixed-layer scaling. For passive scalars (and similar for temperature), the scaling quantities are $c_* = \overline{wc_0}/u_*$ for the surface layer and $c_{**} = \overline{wc_0}/w_*$ in the mixed layer. Here, c_* and c_{**} are concentrations of the scalar per unit mass in the surface and mixed layer, respectively; $\overline{wc_0}$ is the surface flux of the scalar. The scaling variables are also useful to describe profiles of mean quantities in the surface layer, similar to (9) for wind speed. Above the surface layer, the scaling of mean profiles becomes more complicated, except on convective days when many variables remain constant with height throughout the ABL or vary slowly.

Let us now turn to stable conditions in the ABL. A stable stratification typically leads to small turbulent eddies because of the opposing buoyancy effects. In such cases, $B < 0$ in Eq. (3), and together with the dissipation term ε it typically balances with the shear production term S (the other terms are usually small in the stable case for horizontally homogeneous and stationary conditions). Consequently, the structure of the stable boundary layer is completely different from that of the unstable boundary layer. Nevertheless, in the stable surface layer, Monin-Obukhov theory still applies for many quantities (as long as the stratification is not too strong). Thus in principle, turbulent quantities can be normalized by their surface fluxes, u_{*0} , L , and z . Above the surface layer ($z/h > 0.1$), it follows from theory and observations that the turbulent structure is determined by the *local* values of the momentum and heat fluxes. This is the basis of “local scaling.” In that case, the relevant length scale becomes the so-called local Obukhov length Λ , defined by

$$\Lambda \equiv -\frac{u_*^3}{\kappa g \overline{w\theta_{v0}}/\Theta_v}. \quad (13)$$

Note that in the stable surface layer $\Lambda \cong L$.

With local scaling we have $\sigma_w/u_* = f(z/\Lambda)$. As such, local scaling can be regarded as a natural extension of the scaling of turbulence in the stable surface layer. For large values of z/Λ , we expect that the dependence on z must disappear. This is the case because for strong stability the turbulent eddies will be so small that the presence of the surface is not felt locally. Consequently, dimensionless quantities approach a constant value. Thus, for instance, σ_w becomes proportional to the local friction velocity u_* . This is known as z -less scaling. It appears that z -less scaling applies for $z/\Lambda > 1$ in homogeneous and stationary cases. In more complicated cases, it has been found that scaling with σ_w sometimes provides more useful results than scaling with u_* .

[Figure 3](#) shows a diagram with the scaling regions for stationary and horizontally homogeneous turbulence in stable conditions. Here, we used linear scales for the axis to indicate that the surface layer is rather thin in stable conditions. For example, often the depth of the turbulent layer h is only 200 m or less. Consequently, the surface layer extends to a height of only 20 m or less. We stress that the diagram is based on simplifications and we refer to the cited literature for more details. Nevertheless, the scaling diagram of [Fig. 3](#) indicates the relative size of the different scaling regimes and their approximate separation. For $h/L < 1$, the normalized turbulent variables are described with z/h and h/L . For $h/L > 1$, we encounter the different scaling regimes with their scaling variables as described above.

So far, the discussion has been on turbulence, which is able to maintain a sufficient level of kinetic energy in the ABL. However, when stability (and consequently the Richardson number and the ratio z/Λ) increases, turbulence may become very weak and isolated in patches. Then turbulence is not continuously present anymore in space and time. This is known as intermittent turbulence, which is caused by the strong opposing effects of buoyancy against the shear production. Often, the standard deviations of wind increase in cases with intermittency, which is believed to be a result of gravity waves rather than turbulence. As a result, smoke plumes meander horizontally but spread little in the vertical direction.

The diagram of Fig. 3 indicates that beyond $h/L \approx 5$, an “intermittency” region becomes relevant and that this re-

gion increases strongly with stability. However, note that the line between the z -less and intermittency regions is rather tentative and only intended to provide a rough indication. Nevertheless, for strong stability, intermittent turbulence is even expected close to the surface. This makes the similarity description of turbulence very difficult. For these strong stability cases, the depth of the turbulent layer h is typically limited by $h \approx 10L$. Thus, h may adopt very small values (50 m or less) with decreasing L in very stable conditions. Such conditions appear typically for light winds in clear nights over land.

In this section, we have limited ourselves to the scaling of turbulence in the clear boundary layer without considering the effects by fog or boundary layer clouds. Incorporation of the latter will introduce additional parameters.

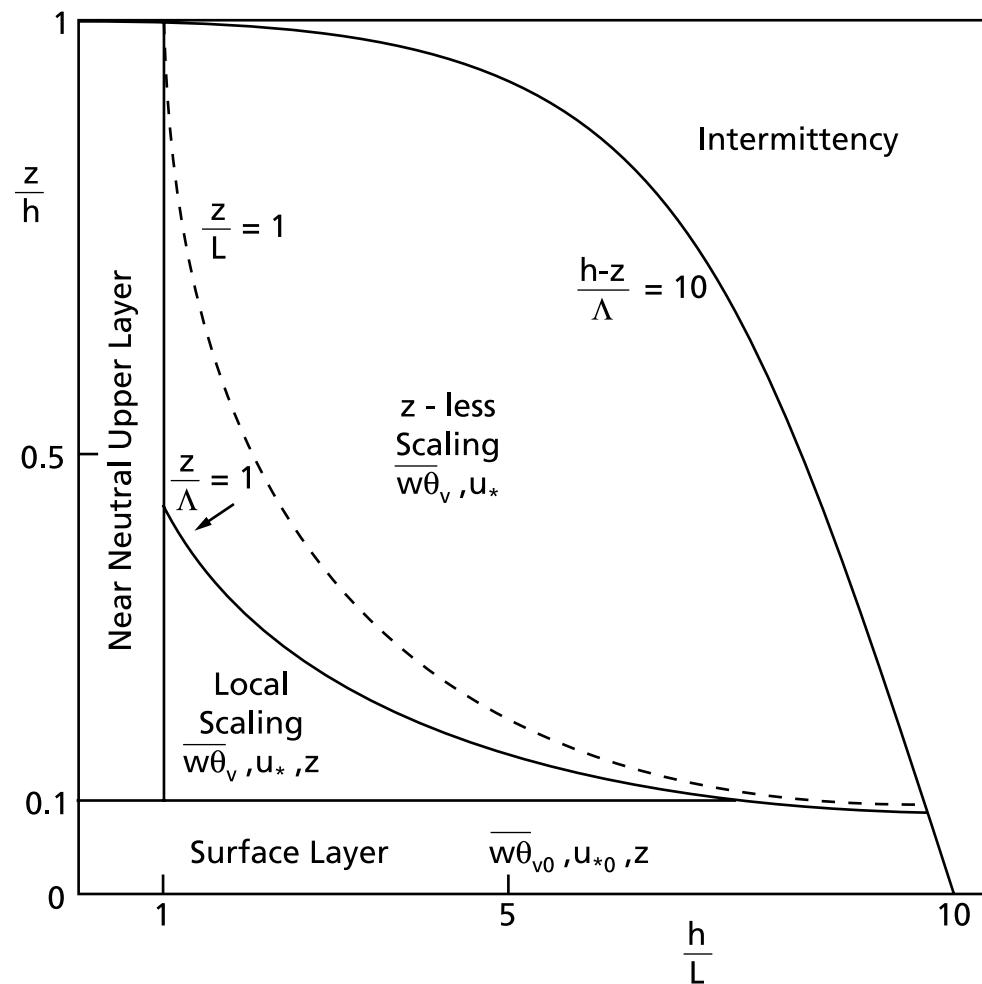


FIGURE 3 A diagram indicating the various similarity regions in the stable atmospheric boundary layer. The indicated scaling parameters are defined in the text. [After Fig. 2 in Holtslag, A. A. M., and Nieuwstadt, F. T. M. (1986). *Boundary-Layer Meteorol.* **36**, 201–209.]

VI. SPECTRUM OF ATMOSPHERIC TURBULENCE

It is often not sufficient to characterize the magnitude of small-scale fluctuations, but also to understand the “scale” of the fluctuations or, more precisely, the distribution of the energy of the fluctuations over different frequencies or wave numbers. This is known as the spectrum of turbulence. For example, in the case of “shaking” of towers or bridges, only relatively high frequencies are important; in contrast, high frequencies are relatively inefficient for mixing air masses with different properties.

Three types of statistics give information about the scale characteristics of small-scale motion: correlation functions, structure functions, and spectra. These three types of statistics give the same type of information and can be computed from each other. For example, spectra are cosine transforms of correlation functions. We describe here the properties of spectra only because they have several advantages over the other two types of statistics. First, spectral estimates at different frequencies or wave numbers are statistically independent. Second, the response of structures to atmospheric forcing can be calculated separately at different frequencies, given the transfer functions of the system. However, this scheme is limited to linear interactions between the atmospheric motions and the systems.

Spectra describe the distribution of variances over different frequency (f) or wave numbers (k). Let $S_k(k)$ be the wave number spectrum such that

$$\int_0^\infty S_k(k) dk = \int_0^\infty k S_k(k) d(\ln k) \equiv \sigma^2. \quad (14)$$

Here, subscripts are often used to indicate the variables to which the spectral densities and variables apply. Note that when $S_k(k)$ is plotted as function of the wave number k or $k S_k(k)$ as function of $\ln(k)$, the area between two frequencies represents the contribution of this frequency interval to the variance σ^2 . The logarithmic scale of frequency is most often used in atmospheric science. Note that the combination $k S_k(k)$ has the same dimensions as σ^2 .

There is a significant difference between spectra of the horizontal and vertical velocity components. Because the atmosphere is quasi-two-dimensional, very large “eddies” are quasi-two-dimensional; that is, at low frequencies, horizontal eddies have much more energy than vertical eddies. In contrast, high-frequency turbulence has more nearly equal energy for all components. This is why the variances of the horizontal components are generally much larger than those of the vertical component.

Following Kolmogorov, we may divide spectra into three regions. First, the portion of the spectrum where most of the energy is located and also where energy is in-

troduced by wind shear or buoyancy, is called the energy-containing region. This part of the spectrum occurs at relatively low wave numbers, and the related low frequencies are typically of the order of cycles per minute. Second, at very high wave numbers and high frequencies, the turbulence is dissipated into heat. The frequencies involved are of the order of hundreds of cycles per second. This region is called the viscous subrange. In between is the third region, in which energy is neither created nor destroyed. This region just serves as a conduit for energy from the energy-containing region to the viscous subrange. Therefore, this region is called the *inertial subrange*. Both the viscous and inertial ranges are nearly isotropic; this means that the statistics of the flow are rather invariant with rotation of the axes.

The theory of turbulence characteristics in the inertial range is particularly simple. Kolmogorov suggested that spectral density should depend only on turbulence dissipation rate ε [see Eq. (3)] and wave number k . Hence, by dimensional reasoning we have for the velocity components

$$S_k(k) = \alpha \varepsilon^{2/3} k^{-5/3}. \quad (15)$$

Here, α represents the so-called Kolmogorov constant, which is in the order of 1.6. Note that k at any height is normally proportional to f/\bar{U} , where f is frequency and \bar{U} is mean horizontal wind at that height.

Equation (15) has several practical applications. For example, structural engineers are interested in the spectral density of the horizontal velocity components in the frequency range 0.1–1 cycles per second (Hz), in which many structures have free periods. In this range, Eq. (15) applies. The value of ε can be estimated by assuming equilibrium between energy production and dissipation (for strong winds), as discussed in the derivation of Eq. (5). Subsequently, Eq. (15) then provides $S_k(k)$. Conversely, spectral densities $S_k(k)$ can be measured on masts or from aircraft yielding estimates of ε . Then given ε , stress can be estimated from the turbulent kinetic energy budget.

The spectra of scalars also have the same three regions: energy-containing, inertial, and dissipation ranges. The energy-containing range of the spectra of the vertical velocity follows surface-layer scaling, but the spectra of the horizontal velocity components and of scalars are more complex. In fact, most of the energy of the vertical velocity spectra near the ground occurs at frequencies of the order of 1 Hz. Eddies with such high frequencies are quite small and adjust rapidly to changing terrain. Therefore, statistics of it represent local terrain features. In contrast, horizontal velocity components have much more low-frequency energy, which changes slowly as the air moves over changing terrain, therefore, they have greater memory. As a result, variances of horizontal velocity components generally represent terrain up to hundreds of meters upstream.

Just as the spectra describe the distribution of variance over different frequencies or wave numbers, the cospectra describe the distribution of covariances over different frequencies or wave numbers. Vertical fluxes of momentum, heat, moisture, and pollutants are proportional to the covariances and are important in atmospheric dynamics and thermodynamics. They also have important applications to agricultural problems and to oceanography. In order to measure these quantities, instruments must be designed with the proper frequency response to give correct estimates. Area estimates of surface fluxes can be obtained by using remote-sensing sensors in the atmospheric surface layer on a scale of kilometers.

For some engineering applications, such as consideration of the stability of bridges or towers, the relation between fluctuations at different points along the structure is of interest. The two quantities usually needed are the coherence between fluctuations at points separated by distance and the phase difference between the fluctuations at such places. Coherence acts as the square of a correlation coefficient but is a function of frequency, usually in the sense that low-frequency fluctuations are more coherent than high-frequency fluctuations. Generally, coherence is large in convective turbulence and in gravity waves and becomes small close to the ground and in mechanical turbulence. Furthermore, it is larger along the mean wind direction than across the mean wind direction.

VII. ATMOSPHERIC TURBULENCE AND MODELING

Atmospheric models for research and forecasting of weather, climate, and air quality are all based on numerical integration of the basic equations governing atmospheric behavior. These equations are the gas law, the equation of continuity (mass), the first law of thermodynamics (heat), the conservation equations for momentum (Navier-Stokes equations), and usually equations expressing the conservation of moisture and air pollutants. At one extreme, atmospheric models deal with the world's climate and climate change; at the other extreme, they may account for the behavior of local flows at coasts, in mountain-valley areas, or even deal with individual clouds. This all depends on the selected horizontal scale and the available computing resources!

[Figure 1](#) indicates the enormous range of temporal and horizontal scales. In order to apply the governing equations on the regional and global scale, the variables are split into mean (larger scale) motions and small-scale fluctuations (as in Section II). Inserting this into the basic equations and averaging provides a set of equations for the behavior of the mean variables. Because the basis equations are nonlinear, the mean equations contain terms involving

small-scale motions. These are of the form of a divergence of fluxes produced by such motions. On the global scale, the atmospheric model equations are usually applied to fairly large air "boxes," which are typically a few hundred kilometers wide and a few hundred meters thick, but small-scale motions interact with the air parcels in these large boxes by their capacity to mix. For example, if a hot parcel is located next to a cold parcel, turbulent motion at their boundaries will heat the cool and cool the hot parcel. Thus, a closure formulation is needed to introduce mixing by the small-scale motion into the equations for the mean motion.

The most commonly used form of closure is known as first-order closure, also called K -theory. For instance, in this theory it is assumed that the flux \overline{wc} of a property C in the z -direction is down the gradient of mean concentration of \overline{C} per unit mass in that direction. Thus,

$$\overline{wc} = -K_c \frac{\partial \overline{C}}{\partial z}. \quad (16)$$

Here, K_c is known as the "eddy diffusivity" or mixing coefficient for the variable C . It can be different for different variables and for different directions.

The dimension of the eddy diffusivity is a length scale times a velocity scale. These are proportional to products of eddy size and eddy velocity in the corresponding directions, and they typically vary with height, wind speed, stability, etc. Generally, the eddy diffusivities can be modeled by using the kinetic energy equation of Eq. (3) and an appropriate choice for the length scale. As a simplification, it appears that the eddy diffusivity is often well described by a profile function such as

$$K_c = \frac{\kappa u_{*0} z}{\phi_c(z/L)} \left(1 - \frac{z}{h}\right)^2, \quad (17)$$

where $\phi_c(z/L)$ is a stability function for the variable of interest [such as $\phi_m(z/L)$ of Eq. (9) for momentum]. This equation appears to be quite successful for the surface layer and for the clear ABL.

In convective conditions, however, the flux is not proportional to the local gradient alone. In fact, in a large part of the ABL the gradients are small in convective conditions. Therefore, the fluxes depend mostly on the mixing characteristics of the large eddies across the ABL. Theories are available which have generalized K -theory to allow for this type of situation, for example, by including additional terms at the right-hand side of Eq. (16) representing the large-eddy mixing. In second-order closure, new equations are developed for the fluxes and variances themselves. Such equations have a very similar structure as Eq. (3) for kinetic energy. These equations also contain third-order terms, pressure terms, and molecular terms on the right-hand side, which are new unknowns. These must be related to the other variables in the equation, always involving assumptions. Second-order closure involves many

more than the original equations and is therefore more time consuming and expensive than first-order closure.

Finally, the atmospheric model equations can also be applied on the much smaller horizontal scale of the order of 1 to 10 km. In such cases, the horizontal resolution is only 100 m or less, which means that most of the turbulent fluctuations by convection are resolved by the model. This type of modeling is nowadays known as “large-eddy simulation (LES).” This has become a powerful and popular tool in the last decade to study turbulence in clear and cloudy boundary layers under well-defined conditions.

VIII. TURBULENT DISPERSION OF AIR POLLUTANTS

Atmospheric turbulence plays an important role in the estimation of concentrations of air pollutants at locations near the pollution source and on larger scales. This problem can be broken down into four parts: estimation of plume rise, transport, dispersion, and transformation. Plume rise depends partly on meteorological variables such as wind speed and stability, partly on exit speed and temperature of the pollutants, and partly on the source diameter. Many techniques exist for estimating plume rise, and turbulent entrainment into the plume is one factor that must be considered. However, in this case, the turbulence arises from the motion of the plume itself and is not natural turbulence. Natural turbulence may eventually play a role in mixing the plume to cause it to level off.

Transport involves the mean wind speed and direction. Dispersion is produced primarily by atmospheric turbulence and is discussed in more detail below. Transformations of the pollutants are due to chemical reactions, deposition, and interactions with liquid water. These constitute separate problems treated elsewhere. Most air pollution arises from continuous emitting sources, such as stacks or highways. They may be ground or elevated sources. An important parameter in the determination of concentration is the “effective source height,” which is the sum of physical source height and plume rise.

To estimate turbulent dispersion and the connected concentrations of air pollutants, no completely satisfactory techniques are available at present. For vertical mixing, first-order closure [K -theory; see Eq. (16)] gives reasonably good results in the case of ground sources. The problem here is the proper description of the eddy diffusivity K . If K is assumed to be independent of downward distance from the source x , the plume width varies as \sqrt{x} . This differs from observed plume widths. The difficulty is that a constant K describes a situation in which all diffusing “eddies” are small compared with the scale of the plume. This is true for molecular dispersion but not for turbulent dispersion. If the eddy diffusivity K is allowed to

vary with height (in the same manner as K for heat transport, for example), the effective K will grow with distance from a ground source. The results are then quite realistic for estimating vertical diffusion from a continuous ground source, except in the case of strong convection. K -theory is generally unsatisfactory for lateral dispersion.

The most popular method for determining both vertical and lateral dispersion does not involve the solution of any differential equation. Instead, the basic mathematical condition is one of continuity. For example, in the case of a continuous point source, the emission rate Q (in kg/s) must equal the integrals over the fluxes through any plane at right angles to the mean wind. Assuming the mean wind \bar{U} at the (effective) stack height to be in the x -direction, then the shape of the distributions in the y - and z -directions is needed. Close to the source, both vertical and lateral distributions are often assumed to be Gaussian with standard deviations σ_y and σ_z (both in meters). Then the mean maximum concentration \bar{C}_{\max} (in kg/m³) of a pollutant at plume height is given by

$$\bar{C}_{\max} = \alpha \frac{Q}{\bar{U} \sigma_y \sigma_z}, \quad (18)$$

where α is a constant.

The Gaussian assumption is well documented in the case of lateral dispersion but not for vertical dispersion and is particularly unsatisfactory for ground sources. Nevertheless, Gaussian methods, including reflection at the ground, are commonly used in practice. Once Gaussian forms are postulated, it remains to estimate the standard deviations. The standard deviations measure, approximately, one-quarter of the width of the plume in the horizontal and vertical, respectively. The deviations grow with distance from the source and depend on meteorological parameters such as the standard deviations of the horizontal and vertical wind directions. Theories are available to estimate these variables from routinely available observations. The fluctuations of horizontal and vertical wind direction are especially sensitive to atmospheric stability. Overall, mixing and dispersion are strongest in convective turbulence.

Far from the source the vertical distribution is taken as uniform because of the turbulent mixing in the ABL. As a result, the depth of the turbulent ABL (h) replaces σ_z in Eq. (18) for concentration. Thus, in such cases, the smaller the wind speed and mixing depth, the larger is the concentration. The product of wind speed and mixing depth is sometimes referred to as a ventilation factor. Maps of this quantity have been constructed. The factor is especially low in stagnant, high-pressure regions, leading to persistent high concentrations of pollutants in the ABL.

For air quality modeling, other techniques are also being tried, some involve Monte Carlo techniques and others similarity methods with the scaling parameters of Figs. 2

and 3. However, no consensus has emerged on the generally best techniques to estimate air pollution concentrations, not even for single sources. There is also a need for constructing and evaluating multiple-source models, particularly for cities. Typically, these models consist of superposition of single-source models with areas of finite size represented as single sources. Such models can be very complex in cases with important chemical transformations (e.g., photochemical smog). These complex models are used for planning future growth and designing strategies for rolling back undesirably high concentrations. Special models are required for the estimation of air pollution by long-range transport, for the understanding of acid-rain patterns, and for other subjects dealing with air quality on large atmospheric scales. In fact, such models are closely related to the atmospheric models for the forecasting and research of weather and climate.

ACKNOWLEDGMENTS

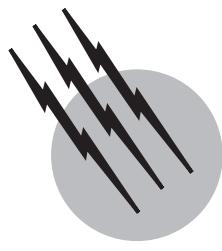
It has been a great pleasure and honor to edit and update the original contribution in this book by the late Professor H. Panofsky. The manuscript improved due to the comments of an unknown reviewer.

SEE ALSO THE FOLLOWING ARTICLES

ATMOSPHERIC DIFFUSION MODELING • CLOUD PHYSICS
• FLUID DYNAMICS • POLLUTION, AIR • WAVE PHENOMENA

BIBLIOGRAPHY

- Ahrens, C. D. (1999). "Meteorology Today: An Introduction to Weather, Climate and the Environment," 6th ed., Brooks/Cole, Pacific Grove.
- Arya, S. P. (1999). "Air Pollution Meteorology and Dispersion," Oxford Univ. Press, Oxford.
- Garratt, J. (1992). "The Atmospheric Boundary Layer," Cambridge Univ. Press, Cambridge, UK.
- Graedel, T. E., and Crutzen, P. J. (1993). "Atmospheric Change: An Earth System Perspective," Freeman, New York, USA.
- Holtslag, A. A. M., and Duynkerke, P. G., eds. (1998). "Clear and Cloudy Boundary Layers," Roy. Acad. Arts Sciences, Amsterdam.
- Panofsky, H. A., and Dutton, J. A. (1984). "Atmospheric Turbulence," Wiley, New York.
- Pasquill, F., and Smith, F. B. (1983). "Atmospheric Diffusion," 3rd ed., Ellis Horwood, Halsted, New York.
- Stull, R. B. (1999). "An introduction to Boundary-Layer Meteorology," Kluwer Academic, Dordrecht.
- Tennekes, H., and Lumley, J. L. (1982). "A First Course in Turbulence," 2nd ed., MIT Press, Cambridge, MA.



Aurora

S.-I. Akasofu

University of Alaska–Fairbanks

- I. Auroral Distribution
- II. Auroral Forms
- III. Auroral Spectra
- IV. Dynamics of the Aurora and Auroral Substorms
- V. Power Supply of the Aurora
- VI. Auroral Discharge Current System
- VII. Solar Activities and the Aurora
- VIII. Auroral Effects

GLOSSARY

Auroral electrojets Narrowly concentrated ionospheric currents; one flows eastward along the evening part of the auroral oval and the other flows westward in the morning part.

Auroral kilometric radiations Intense kilometric radio emissions emitted by auroral plasmas in the auroral potential structure.

Auroral oval The annular-like belt around the geomagnetic pole in which various auroral forms lie.

Auroral potential structure Electric structure formed by interaction between field-aligned electric currents and plasma at an altitude of about 10,000 km; accelerates the current-carrying electrons that excite or ionize upper atmospheric atoms and molecules. At the same time, it accelerates upward positive ions, such as O⁺, from the ionosphere.

Auroral substorm Global auroral activity with a lifetime of ~3 hr, consisting of three phases—growth, expansion, and recovery. It is the only visible manifestation of the magnetospheric substorm.

Auroral zone Narrow belt around the geomagnetic pole in which the seeing frequency (the number of nights per year) is maximum. Unlike the auroral oval, it is not the belt along which the aurora lies.

Coronal hole Relatively dark regions of the corona in X-ray photographs. A high-speed solar wind flows out from coronal holes. It tends to appear a few years after the year of sunspot maximum. The aurora becomes active when the Earth is.

Field-aligned currents Part of the auroral discharge current that flows along the geomagnetic field lines.

Geomagnetic pole Point where the axis of the Earth's dipole magnetic field reaches the Earth's surface.

Green line The line emission of wavelength 5577 Å (or 557.7 nm) emitted by oxygen atoms; this light has the most commonly observed color of the aurora.

Ionospheric current Part of the auroral discharge current that flows in the ionosphere, consisting of two narrowly concentrated currents—the eastward and westward electrojets.

Isochasm The line of equal auroral visibility.

Magnetosphere A cone-like cavity carved in the solar wind in which the Earth's magnetic field is confined.

Magnetospheric substorm The most basic type of disturbances of the magnetosphere, which manifests as the auroral substorm and other polar upper-atmosphere phenomena.

Plasma Fully or partially ionized gas consisting of an equal number of positive ions and electrons.

Polar magnetic disturbance (or substorm) Intense geomagnetic disturbances associated with the auroral substorm caused by the auroral discharge current.

Solar wind Solar particles consisting mainly of protons and electrons that flow out from the sun with a supersonic speed, reaching the outer fringe of the solar system.

THE AURORA is the visible manifestation of a large-scale electrical discharge process in the polar upper atmosphere and in space around the Earth. The discharge is powered by the generator process resulting from interaction between the solar wind and the Earth's magnetic field. The power involved is $\sim 10^6$ MW (megawatts). The discharge current-carrying electrons collide with atoms and molecules in the polar upper atmosphere, ionizing and/or exciting them. These atoms and molecules emit the visible lights, as well as extreme ultraviolet (EUV), ultraviolet (UV), and infrared (IR) emissions. The aurora is also associated with very low frequency (VLF) and ultra low frequency (ULF) radio emissions. The aurora is a very dynamic phenomenon and exhibits systematic motions on a global scale, particularly when solar wind is intense and the resulting generator power is high (see section VI). A typical auroral activity, called the auroral substorm, lasts for ~ 3 hr and is accompanied by magnetic field disturbances produced by the discharge currents that flow along the geomagnetic field lines and through the ionosphere. Auroral activity is related to solar activity in very complicated ways. Intense solar activities (solar flares, coronal mass ejections) associated with a large sunspot group cause gusty winds, while spotless regions called "coronal holes" tend to cause a wide, high-speed spectrum. Both cause intense auroral activities. Since the aurora is an electrical discharge phenomenon, it can cause interference on radio communications, radar, power transmission lines, oil and gas pipelines, and satellites.

I. AURORAL DISTRIBUTION

A. Auroral Oval

When one looks down on the Earth from far above the northern polar region, there is an annular belt of glow

surrounding the geomagnetic pole. This belt is called the auroral oval and consists of curtain-like forms of luminosity in the poleward half of the oval and a wide, diffuse glow in the equatorward half. Figure 1 shows an auroral image taken from the *Dynamics Explorer A* satellite at a distance of about 20,000 km from above the northern polar region. There is a similar belt in the Southern Hemisphere. The aurora in the Northern and Southern Hemispheres are called the *aurora borealis* and the *aurora australis*, respectively. The aurora borealis is commonly called the *northern lights*.

The size of the auroral oval changes considerably on a continuous basis. For an average-sized oval, the midday part is located at $\sim 76^\circ$ in geomagnetic latitude and the midnight part is located at $\sim 67^\circ$ (the geomagnetic latitude-longitude system based on the geomagnetic poles), with the average diameter being ~ 4000 km. At times, the oval can contract poleward (the midday and midnight parts being located at $\sim 82^\circ$ and 72° , respectively) or can expand considerably equatorward about 40 hr after an intense solar activity (the midday and midnight parts being located at $\sim 70^\circ$ or less and $\sim 55^\circ$, respectively).

B. Auroral Zone

When the aurora is moderately active for several days, the auroral oval is approximately fixed with respect to the sun, and the Earth rotates under it once a day. The locus of the midnight portion of such an oval on a geographic map is called the *auroral zone*. This zone coincides with the belt of the maximum frequency of auroral visibility, and it passes over central Alaska, northern Canada, the southern tips of Greenland and Iceland, the northern tip of the Scandinavian Peninsula, and the Arctic Ocean coast in Siberia. The center line of this belt lies close to the geomagnetic latitude circle of 67° . The occurrence frequency of the aurora decreases away from this belt toward both higher and lower latitudes.

When maps of the auroral occurrence frequency were first constructed by E. Loomis in 1860 and H. Fritz in 1873, the frequency was expressed in terms of the number of nights of visible auroras per year. Thus, in their maps several lines of equal average annual frequency of auroral visibility (the so-called isochasms) were indicated. Some of the cities lying near the line of 10 nights per year are New York City, Seattle, Leningrad, and London. Some of the cities near the line of 1 night per year are San Francisco, Moscow, and Vienna. The line of 0.1 night per year (i.e., once in 10 years) lies over Mexico, the northern tip of Japan, southern Italy, southern Spain, and Cuba. Note that such maps were constructed on the basis of several hundred years of observations made from the subauroral zone and from lower latitudes. The occurrence frequency is generally higher during years of frequent sunspots.

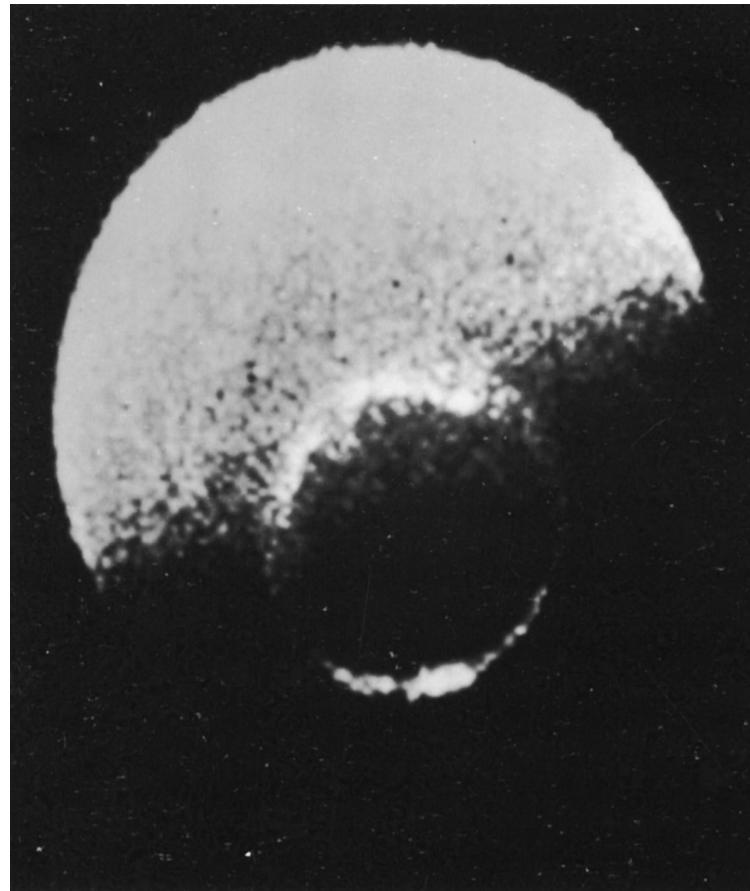


FIGURE 1 Image of the auroral oval taken from above the north pole by the *Dynamics Explorer A* satellite. (Courtesy of L. Frank, University of Iowa–Ames.)

Exceptionally high solar activity in the past caused the aurora to appear in low latitudes. For example, auroras were sighted in Honolulu on 1 September 1859, in Samoa on 13 May 1921, and in Mexico on 12 and 23 September and 11 February 1958.

II. AURORAL FORMS

A. Discrete Form

The most common auroral form has a discrete, curtain-like shape. The altitude of its bottom edge is $\sim 100\text{--}110$ km, and the altitude of its upper edge can vary greatly, from ~ 400 to >1000 km. The auroral curtain develops folds of various scales, from a few kilometers to a few hundred kilometers. [Figure 2](#) shows the altitude of the aurora in relationship to some other atmospheric phenomena. Active auroras tend to have higher upper heights and larger scale folds. When an auroral curtain develops the smallest scale folds on the order of a few kilometers in size, the curtain

appears to have vertical striations called the *ray structure*. In early literature, large-scale folds were called “horse-shoe” or “drapery” forms. The curtain-like form is traditionally called the *auroral arc*. This is because when the curtain-like form is located near the poleward horizon, it looks like an arc because of the perspective effect ([Fig. 3B](#))

If the aurora is located too far away in the northern sky, the bottom part of the auroral curtain cannot be observed and only the upper part of it can be seen above the horizon, giving the impression of dawn, as described in M. V. Lomonosov’s poem: “But, Where O Nature is the Law? From the Midnight Lands Comes Up the Dawn!”. The term *aurora* was adopted by P. Gassendi from Aurora—the rosy-fingered goddess of the dawn in Roman mythology. When an auroral curtain with the ray structure is located near the zenith, the auroral rays appear to converge from a point, and such a form is called the *corona*. This is because the parallel rays of a few hundred kilometers length appear to converge as a result of the perspective effect ([Fig. 3C](#)). Thus, the appearance of the corona form simply indicates that the curtain form is located near the zenith and is not

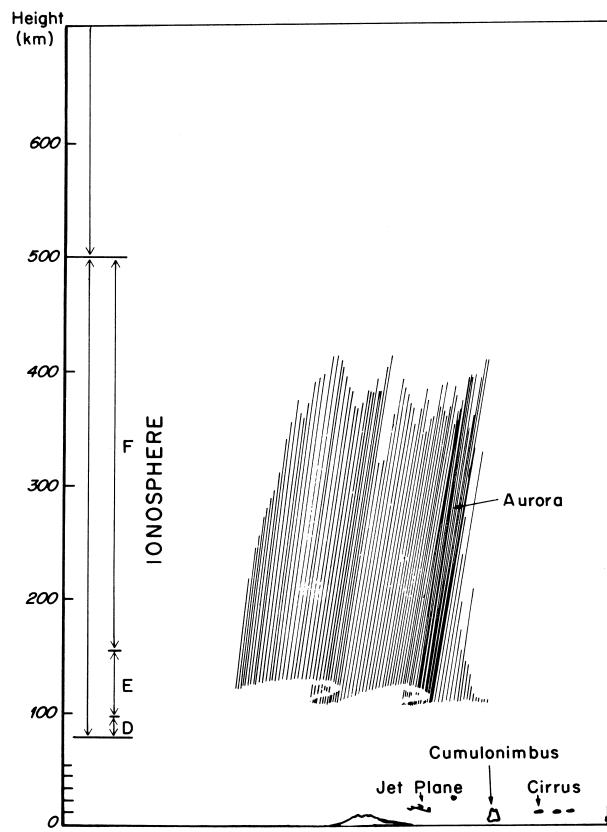


FIGURE 2 Altitude of the aurora in the atmosphere.

a different type of aurora (as many popular books on the aurora mistakenly describe it).

B. Diffuse Form

The other form of the aurora is a faint, diffuse, Milky Way-like glow that covers a part or all of the sky (as seen from the ground); it is located equatorward of the curtain-like auroras. This glow is fairly uniform in the evening sky, but tends to develop delicate east–west structures in the morning sky. Often, the diffuse glow disintegrates into patchy luminosity (looking like cumulous clouds) that drifts eastward with a speed of about 300 m/sec and tends to pulsate with a period of 5–10 sec.

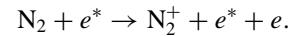
III. AURORAL SPECTRA

The aurora is associated with electromagnetic waves over a wide frequency range, including X-rays, visible, EUV, UV, and IR emissions as well as VLF and ULF radio emissions. Thus, the visible emissions constitute only a small part of the spectrum. In addition, the aurora is associated with intense VLF radio waves that can be observed only

from above the ionosphere by satellites and rockets. These VLF emissions are called *auroral kilometric radiations*.

A. Visible, EUV, UV, and IR Emissions

Until about the middle of the 19th century, it was generally believed that auroral lights arose from the reflection of sunlight by ice crystals in the air. It was A. J. Angstrom who showed that the auroral spectra consist of many lines and bands instead of only one continuous spectrum of sunlight. Some of the most familiar visible emissions from the aurora are caused by the discharge-current-carrying electrons in the following manner. When the discharge-current-carrying electrons (e^*) descend through the auroral potential structure, they are accelerated, acquiring energies of a few kiloelectron volts or greater. They collide with neutral atoms and molecules during their passage through the upper atmosphere. There they collide with nitrogen molecules (N_2) and ionize them:



The ionized nitrogen molecules (N_2^+) emit a series of band emissions, called the first negative band, in the violet–green color range. The energetic electrons (e^*) lose only a very small fraction of their energy in each collision, thus they can ionize hundreds of molecules and atoms along

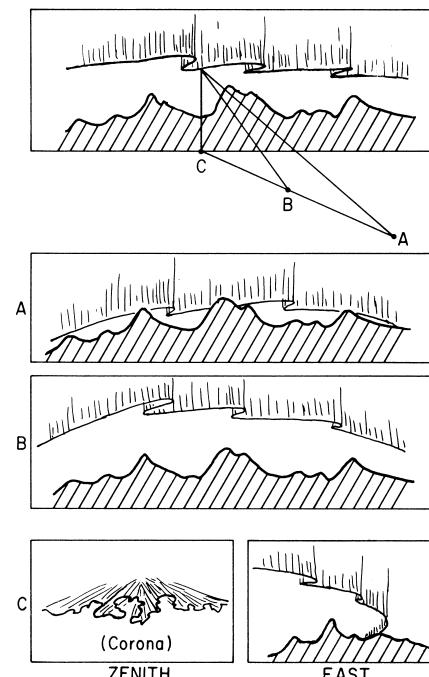


FIGURE 3 Different appearances of an auroral curtain depending on the relative location of an observer. (A) a distant (~300 km) observer, (B) a medium-distance (~200 km) observer, and (C) an observer directly under the aurora.

their pass before they completely lose their energy. For electrons having energies of a few kiloelectron volts, their stopping height is about 100–110 km. This is why the bottom of the auroral curtain is at this altitude.

Each collision produces a secondary electron (e), which has sufficient energy to excite oxygen atoms. Some of the excited oxygen atoms emit the most familiar light of the aurora, the so-called green line; its wavelength is 5577 Å (557.7 nm). The excitation potential of this state is 4 eV. There are also, however, complicated chains of photochemical processes that produce excited oxygen atoms as the energetic electrons produce a variety of ions in the polar upper atmosphere. Less energetic secondary electrons can excite oxygen atoms to another state requiring 2 eV. Oxygen atoms thus excited tend to emit the so-called red line; its wavelength is 6300 Å (630.0 nm). It takes about 200 sec for them to emit this light (actually a doublet line) after the excitation. If the excited atoms are collided by other atoms or molecules before emission, they lose the excitation energy; in such a situation they are said to be quenched. For this reason, this emission tends to take place at altitudes greater than 300 km, where the quenching collision is not frequent. As described in Section VII, the greatest auroral displays in history were characterized by a great enhancement of the red line, particularly in the middle latitudes. This enhancement is caused partly by secondary electrons and partly by intense heating of the upper atmosphere during major magnetic storms.

Excited oxygen atoms and nitrogen molecules are known to emit intense EUV and IR emissions that cannot be observed from the ground (because they are absorbed in the atmosphere). Detectors aboard both rockets and satellites have been used extensively to study such emissions.

B. X-Rays

As the penetrating energetic electrons are decelerated by atoms and molecules, they emit Bremsstrahlung X-rays, which can penetrate down to an altitude of \sim 30 km. A balloon-borne X-ray detector is needed to observe it. Satellite-borne X-ray detectors have recently succeeded in obtaining global X-ray images of the aurora from above the north polar region.

C. Radio Emissions

The aurora is accompanied by various VLF radio emissions that cannot be observed by radio devices on the ground. The discharge-current-carrying electrons are accelerated in the auroral potential structure, which is located at an altitude of \sim 10,000 km above the aurora (see Section VI). This potential structure was discovered

in 1977 and was found to be the source of very strong radio emissions called the auroral kilometric radiations (AKR). The reason for such a late discovery of the AKR is that such low-frequency electromagnetic waves are reflected by the ionosphere and cannot penetrate down to the ground. Thus, it was a satellite-borne radio device that first detected the waves. It is said that the AKR is so intense that it would be detected by “visitors from outer space” well before the Earth would become visible to them.

The aurora is also associated with various ULF waves (or Alfvén waves) having periods of 1 sec to 15 min. These waves are called geomagnetic micropulsations. When the aurora becomes suddenly active, a particular type of micropulsation, Pi2, is observed not only in the auroral latitude, but also in middle and low latitudes; it is a train of pulsations with decreasing amplitude.

IV. DYNAMICS OF THE AURORA AND AURORAL SUBSTORMS

The aurora is a very dynamic phenomenon and exhibits systematic motions on a global scale. Such auroral activity can be described in terms of the auroral substorm, which is a manifestation of the magnetospheric substorm, electromagnetic disturbances around the Earth. The morphology of auroral dynamics is described here. It should be noted that auroral motions are caused by the shifting of impact points of the discharge-current-carrying electrons, which are caused by changes in the electromagnetic field in the magnetosphere (not by motions of light-emitting atoms or molecules). The principle involved is the same as that of a cathode-ray tube.

The first sign of a typical auroral substorm is a sudden brightening of the auroral curtain in the midnight sector. The brightening spreads rapidly westward (toward the dusk sector) and eastward (toward the dawn sector); in <10 min an auroral curtain in the entire dark hemisphere brightens. This brightening is associated with a poleward advance of the curtain in the midnight sector having a speed of about a few hundred meters per second, forming a bright bulge in the auroral oval. [Figure 4](#) shows schematically the main characteristics of auroral displays, as would be seen from above the north polar region.

The formation of the bulge is associated with a large-scale wavy structure of the auroral curtain in the evening sector. This structure, the westward traveling surge, propagates westward with a speed of a few kilometers per second. In this surge, the auroral curtain develops a large-scale fold, making it the most spectacular display that can be observed from the ground. Some of the surges propagate all the way to the midday sector. [Figure 5](#) shows an example of an auroral substorm observed from the

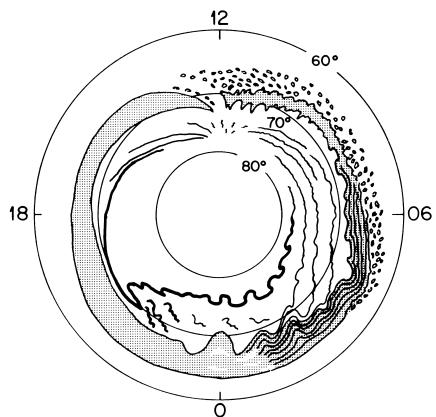


FIGURE 4 Schematic illustration of auroral activity. Solid lines indicate curtain-like discrete forms; shaded parts indicate the diffuse aurora.

Dynamics Explorer A satellite in a series of images taken 12 min apart.

During an auroral substorm, the diffuse aurora also undergoes distinct changes. In general, its brightness is considerably increased. In the evening sector, however, it

tends to maintain its uniformity in brightness; it is only during very intense substorms that it is disturbed. In the midnight sector, the poleward boundary of the diffuse aurora develops large-scale wavy structures called torches and omega (Ω) bands (see Fig. 4). The diffuse aurora also develops a number of striated structures in an east–west direction. Near the equatorward boundary, the diffuse aurora disintegrates into patchy (cumulus-cloud-like) structures that extend well into the midday sector or even into the early afternoon sector (see Fig. 4).

Eventually, the poleward advancing aurora reaches the highest latitude. This is the maximum epoch of the auroral substorm. The period between the sudden brightening of the auroral curtain in the midnight sector and the maximum epoch is called the expansive phase. A typical duration of this phase is 30–60 min. As the auroral curtain begins to recede toward lower latitudes and the bulge begins to contract, the recovery phase sets in. This phase typically lasts for about 2 hr; thus, a typical substorm lasts for about 3 hr.

Auroral motions are caused by the shift of impact points of the discharge-current-carrying electrons. In this respect, the polar upper atmosphere is analogous to the screen of

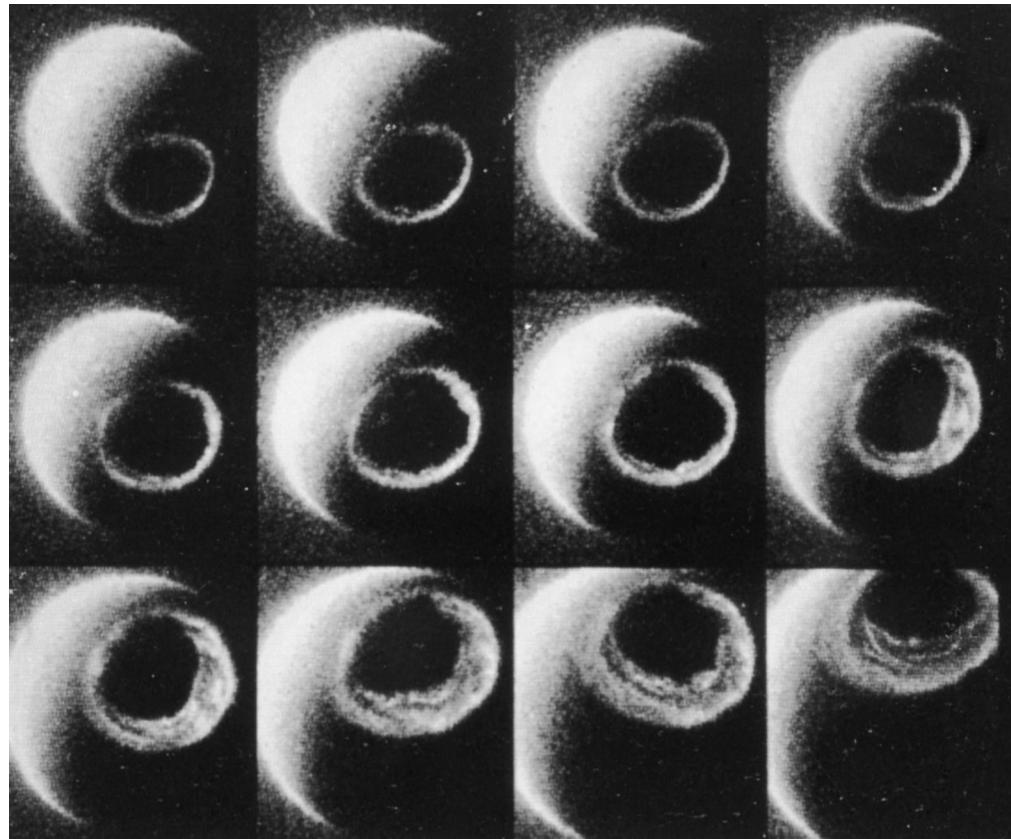


FIGURE 5 Successive images of the auroral oval (taken 12 min apart from the *Dynamics Explorer A* satellite) showing the development of auroral progress. (Courtesy of J. Craven and L. Frank, University of Iowa–Ames.)

a cathode-ray tube. The aurora is thus analogous to an image on the screen, which is produced by impact of the electron beam shot from the cathode. Thus, a study of auroral dynamics can provide important information on the locations where upward currents carried by downcoming discharge-current-carrying electrons flow down to the polar upper atmosphere (more specifically, the ionosphere) and how such locations vary during the course of an auroral substorm. That is to say, the shift of the electron beams is an indication that electric and magnetic fields around the Earth are changing and thus causing electromagnetic disturbances (magnetospheric substorms) around the Earth. As described in Section V, auroral substorms occur when the discharge power supply exceeds 10^5 MW. Therefore, auroral activity is a good (in fact, the only visible) indicator of the level of this power supply.

V. POWER SUPPLY OF THE AURORA

One of the central problems in auroral physics is to explain the processes that generate the auroral discharge currents, feed the auroral electrojets in the ionosphere, and drive the large-scale vortex motions of plasma in the magnetosphere. In other words, one must explain the generator processes that supply the power for the auroral discharge. It is not difficult to infer the total power required for the discharge, since one can estimate the total-energy dissipation rate in auroral processes. The joule heat loss in the lower ionosphere alone requires about 10^5 MW. During an intense auroral substorm, the total power required in the polar ionosphere is about 10^6 MW.

A. Solar Wind–Magnetosphere Generator

The sun continuously blows away its atmosphere with a speed of a few hundred kilometers per second. This gas flow, called the solar wind, consists mainly of protons and electrons. Deep-space observations have indicated the presence of the solar wind at a distance of about 30 astronomical units (AU), far beyond the distance of Uranus and Neptune. The solar wind tends to blow around strongly magnetized planets, such as Mercury, Earth, Jupiter, Saturn, Uranus, and Neptune. This is because their magnetic fields act as barriers. As a result, a “cavity” is formed around the magnetized planets. This cavity is the magnetosphere, and it has the shape of a long cylinder with a blunt nose pointed in the upstream direction of the solar wind (Fig. 6). In the case of the Earth’s magnetosphere, the distance from the Earth’s center to the nose is ~ 10 Earth radii. The diameter and length of the cylinder are approximately 30–40 Earth radii and 500–1000 Earth radii, respectively. The downstream portion of the magnetosphere is called the magnetotail. The situa-

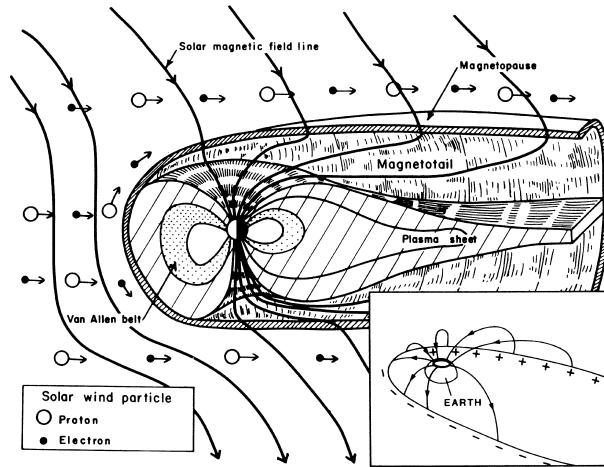


FIGURE 6 Schematic illustration of the noon–midnight meridian cross section of the magnetosphere. Solar wind particles flow around the boundary of the magnetosphere and across the geomagnetic field lines that are connected to the solar wind magnetic field lines. The small inset shows a part of the electric current circuit that is generated by the solar wind–magnetosphere dynamo.

tion is completely different for a nonmagnetized celestial body without an atmosphere (for example, the moon); in this case, the solar wind particles directly hit the surface of the body.

The solar wind is a magnetized plasma flow resulting from the solar wind carrying away some of the magnetic field of the sun. This magnetic field is of fundamental importance in the solar wind–magnetosphere interaction and in auroral physics, because some of the magnetic field lines from the polar ionosphere are connected to solar wind magnetic field lines across the boundary of the magnetosphere (Fig. 6).

The two basic elements of a generator are a magnetic field and an electrical conductor that moves rapidly in it, thus generating an electromotive force. The solar wind is a conductor, since it consists of charged particles, mainly protons and electrons. As the solar wind blows around the boundary of the magnetosphere, it has to blow across the connected magnetic field lines (Fig. 6), thus generating an electromotive force as in a generator. Actually, the basic mechanism involved in this process is the same as that of a magnetohydrodynamic (MHD) generator.

Therefore, the entire boundary surface of the magnetosphere acts as the generator. As a result, the dawn side of the boundary in the equatorial plane becomes the positive “terminal” and the dusk side becomes the negative “terminal.” The total voltage between the terminals is about 100 kV, and it is this potential drop that drives the two vortex motions of plasmas in the magnetosphere and ionosphere.

As a first approximation, the power P (in megawatts) of the solar wind–magnetosphere generator is given by

$$p = 20VB^2 \sin^4(\Theta/2),$$

where V (km/sec) is the solar wind velocity, B (nT) is the solar wind magnetic field magnitude, and the angle Θ is approximately the polar angle of the solar wind magnetic field ($\Theta = 0^\circ$ for a northward-directed field and 180° for a southward-directed field).

Much of the current generated flows across the magnetotail along the midplane separating the northern and southern halves of it. However, a small portion of the current ($\sim 2 \times 10^6$ Å) is discharged through the ionosphere (see the small inset in Fig. 6). It is this portion that is connected to the inward field-aligned currents in the morning part of the oval and the outward field-aligned currents in the evening part of the oval.

The auroral phenomena are not unique to Earth. The aurora is found on magnetized planets such as Jupiter, Uranus, and Neptune. On the other hand, nonmagnetized planets such as Venus and Mars have no aurora.

B. Auroral Discharge Currents

Many popular books on the aurora explain that the aurora is caused by a direct impact of solar wind particles on the polar region that are deflected by the Earth's magnetic field. Although such a direct entry of solar wind particles does occur in the magnetosphere, their "impact" region is limited to only a small portion of the midday part of the auroral oval called the cusp. This mechanism does not, however, explain why the aurora appears along an annular belt around the pole in the form of the auroral oval. Furthermore, the energy of solar wind protons and electrons is not high enough to penetrate to the lower ionosphere and produce the auroral luminosity.

As mentioned earlier in this section, some of the geomagnetic field lines from the polar ionosphere are connected to the solar wind magnetic field lines. These geomagnetic field lines are called open field lines; other field lines cross the equatorial plane and reach the opposite hemisphere (closed field lines). The open field lines originate from the area bounded by the auroral oval. Consider the bundle of these open field lines and those which constitute the surface of this bundle in the morning and evening sectors. The morning sector field lines are connected to the positive terminal of the solar wind-magnetospheric generator and those evening sector field lines are connected to the negative terminal (see the inset in Fig. 6). It is for these reasons that the primary discharge (field-aligned) current flows from the morning side of the magnetotail to the poleward boundary of the morning half of the oval. A part of the current then flows across the polar cap and finally flows back to the evening side of the magnetotail from the poleward boundary of the evening half of the oval (see Fig. 8). The actual discharge current circuit is much more complicated, as discussed in Section VI.

VI. AURORAL DISCHARGE CURRENT SYSTEM

A. Auroral Discharge Circuit

The "anchoring points" of field lines which connect the positive and negative terminals of the solar wind-magnetosphere generator form an approximate circle in the polar region (see the insert in Fig. 6). The morning half of the circle is positively charged, while the evening half is negatively charged. The electric potential associated with these charges is schematically shown in Fig. 7; the associated electric field is perpendicular to the equipotential lines. The potential difference between the two half-circles is about 100 kV, which is the generator's voltage.

In a tenuous plasma in the magnetosphere, where collision among charged particles is rare, the charged particles can move in two ways, in addition to circular motions around the magnetic field lines. The first is a motion along the magnetic field lines. Thus, the magnetic field lines which are connected to the terminals act like "invisible wire," carrying the electric current generated by the generator from the positive terminal to the negative terminal through the ionosphere, which is the resistor.

This primary circuit generates a complicated current system (Fig. 8). The incoming (downward) primary current from the positive terminal to the ionosphere is separated into two parts, one flowing across the polar cap and the other flowing equatorward across the auroral oval and upward at the equatorward boundary (the secondary current). The outgoing (upward) primary current is the combination of two parts, one coming from the polar cap and the other coming from the equatorward boundary of the oval (the secondary current). The secondary currents are closed in the equatorial plane.

The upward field-aligned portions of the current system, both the primary incoming (downward) portion from the

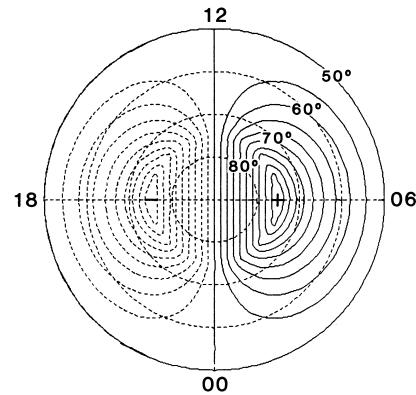


FIGURE 7 Distribution of the electric potential in the polar ionosphere in geomagnetic latitude and magnetic local time coordinates.

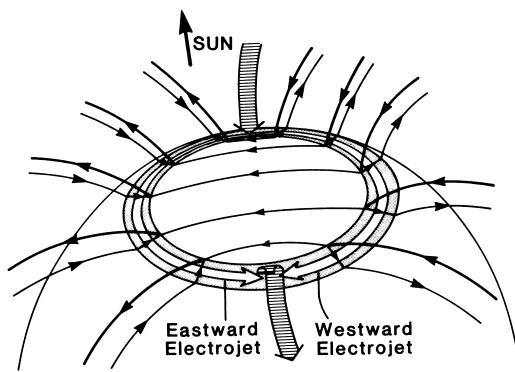


FIGURE 8 Connection of the ionospheric currents to the currents that flow along the geomagnetic field lines (field-aligned currents).

positive terminal in the morning sector and the secondary (downward) portion, also form a vacuum discharge tube ([Fig. 9](#)).

B. Auroral Potential Structure

There is one more factor to consider in an explanation of the generation of the aurora. To close the field-aligned currents, the discharge-current-carrying electrons must be able to reach the E region of the ionosphere where the electric conductivity perpendicular to the magnetic field is high. However, since the electrons lose energy during each collision, they must have at least a few kiloelectron volts to be able to reach the lower ionosphere. Most solar wind and magnetospheric electrons do not have such high energies (only a few hundred electron volts). Although it is not well understood, an interesting electric potential

structure (with a potential drop of a few kilovolts) appears to develop at an altitude of 10,000 to 20,000 km, where the solar wind–magnetosphere generation power exceeds 10^5 MW; it is speculated by some researchers that this potential structure is a sort of double layer. The electric field associated with this potential structure is directed upward along the geomagnetic field lines, so that the discharge-current-carrying electrons become accelerated downward as they go through the structure, enabling them to reach the lower ionosphere. They ionize and/or excite atoms and molecules that emit visible light, thus creating the aurora.

C. Ionospheric Currents

As mentioned earlier, charged particles in a collisionless environment are constrained to move along magnetic field lines. However, when there is an electric field, both positive ions and electrons tend to move together in the direction perpendicular to both the electric field \mathbf{E} and the magnetic field \mathbf{B} with the velocity $\mathbf{V} = \mathbf{E} \times \mathbf{B}/B^2$. Since the electric field \mathbf{E} is perpendicular to the equipotential lines, both ions and electrons drift along the equipotential lines. However, such a condition is possible where both positive ions and electrons are collision free. This condition holds down to the F region of the ionosphere (>200 km in altitude). However, in the E region of the ionosphere, electrons are collision free, but positive ions collide frequently with neutral particles and thus cannot participate in the $\mathbf{E} \times \mathbf{B}$ drift motion. Therefore, only electrons drift with the velocity $\mathbf{V} = \mathbf{E} \times \mathbf{B}/B^2$ along the equipotential contour lines, and thus, the current flows along the equipotential contour lines (opposite to the direction the drift motion of electrons and perpendicular to \mathbf{E}). This component of the current is called the Hall current. Actually, the positive ions can move in the direction of \mathbf{E} , and this component is called the Pedersen current.

As a result of all these components, the currents in the ionosphere have a complicated distribution. The most intense current flows westward in the night sector and causes intense magnetic disturbances on the ground ([Fig. 10](#)).

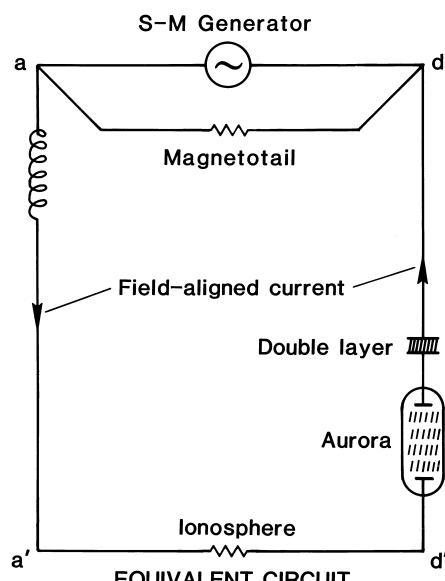


FIGURE 9

VII. SOLAR ACTIVITIES AND THE AURORA

It is generally known that auroral activity is closely related to solar activities; specifically, sunspots are believed to be the basic cause of the aurora. However, we are far from a good understanding of this “well-known” fact.

Auroral activity depends on the power of the solar wind–magnetosphere generator, which in turn depends on the solar wind velocity (V), the magnitude of the solar wind magnetic field (B), and the polar angle (Θ) of the solar wind magnetic field. Thus, an important problem is

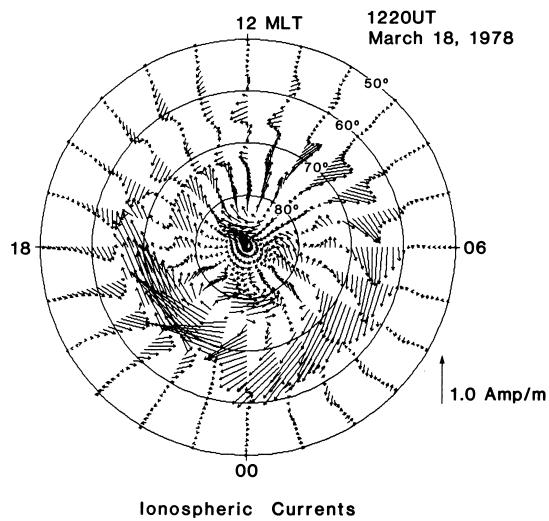


FIGURE 10 Distribution of electric currents in the ionosphere during an auroral substorm in geomagnetic latitude and magnetic local time (MLT) coordinates.

to explain how solar activities are related to these three quantities. Using the solar wind velocity as an example, we can see that there is no increase in V when the sunspot number is high. In fact, sunspots tend to suppress a high-speed flow of the solar wind, and some of the highest solar wind actually blows out from spotless areas called coronal holes. Sunspots are, however, closely related (at least statistically) to auroral activity. This is partly because large sunspot groups tend to produce explosive phenomena, solar flares, which produce a gutsy flow.

A. Solar Flares and Auroral Activity

As a gusty flow penetrates into the slow solar wind, it generates a shock wave in which the solar wind magnetic field is compressed and B generally becomes larger. Figure 11 shows a hypothetical situation in which a large sunspot group produced two flares that occurred 48 hr apart; for two large-scale spiral structures, see Section VII.B. The figure shows a “snap shot” of the corresponding two shock waves 5 days after the first flare. When the shock wave (in which V and B tend to be large) collides with the solar wind-magnetospheric generator, there is a high probability that the power will increase (according to the formula given in Section V.A). Some flares, however, produce a large Θ ($\sim 180^\circ$), while others produce a small Θ ($\sim 0^\circ$). When all three quantities V , B , and Θ become large, the generator power will increase considerably, by as much as 100. Major variations in these quantities last for about 6 hr; thus, if the peak times of V , B , and Θ approximately coincide, there will be a major auroral display. The auroral oval will then expand and descend to the U.S.–Canadian border or even lower. The upper part of the auroral cur-

tain will extend high up to >1000 km and will be rich in the oxygen red line. The magnetic fields produced by the discharge current will produce intense magnetic fields and thus a major magnetic storm. This is the reason why a great auroral display is always accompanied by a great magnetic storm and why both are sometimes associated with an intense solar flare in a large sunspot group.

Such an occurrence, however, is actually rare. Very often, the three peaks do not coincide. Furthermore, occasionally when V and B increase, Θ may become very small, so that the resulting power is low. It is for these reasons that a great solar flare does not necessarily produce a great auroral display and a great magnetic storm. The situation is also complicated by solar wind disturbances being most intense along the line connecting the center of the sun and the flare location. Thus, if a flare takes place near the center of the solar disk, the Earth will be very close to an extension of this line along which the most intense disturbance propagates. If a flare takes place near the limb of the solar disk, the most intense disturbance propagates along the line 90° longitude away from it. The resulting increase in the dynamo power is small.

The high correlation between the large number of sunspots and great auroral displays (as well as great magnetic storms) arises partly from the fact that great auroral displays in populated areas (the middle latitude belt) occur after intense solar flares in unusually large sunspot groups.

B. Coronal Holes and Auroral Activity

Sunspots tend to suppress a high-speed solar wind, and some of the highest speed winds flow out from the

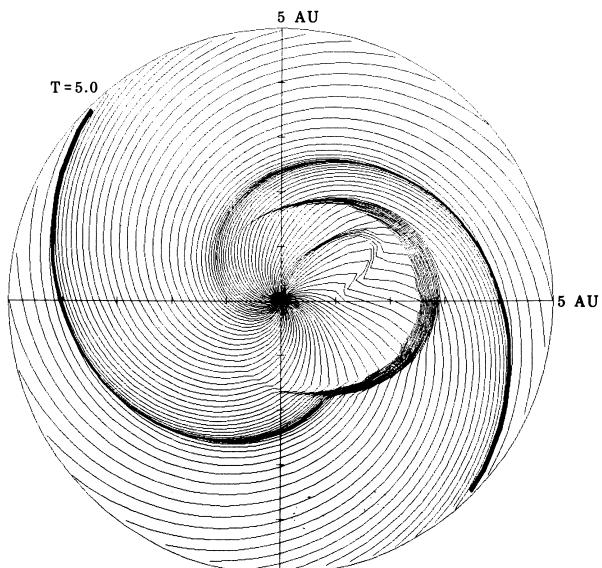


FIGURE 11 Schematic illustration of two interplanetary shock waves that are propagating in interplanetary space.

coronal holes. It is believed that the magnetic field lines from coronal holes emanate into interplanetary space and are carried away by the solar wind in a spiral form (Fig. 11). In the vicinity of sunspots, magnetic field lines tend to loop back to the solar surface, suppressing the solar wind. Furthermore, many coronal holes are long lived, some more than 1 year. Therefore, a high-speed stream from a coronal hole can last more than 1 year. Because the sun rotates once every 27 days, the solar wind–magnetosphere generator encounters the high-speed stream every 27 days, similar to a beacon. The stream is wide, and so it takes about 1 week for the stream to go by at the distance of the Earth. The variability of the angle Θ is generally high in the stream. For this reason, high auroral activity tends to occur every 27 days.

The magnetic field magnitude in a high-speed stream is generally not very large, and the resulting power is usually $<10^6$ MW. Thus, the resulting auroral activity is not as intense as some of the flare-induced ones and is confined mostly to latitudes north of 65° . On the other hand, since the width of the stream is wide, moderate auroral activity can continue for about 1 week for each encounter. During the declining epoch of the sunspot cycle, coronal holes in both the northern and southern polar regions (the polar coronal holes) tend to extend to lower latitudes in limited longitude sectors (separately by $\sim 180^\circ$). Thus, two high-speed streams (one from the northern hole and the other from the southern hole) emanate from the sun for an extended period. For these reasons, at latitudes above 65° the occurrence frequency of the aurora may peak during the declining period of the sunspot cycle at ~ 2 to 3 years after the maximum year. In auroral latitudes, the sunspot number is thus not the most reliable indicator in predicting auroral activity.

VIII. AURORAL EFFECTS

As human activity advances northward, its supporting systems, such as power transmission lines and oil and gas pipelines, extend across the latitude of the auroral oval. Auroral activity has significant effects on such manmade systems. Since the aurora is an electrical discharge process, it tends to induce a potential drop of about 1 V/km on the ground under the auroral oval. If two points separated by 1000 km are connected by a conductor with a total resistance of 10Ω , there will be an electric current of $\sim 100 \text{ A}$. This situation is closely approximated by the trans-Alaskan oil pipeline. Power transmission lines are affected in more complicated ways. Since they are set up for alternative currents, any direct currents (or voltages) cause transformers to generate higher harmonic components (other than 60 Hz), which in turn generate heat in

the core. There are elaborate relay systems to avoid such interference, but a resulting blackout would be very costly.

It has long been known that auroral activity causes serious interference to radio systems, such as high-frequency (HF) communication systems and HF radars systems. Deflection of a compass during auroral activity is, however, often exaggerated in popular books on the aurora. It is rare that a compass needle can be deflected more than a few degrees by the aurora. However, there are a number of operational systems in arctic regions that require accuracy of $1'$ or less in determining the direction by a compass, and such operations suffer from auroral activity. Intense auroral activity is associated with heating of the upper atmosphere, which causes an expansion of the atmosphere. The resulting increase in atmospheric density has been a serious problem for satellites, since it increases the drag, decelerating them and shortening their lives. Intense field-aligned currents are believed to be a cause of some malfunctions (phantom commands) of satellites.

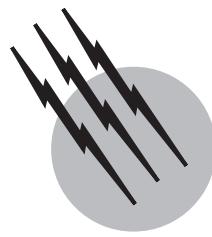
In this respect, accurate forecasting of auroral activity is an important subject. Since auroral activity depends on the solar wind–magnetosphere generator power, auroral forecasting involves prediction of V , B , and Θ or observation of these quantities at an upstream point. There has been significant progress in this endeavor during the past several years.

SEE ALSO THE FOLLOWING ARTICLES

GEOMAGNETISM • IONOSPHERE • SOLAR PHYSICS • SOLAR SYSTEM, MAGNETIC AND ELECTRIC FIELDS

BIBLIOGRAPHY

- Akasofu, S.-I. (1977). "Physics of Magnetospheric Substorms," Reidel, Dordrecht, Netherlands.
- Akasofu, S.-I. (1979). "Aurora Borealis, the Amazing Northern Lights," Alaska Geographic Society, Anchorage.
- Akasofu, S.-I., and Chapman, S. (1972). "Solar-Terrestrial Physics," Oxford Univ. Press, London and New York.
- Akasofu, S.-I., and Kan, J. R., eds. (1981). "Physics of Auroral Arc Formation," Geophys. Monogr. No. 25, Am. Geophys. Union, Washington, DC.
- Alfvén, H. (1950). "Cosmical Electrodynamics," Oxford Univ. Press, London and New York.
- Brekke, A., and Egeland, A. (1983). "The Northern Light: From Mythology to Space Research," Springer-Verlag, Berlin and New York.
- Chamberlain, J. W. (1961). "Physics of the Aurora and Airglow," Academic Press, New York.
- Eather, R. H. (1980). "Majestic Lights," Am. Geophys. Union, Washington, DC.
- Jones, A. V. (1974). "Aurora," Reidel, Dordrecht, Netherlands.
- Potemra, T. A., ed. (1984). "Magnetospheric Currents," Geophys. Monogr. No. 28, Am. Geophys. Union, Washington, DC.
- Stormer, C. (1955). "The Polar Aurora," Oxford Univ. Press, London and New York.



Carbon Cycle

Inez Fung

University of California, Berkeley

- I. Background
- II. The Atmospheric Carbon Cycle
- III. The Terrestrial Carbon Cycle
- IV. The Oceanic Carbon Cycle
- V. The Geologic Carbon Cycle
- VI. Unraveling the Contemporary Carbon Budget
- VII. Mechanisms for the Contemporary Carbon Sink
- VIII. Interannual Variability of the Contemporary Carbon Cycle
- IX. Climate Interactions
- X. Outlook

GLOSSARY

CO₂ fertilization The stimulation of net primary productivity by increased ambient CO₂ concentration.

Lithosphere The outer layer of the solid earth, consisting of the crust and upper mantle.

Net primary productivity The rate of total energy acquisition by green plants during photosynthesis, minus the rate of energy loss through plant respiration.

Revelle factor A dimensionless number that expresses the relative sensitivity of the partial pressure of CO₂ and total dissolved inorganic carbon in a solution to the incremental additional or removal of CO₂, bicarbonate or carbonate ions.

Suess effect Decrease in the ratio of ¹³C:¹²C or ¹⁴C:¹²C in a sample of seawater or plant material as a result of the addition of fossil fuel carbon to the atmosphere.

Total alkalinity The amount of acid required to neutralize

all the weak bases in the solution (mainly bicarbonate and carbonate ions).

Volume mixing ratio The dimensionless ratio of the number of molecules of a specific gas to the total number of molecules of all gases in the volume.

CARBON CYCLE refers to the continual transformation of carbon from one form to another, and its ceaseless transport from one place to another. Schematically, the transformation can be represented as conversions between inorganic and organic forms of carbon; and the transport as redistribution within and exchanges among different reservoirs: the atmosphere, vegetation and soils, oceans, and the lithosphere. This is presented in Fig. 1. CO₂ is chemically inert in the atmosphere, and so its abundance is determined principally by its carbon exchanges with the land, ocean and lithosphere.

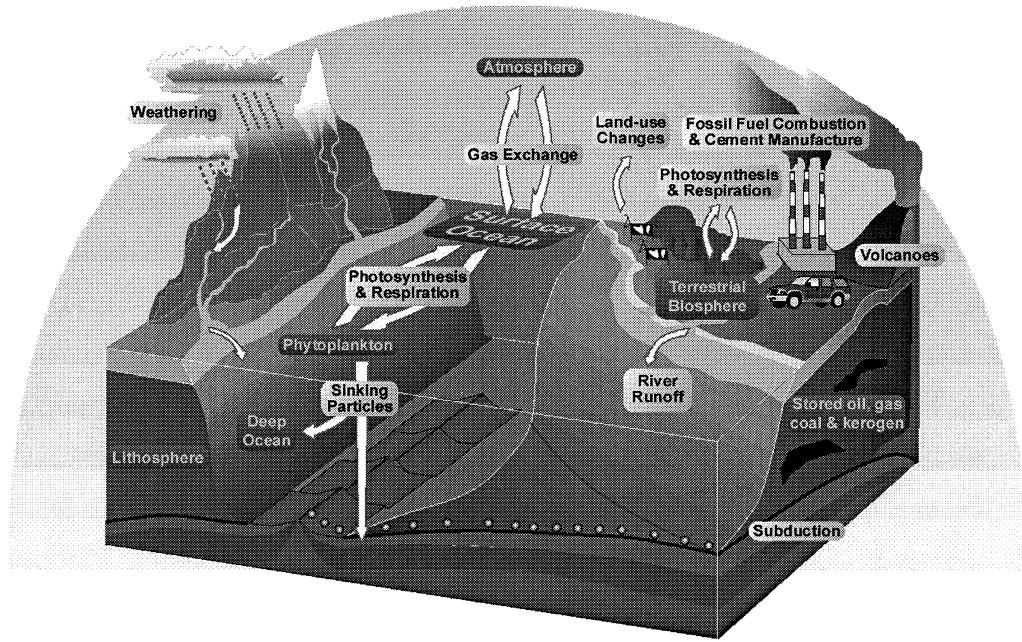


FIGURE 1 Schematic diagram of the global carbon cycle, showing carbon exchanges among the different reservoirs.

I. BACKGROUND

Table I lists the sizes of the carbon reservoirs and the fluxes among them. On Earth, there are over 75 million PgC ($1 \text{ PgC} = 10^{15} \text{ g}$). Most of the carbon is in the lithosphere,

with the remainder distributed in approximate ratios of 1:3:50 in the atmosphere, vegetation and soils, and oceans.

Annually approximately 100 Pg of C (as CO_2) is absorbed by vegetation as photosynthesis converts CO_2 into organic carbon. About half the amount of carbon

TABLE I

Carbon reservoir	Inventory (PgC)	Annual fluxes into and out of reservoir (PgC/year)	Turnover time (years)
Atmosphere	720		
Land biosphere	2,000	± 50	40
Live biomass	800		15
Dead biomass	1,200		25
Ocean	38,400	± 90	700
Total inorganic C	37,400		
Surface layer	670		
Deep ocean	36,730		
Total organic C	1,000		
Aquatic Biosphere	1–2	± 50	<0.04
Lithosphere	>75,000,000	$\pm 0.1\text{--}0.2$	>400 million
Sedimentary carbonates	>60,000,000		
Kerogens	15,000,000		
Fossil fuels	4,130		–5–10
Coal	3,510		
Oil	230		
Gas	140		
Peat	250		

is returned to the atmosphere immediately as CO₂ via plant respiration, so that net primary productivity (NPP) is about 50 PgC/year. Annually, an amount of carbon equal to the NPP is returned to the atmosphere via microbial respiration during decomposition of dead organic matter. The atmosphere and surface oceans exchange approximately 90 PgC/year, with the exchange driven by the difference in CO₂ partial pressure across the air-sea interface. The marine biological cycle starts with photosynthesis in the surface ocean, where inorganic carbon dissolved in seawater is converted into organic matter and shells of phytoplankton. The phytoplankton are consumed and their carbon and nutrients recycled to support further production. The biological detritus sinks and is remineralized and returned to dissolved inorganic forms which eventually are upwelled to the surface by the ocean circulation.

The long-term carbon cycle has two subcycles. The first involves chemical weathering of calcium and silicate rocks, and the eventual outgassing of CO₂ to the atmosphere in volcanic eruptions and hot springs. The second involves residual undecomposed organic matter. Their burial and transformation at high pressures and temperature ultimately forms coal, oil, natural gas, and dispersed organic deposits known as kerogen.

A measure of the rate of cycling in the reservoirs is the turnover time: the average time a carbon atom spends in the reservoir before exiting. Mathematically, it is estimated by dividing the reservoir size by the total exit fluxes. Turnover time of carbon is ~15 years in vegetation and ~20 years in soils. There is at least a factor of ten variation in these turnover times between the equator and pole, as both photosynthetic and decomposition rates are climate sensitive. Turnover time of carbon is ~1 year in the surface oceans (because of the buffer factor, to be discussed further below) and ~10³ years for the global ocean. However, the turnover time of carbon in the marine biota is of order days only. Turnover time of carbon in the lithosphere is >400 million years, and hence the geologic reservoir is referred to as the long-term reservoir.

Insight into the controls of atmospheric CO₂ levels is provided by the sizes and turnover times of the different carbon reservoirs. Lithospheric processes control atmospheric CO₂ variations on time scales of millions of years. With its large carbon inventory and air-sea fluxes, the oceanic processes dominate atmospheric CO₂ variations on millennial time scales. On seasonal to interannual time scales, the terrestrial biosphere must play the leading role.

II. THE ATMOSPHERIC CARBON CYCLE

Chemical measurements of atmosphere CO₂ were made in the nineteenth century at a few locations in Europe. Modern high precision record of CO₂ in the atmosphere

did not begin until 1958, the International Geophysical Year (IGY), when C. D. Keeling of Scripps Institution of Oceanography pioneered measurements of CO₂ using an infrared gas analyzer at Mauna Loa Observatory in Hawaii and at the South Pole. Since 1974, continuous measurements of background atmospheric CO₂ have been made by the Climate Monitoring and Diagnostics Laboratory (CMDL) of the National Oceanic and Atmospheric Administration (NOAA) of the U.S. Department of Commerce at four stations (Pt. Barrow, Alaska; Mauna Loa, Hawaii; American Samoa; and the South Pole). In addition to the continuous monitoring stations, NOAA/CMDL also operates a cooperative sampling network. Flask samples of air are collected weekly or biweekly from these sites, and are shipped to the CMDL facility in Boulder, Colorado, for analysis. The sampling network began before 1970 at a few initial sites, and by 2000 AD consists of over 70 sites worldwide. Besides the U.S. program, surface measurements of atmospheric CO₂ are made by many countries including Australia, Canada, France, Hungary, Italy, Japan, New Zealand, Spain, Germany, and Switzerland.

Atmospheric concentration of CO₂ was 315 parts per million by volume (ppmv) in 1958, and has increased to 368 ppmv in 1999. With a total of 1.7×10^{20} molecules in the atmosphere, a CO₂ volume mixing ratio of 300 ppmv translates to 5.2×10^{16} molecules of CO₂ or 620 PgC. The latitudinal and temporal variations in atmospheric CO₂ for 1990–1999 are shown in Fig. 2. The CO₂ abundance in the atmosphere increases steadily with an average of 0.5–1%/year. Also prominent in the measurements is a repetitive seasonal cycle in CO₂ in the northern hemisphere, which peaks in May–June and reaches a minimum in September–October. The peak-trough amplitude

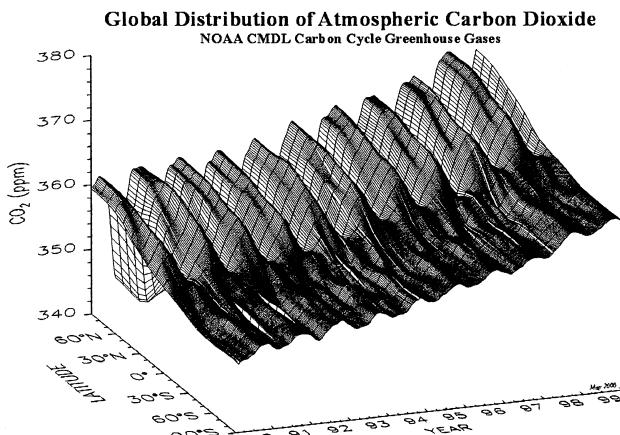


FIGURE 2 Latitudinal and temporal variations in atmospheric CO₂ from 1990 to 1999. The data are from the NOAA/CMDL cooperative air sampling network, consisting of stations in the remote marine locations. (Data and figure were from <http://www.cmdl.noaa.gov>).

decreases from \sim 18 ppmv at Pt. Barrow, Alaska, to \sim 1 ppmv at the equator. The southern hemisphere CO₂ seasonal cycle has a small amplitude and is six months out of phase with that in the northern hemisphere. The annually averaged concentrations are higher, by 1–3 ppmv in the 1990s, in the northern than in the southern hemisphere. The direct measurements document unambiguously the ongoing rise of CO₂ in the atmosphere.

Air bubbles in glaciers contain samples of ancient air. Analysis of gases occluded in air bubbles in polar ice has provided a unique reconstruction of atmospheric CO₂ history prior to the modern high-precision instrumental record. The first long (160,000 years) history of atmospheric CO₂ was obtained from the 2202-m deep ice core drilled in 1984 at Vostok Station in East Antarctica by Russia, France, and the U.S. (Barnola *et al.*, 1987). Since then, ice cores of different lengths have been obtained from Greenland and other locations on Antarctica. The longest (420,000 years) CO₂ record comes from the 3623-m ice core at Vostok Station (Petit *et al.*, 1999). Over the past four 420,000 years, CO₂ varied between \sim 180 ppmv during glacial periods to \sim 280 ppmv during interglacials (Fig. 3), with the CO₂ rise more rapid during deglaciation (100 ppmv in \sim 20,000 years) than the decrease during glaciation (100 ppmv in \sim 80,000 years). The distinct natural cycle in atmospheric CO₂ is linked to natural climate cycles: the climate-sensitive rates at which the oceans and terrestrial biosphere absorb, release, and store carbon determine the CO₂ variations in the atmosphere.

CO₂ variations in the past 200 years are unique when compared to variations in the past 420,000 years. Contemporary CO₂ has increased at a rate (90 ppmv in 200 years) that is faster than any time in the historical record, and has reached a level that is higher than has been observed in the last 420,000 years.

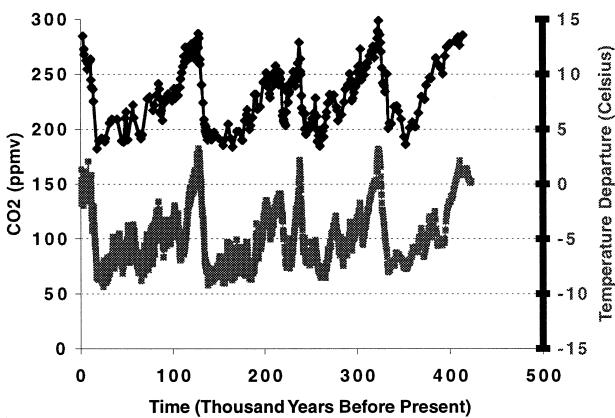
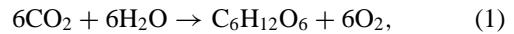


FIGURE 3 Variations of CO₂ and temperature anomalies in the past 420,000 years, as determined by analysis of the Vostok ice core. [Data are from Petit *et al.* (1999). *Nature* **399**, 429–436].

III. THE TERRESTRIAL CARBON CYCLE

Atmospheric CO₂ diffuses through leaf stomata and is assimilated, in the presence of sunlight, phosphate, nitrate, and other nutrients, into organic compounds via photosynthesis. The photosynthetic reaction can be summarized as follows:



where C₆H₁₂O₆ is glucose, and is a shorthand representation of organic compounds. The photosynthate is allocated into leaves, roots, and woody parts of plants. In order to maintain and synthesize their tissue, green plants respire, i.e., break down and release the by-products of part of the organic matter they create. NPP is the rate of production of organic matter, after autotrophic respiration has been accounted for. When plants die or when leaves are shed, carbon is transferred to the dead organic carbon pool where it is decomposed by microbes. CO₂ is released by microbial respiration. The decomposition rate and hence CO₂ flux to the atmosphere vary with the composition of soil organic matter, and with temperature and moisture and other conditions of the soil. At steady state over a large area, the long-term averaged photosynthetic uptake of CO₂ balances the autotrophic (plant) and heterotrophic (microbial) respiratory release, with no net change in atmospheric CO₂ or biospheric carbon inventory.

The inventory in the terrestrial biosphere for the present day is summarized in Table II. Because of the heterogeneity of the landscape and the sparsity of measurements, the values in Table II are necessarily estimates. In general, above-ground carbon density is greatest in tropical rain forest where the temperature and abundant precipitation favor NPP. By contrast, soil carbon is highest in the tundra where decomposition rates are slow.

The annual NPPs for the different vegetation types given in Table II are extrapolated from direct measurements. Global observations of an index of terrestrial photosynthesis have been made using spectral reflectances of the land surface measured by space-borne instruments. Green leaves absorb solar irradiance in the visible wavelengths and reflect that in the near infrared. An index of terrestrial NPP that exploits this spectral signature is the Normalized Difference Vegetation Index (NDVI), defined as the difference divided by the sum of the reflectances in these spectral regions. Since the early 1980s the NDVI has been estimated from reflectances measured by the Advanced Very High Resolution Radiometer (AVHRR) on board the NOAA series of polar-orbiting weather satellites. The AVHRR NDVI time series will be merged with NDVIs derived from measurements by successive generations of satellite instruments, such as Sea-viewing Wide

TABLE II Typical Carbon Densities and Areal Extents for 14 Vegetation Types for the Present Day

Vegetation type	Area (10^{12} m^2)	Typical Biomass (kg C m^{-2})	Soil carbon (kg C m^{-2})	Total inventory (PgC)	NPP (gC m^{-2} year^{-1})	Turnover time (year)
Evergreen tropical forest	20	15	10	475	950	25
Drought—deciduous woodland	6	5	7	72	800	15
Savanna	5	3	6	45	490	20
Arid grassland and shrubland	30	1	6	210	150	50
Desert	14	9	3	168	30	400
Mesic grassland	2	3	11	28	330	40
Mediterranean forest and woodland	2	3	8	22	500	20
Temperate evergreen seasonal broad leaved forest	2	8	8	32	650	25
Drought deciduous and drought seasonal broadleaved forest	9	6	7	117	650	20
Cold-deciduous broad leaved forest and woodland	12	10	10	240	550	40
Cold-deciduous needle-leaved forest and woodland	16	9	14	368	350	70
Evergreen needle-leaved forest and woodland	6	6	12	108	250	70
Tundra	7	1	17	126	100	180
Polar desert and ice	3	0	0	0	0	0
Total	$132 \text{ } 10^{12} \text{ m}^2$	800 PgC	1200 PgC	2000 PgC	50 PgC/year	40 year

Field-of-view Sensor (SeaWiFS) and Moderate-resolution Imaging Spectroradiometer (MODIS). The geographic and seasonal variations in NPP are evident in the satellite observations (Fig. 4). There is a progressive “green wave” advancing poleward in the spring–summer hemisphere, and retreating equatorward in the autumn–winter hemisphere. The tropical biosphere remains photosynthetically active throughout the year, with modulations associated with rainfall seasonality. The satellite observations, with uniform global and repeated temporal coverage, thus document the variability of terrestrial photosynthesis on seasonal to interannual and interdecadal time scales.

At steady state, the annually averaged NPP of an ecosystem equals the annually averaged litterfall and mortality, and equals the annually averaged decomposition. Within the year, however, the timing of growth, death, and decay are not synchronous, so that there is on average a net flux of CO₂ from the atmosphere to the biosphere during the growing season, and a compensating flux from the biosphere the rest of the year. The asynchrony gives rise to the very distinctive seasonal cycle in atmospheric CO₂ (cf. Fig. 2): the seasonal cycle is greatest in the middle to higher latitudes in the northern hemisphere where land–sea contrast gives rise to a very strong summer–winter contrast in climate and biospheric functioning.

Climate variations and changes in atmospheric composition itself alter the rates of photosynthesis and decomposition, and may lead to net sequestration or release of carbon in the land. During the Last Glacial Maximum, ~20,000 years ago, for example, vegetation shifts associated with the colder and drier climate resulted in ~25% less carbon stored on land compared to the present day. Other factors alter the balance between the uptake and release of CO₂, the most important one being disturbance. Fires cause an immediate release of terrestrial carbon to the atmosphere. Regrowth of vegetation would transfer carbon from the atmosphere to the land. Disturbance and recovery, like growth and decay, are part of the natural cycle. When integrated over very long periods and over large areas, a steady state may obtain.

Humans modify the landscape, such as by deforestation, forest management, reforestation, and agriculture. These modifications alter the age class distribution of the vegetation, turnover times of carbon in vegetation and soils, and in turn the net carbon balance. With deforestation, the reduction in (or elimination of) photosynthesis, enhancement of decomposition due to the additional detritus, and the accelerated oxidation of soil carbon all lead to a net flux of carbon to the atmosphere. Subsequent agriculture and/or regrowth of vegetation may lead to a slow

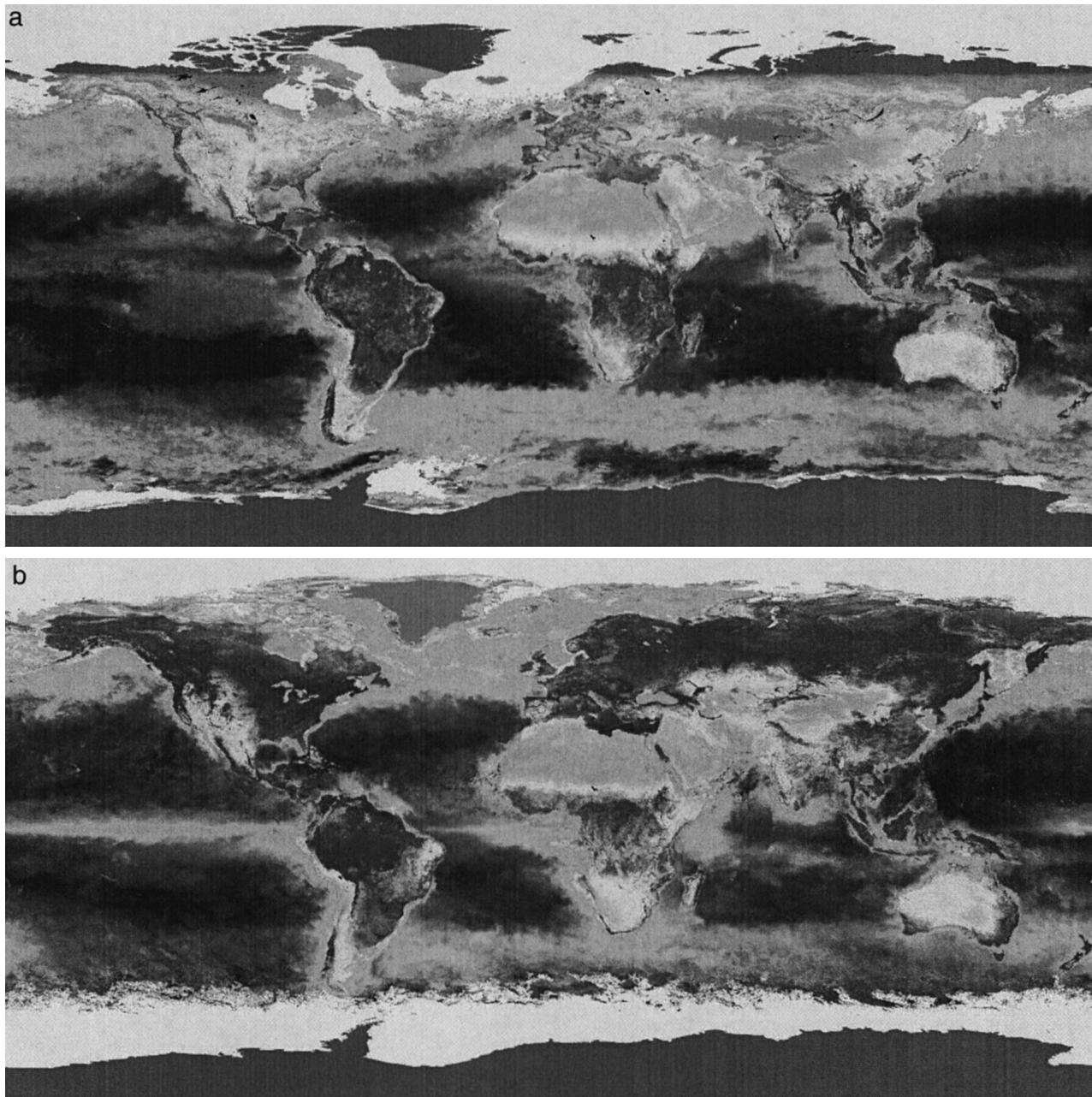


FIGURE 4 Seasonal variations in the terrestrial and marine biospheres as observed by SeaWiFS satellite for (a) December 1997–February 1998, and (b) June 1998–August 1998. (Data are from <http://seawifs.gsfc.nasa.gov/SEAWIFS>.)

accumulation of carbon on land. Thus, the net CO₂ flux between the biosphere and the atmosphere change in both magnitude and direction with time, and depends strongly on the history of land cover modification and land use practice.

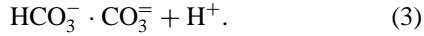
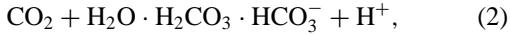
Natural and human disturbances are either episodic in time, or highly heterogeneous in space; direct estimates of the fluxes and changes in carbon inventories are highly un-

certain. There was a net loss of carbon from the land to the atmosphere around the turn of the twentieth century, when mechanization facilitated the expansion of agriculture and the exploitation of new lands. By the end of the twentieth century, some of these areas have shifted from being a carbon source to a carbon sink due to the subsequent abandonment of agriculture and regrowth of forests in Europe and North America. Near the end of the twentieth

century, deforestation has shifted from middle latitudes in the northern hemisphere to the tropics, with a net loss of 1–2 PgC/year from the tropical biosphere.

IV. THE OCEANIC CARBON CYCLE

Carbon in the oceans exists as inorganic and organic species, and in both particulate and dissolved forms. Dissolved inorganic carbon (DIC) comprises aqueous CO_2 , the bicarbonate (HCO_3^-) and carbonate ($\text{CO}_3^{=}$) ions. CO_2 dissolves in water to form a weak acid, which reacts with carbonate ions to form bicarbonates:



The reactions are reversible and are governed by equilibrium dissociation constants that are known functions of temperature and salinity of the solution. At 18° , the first and second apparent dissociation constants ($K_1^* = [\text{H}^+][\text{HCO}_3^-]/[\text{CO}_2]$, and $K_2^* = [\text{H}^+][\text{CO}_3^=]/[\text{HCO}_3^-]$) are 9.079×10^{-7} and 6.168×10^{-10} , respectively. The reversible reactions conserve mass (DIC) and charge, represented by the total alkalinity TALK:

$$\text{DIC} = [\text{CO}_2] + [\text{HCO}_3^-] + [\text{CO}_3^=], \quad (4)$$

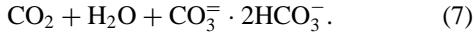
$$\text{TALK} = [\text{HCO}_3^-] + 2[\text{CO}_3^=]. \quad (5)$$

The partial pressure of CO_2 in water, pCO_2 , is governed by Henry's Law:

$$[\text{CO}_2] = K_{\text{H}} \text{pCO}_2, \quad (6)$$

where K_{H} is the solubility of CO_2 . At $T = 18^\circ$ and $P = 1 \text{ atm}$, $K_{\text{H}} = 3.429 \times 10^{-2} \text{ mol kg}^{-1} \text{ atm}^{-1}$. On average, the relative proportions of $[\text{CO}_2]:[\text{HCO}_3^-]:[\text{CO}_3^=]$ are $1:150:15$ in the surface ocean. At present oceanic values, surface water pCO_2 changes by ~ 4.3 percent per degree Celsius change in temperature. This is generally referred to as the "solubility pump."

The reversible reactions Eqs. (2) and (3) combine to yield:



Dissolved carbonate ions present in seawater partially neutralize the carbonic acid, and so the addition of carbonate ion decreases the pCO_2 in the surface water. The carbonate system thus buffers changes in pCO_2 of seawater, and permits a greater uptake of atmospheric CO_2 than of an inert gas. The sensitivity of pCO_2 to changes in DIC is summarized by the buffer (or Revelle) factor, defined as

$$R = \frac{\Delta \text{pCO}_2/\text{pCO}_2}{\Delta \text{DIC}/\text{DIC}}. \quad (8)$$

The Revelle factor ranges from a value of ~ 14 at 0° to ~ 8 at 30° , with a globally averaged value of 10 in seawater. The Revelle factor is the reason that the turnover time for carbon in the upper 100 m of the ocean is 1 instead of 10 years.

Marine photosynthesis takes place where there is sunlight and supply of macronutrient and micronutrients, i.e., in the surface ocean. Photosynthesis by phytoplankton transforms dissolved inorganic carbon into organic carbon. Some phytoplankton, e.g., coccolithophores, also precipitate solid calcium carbonate plates as a by-product of their metabolism. Grazing by zooplankton and other consumers, and the formation of skeletal parts further transform the dissolved and particulate carbon species. Some of the biological detritus is decomposed in the surface ocean and recycled to support further production. The remainder of the detritus sinks and is remineralized (converted back to dissolved inorganic forms) at depth. Marine biology thus reduces DIC near the surface and increases DIC at depth. The DIC is redistributed by the ocean circulation, with upwelling returning DIC and nutrients to the surface, where the nutrients continue to fuel photosynthesis. Marine photosynthesis averages $\sim 50 \text{ PgC/year}$ while the biomass totals only 1–2 PgC; the turnover time of the marine biological cycle is hence on the order of days. The strength of marine productivity and hence biological pump is intimately tied to the ocean circulation, which varies on time scales of hours to millennia.

Marine productivity has a complicated effect on surface pCO_2 ; the effect is often referred to as the "biological pump." Primary production decreases DIC and hence pCO_2 in the surface ocean [cf. Eq. (8)]. At the same time, the formation of carbonate shells decreases both total alkalinity and DIC, resulting in a net increase in pCO_2 [cf. Eq. (7)]. Furthermore, the upwelling that supplies nutrients for photosynthesis brings excess DIC as well as cold waters to the surface, both tending to increase surface pCO_2 . Indeed, analysis of the available pCO_2 time series shows summer highs in subtropical ocean gyres where temperature effects dominate, and winter highs in upwelling or biologically productive regions. On the other hand, during intense spring blooms, such as those occurring off the coast of Iceland, pCO_2 of the surface waters has been observed to decrease by $> 100 \mu\text{atm}$ in a week because of the intense biological uptake.

Air-sea exchange of CO_2 is driven by the gradient in CO_2 partial pressures across a thin film at the ocean surface. In the annual mean, approximately 90 PgC is exchanged between the atmosphere and oceans. The exchange is not balanced locally. The pCO_2 in the surface waters has a range $\sim 150 \mu\text{atm}$ between the equator and high latitudes. In comparison, atmospheric CO_2 is relatively well-mixed, so that there is net outgassing of CO_2 .

from the equatorial waters, and a compensating invasion of atmospheric CO₂ into the middle and high latitude oceans (Takahashi *et al.*, 1997).

The distribution of carbon in the ocean is thus a result of local biology, air-sea exchange, and transport by the global scale ocean circulation. A global scale high-precision measurement of the distribution of DIC in the ocean interior was undertaken by over 20 countries in the

World Ocean Circulation Experiment (WOCE) and the Joint Global Flux Study (JGOFS). The observations (Fig. 5) show not only the DIC increasing with depth, as would be expected by the

biology, but also a lower DIC in the interior Atlantic compared to the North Pacific. The DIC difference is consistent with southward export of carbon out of the North Atlantic, and northward transport into the North Pacific, following the large-scale thermohaline circulation (the “Conveyor Belt”).

V. THE GEOLOGIC CARBON CYCLE

Limestone rocks are mainly calcium carbonate (CaCO₃), and are the skeletal remains of marine organisms and

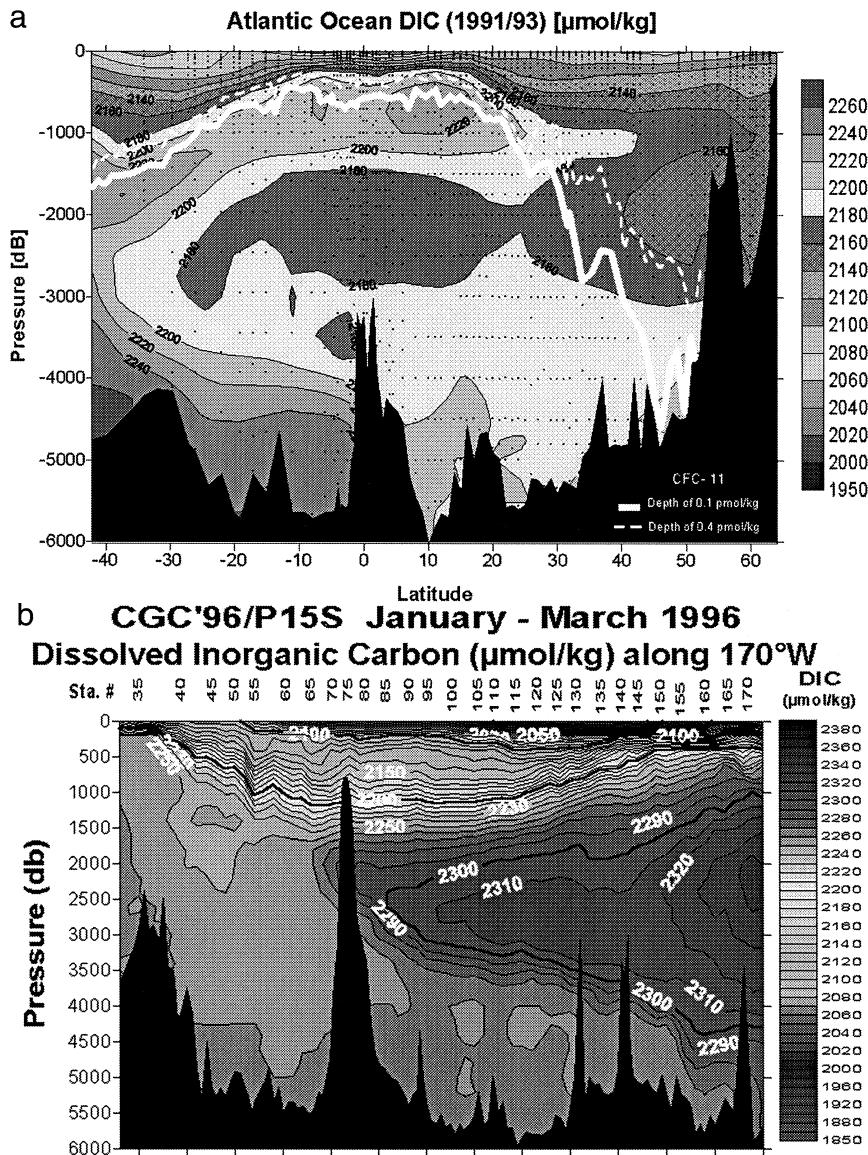
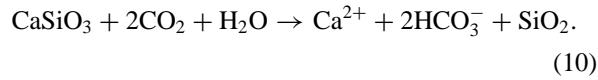
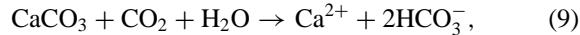


FIGURE 5 Latitudinal-depth distribution of dissolved inorganic carbon in the (a) Atlantic and (b) Pacific oceans. The Pacific data were obtained along 170°W between January and March 1996 along WOCE lines P15S. (Data are from <http://www.pmel.noaa.gov/co2/co2-home.html>.)

chemical precipitates of CaCO_3 . The principal agent of chemical weathering is carbonic acid (H_2CO_3) in the soil solution. Because plant roots and decomposing soil organic matter release CO_2 to the soil pore space, the concentration of H_2CO_3 in soil waters is generally greater than that in equilibrium with atmospheric CO_2 . Weathering and erosion of the earth's surface leaches dissolved calcium, carbon, and silica (SiO_2) from rocks containing calcium carbonate and calcium silicate (CaSiO_3):



The dissolved substances are transported to the oceans by rivers to become part of the marine carbon cycle. A small fraction of the dissolved Ca^{2+} and HCO_3^- are precipitated as CaCO_3 in sediments [reverse of reaction in Eq. (9)].

Seafloor spreading as a result of plate tectonics carries the sediments to subduction zones, where they are transported down into the mantle. At the high temperatures and pressures of the subduction zone, the residual calcium carbonate and organic matter are transformed by volcanism, metamorphosis, or deep diagenesis; the resultant CO_2 and water vapor are eventually outgassed to the atmosphere from volcanic eruptions and hot springs. At steady state, the erosion processes consume on average $\sim 0.1 \text{ PgC/year}$ from the atmosphere. This amount is approximately that deposited in the ocean sediments and that outgassed to the atmosphere.

VI. UNRAVELING THE CONTEMPORARY CARBON BUDGET

In the past 200 years, the abundance of CO_2 in the atmosphere has increased by over 30%, from a concentration of 280 ppmv during the preindustrial era, to 368 ppmv by 1999 AD. The annual increase in atmospheric CO_2 , while large from the perspective of the atmospheric carbon inventory, is equivalent in magnitude to only <5% of the gross fluxes. While it is not implausible that the increase could have been caused by small transient imbalances in the natural CO_2 cycle, there is unambiguous evidence that the principal cause for the contemporary increase atmospheric CO_2 is the combustion of fossil fuels, with a small contribution from cement manufacturing. The relative abundance of both the stable and radioactive isotopes of carbon have been decreasing in the atmosphere while the total CO_2 is increasing, thus reflecting the addition of ancient, ^{14}C -free, carbon of plant origin depleted in ^{13}C , i.e., carbon in fossil fuels. Furthermore, a slow steady decline (<0.002% per year) in atmospheric oxygen has been revealed by high-precision measurements of

TABLE III Carbon Balance for 1980–1989 and for 1990–1997

	Carbon sources/sinks (PgC/year)	
	1980–1989	1990–1997
Fossil Fuel Combustion	5.5 ± 0.3	6.3 ± 0.4
Atmospheric accumulation	3.3 ± 0.1	2.9 ± 0.1
Net Sink	2.2 ± 0.4	3.4 ± 0.5

the atmospheric O_2/N_2 ratio. The measurements pioneered by R. F. Keeling in 1988 (Keeling and Shertz, 1992) use interferometry and those by M. Bender in 1993 (Bender et al., 1995) use mass spectroscopy. The oxygen decrease has occurred at a rate consistent with oxygen consumption by the combustion of the fossil carbon and oxygen release during net uptake of CO_2 by vegetation. Furthermore, the increasing hemispheric gradient in atmospheric CO_2 concentration reflects the industrial source of the increase: over 90% of the fossil fuel combustion is in the industrialized countries in the northern hemisphere.

Comparison of the annual increase in atmospheric CO_2 with the annual inputs from fossil fuel combustion reveals that on average only 50–60% of the CO_2 from the fossil fuel CO_2 has remained in the atmosphere, with residual absorbed by the land and oceans (Table III). This residual by definition includes other perturbations to the carbon cycle, such as CO_2 release that due to land-use modification, as well as transient imbalances between the gross land and ocean fluxes into and out of the atmosphere. The flux imbalances, or net fluxes, are referred as sources and sinks, adopting the perspective of the atmosphere.

The location of the sink for anthropogenic carbon is under intense study. The perturbation fluxes required to balance the carbon budget are smaller than the background fluxes by a factor of 30–50; e.g., the net ocean uptake is $\sim 2 \text{ PgC/year}$ in the 1990s compared with gross fluxes of 90 PgC/year . Their direct detection is hence difficult. Also, because of the heterogeneity of the land surface, year-to-year variations in climate and ocean circulation, intense measurements at a few locations for short periods may not be readily extrapolated to yield meaningful sums on regional or global scales. Hence the identification of the locations of the carbon sink was first inferred from atmospheric measurements of CO_2 and other proxies.

Variations in atmospheric oxygen have provided unique insight into carbon sources and sinks. Between 1990 and 1997, the mixing ratio of atmospheric O_2 decreased by 25 ppmv, when that of atmospheric CO_2 increased by 9 ppmv. The decrease in O_2 is less than expected from fossil fuel combustion, which has an $\text{O}_2:\text{CO}_2$ ratio of 1.43 ± 0.02 . As air-sea exchange of CO_2 has little impact on atmospheric oxygen, and net land uptake has a $\text{O}_2:\text{CO}_2$

ratio of 1.1 ± 0.05 , the combined analyses of the O₂:CO₂ budgets demand a net land carbon sink (and O₂ source). The land sink is thus estimated to be 1.0 ± 0.6 PgC/year, so that the remaining sink of 2.4 ± 0.5 PgC/year has to be absorption by the oceans.

Information on the sources and sinks of CO₂ can be obtained from variations of carbon isotopes. Carbon-13, a stable isotope, comprises approximately 1% the total inventory of carbon. Different carbon exchange processes have different degrees of fractionation or discrimination against the heavier isotope, so that the variations in the ratio of ¹³C:¹²C in the atmospheric terrestrial and oceanic carbon reservoirs provide additional information about the sources and sinks of atmospheric CO₂.

The ratio of ¹³C:¹²C is commonly expressed as

$$\delta^{13}\text{C} = \{(\text{sample}/\text{standard}) / (\text{sample}/\text{standard})_{\text{standard}} - 1\},$$

where ¹³C/¹²C_{standard} = 0.0112372 is the ¹³C/¹²C ratio of Pee-Dee Belemnite, the reference material. High-precision measurements of the ¹³C/¹²C ratio have permitted the potential separation among the exchange processes that contribute to atmospheric CO₂ variations.

Terrestrial carbon has a $\delta^{13}\text{C}$ of -25 permil (¹³C/¹²C $\sim 1.096\%$), while DIC has an average $\delta^{13}\text{C}$ of $+2$ permil (¹³C/¹²C $\sim 1.126\%$). In the atmosphere, $\delta^{13}\text{C}$ has decreased from -6.4 permil in the preindustrial era to ~ -8 permil in 2000 AD, or the ¹³C/¹²C ratio has decreased from 1.117 to 1.115%. The atmospheric decrease reflects the changing balance between sources and sinks in the contemporary carbon cycle. Fossil fuels are derived from organic matter, whose relative ¹³C abundance is lower than that in the atmosphere, as photosynthesis preferentially discriminates against the heavier isotope. As with O₂/N₂, the atmospheric decrease in $\delta^{13}\text{C}$ is slower than would be expected from the addition of fossil fuel CO₂ alone, thus reaffirming in the role of the ocean and land carbon sinks in maintaining contemporary atmospheric CO₂ growth rate. Also, like O₂/N₂, terrestrial uptake has a greater atmospheric $\delta^{13}\text{C}$ signature than oceanic uptake of the same strength. Hence the atmospheric $\delta^{13}\text{C}$ decrease rate supports both a land and an ocean sink of the anthropogenic carbon (Fig. 6).

The use of $\delta^{13}\text{C}$ to constrain the contemporary carbon budget is complicated, even though the fractionation coefficients associated with terrestrial and marine photosynthesis and respiration, and with dissociation of DIC in the ocean are relatively well-known. The complication arises from the Suess effect, or the isotopic disequilibrium associated with CO₂ gross, or steady-state background, exchanges between the atmosphere and land and between the atmosphere and oceans (Fung *et al.*, 1997; Sonnerup *et al.*, 1999; Gruber and Keeling, 2000). On average, these background exchanges cancel with no impact on atmospheric CO₂. However, as the $\delta^{13}\text{C}$ in the

C₁₃ Constraint on the Land-Sea Partitioning of the Carbon Sink

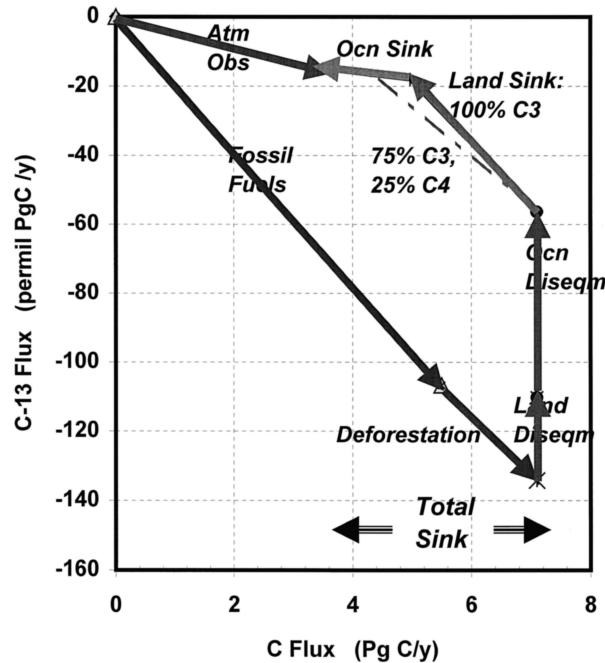


FIGURE 6 Annual ¹³C budget (expressed in PgC-permil/year) versus annual CO₂ budget (expressed in PgC/year). The slopes represent the relatively well-known fractionation coefficients.

atmosphere decreases due to the addition of fossil fuel carbon, the flux leaving the atmosphere is relatively depleted in ¹³C compared with the flux entering the atmosphere, with the difference given by the residence time of carbon in the terrestrial and oceanic reservoirs. Determination of these ages or residence times is still rudimentary, and has been based on models or on extrapolation of very sparse data. The land-sea partitioning of the carbon sink obtained using ¹³C in addition ¹²C is not inconsistent with that obtained from changes in the O₂/N₂ ratio.

The geographic variations in atmospheric CO₂ also contain information about carbon cycle. The atmospheric circulation mixes, but not completely, CO₂ and other trace substances exchanged at the surface. The smoothed but detectable atmospheric CO₂ gradients thus reflect the locations of the sources and sinks. In the 1990s, the countries at mid latitudes of the northern hemisphere are responsible for over 90% of the fossil fuel combustion, so that CO₂ is higher in the northern than in the southern hemisphere. In the 1990s, the hemispheric CO₂ difference is about 1% of the global mean value. This observed north-south gradient is smaller than that would be expected if all the fossil fuel CO₂ remained airborne. The atmospheric hemispheric CO₂ gradient thus demands a combined land

and ocean carbon sink that is greater in the northern than in the southern hemisphere (Tans *et al.*, 1990). The available surface ocean pCO₂ data suggests that the northern hemisphere ocean sink is less than that required, hence requiring the land surface of the northern hemisphere to also be acting as carbon sinks. The location and mechanism of the land sink with the northern hemisphere remain elusive. The atmosphere mixes faster in the east–west direction (in 1–3 months) than in the north–south direction across the equator (\sim 1 year). As a result, the east–west gradient of atmospheric CO₂ is small (<0.5 ppmv between the Atlantic and Pacific in 2000 AD), and its interpretation in terms of partitioning of the land sink between North America and Eurasia is highly uncertain (Rayner *et al.*, 1999; Fan *et al.*, 1998, Bousquet *et al.*, 2000).

VII. MECHANISMS FOR THE CONTEMPORARY CARBON SINK

A. Ocean Sink

Increasing partial pressure of CO₂ in the atmosphere increases the partial pressure difference across the air–sea interface and the invasion of CO₂ into the ocean. The buffering effect of carbonate chemistry [cf. Eq. (6)] essentially enables the storage of carbon in the ocean at a level significantly greater than that possible if CO₂ were an inert gas. The upper 100 m ocean is well-mixed and exchanges directly with the atmosphere on time scales of a year. Hence, the effectiveness of the ocean as a carbon sink depends critically on the rate the anthropogenic carbon is mixed into the ocean interior. The existence of a thermocline (region of steep temperature gradient) between 200–1000 m bespeaks a mixing barrier between the turbulently mixed upper ocean and the stably stratified ocean below. Observations of the slow penetration of CFC and other pollutants into the ocean interior allow the estimation of the exchange rate, and the estimation of the transient uptake of anthropogenic carbon by the oceans.

Over the past 200 years, the cumulative ocean sequestration of anthropogenic carbon is estimated to be <100 PgC. This sequestration is small (<0.5%) compared to the total carbon inventory in the ocean. To the lowest order, oceanic uptake of anthropogenic CO₂ is a small physical–chemical perturbation and does not involve alteration of marine biology, as long as the ocean circulation remains the same.

B. Land Sink

Net carbon sequestration by the land necessarily implies a net enhancement of photosynthesis compared to respiration, i.e., an increase in photosynthesis that exceeds an increase in decomposition, or a retardation of decomposi-

tion compared to photosynthesis. Many mechanisms may contribute to this net imbalance between the fluxes.

The rate of photosynthesis is governed by the rate of CO₂ diffusion across the stomatal opening. It is therefore expected that increasing CO₂ concentrations in the atmosphere would enhance photosynthesis (CO₂ fertilization) (Lloyd and Farquhar, 1996). Deliberate enhancement of CO₂ in an open forest has demonstrated enhanced photosynthesis during a drought year, as water use efficiency (molecules of carbon assimilated per molecule of water lost) increases (DeLucia *et al.*, 1999). Whether CO₂ fertilization operates in nature depends on the availability of other resources (nutrient, water and light) to support the enhanced photosynthesis (Field *et al.*, 1992). Other hypotheses for the terrestrial carbon sink include favorable climate (Dai and Fung, 1993), enhanced productivity from deposition of atmospheric nitrogen compounds (Holland *et al.*, 1997), and carbon burial in sediments (Stallard, 1998).

A likely strong candidate for the contemporary land carbon sink, especially in the middle latitudes of the northern hemisphere, is land use history (Kauppi *et al.*, 1992). In the industrialized countries, forests were converted for agriculture in the early twentieth century. With the subsequent abandonment of agriculture, and regrowth and management of forests, the carbon source at the beginning of the century becomes a carbon sink at the end of the century. Furthermore, the adoption of soil conservation measures in recent years restores carbon lost in the intensive agricultural period, further enhancing the carbon sink. The relative contributions of these mechanisms to the contemporary land uptake are uncertain and difficult to quantify. It is expected that they would vary from region to region.

VIII. INTERANNUAL VARIABILITY OF THE CONTEMPORARY CARBON CYCLE

Both rates of carbon exchange between the atmosphere and the biosphere and between the atmosphere and the ocean are sensitive to climate variations. It is therefore not surprising that over the past 20 years, the growth rate of atmospheric CO₂ has varied by a factor of 2 even though fossil fuel source has increased steadily (Fig. 7).

Interannual variability of CO₂ sources and sinks have been inferred from the variability in the atmospheric abundance of CO₂, δ¹³C, and O₂/N₂ (Francey *et al.*, 1995; Battle *et al.*, 1999). These time series suggest that the ocean carbon sink, when averaged over the globe, is less variable than the terrestrial sink. For high latitude vegetation in the northern hemisphere, warmer temperatures enhance both photosynthesis and decomposition, so that the net carbon balance may be a net source for a few years and shift to a sink for a few years (Randerson *et al.*, 1999). At

Contemporary Carbon Mass Balance

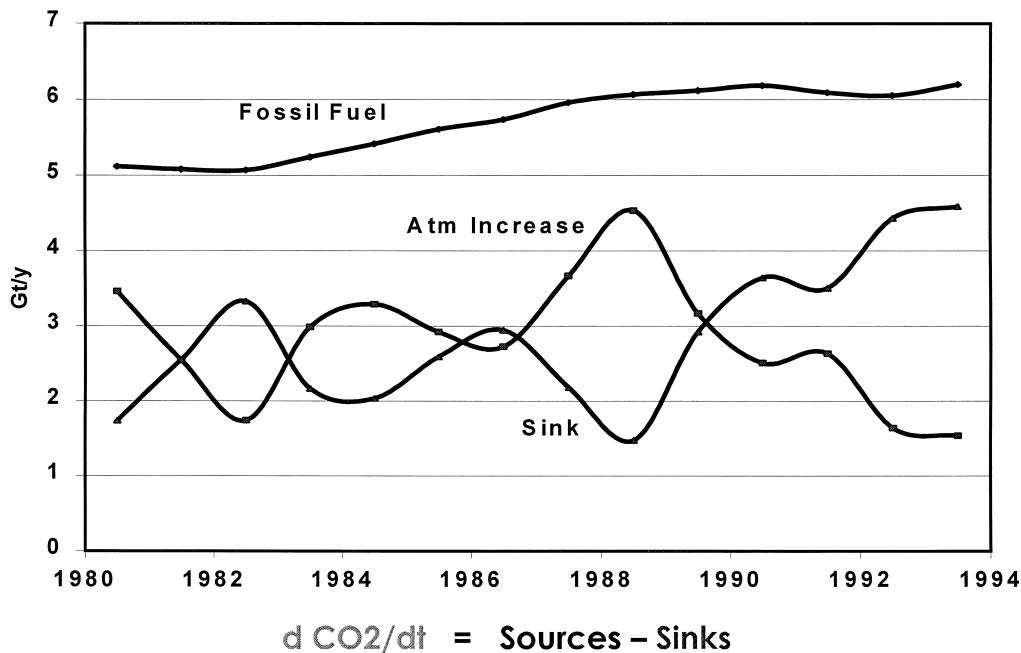


FIGURE 7 Interannual variations in the fossil fuel emission and atmospheric increase in CO₂. The difference between the two is the carbon sink. The varying rate of CO₂ increase in the atmosphere reflects the response of the terrestrial biosphere and ocean to interannual climate variations. After 1988, global sink strength increased, while the atmosphere CO₂ increase slowed down.

mid latitudes and the tropics, there is also great variability in the net carbon exchange with the terrestrial biosphere (Bousquet *et al.*, 2000). However, with less seasonal overlap between the uptake and release of CO₂, the separation of the net flux into its components is not straight-forward.

Interannual variability in terrestrial photosynthesis is clearly evident in the 20-year satellite NDVI time series. Much of the variability can be related to the El-Nino/Southern Oscillation and other aspects of climate variability. For example, there is a dramatic suppression of photosynthesis following dramatic cooling after the eruption of Mt. Pinatubo. At this writing, it appears that the NDVI data also suggest an increasing trend in NPP at middle to high latitudes (Myneni *et al.*, 1999). Temporal variations in decomposition rates cannot be observed directly on the global scale, and have been inferred from variations in temperature and precipitation.

Interannual variability of the marine carbon cycle is difficult to document by direct measurement, just because of the nature of shipboard measurements. A unique view is captured by two observational campaigns of the eastern equatorial Pacific, one during an El-Nino, and the other during a “normal” year (Chavez *et al.*, 1999). In the surface waters, the pCO₂ reduction from reducing upwelling supply of DIC exceeded the pCO₂ enhancement from higher

sea surface temperatures, resulting in a net reduction of outgassing of CO₂ to the atmosphere during the El-Nino year. In the western equatorial Pacific, changes in surface salinity associated with the shifts in the convective region resulted in pCO₂ changes opposite those in the east. The near cancellation between the anomalous air-sea CO₂ fluxes partially explain the small oceanic contribution on the interannual time scale inferred from the atmospheric signal.

IX. CLIMATE INTERACTIONS

The earth, with a globally averaged surface temperature of ~ 288 K, emits radiation in the infrared region of the electromagnetic spectrum. The energy at some of these wavelengths excites CO₂ and other trace species (water vapor, methane, ozone) into higher energy states, resulting in absorption of the emitted radiation. The absorption leads to a retention of the energy that would otherwise have been lost to space, and hence a warming of the atmosphere and surface. This is often referred to as the “greenhouse effect.” In the earth’s atmosphere, CO₂ is the second-most important greenhouse gas, after water vapor. The role of atmospheric CO₂ in regulating earth’s temperature was recognized over 100 years ago by Svante Arrhenius (1859–1927).

The variations in CO₂ and temperature retrieved from the deep ice core at Vostok, Antarctica, illustrate the interaction between CO₂ and climate (Petit *et al.*, 1999). Over the past 420,000 years, CO₂ and temperature have covaried, though not in lock-step fashion, over the four glacial-interglacial cycles. CO₂ concentrations were ~280 ppmv during the warm interglacials and declined to ~180 ppmv during the glacial periods.

Explanations for the variations in the carbon cycle over the past 420,000 years remain elusive. Analyses of the pollen record from the Last Glacial Maximum suggest a shift to vegetation with low biomass density, as is expected from a cooler climate. The inventory reduction in both the atmosphere and terrestrial biosphere would demand an increased storage of carbon in the oceans to maintain mass balance. The mechanisms for such an oceanic increase remain under debate. One hypothesis involves changing ocean circulation and the redistribution of DIC in the ocean interior. Another involves changes in the ocean carbonate system. A third hypothesis involves changes in the supply nutrient in the ocean, especially of micronutrients such as iron, which is essential for primary production and nitrogen fixation. Enhanced dust deposition during each of the four glacial periods has been found in the Vostok ice core data and has been hypothesized to fertilize the oceans (Martin 1990; Broecker and Henderson, 1998). There is as yet no adequate explanation for the apparent ceiling of 280 ppmv in atmospheric CO₂ in the 400,000 years before anthropogenic perturbations.

X. OUTLOOK

The carbon cycle is dynamic, and is fully interactive with the climate. The earth's climate changes as a result of changes in the abundance of CO₂ in the atmosphere; climate change in turn changes the dynamics of carbon exchange among the reservoirs and causes shifts in the atmospheric CO₂ levels.

Humans have perturbed the carbon cycle in significant ways. The demand for energy depletes, in two hundred years, the reservoir of fossil fuel carbon formed over millions of years. Agriculture and land use alter the turnover rates of carbon in the terrestrial biosphere. As the earth evolves into a climate and CO₂ space with no known ecological analog, historical changes do not provide clues or warning about how the terrestrial and marine biota may respond. The outcome of the "great geophysical experiment" remains an enigma.

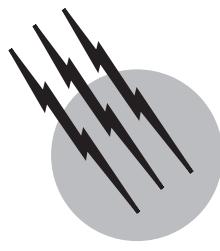
SEE ALSO THE FOLLOWING ARTICLES

BIOENERGETICS • CHEMICAL COMPOSITION AND ELEMENT DISTRIBUTION IN THE EARTH'S CRUST • CLIMA-

TOLOGY • ENERGY RESOURCES AND RESERVES • ENVIRONMENTAL GEOCHEMISTRY • GEOCHEMISTRY, ORGANIC • GREENHOUSE EFFECT AND CLIMATE DATA • NITROGEN CYCLE, ATMOSPHERIC • NITROGEN CYCLE, BIOLOGICAL • OCEAN-ATMOSPHERIC EXCHANGE • RADIATION, ATMOSPHERIC

BIBLIOGRAPHY

- Barnola, J. M., Raynaud, D., Korotkevich, Y. S., and Lorius, C. (1987). *Nature* **329**, 408.
- Battle, M., Bender, M. L., Tans, P. P., White, J. W. C., Ellis, J. T., Conway, T. J., and Francey, R. J. (2000). *Science* **287**, 2467.
- Bender, M. *et al.* (1995). *Geochem. Cosmochim. Acta* **58**, 4751.
- Bousquet, P., Peylin, P., Ciais, P., LeQuere, C., Friedlingstein, P., and Tans, P. (2000). *Science* **290**, 1342.
- Broecker, W. S., and Henderson, G. M. (1998). *Paleoceanography* **13**, 352.
- Casperson, J. P., Pacala, S. W., Jenkins, J. C., Hurt, G. C., Moorcroft, P. R., and Birdsey, R. A. (2000). *Science* **290**, 1148–1151.
- Chavez, F. P., Strutton, P. G., Friederich, G. E., Feely, R. A., Feldman, G. C., Foley, D., and McPhaden, M. J. (1999). *Science* **286**, 2126.
- Dai, A. G., and Fung, I. Y. (1993). *Global Biogeochem. Cycles* **7**, 599.
- DeLucia, E. H., Hamilton, J. G., Naidu, S. L., Thomas, R. B., Andrews, J. A., Finzi, A., Lavine, M., Matamala, R., Mohan, J. E., Hendrey, G. R., and Schlesinger, W. H. (1999). *Science* **284**, 1177.
- Fan, S. M., Gloor, M., Mahlman, J., Pacala, S., Sarmiento, J., Takahashi, T., and Tans, P. (1998). *Science* **282**, 442.
- Field, C. B., Chapin, F. S., Matson, P. A., and Mooney, H. A. (1992). *Ann. Rev. Ecol. Systematics* **23**, 201.
- Francey, R. *et al.* (1995). *Nature* **373**, 326.
- Fung, I., Field, C. B., Berry, J. A., Thompson, M. V., Randerson, J. T., Malmstrom, C. M., Vitousek, P. M., Collatz, G. J., Sellers, P. J., Randall, D. A., Denning, A. S., Badeck, F., and John, J. (1997). *Global Biogeochem. Cycles* **11**, 507.
- Gruber, N., and Keeling, C. D. (2001). *Geophys. Res. Lett.* **28**, 555.
- Holland, E. A., Braswell, B. H., Lamarque, J.-F., Townsend, A., Sulzman, J., Müller, J.-F., Dentener, F., Brasseur, G. H. L. II, Penner, J. E., and Roelofs, G.-J. (1997). *J. Geophys. Res.* **106**, 15,849.
- Kauppi, P. E., Mielikäinen, K., and Kuusela, K. (1992). *Science* **256**, 70.
- Keeling, R. F., and Shertz, S. R. (1992). *Nature* **358**, 723.
- Lloyd, J., and Farquhar, G. D. (1996). *Functional Ecol.* **10**, 4.
- Martin, J. (1990). *Paleoceanography* **5**, 1.
- Myneni, R., Tucker, C. J., Asrar, G., and Keeling, C. D. (1998). *J. Geophys. Res.* **103**, 6145.
- Petit, J. R., Jouzel, J., and Raynaud, D., *et al.* (1999). *Nature* **399**, 429.
- Prentice, I. C., and Webb, T. (1998). *J. Biogeogr.* **25**, 997.
- Randerson, J. T., Field, C. B., and Fung, I. Y., *et al.* (1999). *Geophys. Res. Lett.* **26**, 2765.
- Rayner, P. J., and Law, R. M. (1999). *Tellus* **51B**, 210.
- Stallard, R. (1998). *Global Biogeochem. Cycles* **12**, 231.
- Sonnerup, R. E., Quay, P. D., and McNichol, A. P., *et al.* (1999). *Global Biogeochem. Cycles* **13**, 857.
- Takahashi, T., Feely, R. A., and Weiss, R. F., *et al.* (1997). *Proc. Natl. Acad. Sci.* **94**, 8292.
- Tans, T. T., Fung, I. Y., and Takahashi, T. (1990). *Science* **247**, 1431.
- Yu, G., Chen, X., and Ni, J., *et al.* (2000). *J. Biogeogr.* **27**, 635.



Climatology

J. E. Oliver

G. D. Bierly

Indiana State University

Hans A. Panofsky

University of California, San Diego

- I. Introduction
- II. Fundamentals of Climate Theory
- III. Regional Climatology
- IV. Microclimatology
- V. Climate and General-Circulation Models
- VI. Past Climates
- VII. Future Climates

GLOSSARY

- Albedo** Fraction of solar radiation reflected; earth-atmosphere system albedo is approximately 30%.
- Doldrums** Calms of the intertropical convergence located between the northeast and southeast trade winds.
- Eccentricity** Ratio of focus-center distance to semimajor axis of ellipse.
- Ecliptic** Plane of the earth's orbit about the sun.
- El Niño** Almost periodic phenomenon involving warming of the equatorial Pacific.
- Front** Surface separating air masses of differing properties.
- GCM** Computer-derived general-circulation model of the earth's atmosphere.
- Hadley cell** Circulation cell with updrafts near the Equator and downdrafts near 30° latitude.

Horse latitudes Regions of sinking air and light winds located around 30° to 35° latitude, the subtropical high-pressure belts.

Isopleth Lines joining equal values on a map or chart, e.g., isotherm—equal temperatures; isobar—equal pressures.

Jet stream Rapidly flowing, narrow airstream in the upper troposphere or lower stratosphere.

Latent heat Energy released or required by change of phase (usually for water in the climate context).

Obliquity of the ecliptic Angle between the earth's orbit and the Equator.

Precession Movement of the earth's axis with a period of 26,000 years.

Solar constant Solar energy received at right angles to solar beam at the earth's mean distance from the sun. Current value: 1370 W/m².

Trades Tropical easterly winds, northeast in the Northern Hemisphere and southeast in the Southern, between the subtropical highs and the intertropical convergence.

Tropopause Top of the troposphere.

Troposphere Region of the atmosphere from the surface to 9 to 16 km, which is heated from below and is fairly well mixed.

Westerlies Mid-latitude westerly winds which flow as a train of waves perpendicular to the hemispheric pressure gradient force in the upper atmosphere.

WE SHALL DEFINE CLIMATE as weather statistics over periods of the order of 30 years as measured at particular points on the globe. Thus, climate includes not only averages of weather elements, but also measures of their variation. These might include variations throughout the day, throughout the year, from year to year, and extremes that may occur only once in several years or decades. Other definitions of climate involve statistics over different periods.

Temperatures and precipitation are the most common elements used to define climate, but wind, amount of snowfall, and other variables are also included. Climate information of all types is available from the National Climate of the National Oceanographic and Atmospheric Administration in Asheville, North Carolina, and also from the World Meteorological Office in Geneva, Switzerland.

I. INTRODUCTION

Climate data are required for planning many types of human activity, such as agriculture, architecture, and transportation. For example, in the construction of a house information is needed about the averages and extremes of temperature, precipitation, wind, and snowfall. Also, obviously, different types of agriculture are possible in different climates.

A. Brief History of Climate Study

An appreciation of climate existed long before any formal history of the discipline was initiated. People have always been aware of their environment, and if a person resides in a place for a decade or more, an image of the climate of that location is formed. Given the large seasonal changes in many world areas and their significance to human survival, such an image must have been formed in early times. However, it remained for the ancient Greeks to formalize its study; in fact, the word *climate* is derived from the Greek *klima*, meaning “slope.” In this context, the word applies to the slope or inclination of the earth’s axis and is

applied to an earth region at a particular elevation on that slope, that is, the location of that place in relation to the parallels of latitude and the resulting angle of the sun in the sky.

Apart from establishing the geometric relationships of climatology, Greek scholars wrote treatises on climate. The first climatography (descriptive study of climate) is attributed to Hippocrates, who wrote “Airs, Waters, and Places” in 400 BC. In 350 BC, Aristotle wrote the first meteorological treatise, “Meteorologica.” The Greeks gave names to winds, described marine climates, and examined the role of mountains in determining the climate of a location.

With the decline of ancient Greece, it remained for Chinese scholars to make major efforts to understand their atmospheric environment. For example, they used seasonal rainfall to estimate harvests and tax revenues and noted the relative severity of floods and cold winters. While the Chinese continued their efforts, the Western world entered a period in which scientific inquiry was not encouraged, and climatic observations were not accorded importance. It was not until the middle of the 15th century, the Age of Discovery, that long-term observation of the atmosphere was again of interest. With the extended sea voyages and development of new trading routes, descriptive reports of world climates became available, especially those concerning prevailing wind systems.

During the 17th century, scientific analysis of the atmosphere got underway when instruments were developed. This set the scene for the advent of the modern era. The availability of instruments and the formulation of basic laws of gases led to a new era of climatic observation and analysis. **Table I** shows a partial listing of some of the significant events.

B. Subdisciplines of Climatology

Descriptive climatology describes the climates of the world. It is subdivided into regional climatology, which deals on a broad scale with the climates of large portions of the world, and microclimatology, the modifications of local climates by local factors, such as topography and land-water contrasts. Physical climatology attempts to explain the properties of climates by the earth–sun geometry, atmospheric composition, ground characteristics, and the laws of physics. Synoptic climatology draws linkages between physical and dynamic aspects of the large-scale atmospheric circulation and surface weather tendencies at regional and local scales. Historical climatology deals with climate change and the many hypotheses suggested to account for them. It also attempts to predict future climate change, produced both naturally and by human action.

TABLE I Significant Events in the Development of Climatology

ca. 400 BC	Influence of climate on health is discussed by Hippocrates in "Airs, Waters, and Places."
ca. 350 BC	Weather science is discussed in Aristotle's "Meteorologica."
ca. 300 BC	"De Ventis" by Theophrastus describes winds and offers a critique of Aristotle's ideas.
ca. 1593	Thermoscope is described by Galileo; the first thermometer is attributed most likely to Santorre, 1612.
1622	Significant treatise on the wind is written by Francis Bacon.
1643	Barometer is invented by Torricelli.
1661	Boyle's law on gases is propounded.
1664	Weather observations begin in Paris; although often described as the longest continuous sequence of weather data available, the records are not homogeneous or complete.
1668	Edmund Halley constructs a map of the trade winds.
1714	Fahrenheit scale is introduced.
1735	George Hadley's treatise on trade winds and effects of earth rotation is written.
1736	Centigrade scale is introduced. (It was formally proposed by du Crest in 1641.)
1779	Weather observations begin at New Haven, CT; they represent the longest continuous sequence of records in the United States.
1783	Hair hygrometer is invented.
1802	Lamark and Howard propose the first cloud classification system.
1817	Alexander von Humboldt constructs the first map showing mean annual temperature over the globe.
1825	Psychrometer is devised by August.
1827	The period begins during which H. W. Dove developed the laws of storms.
1831	William Redfield produces the first weather map of the United States.
1837	Pyrheliometer for measuring insolation is constructed.
1841	Movement and development of storms are described by Espy.
1844	Gaspard de Coriolis formulates the "Coriolis force."
1845	First world map of precipitation is constructed by Bergaus.
1848	First of M. F. Maury's publications on winds and currents at sea is written.
1848	Dove publishes the first maps of mean monthly temperatures.
1862	First map (showing western Europe) of mean pressure is drafted by Renou.
1869	Supan publishes a map showing world temperature regions.
1892	Systematic use of balloons to monitor free air begins.
1900	Term <i>classification of climate</i> is first used by Köppen.
1902	Existence of the stratosphere is discovered.
1913	Ozone layer is discovered.
1918	V. Bjerknes begins to develop his polar front theory.
1925	Systematic data collection using aircraft begins.
1928	Radiosondes are first used.
1940	Nature of jet streams is first investigated.
1956	First computer model explains general-circulation techniques.
1960	First meteorological satellite, <i>Tiros I</i> , is launched by the United States.
1966	First geostationary meteorological satellite is launched by the United States.
1968	Global atmospheric research program begins.
1978	National Climate Program is adopted by the U.S. government.
1974	Rowland and Molina formulate the theoretical linkage between CFCs and stratospheric ozone depletion.
1985	British Antarctic Survey at Halley Bay discovers Southern Hemisphere ozone hole.
1987	Montreal Protocol established.
1992	Rio Framework Convention on Climatic Change.
1997	Kyoto Climate Conference.

Because of the profound effect of climate on many aspects of human existence, there are myriad applications of climatology: bioclimatology, the effects of climate on human, animal, and plant life; architectural climatology,

the impact of climate on architecture; and so on. Since climatologists cannot foresee all the applications of their knowledge, applied climatology is best left to the users of climate information.

II. FUNDAMENTALS OF CLIMATE THEORY

A. Reasons for Climate Differences

Weather and climate are produced by solar heating. The reasons for variations in climate around the earth are primarily of four types:

1. Incoming solar radiation (insolation) varies with latitude, being largest at the equator and smallest at the poles; hence, climate varies with latitude.
2. Energy transformations at the ground are completely different over water, ice, and land; hence, climates depend critically on the proximity of such surfaces.
3. Hills and mountains modify atmospheric variables.
4. Differences in land use and ground cover affect climate.

B. Solar Source

The *solar constant* is defined as the radiation intensity received at right angles to the sun's radiation at the earth's mean distance, 1.49×10^8 km. This is the average between largest distance (aphelion), in July, and smallest distance (perihelion), on January 4. Aphelion distance exceeds perihelion, at present, by $\sim 3\%$. The solar constant is ~ 1370 W/m². It is a measure of the sun's heat output

and has been remarkably constant during the period of its measurement, the 20th century. Only a decrease of 0.1% from 1981 to 1986, observed by satellites, may be significant, and there is no reason why this small trend should continue.

Given that the solar constant has not been noted to vary much, some of the basic properties of climate can be inferred from the simple geometric laws for insolation I (radiation intensity on a horizontal surface) without an atmosphere:

$$I = (S/r^2) \cos z \quad (1)$$

Here, S is the solar constant, r is the earth–sun distance divided by the mean distance, and z is the zenith angle of the sun—that is, the angle between the sun and the zenith.

According to a basic formula of practical astronomy, $\cos z$ depends only on time of day, time of year, and latitude. The zenith angle, and hence I , depends on time of year (seasonal variation) because the angle between the earth's Equator and its orbit around the sun, the obliquity of the ecliptic ε , is not zero. Its present value is 23.5° (Fig. 1). The larger ε is, the larger are the seasonal contrasts.

As mentioned earlier, the distance r currently varies by only 3% throughout the year. This variation is not of great importance in explaining the properties of the current climate.

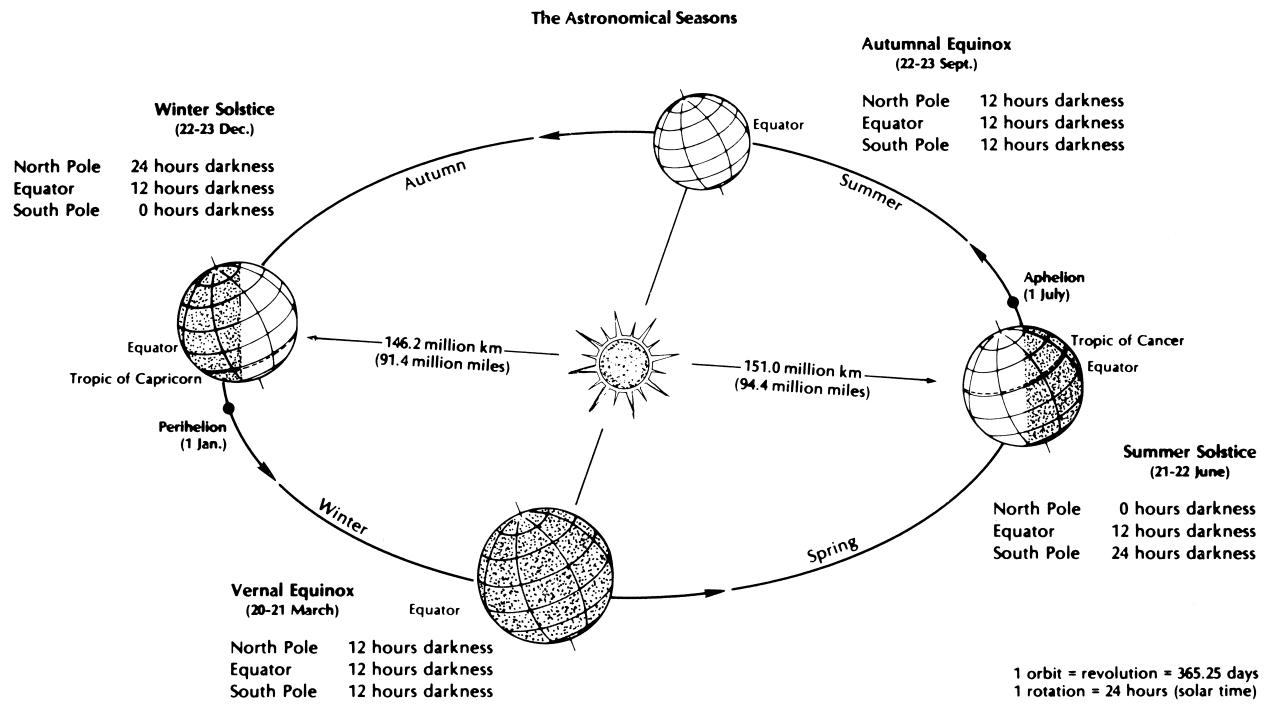


FIGURE 1 Earth's orbit, showing positions of the earth at solstices and equinoxes.

The presence of the atmosphere means that the actual insolation is less than that given by Eq. (1), even without clouds. The difference is largest at high latitudes, where the light path through the air is longest. The difference also depends on the transparency of the atmosphere, which is quite variable. Clouds are the most variable factors that affect the transparency of the atmosphere, since they both reflect and absorb sunlight. The net result of the presence of the earth's atmosphere is that the actual energy available at the surface is only about half that indicated by Eq. (1).

Even when we are given accurate radiation intensity estimates everywhere, we still need to calculate the meteorological variables such as temperature, wind, and precipitation. This is done by climate models and general-circulation models.

C. Energy Budget

Given the distribution of insolation at the surface, we shall briefly follow the fate of the energy after it has been absorbed at the ground—but averaged over the entire earth's surface. Energy received and released must almost balance on a globally and yearly averaged basis, since climate changes are slow and small in magnitude.

In this brief survey, we shall consider the average insolation at the atmosphere to be one-fourth of the solar constant S ; the reason is that the earth intercepts $\pi R^2 S$ of solar radiation (R is the earth's radius), but the earth loses radiation to space over its entire surface of area $4\pi R^2$.

For convenience, we shall consider $S/4$ to be 100%. As mentioned before, about 50% is absorbed at the surface, 30% is reflected by the atmosphere (mostly by clouds), and 20% is absorbed. The ground loses energy through evaporation (~22%) and direct conduction to the atmosphere (6%). The evaporated water vapor condenses in the lower atmosphere and is an important heat source for it.

In addition, the ground loses more than 100% in infrared radiation. If this were emitted into space, the ground would lose more energy than it receives and would cool rapidly. Actually, much of this radiation is intercepted by clouds, water vapor, and carbon dioxide in the atmosphere and returned to the ground, leading to a near balance of energy at the ground. This “trapping” of infrared radiation is often called the “greenhouse effect” even though greenhouse glass also inhibits convection, which the atmosphere does not.

To complete the balance at the top of the atmosphere, the atmosphere radiates ~50% to space, while ~20% is emitted directly from the earth's surface; 30% is reflected solar radiation.

Energy is approximately balanced only for the globe averaged for the year, not separately for each latitude. [Figure 2](#) shows the latitudinal distribution of incoming and outgoing radiation at the top of the atmosphere. The incoming radiation varies rapidly with latitude, being by far the largest at the Equator. The outgoing radiation is much more uniform. This is because part of this radiation originates near the tropopause (~9 km height at the

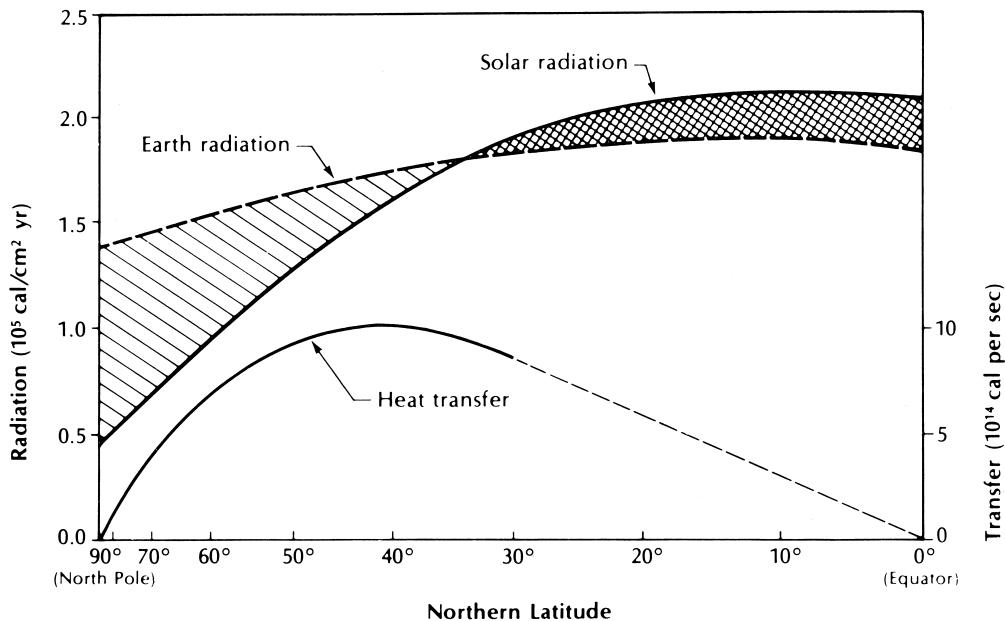


FIGURE 2 Latitudinal variation of incoming and outgoing radiations at the top of the atmosphere. Hatched area, surplus; striped area, deficit.

poles to (~ 16 km at the equator), which is actually cooler near the Equator than either to the north or south. As a result, low latitudes receive an excess of energy, and high latitudes a deficit. These inequalities must be made up by north–south heat exchange. This occurs in three ways:

1. Heat transport in the atmosphere
2. Heat transport in the oceans
3. Latent heat transport (evaporation at one latitude and condensation at another)

According to current estimates, factors 1 and 2 are about equally important, both transporting heat poleward. Latent heat is transported poleward and equatorward from latitudes $\sim 30^\circ$ north and south, latitudes of little rainfall and large evaporation.

The strong latitudinal variation of insolation is responsible for the “general circulation of the atmosphere,” which in turn interacts with the temperature field. The atmosphere can be viewed as a giant heat engine: Heat is added preferentially in the tropics and removed at high latitudes, thereby driving the winds (providing work). Climate models and general-circulation models treat these subjects quantitatively. Since the models have become progressively more complex and are the subject of a huge worldwide research effort, they will be treated in a separate section (see Section V).

D. Hydrological Cycle

World-average precipitation and evaporation must approximately balance on an annually averaged basis. Each averages ~ 1 m, but whereas $\sim 90\%$ of the evaporation takes place over the ocean, only $\sim 67\%$ of the precipitation falls there. The difference is made up by river flow.

Evaporation and precipitation show large latitudinal, seasonal, and diurnal variations, which are closely related to the characteristics of the general circulation. Over the oceans, air generally sinks near latitudes 30° and rises near the Equator and near latitudes 60° north and south. Hence, most evaporation occurs into the dry air near latitudes 30° , and most precipitation occurs near the Equator and at high latitudes. All those zones move northward and southward with the sun throughout the year. Superimposed on these latitudinal migrations, evaporation over the oceans also shows seasonal cycles, with strongest evaporation in the winter, particularly in the western oceans.

Over continents, the general circulation shows strong seasonal reversals. This leads to strong seasonal variations in continental interiors, with maximum precipitation in the summer. Coastal areas are dominated by winter storms, except in regions of strong hurricane activity, which are basically late-summer phenomena.

III. REGIONAL CLIMATOLOGY

Mean fields of atmospheric variables are usually represented by isopleths, that is, lines along which a given variable is constant. For example, Figs. 3 and 4 show global isotherms, lines of constant temperature in July and January, respectively. The lines run basically east–west, showing the prime importance of latitude. The region of warmest air has a tendency to move north of the Equator in the northern summer. Otherwise, seasonal contrasts are small over water and very large over land. Also, seasonal contrasts over land grow from the Equator toward the poles. Owing to the prevalence of water in the Southern Hemisphere, seasonal differences are smaller in this hemisphere.

Over the oceans, in middle latitudes, the isotherms tend to bend poleward on the western side of the oceans and equatorward on the eastern side. These bends are caused by ocean currents, which are warm in the west and cold in the east of the ocean basins.

Variability of temperature from time to time cannot be shown in any generality. Of course, as we have seen, seasonal variation is indicated by maps drawn for different months or seasons. The largest diurnal variations in temperature are found at low latitudes in midcontinental and desert areas.

Year-to-year changes in ocean climate have been of special interest since the 1970s. One aspect of these changes in the climate system is that large portions of the oceans may develop temperature anomalies in the order of 1° to 2° (sometimes more), which are quite persistent and associated with anomalous circulation patterns. A special case of this is the El Niño–southern oscillation (ENSO) phenomenon, which occurs irregularly, roughly every 3 to 7 years. It was especially strong in the period 1997–1998 when the event received widespread coverage in the media. All ENSO events result from a weakening of the easterly trade winds in the equatorial Pacific Ocean. When these weaken, there is a build up of warm surface water and a sinking of the thermocline in the eastern Pacific. Modified circulation patterns result in drought in Indonesia and Australia and storms and floods in Peru, Columbia, and Bolivia. Worldwide repercussions also occur. For example, in the 1997–1998 event, the United States experienced stormy weather on the west coast and drier weather in the east; over most areas, warmer than normal temperatures occurred, with resulting diminished snowfall totals.

Pressure patterns are related to patterns of wind and precipitation, but the latter is more affected by local topography and other local peculiarities. Thus, pressure maps are often shown to describe the circulation. Wind is nearly parallel to the isobars (lines of constant pressure) in mid-latitudes, but blows across the isobars near the Equator

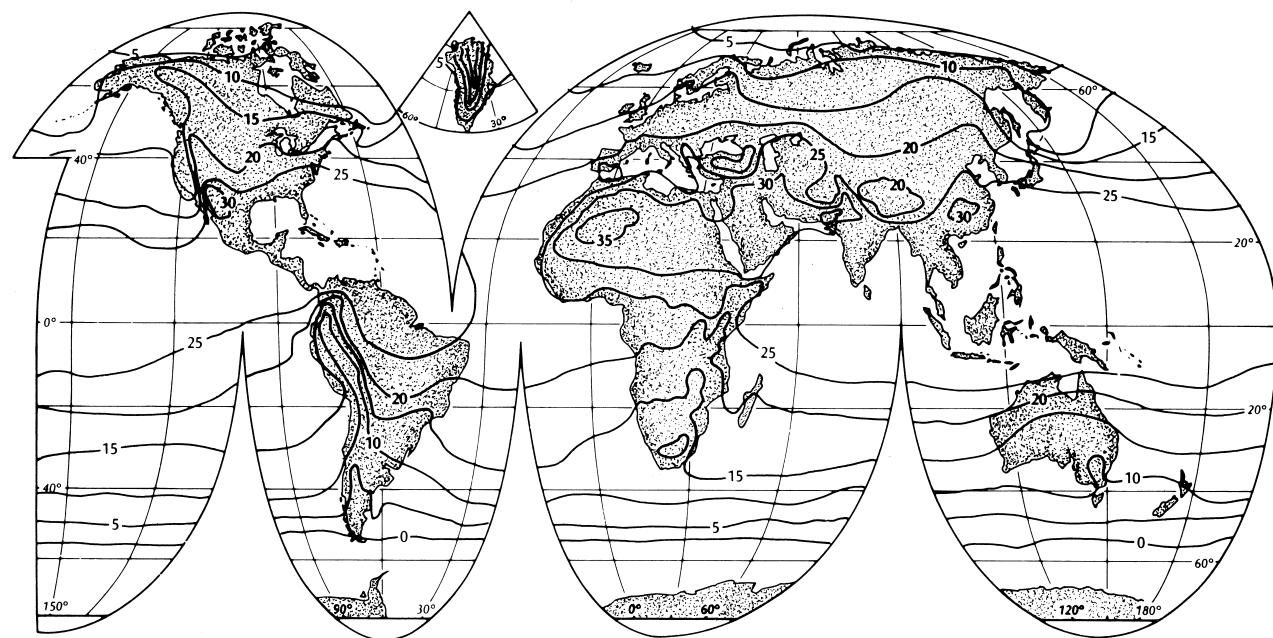


FIGURE 3 Distribution of average July temperature over the world. [From Oliver, J. E., and Hidore, J. J. (1983). "Climatology: An Introduction," by permission of Merrill, Columbus, OH.]

toward low pressure. In the Northern Hemisphere, if one puts one's back to the wind, low pressure is to the left. The reverse is true in the Southern Hemisphere. Close spacing of isobars indicates strong winds and vice versa.

Winds tend to converge into low-pressure centers (lows). Therefore, the air rises there, yielding precipitation. High-pressure ridges indicate sinking air and generally little precipitation over the sea level

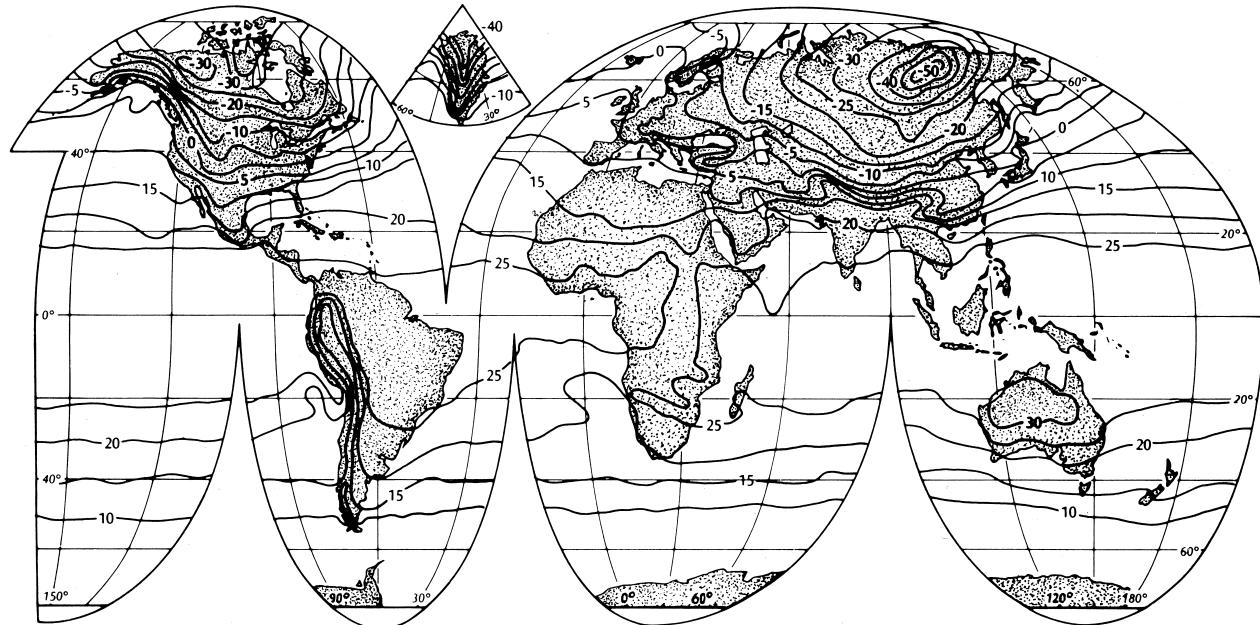


FIGURE 4 Distribution of average January temperature over the world. [From Oliver, J. E., and Hidore, J. J. (1983). "Climatology: An Introduction," by permission of Merrill, Columbus, OH.]

pressure pattern changes little between seasons, except for a northward movement in summer.

North and south of the Equator, the winds are generally from the east and are called the trades. In the Northern Hemisphere, the trades blow from the northeast; to the south, from the southeast. Thus, a convergence zone is formed near the Equator, the Intertropical Convergence Zone (ITCZ). Here air tends to rise, and precipitation is plentiful. It is a region of tropical rain forests and is also called the doldrums. Near latitude 30° , there are strong pressure ridges over the oceans with little wind and negligible precipitation. This is the belt of the horse latitudes, where most of the world's deserts are found. Between latitudes 30° north and south and 60° north and south lies the region of prevailing westerlies. Here, winds have more often westerly than easterly directions, and pressure systems generally move from west to east. This is a region of active weather, where cold air masses from polar regions meet warm air masses from the tropics at fronts. New low-pressure centers are formed at fronts, and thus precipitation is quite plentiful in this region. Near latitude 60° is the location of permanent, intense low-pressure systems with especially stormy and wet weather.

The permanent pressure and wind systems are most clearly visible over the ocean. Over the continents, there is seasonal reversal of pressure patterns. In the winter, with cold, heavy air masses over the continents, high pressure predominates in these areas, with diverging winds (clockwise in the Northern Hemisphere, counterclockwise in the Southern Hemisphere). Winter precipitation is light. In summer, pressure over the continents is low, with pre-

cipitation in midcontinent at a maximum. However, precipitation then is more spotty, in the form of showers and thunderstorms. Local topography affects some of the precipitation.

At the edges of the continents, then, there often is a reversal of winds from summer to winter; for example, in East China, winds are from the northwest in winter and southeast in summer. Such a reversal is called a monsoon and is often associated with an equally dramatic change in precipitation patterns. For example, the summer monsoon in India drives hot, moist air against the Himalayas, causing heavy precipitation. In contrast, the offshore, cold and dry winds of winter lead to little precipitation.

Although pressure patterns give an indication of general precipitation patterns, precipitation regions are often indicated separately, since not only pressure distribution, but also topography and ground characteristics contribute to the variability of precipitation. [Figure 5](#) gives a more detailed picture of global precipitation climatology.

Sometimes, climatologists are interested in combinations of variables—for example, the simultaneous occurrences of certain temperature and precipitation ranges, which make certain types of life possible. This leads to the definitions of climate classes, the distribution of which around the globe is then shown. The most famous of these is the Köppen classification system, the details of which are given in many climatology texts.

Although regional climatology has traditionally concentrated on conditions close to the ground, upper-air climatology has increased in importance in the second half of the 20th century. There are several reasons

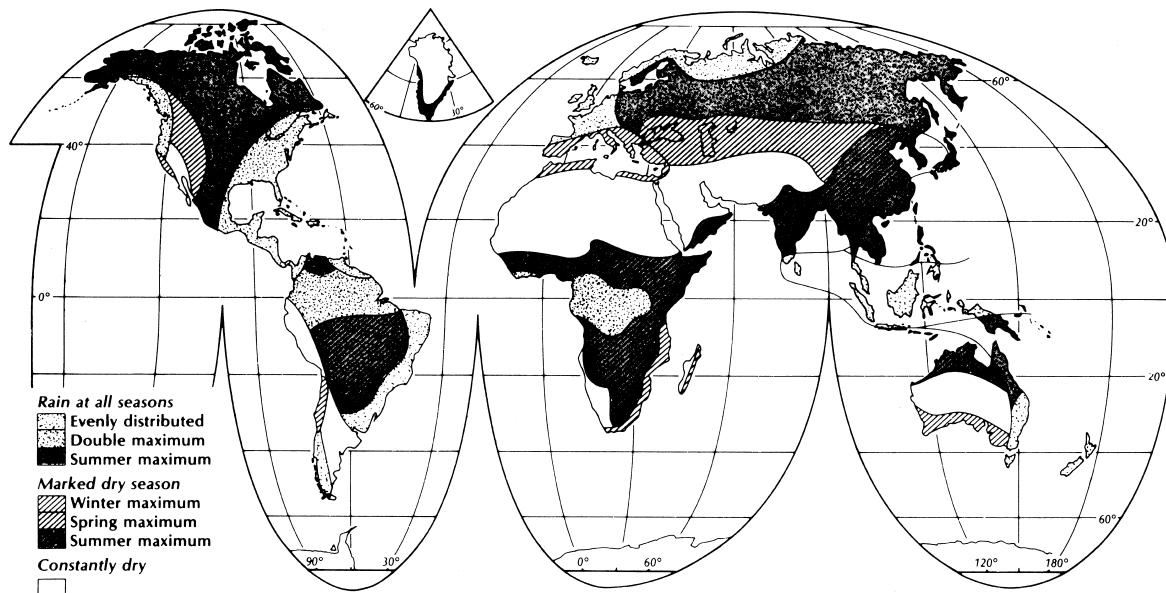


FIGURE 5 Generalized map showing seasonal distribution of precipitation.

for this: increased availability of observations; the importance of upper-air flow for steering surface storms and for weather forecasting generally; and increased air traffic.

Pressure falls more rapidly with height in cold than in warm air. Hence, up to ~ 20 km height, pressure tends to be lower at the poles than at the Equator. The pressure differences are largest at about tropopause height and decrease above. This is due to the fact that the horizontal temperature differences reverse near the tropopause.

In the upper air, friction is unimportant. Newton's second law shows that, in unaccelerated air, the wind is parallel to the isobars, with low pressure to its left in the Northern Hemisphere and to its right in the Southern Hemisphere.

[Figure 6](#) shows the equivalent of a mean January pressure map at an average of 5500 m height. Actually, it is a map for conditions at 500 mbar; about half the atmosphere is above that surface and half is below. This surface is not horizontal. Here, the lines are "contour lines," of equal geopotential height, which, for practical purposes, equals ordinary geometric height. Heights are given in tens of meters. It turns out that these contour lines have the same relation to the direction of frictionless, unaccelerated flow as isobars on horizontal charts. Also, the closer the contours, the stronger the wind. Hence, [Fig. 6](#) shows that the flow is generally from west to east, except near the Equator and the poles. This is why airplanes generally travel faster eastward than westward.

The contours are not exactly east–west, and thus average winds are not everywhere exactly from the west but have small components from the north or south. Such deviations from a strict westerly flow are associated with east–west variations of temperature in the lower atmosphere. Above 3 km, isotherms are generally approximately parallel to the contours.

Summer circulations tend to be weaker than winter circulations, since north–south temperature variations are weaker in summer. Hence, surface pressure systems move eastward more slowly in summer than in winter, and so do aircraft. Otherwise, wind patterns aloft are similar in winter and summer, except that the westerlies do not extend quite so far to the south.

As we get to the top of the troposphere (the tropopause), winds get stronger, without change of pattern. Of course, this description covers average flow only, but even winds on individual days show the same dominant west–east flow; however, the strongest winds may occur at different latitudes and may be strongly concentrated, especially above surface fronts (jet streams). Wavelike perturbations may have much larger amplitudes, with occasional closed circulation patterns even in middle latitudes.

IV. MICROCLIMATOLOGY

Regional climatology can delineate only the average climate over large areas, while local climates can be extremely varied due to local topography, local water–land and ice–snow distribution, or land use. The area of microclimatology is large, and huge books, most notably the classical work by Geiger, have been written about the subject. We shall give only a few examples here.

Even small hills influence the temperature patterns dramatically, particularly on nights with light winds. The temperatures over the slopes of shallow valleys may be 10°C or more lower than at the same heights over the center of valleys, so that cold air drains into the valley. Thus, even on calm nights there is downslope flow and general drainage along the valleys. In general, winds are channeled by valleys, so that winds in valleys often deviate from winds on nearby plateaus.

Even small water bodies moderate temperatures, particularly on light-wind winter nights. Before the water bodies freeze, the temperatures at their shores remain near freezing even though temperatures only a few kilometers away may be 10°C , even 20°C colder.

Of special interest in the 1970s was the impact of cities in local climate. First, there is the heat-island effect. Surface temperatures in cities are always higher than those in the surrounding countryside. The difference is largest in big cities, depending approximately on the fourth root of the population; it is largest at night, in winter, and with light wind speeds, depending inversely on the square root of wind speed. Cities produce a warm plume of air, slightly elevated, downstream. This is responsible for a local maximum of showers and thunderstorms on the lee side of cities.

Also, as part of microclimatology, we should mention the extremely rapid vertical variation of most meteorological variables close to the ground (except for pressure). The greatest variation is in the lowest meter, and then the variables always change more slowly. The reason is that the air very close to the surface is poorly mixed because the turbulent "eddies" there are small. In fact, in the lowest millimeter or so, eddies have no room, and vertical exchange is through molecules, which produce little transport because of their small mean free paths. Only in the area of vertical transport of momentum (wind) over rough terrain is turbulent transfer possible right into the ground. As an example of extreme vertical variation, it is possible for the surface temperature of a black asphalt road to be 80°C , whereas it is only 45°C above the surface. Similarly, cold layers can form on the ground on cold, light-wind nights. The wind speed is zero just at the surface, but it may rise to

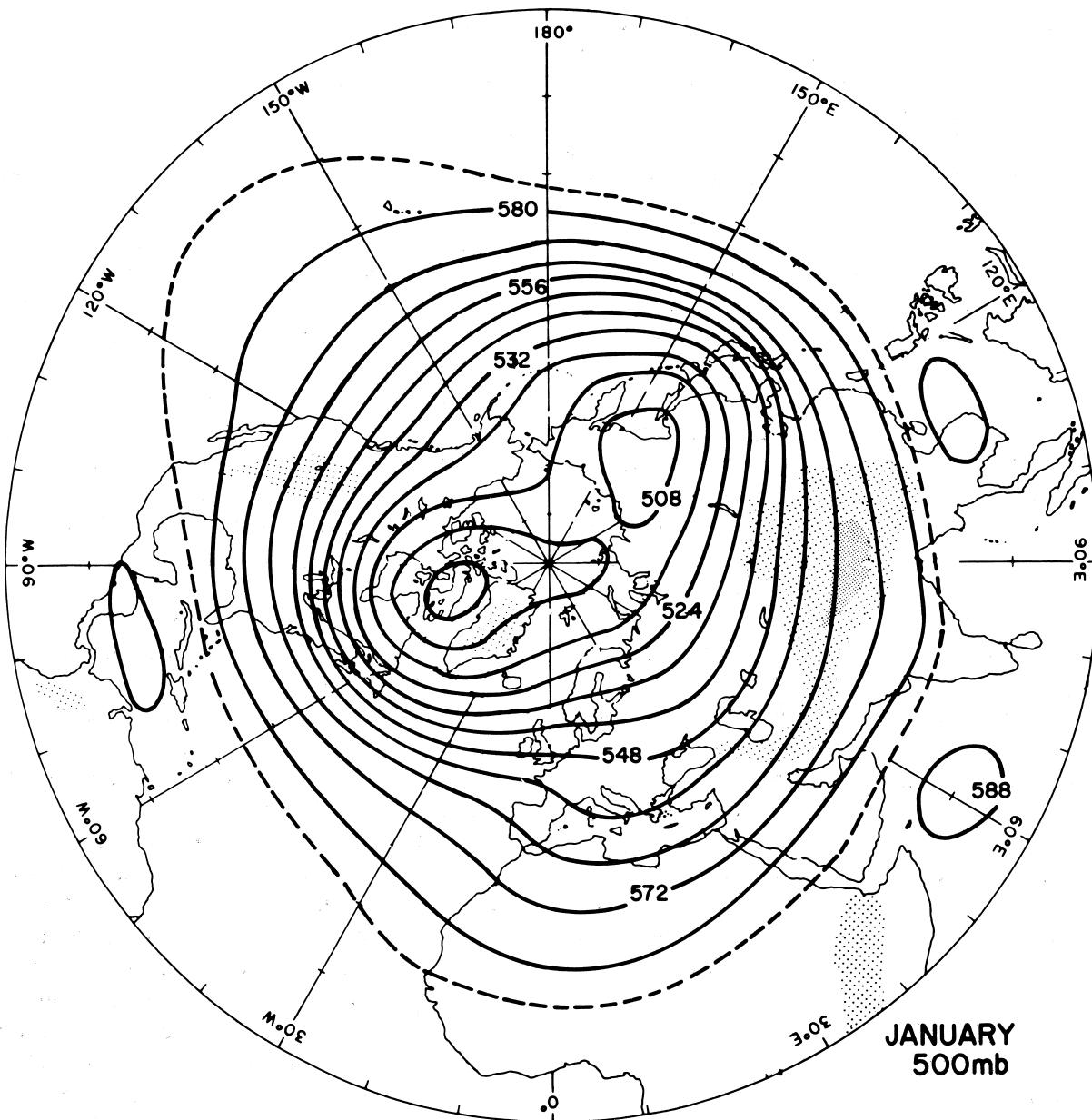


FIGURE 6 Distribution of mean January geopotential height (almost the same as geometric height) above sea level at 500 mbar. Contours labeled in tens of meters. [Based on Palmen, E., and Newton, C. W. (1969). "Atmospheric Circulation Systems," copyright Academic Press, New York.]

many meters per second in the first meter and then vary gradually.

V. CLIMATE AND GENERAL-CIRCULATION MODELS

The basic purpose of climate models originally was to explain features of current climate and general circulation

in terms of the geometry of the earth-sun systems and basic physical principles. Once this was accomplished, it was possible to experiment with climate models by changing various input parameters to account for the features of past climates and to speculate about future climates.

Relatively simple "climate models" are concerned entirely with explaining the temperature distribution; the effect of circulation, if included at all, is prescribed

in terms of the temperature distribution. Such climate models range from simple zero dimensional (averaged over the whole atmosphere) to models allowing for vertical and horizontal variations of input parameters. The models differ in the way boundary conditions and physical processes are handled. Sometimes clouds are prescribed, or they may be generated by the models. Ocean temperatures may be given, or the atmospheric model may be combined with an oceanic model. One of the important effects in many climate models is the ice-albedo feedback: If the model predicts a cooling, the ice sheets expand, causing increased albedo and more cooling—a “positive” feedback. In early models, this feedback was so severe that only a small cooling led to an ice-covered earth. In spite of such excesses, climate models have explained most features of the vertical temperature distributions and effects of land-water differences and topography.

General-circulation models (GCMs) attempt to duplicate the distribution of wind as well as of temperature and moisture. They are usually three dimensional; however, the earliest models had little resolution in the vertical; for example, the pioneering model by Norman Phillips in 1956 consisted of only two layers. In the meantime, much more complete models have been developed, many requiring the fastest computers available.

In the models, generally seven basic equations are solved for the seven basic variables of meteorology: pressure, density, temperature, moisture, and three velocity components. The seven equations are the gas law, the first law of thermodynamics, equations of continuity for air and water substance, and three components of Newton's second law. If the stratosphere is modeled explicitly, ozone concentration must be added as a variable and at least one equation must be included to describe the ozone budget. Since the ozone budget depends on concentrations of many other trace gases, many more variables and equations are sometimes added.

Typically, the initial state of the atmosphere is simple, with no motion. Then radiation is introduced, and the models are integrated numerically for several model years. Eventually, the models acquire many of the properties of the actual atmosphere.

In the United States alone, there are at least five centers studying the properties of GCMs. The models differ in horizontal and vertical resolution, in vertical extent, in whether clouds are assumed or generated by the model, and in the characteristics of the lower boundary. Some of the atmospheric models are coupled with GCMs of the ocean. Such coupled models produce problems because of the differences in time and space scales of important atmospheric and oceanic circulation characteristics. Other

differences involve the prescription (parameterization) of motion on scales too small to be resolved but that nevertheless affect the resolved motions by their capacity to mix air with different properties.

The first real GCM in 1956 proved successful in explaining some of the basic features of the general circulation and of climate: the basic three-cell structure including the strong Hadley cell with its rising motion at the ITCZ, the sinking near latitudes of 30° , and the equatorial flow at the surface in between. In addition, the model duplicated a persistent characteristic of atmospheric flow: the nonuniform temperature gradients, with relatively homogeneous air masses separated by sharp fronts and, associated with the fronts, narrow streams of strong west winds aloft, the jet streams.

Since Phillips's original model, more recent GCMs have become more realistic in many details: Vertical and horizontal resolution has increased; oceanic models have been combined with the GCMs; realistic hydrological cycles have been added, as have realistic prescriptions for changing clouds and snow cover; and the models have been extended to include the stratosphere with its relatively high ozone concentration.

Although the GCMs do not represent climate perfectly, they are sufficiently realistic to make experimentation with past and future climates very promising indeed.

VI. PAST CLIMATES

Climate has varied in the past on many time scales. There have been long periods (~ 50 million years) of relatively undisturbed climates, generally warmer than the current climate, interrupted by shorter periods (a few million years) of quite variable climates. For about the past 2 million years, we have been in such a disturbed period, with ice ages alternating with somewhat milder interglacials. The last ice age ended $\sim 10,000$ years ago; since then there have been minor climate fluctuations—for example, a warm period ~ 1000 AD and a “little ice age” in the second half of the 17th century. In the 20th century, the atmosphere warmed $\sim 0.5^{\circ}\text{C}$ from 1880 to 1940 and cooled from 1940 to at least 1970.

Many reasons for historical climate changes have been suggested: changes in the earth's crust (e.g., migration of the continents, mountain building, and volcanic eruptions); changes in atmospheric composition, particularly the amount of carbon dioxide; changes in earth-sun geometry; and changes in solar heat output.

Of course, it is quite possible that some of these factors worked together, but most important climatic variations on different time scales do not have the same causes. At least,

past climates have not been affected by human activity; the same cannot be said about future climates.

Climate changes on the longest time scales considered here (50 million years) almost certainly were affected by changes in the earth's crust; for example, it is possible that the active periods coincided with periods of mountain building; and perhaps, ice ages occurred when there were landmasses near the poles.

The cause of climate changes *within* the disturbed periods, which have time scales of 10,000 to 100,000 years, is now fairly well understood. We have good records from undersea and underice cores, and we have a good quantitative theory that agrees with these records. This theory was proposed in 1926 by Milankovich but has been taken seriously only since about 1970. It is known from quantitative celestial mechanics (and from observations) that several characteristics of the earth–sun geometry are variable. For example, the earth's axis does not point in the same direction in space at all times, but precesses with a period of 26,000 years. At the same time, the major axis of the earth's orbit spins in the opposite direction. The net result is that, about every 20,000 years, the sun and earth are closest in northern winter (as they are now). So right now, northern winters are relatively mild and southern winters are relatively cold. Ten thousand years from now, this situation will be reversed. A second factor is the eccentricity of the earth's orbit. Right now, the difference between aphelion and perihelion distance is $\sim 3\%$, but the distance difference has been as large as 10% and at other times the orbit has been circular. The period of this variation is $\sim 100,000$ years. Finally, the angle between the earth's Equator and the earth's orbit has varied from about 22° to 24.5° , with a period of $\sim 40,000$ years. When this angle is largest, seasonal contrasts are largest.

None of these factors affect significantly the total radiation received by the earth, but they all affect seasonal contrast. Milankovich suggested that ice ages result when there are cool summers and mild winters, for snowfall can be heavy in mild winters, and less snow melts in cool summers.

The chronology of observed ice ages in the past million years agrees well with predictions of the Milankovich theory; in particular, the various periodicities have been found in the geological records. Climate models have some difficulties with the phases of the largest cycles, but there are satisfactory theories to explain these difficulties. As a result, the basic theory is generally accepted as the explanation for climate changes on the scale of 10,000 to several hundred thousand years.

There is no generally accepted hypothesis to explain more recent climate changes. Solar activity has periods of 11, 21, and 80 years, but statistical analyses of relationships between solar activity and atmospheric varia-

tions have not been convincing, even though a period of low sunspot activity coincided with the little ice age, and U.S. western droughts are correlated somewhat with the sunspot cycle. However, there are no known mechanisms connecting solar activity with changes of climate.

There is some indication of cooling at the surface following major volcanic eruptions, but the effects are not large and it is not clear whether volcanoes produce long-lasting changes. In short, there is no agreement concerning the causes of recent climate changes.

VII. FUTURE CLIMATES

Since the reliability of the Milankovich theory has been established from past records, it should also provide some indication of future climate change, in the absence of complications produced by human activity. According to this theory, a cooling trend has already set in, will intensify several thousand years from now, and will produce peak glaciation in $\sim 20,000$ years.

Superimposed on this scenario are natural short-period fluctuations, which we cannot predict because we do not understand their causes, and man-made changes. The most important of these are produced by increased CO₂ in the atmosphere, which is due to burning of fossil fuels and clearing of forests. Increases in atmospheric CO₂ have been measured and amount to about half of the CO₂ known to be emitted into the atmosphere; the rest is presumably absorbed by the ocean. Some time after the year 2050, the amount of CO₂ in the atmosphere is expected to be double that before the onset of the industrial age.

Increasing atmospheric CO₂ concentration enhances the natural “greenhouse effect,” thus causing surface warming. As soon as warming begins, increasing evaporation increases the amount of water vapor in the atmosphere, producing even more warming. Many climate models have been run with increased CO₂ concentrations, typically double the normal. A consensus is starting to emerge among modelers that doubling CO₂ concentration results in a global average warming of 2° to 3°C and polar warming of more than twice that amount. Warming due to increased CO₂ is not yet large enough to be observed in view of irregular temperature variations. It is expected that CO₂ warming will be detected in the future, if the models are correct.

However, most of the models accounting for increased CO₂ are unrealistic in the specification of clouds and possibly in the lower boundary conditions. For example, it is not known whether warming will increase cloudiness or change the physical characteristics of clouds. In particular, clouds may thicken and reflect more sunlight, limiting the warming effect.

Other long-term atmospheric impacts resulting from human activity are also being monitored. Of note is the role of chlorofluorocarbons (CFCs), multiuse chemicals best known as refrigerants. Chemical breakdown of CFCs in the ozone layer of the stratosphere results in a chemical reaction leading to the diminution of ozone. The decrease modifies the role of ozone in screening solar radiation, resulting in an increase in shortwave ultraviolet radiation reaching the earth's surface.

Few environmental problems have received the media attention of the problem of global warming. The continued rise of atmospheric CO₂ content, together with the warmth of the 1980s and particularly the 1990s, which had the three warmest years on record (1998, 1997, 1995), caused some climatologists to suggest that the global warming suggested by computer models is already underway. Forecasts of increasing warmth, melting ice caps, rising sea levels, and modified environments have been widely reported. Given such forecasts, attempts have been made to introduce legislature to limit the burning of fossils fuels and, hence, CO₂ production. Notably, the Earth Summit in Rio de Janeiro, Brazil, marked an important global moment of recognition of the significance of potential human impacts on climate. Further, the Climate Conference in Kyoto, Japan, in 1998 provided a substantive basis for the reduction of greenhouse emissions and a procedure for monitoring the success of mitigation efforts. The latter document lists the Intergovernmental Panel on Climate Change (IPCC) as the authority for all scientific decisions. No definitive action has yet been taken.

The relative lack of political action, particularly by the developed nations, reflects the economic costs involved and the varying interpretations of the global warming signal. Models do not provide a uniform interpretation of the future. While most climatologists agree that the earth will experience a warming trend, there is disagreement on how much the temperature will rise and when the impacts will be felt by the human population. While most models predict modest warming, an alternative scenario suggest that high latitude warming could initiate a cooling episode. The mechanism for this scenario is fairly complex, involving adjustments in the North Atlantic Deep Water oceanic circulation, and overall implications are unclear. As global models become more refined, the answers to the problem will become clearer.

In contrast to the paucity of action concerning the CO₂ problem, appreciable headway has been made in combatting the ozone depletion problem. Most industrial nations that produce CFCs have agreed to curb production and examine substitute chemicals. Such action will take time, however, and the depletion of ozone continues to be of concern. Atmospheric scientists researching the problem are now examining the dynamics that cause "holes" that periodically occur in the polar ozone layers.

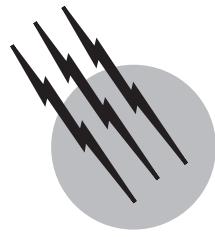
Given the current interpretation of climate, speculations concerning future climates are possible. If the GCMs are correct and no preventative measures are taken, substantial global warming will take place over the next few hundred years. Eventually, perhaps after 1000 years, most fossil fuels will have been used and excess CO₂ will be taken up by the oceans. The Milankovitch mechanisms will then take over to lead, potentially, to a global cooling. Climatic change has occurred in the past and will occur in the future.

SEE ALSO THE FOLLOWING ARTICLES

GREENHOUSE EFFECT AND CLIMATE DATA • GREENHOUSE WARMING RESEARCH • HYDROLOGIC FORECASTING • METEOROLOGY, DYNAMIC (TROPOSPHERE) • METEOROLOGY, DYNAMIC (STRATOSPHERE) • POLLUTION, AIR • WEATHER PREDICTION, NUMERICAL

BIBLIOGRAPHY

- Barry, R. G., and Chorley, R. J. (1998). "Atmosphere, Weather and Climate," Routledge, London.
Hidore, J. J., and Oliver, J. E. (1993). "Climatology: An Atmospheric Science," Macmillan Co., New York.
IPCC. (1991). "Climate Change: The IPCC Scientific Assessment" (Houghton, J. T., Jenkins, G. J., and Ephraums, J. J., eds.), Cambridge Univ. Press, New York.
Linacre, E. (1992). "Climate Data and Resources: A Reference and Guide," Routledge, London.
Oke, T. R. (1987). "Boundary Layer Climates," Methuen, New York.
Oliver, J. E., and Fairbridge, R., eds. (1987). "The Encyclopedia of Climatology," Van Nostrand-Reinhold, New York.
Schneider, S. H., ed. (1996). "Encyclopedia of Climate and Weather," Oxford Univ. Press, New York.
Thompson, R. D., and Perry, R., eds. (1997). "Applied Climatology: Principles and Practice," Routledge, London.



Cloud Physics

Andrew J. Heymsfield

*National Center for Atmospheric Research**

- I. Cloud Formation Mechanisms and Cloud Classification
- II. Cooling of Moist Air
- III. Formation of Cloud Droplets
- IV. Physics of the Growth of Cloud Droplets
- V. Observations of Cloud Droplets and Drops
- VI. Ice Formation Mechanisms
- VII. Observations of the Growth Mechanisms of Ice Particles
- VIII. Physics of the Growth of Ice Particles
- IX. In-Cloud Observations of Ice Particles
- X. Effects of Clouds on Climate
- XI. Cloud Seeding
- XII. Concluding Remarks

GLOSSARY

Adiabatic process Thermodynamic change of state of a system in which there is no transfer of heat or mass across the boundaries of the system.

Cirriform cloud Cloud with a relatively transparent and white or silky appearance that forms at high altitudes and is composed of small particles, typically ice crystals.

Cloud condensation nucleus (CCN) Small, solid particle, typically containing a salt, on which condensation of water vapor begins in the atmosphere.

*The National Center for Atmospheric Research is supported by the National Science Foundation.

Cloud droplet Particle of liquid water from a few micrometers to $\sim 200 \mu\text{m}$ in diameter.

Cloud seeding Any technique carried out with the intent of adding to a cloud certain particles that will alter the natural development of the cloud.

Cumuliform cloud Cloud whose principal characteristic is vertical development in the form of rising mounds, domes, or towers.

Graupel Spherical or conical ice particle about 2–5 mm in diameter, consisting of a white or opaque snowlike structure, formed by the collection of cloud droplets.

Hailstone Spherical, conical, or irregularly shaped ice particle, from ~ 5 mm to more than 5 cm in diameter, formed by the collection of cloud droplets and drops.

Ice crystal Any one of a number of macroscopic crystal forms in which ice appears in the atmosphere.

Ice nucleus Any particle that serves as a nucleus for the formation of ice crystals in the atmosphere.

Stratiform cloud Cloud of extensive horizontal development.

Supercooled cloud Cloud composed of liquid water droplets at temperatures below 0°C.

Supersaturation Condition existing in a given portion of the atmosphere when the relative humidity is greater than 100%.

CLOUD PHYSICS is a discipline within meteorology concerned with the properties of atmospheric clouds and the processes that operate within them, the diversity of phenomena intrinsic to natural clouds, the interactions of clouds with the atmosphere, and the effects of clouds on climate. The discipline covers the range from single clouds to large-scale weather systems and even weather on a global scale.

Cloud physicists draw on the well-developed sciences of chemistry, physics, and fluid dynamics to study these phenomena. Such topics as the thermodynamics of moist air, the physics of the growth of water droplets and ice particles, radiation, effects of clouds on climate, electrification, and chemical conversion processes are all part of this discipline. Major research tools include computers for numerical simulation and aircraft and radars for observation, along with wind tunnels and cold rooms for the study of the properties of cloud and precipitation particles.

I. CLOUD FORMATION MECHANISMS AND CLOUD CLASSIFICATION

A complete description of the many genera, species, and varieties of clouds is given in the “International Cloud Atlas” (1956) of the World Meteorological Organization. This international classification evolved from the experience of ground observers over many years, depending primarily on appearance. From a physical point of view, the distinctions among types of clouds arise from the vertical motion characteristics that produce them and from their microphysical properties—the presence or absence of ice particles and the sizes and concentrations of each type of cloud particle.

There are three fundamental classes of clouds: cirrus, cumulus, and stratus. *Cirrus* are high clouds with a silken appearance, because they are composed of ice crystals. *Cumulus* are detached, dense clouds that rise in mounds or towers from a level base. *Stratus* is the name given to an extensive layer or flat patches of low clouds showing hardly any well-defined detail. These names are some-

times used in conjunction. For example, when cirrus is in layered form, it is termed cirrostratus; a low-level layer cloud that is broken up into a wavy pattern is called stratocumulus; and so on. Similar clouds at intermediate levels are called altocumulus, and a thick-layer cloud at these levels is called an altostratus. A class of lesser fundamental importance comprises the lenticular cloud, having a lens or almond shape.

Most types of clouds are formed as a result of vertical motions produced in the following ways:

1. *Widespread gradual lifting.* Up glide motion of air at a frontal surface occurs in the cyclones of temperate latitudes and gives rise to expansive layers of deep and often layered altostratus and nimbostratus clouds. The vertical component of air velocity is of the order of a few to a few tens of centimeters per second, and the lifting results in steady precipitation of long duration.

2. *Widespread irregular stirring.* When air is cooled at the ground, fogs may form. Over land at night, the cooling may be due to the radiation of heat from the ground, but fogs also occur over land and sea when air flows slowly into regions of lower surface temperature.

3. *Convection.* Heating at the ground either by sunshine or when cool air undercuts warmer air can cause masses of air from the surface layers to ascend through a relatively undisturbed environment, often producing clouds of cumuliform type. Above the level of the cloud base the liberation of latent heat due to the conversion of water vapor to droplets usually increases the buoyancy and vertical velocity of the rising masses. In settled weather, “fair weather” cumulus clouds are well scattered and small, with horizontal and vertical dimensions of only 1 or 2 km and vertical motions from about 1–5 m sec⁻¹. In disturbed weather, cumulus congestus and cumulonimbus (thunderhead) clouds form. The tops of the latter can reach up to 20 km above the ground, spreading out into a flat-topped, anvil shape. Thunderclouds can have updrafts of more than 40 m sec⁻¹ in the most extreme cases and can produce heavy rain, hail, and tornadoes.

4. *Orographic lifting.* When an air mass moves against a mountain barrier, some of the air is forced to rise. This can result in an extensive sheet of deep stratiform and cumuliform clouds. Lenticular clouds can also form high above mountains and in their lee (downwind locations) within mountain-generated waves.

II. COOLING OF MOIST AIR

A. Water Vapor in the Atmosphere

The amount of water vapor present in the atmosphere is dependent in a complex way on (1) the amount that enters the atmosphere by evaporation and sublimation, (2) its

transport by air motions throughout the troposphere (lower portion of the atmosphere) and lower stratosphere, and (3) the amount precipitated in the form of rain, snow, and hail.

Two factors account for the observed decrease in the concentration of water vapor (termed the *water vapor mixing ratio*, grams of water vapor per gram of air) with height. First, the earth's surface is the primary source of water vapor. Second, the air temperature decreases with height in the troposphere. Since the maximum possible mixing ratio, termed *saturation*, decreases with decreasing temperature, water is squeezed out as parcels ascend in the atmosphere.

B. Cooling of Air in the Absence of Water Vapor Condensation

When a parcel of air is lifted in the atmosphere, it expands and cools. Heat may be added to the parcel through such effects as radiation and friction, but, in most cases, the resulting changes in the temperature of the parcel are secondary to the expansion process. It is a reasonable and useful idealization to assume, then, that the expansion is *adiabatic*, that is, that there is no transfer of heat or mass into or out of the parcel, and it is a reversible process. Furthermore, in the absence of water vapor condensation, no heat is added within a rising parcel. From the first law of thermodynamics for an ideal gas, the decrease in temperature with height in the absence of condensation, termed the *dry adiabatic lapse rate*, works out to be $\sim 10^\circ\text{C km}^{-1}$, or about a cooling of $\sim 1^\circ\text{C}$ for every 100 m of lift. The cooling rate in a rising parcel cannot be higher than dry adiabatic.

C. Cooling of Air with Water Vapor Condensation

If a parcel of air being lifted dry adiabatically achieves a relative humidity of 100% (i.e., becomes saturated), water vapor condenses to form a cloud, and latent heat of condensation is released. (This discussion does not consider the latent heat that can be released when ice is present.) Thus, the parcel now cools at a rate less than dry adiabatic, the rate depending on whether all or part of the condensate stays within the parcel. If all of it remains, the first law of thermodynamics can again be used to derive the moist adiabatic lapse rate. The resulting lapse rate is not constant as is the dry adiabatic lapse rate, but is dependent on pressure and temperature. For 1000 kPa and 20°C , this lapse rate is $\sim 4^\circ\text{C km}^{-1}$, while at the same pressure and a temperature of 0°C , it increases to $\sim 6^\circ\text{C km}^{-1}$. At temperatures below about -30°C , the moist adiabatic lapse rate approximates the dry adiabatic rate.

D. Cooling with Entrainment

Numerous in-cloud measurements, along with theoretical and laboratory studies, have indicated that *entrainment* takes place; that is, air in the environment surrounding a parcel of air is mixed into the parcel and becomes part of the rising current. A rising parcel of cloudy air into which dry air is entrained cools at a faster rate than moist adiabatic, because heat is required (1) to evaporate sufficient condensate, increasing the mixing ratio of the air to saturation, and (2) to warm the air from its original temperature to the parcel temperature. The lapse rate in an entrained parcel of air falls somewhere between the moist and dry adiabatic rates.

The adiabatic model of cooling in a rising parcel has been modified by several cloud physicists to take entrainment into account. These researchers have proposed that entrained air originates primarily either from the sides of clouds (lateral entrainment) or from their tops (cloud-top entrainment). The latter model is gaining fairly widespread acceptance.

E. Summary of Cooling Processes During Cumulus Cloud Formation and Resulting Thermal Instability

The vertical temperature distribution in a rising column of air associated with a cumulus cloud is given by the solid line in Fig. 1. The segment *AB* represents dry adiabatic ascent, and the in-cloud segment *BD* represents moist ascent with entrainment. In the absence of entrainment, temperatures would be moist adiabatic, given by dotted line *BC*. The environmental temperature distribution is given by the dashed line. Along segment *AB*, vertical velocities may be of the order of a few meters per second. Along segment *BD*, the parcel becomes increasingly warmer relative to the environment and thus more buoyant, possibly increasing the vertical velocity more than 10 m sec^{-1} . Beyond point *D*, the parcel is negatively buoyant, and the velocities decrease to 0 m sec^{-1} , producing the cloud top.

III. FORMATION OF CLOUD DROPLETS

As a parcel of air is cooled toward saturation, the relative humidity approaches 100%, and water vapor begins to condense, or *nucleate*, on small particles of airborne dust, or *cloud condensation nuclei* (CCNs). These small particles usually contain a soluble component, often a salt such as sodium chloride or ammonium sulfate. CCNs of different sizes and compositions are present at each position in the atmosphere. Some of the nuclei become wet at relative humidities below 100% and form haze, while the

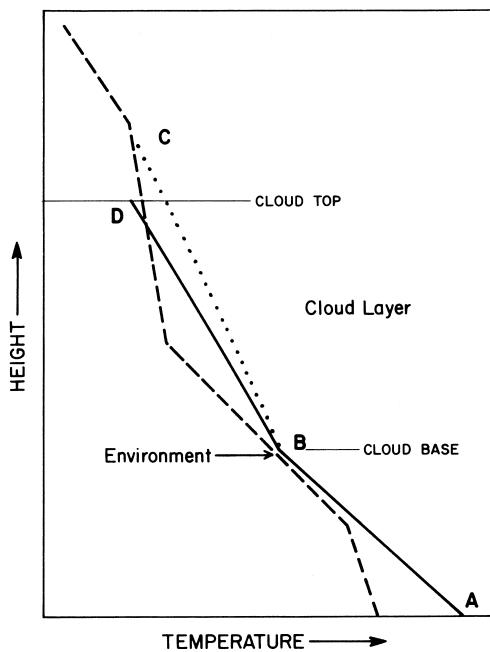


FIGURE 1 Temperature vs height distributions. Solid line, in a rising parcel of air: AB , below cloud base, dry adiabatic; BD , cloud base to cloud top, entrained ascent. Dotted line: BC , moist adiabatic. Dashed line: environmental temperatures.

relatively large CCNs are the most likely ones to grow and become cloud droplets.

This initial stage of droplet formation deserves a careful explanation. Over a flat, pure water surface at 100% relative humidity (saturation with respect to water), water vapor is in equilibrium, which means that the number of water molecules leaving the water surface is balanced by the number arriving at the surface. Molecules at water surfaces are subjected to intermolecular attractive forces exerted by the nearby molecules below. If the water surface area is increased by adding curvature, molecules must be moved from the interior to the surface layer, in which case energy is required to oppose the cohesive forces of the liquid. As a consequence, for a pure water droplet to be at equilibrium, the relative humidity has to exceed the relative humidity at equilibrium over a flat, pure water surface, or be *supersaturated*. The flux of molecules to and from a surface produces what is known as *vapor pressure*. The equilibrium vapor pressure is less over a salt solution than it is over pure water at the same temperature. This effect balances to some extent the increase in equilibrium vapor pressure caused by the surface curvature of small droplets. Droplets with high concentrations of solute can then be at equilibrium at subsaturation.

It follows from the surface energetics of a pure or solution droplet that once it achieves a certain “critical” radius, at a time at which the supersaturation in its environment

achieves a certain “critical” value, the droplet will grow spontaneously and rapidly.

Equations have been developed to express the equilibrium supersaturation of a droplet of a given radius in terms of the composition and size of the CCNs. A family of curves showing the fractional equilibrium relative humidity (relative humidity/100%) of water droplets containing differing masses of sodium chloride salt is shown in **Fig. 2**. A curve for pure water droplets is also shown. Asterisks at the peak of each curve show the critical radius and supersaturation of the droplets and indicate that the higher the mass of the salt, the lower the critical supersaturation. A family of curves can also be generated for other salts found in the atmosphere, such as ammonium sulfate.

Figure 2 can be used to understand the initial formation of droplets. One can see that a spectrum of droplet sizes will be produced in a rising parcel of air because of the diversity of sizes and composition of CCNs. When the supersaturation in the parcel has increased to such an extent that the droplets that have achieved their critical supersaturation are consuming more of the vapor than is being produced by cooling of the air, the supersaturation begins to decrease, and below a certain point no additional droplets are produced. In marine environments, where abundant sodium chloride nuclei of relatively large masses (about 10^{-15} – 10^{-14} g) are produced through sea spray, the condensed water deposits on the nuclei with large masses and the peak saturation ratio achieved in a parcel is comparatively low. In continental areas, where relatively few large sodium chloride nuclei are present, the condensed water deposits on nuclei with small masses

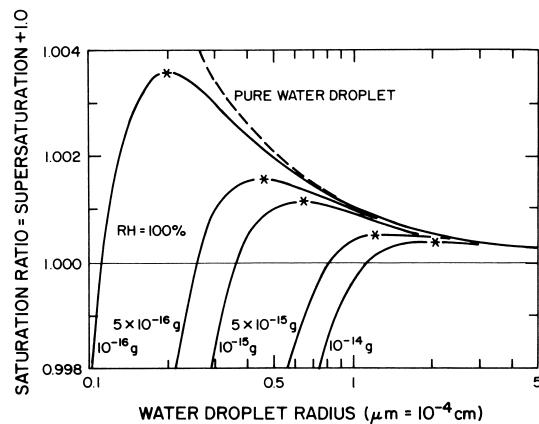


FIGURE 2 Equilibrium saturation ratio versus droplet radius for different masses of sodium chloride nuclei (solid lines). Asterisks show droplet radius where solution drop containing indicated mass of sodium chloride will continue to grow without further increase in the saturation ratio. Dashed curve is for a pure water droplet. RH, Relative humidity. [Adapted from Byers, H. R. (1965). “Elements of Cloud Physics,” courtesy of the University of Chicago Press and the author.]

(about 10^{-15} – 10^{-16} g) and higher supersaturation occurs. Given the same condensed liquid water content, maritime areas will have fewer and larger droplets than continental areas.

IV. PHYSICS OF THE GROWTH OF CLOUD DROPLETS

A. Diffusional Growth

After a solution droplet has been nucleated, it enters a stage of growth by diffusion of vapor to it. This growth is maintained as long as the saturation ratio, or supersaturation for it to be at equilibrium, is exceeded (see Fig. 2).

Consider a solution droplet of radius r in a supersaturated environment in which the concentration of vapor molecules at distance R from the droplet center is denoted by $n(R)$. This vapor diffuses toward the droplet and condenses on it. At any point in the vapor field the concentration of vapor molecules must satisfy the equation representing diffusion, assuming certain approximations that are not discussed here:

$$\nabla^2 n(R) = 0 \quad (1)$$

When $R = \infty$, n must have the value n_0 , the concentration of vapor at a great distance from the droplet. At the droplet's surface n must equal n_r , the equilibrium vapor concentration over the droplet surface. The solution to this equation is

$$n(R) = n_0 + (n_r - n_0)r/R \quad (2)$$

The flux of molecules, each of mass m , on the surface of the droplet is equal to $D(\partial n / \partial R)$, where D is the diffusion coefficient. With the vapor density in the environment being ρ_v , given by mn_0 , and at the droplet's surface ρ_{vr} , given by mn_r , the rate of mass increase of the droplet is

$$dm/dt = 4\pi r D(\rho_v - \rho_{vr}) \quad (3)$$

and since the droplet is spherical, with a radius r and a density ρ_L (1 g cm^{-3}), the time rate of radius increase is

$$dr/dt = D(\rho_v - \rho_{vr})/\rho_L r \quad (4)$$

Equations (3) and (4) can be expressed in terms of the supersaturation since the term ρ_{vr} can be found from the CCN composition and mass.

The growth histories of individual droplets at a constant supersaturation of 1% and an air temperature of 20°C have been calculated in Fig. 3 according to the principles used in deriving Eq. (4). Of considerable importance is that the drops with smaller initial radii catch up to the size of the drops with larger initial radii.

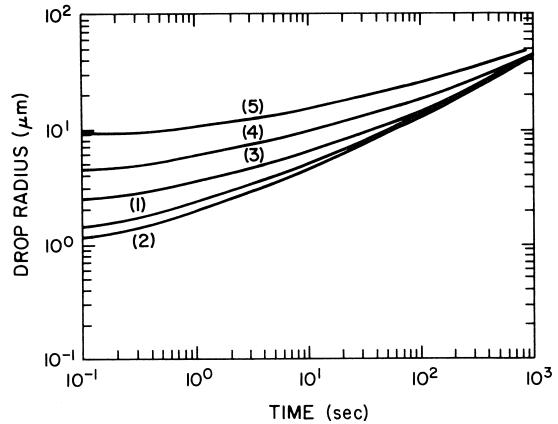


FIGURE 3 Diffusional growth rate of individual solution drops containing indicated salt mass as a function of time at 1% supersaturation and 20°C . (1) Sodium chloride, $m = 10^{-12}$ g; (2) ammonium sulfate, $m = 10^{-12}$ g; (3) sodium chloride, $m = 10^{-11}$ g; (4) sodium chloride, $m = 10^{-10}$ g; (5) sodium chloride, $m = 10^{-9}$ g. [From Pruppacher, H. R., and Klett, J. D. (1978). "Microphysics of Clouds and Precipitation," courtesy of D. Reidel Publishing Co. and the lead author.]

B. Growth through Droplet Collisions

Most of the earth's precipitation reaches the ground as "drops" of radius $100 \mu\text{m}$ or larger, many of which form in clouds whose tops do not extend to a level where ice particles are produced. Figure 3 suggests that drops cannot be produced solely through diffusion in any type of cloud. Growth through diffusion becomes very slow after a radius of $\sim 10 \mu\text{m}$ is reached, and the mechanism that then takes over consists of collisions and merging (coalescence) of cloud droplets.

A spectrum of droplet sizes is formed at each position in a cloud for reasons discussed previously, among others. Collisions between droplets, and between drops and droplets, can occur because droplets and drops of different sizes have differing fall velocities; large droplets or drops fall faster than smaller ones, overtaking and collecting some fraction of those lying in their paths. Electrical fields may promote additional collections.

As a drop falls, it collides with only a fraction of the droplets in its path because some are swept aside in the airstream moving around the drop. Also, some droplets that collide rebound and do not merge. The ratio of the actual number of collisions to the number for geometric sweep-out is called the *collision efficiency*, E , and depends primarily on the size of the collecting drop, R , and the sizes of the collected droplets, r . Collisions first occur for drops with radii as small as $10 \mu\text{m}$, though with small efficiency. Collision efficiency E is generally an increasing function of R and r and has values greater than 0.5 when R is greater than $100 \mu\text{m}$ for most cloud situations.

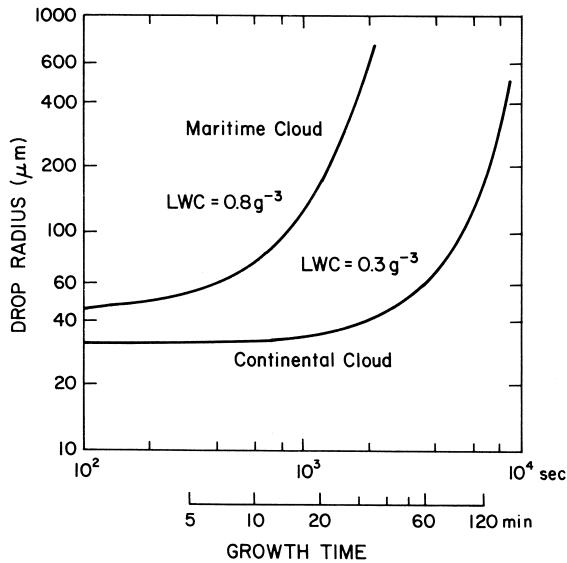


FIGURE 4 Size variation of a drop growing by collision and coalescence in a maritime- and a continental-type cloud. LWC, liquid water content. [From Braham, R. R. (1968). *Bull. Amer. Meteorol. Soc.* **49**, 343–353. Courtesy of the American Meteorological Society and the author.]

It follows from the nature of the geometric sweep-out process that the change in drop radius with time is approximately equal to

$$dr/dt = \bar{E} \times \text{LWC} \times V_T / (4\rho_{LR}), \quad (5)$$

where \bar{E} is the effective average value of collection efficiency for the droplet population, LWC is the liquid water content, and V_T is the fall velocity of the collecting particle. Figure 4 shows the calculated size variations of a drop growing by collision and coalescence in a cloud in a maritime environment (containing relatively few droplets, each of fairly large size) and in a continental environment (containing a relatively large number of droplets, each of relatively small size). Note that this growth by collection is much more rapid than diffusional growth (Fig. 3).

Even if a cloud is spatially homogeneous or “well mixed” with the same average droplet concentration throughout, there will be local variations in droplet concentrations. These follow the Poisson probability law. Equation (5) does not take into account the possibility of statistical fluctuations in the droplet spectrum but applies only to average droplet growth. Some “fortunate” drops fall through regions of locally high droplet concentrations, encountering more than the average number of droplets early in their growth and subsequently can grow more rapidly. Equations have been developed to take into account the statistical, or *stochastic*, nature of the droplet and drop growth process, and it is now recognized as crucial in the early stages of coalescence. It explains why some drops form relatively rapidly and why a distri-

bution of drop sizes is produced in a relatively uniform cloud.

Two major processes limit the growth of drops in otherwise favorable conditions. First, drops larger than ~ 0.1 cm that collide with drops greater than $300\ \mu\text{m}$ often break up and produce a multitude of smaller drops. Second, raindrops are limited to sizes of less than ~ 0.5 cm because at larger sizes they spontaneously break up from aerodynamic forces acting on their surface.

V. OBSERVATIONS OF CLOUD DROPLETS AND DROPS

A. In-Cloud Measurement Devices

Currently, a total of approximately 25 aircraft in the United States, Canada, and Europe are equipped to collect cloud physics data. Three primary instruments on these aircraft acquire data on the concentrations of cloud droplets as a function of droplet diameter (the cloud droplet size spectrum) and the cloud liquid water content. The size spectrum is obtained primarily from an electrooptical device known as the forward-scattering spectrometer probe (FSSP), which sizes droplets over the diameter range $1.5\text{--}46.5\ \mu\text{m}$ in $3\ \mu\text{m}$ increments, although other sizing bounds can be used. The size spectrum is also obtained using another electrooptical device, which produces two-dimensional images of particles; this optical array probe is described in more detail in Section IX. The liquid water content can be obtained indirectly, from the size spectrum data, by integration of the measured droplet spectrum. The liquid water content can be measured directly using “hot-wire” devices. Droplets impinge on a hot wire, and because they evaporate, they cool the wire. The temperature of the wire or the current required to maintain it at a constant temperature is used to derive the liquid water content.

B. Observations of Droplets

Most drop-size distributions, even though measured in many different types of clouds formed under differing meteorological conditions, exhibit a characteristic shape. The concentrations generally increase with size abruptly from a low value to a maximum somewhere between 10 and $20\ \mu\text{m}$, and then decrease gradually toward larger sizes, causing the distribution to be skewed with a long tail toward the larger sizes. A log normal or γ distribution function is used to approximate this characteristic shape. When coalescence growth is involved, the spectra contain a characteristic secondary peak in concentration at diameters usually between 20 and $40\ \mu\text{m}$.

Significant differences are found between spectra formed in maritime air and those formed in continental

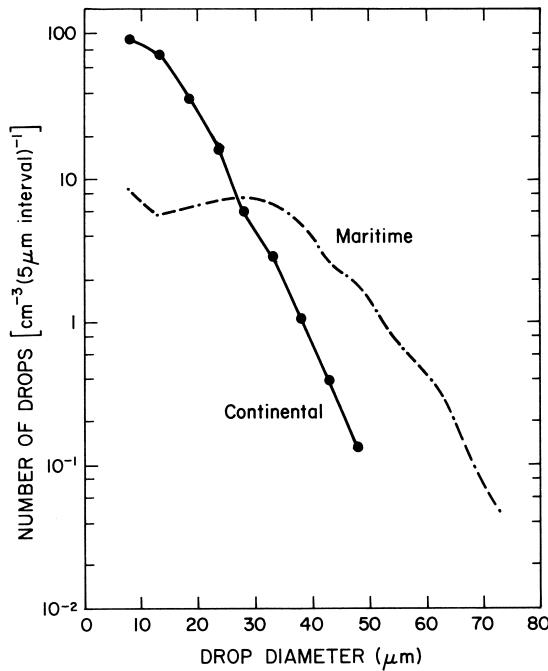


FIGURE 5 Cloud droplet spectra for maritime- and continental-type cumulus clouds. These curves are used for the drop growth rate computations in Fig. 4. [From Braham, R. R. (1968). *Bull. Amer. Meteorol. Soc.* **49**, 343–353. Courtesy of the American Meteorological Society and the author.]

air masses because of the previously discussed nature of the CCNs. Figure 5 illustrates these differences between typical spectra in a maritime and a continental cumulus cloud. Typical concentrations and mean diameters in maritime convective clouds are $50\text{--}100 \text{ cm}^{-3}$ and $15\text{--}20 \mu\text{m}$, respectively, while in continental convective clouds they are $500\text{--}1000 \text{ cm}^{-3}$ and $10\text{--}15 \mu\text{m}$, respectively.

Liquid water contents are dependent on the cloud type, cloud base temperature, and height above cloud base. In stratiform clouds, the values are comparatively low, usually $\sim 0.1 \text{ g m}^{-3}$. In cumulus clouds, typical peak values are $\sim 0.5 \text{ g m}^{-3}$, which increase with increasing cloud intensity to more than 3 g m^{-3} for severe thunderstorms. (As the cloud base temperature increases, a cloud of a given type tends to have a higher liquid water content, increasing with height from cloud base to near cloud top and then falling abruptly to zero at cloud top.)

C. Drop-Size Spectra in Rain

A multitude of measurements of drop-size spectra in rain have been made at the ground. These measurements indicate that drop-size distributions are of an approximately negative exponential form, especially in rain that is fairly steady. The so-called Marshall–Palmer distribution appears to describe the rain spectrum in most meteorological situations. This spectrum has the form

$$N(D) = N_0 e^{-\Lambda D}, \quad (6)$$

where the product $N(D) dD$ is the number of drops per unit volume with diameters between D and $D + dD$. The slope factor Λ of the spectrum is given by

$$\Lambda = 41R^{-0.21} \quad (7)$$

where R is the rainfall rate (in millimeters per hour), and N_0 , the intercept of the spectrum, is given by

$$N_0 = 0.08 \text{ cm}^{-4} \quad (8)$$

VI. ICE FORMATION MECHANISMS

Ice crystals form in the atmosphere in three ways: (1) on solid particles, often airborne soil particles, referred to as *ice nuclei*; (2) by secondary processes that multiply primary ice crystals formed by (1); and (3) by *homogeneous nucleation*, that is, pure water droplets freezing spontaneously when a temperature of -40°C is reached, probably typical of cirrus clouds.

A. Ice Nuclei and Primary Ice Crystals

It was realized many years ago that clouds of liquid water droplets can persist at temperatures below 0°C (super-cooled) unless suitable ice nuclei are present to help ice crystals form. Ice nuclei provide a surface having a structure geometrically similar to that of ice, thereby increasing the probability of formation of the ice structure required for stability and thus causing microscopic droplets to freeze relatively rapidly at temperatures higher than -40°C .

Currently, there is much uncertainty about the mechanisms of ice nucleation in the atmosphere, but it is thought that ice nuclei operate by three basic modes. In one mode, water is absorbed from the vapor phase onto the surface of the ice nucleus, and at sufficiently low temperatures, the adsorbed vapor is converted to ice. In another mode, the ice nucleus, which is inside a supercooled droplet either by collection or as a result of its participation in the condensation process, initiates the ice phase from inside the droplet. In the third mode, the ice nucleus initiates the ice phase at the moment of contact with a supercooled droplet (such nuclei are known as contact nuclei). The relative importance of these different modes of operation is not known with certainty, but the latter two modes are thought to be much more common than the first.

Much effort has been made to measure the concentrations of ice nuclei in the air and the variation with temperature. While considerable variability in concentration is found with time and location, a fairly representative worldwide concentration is ~ 1 per liter active at -20°C , changing with temperature in the way shown in Fig. 6.

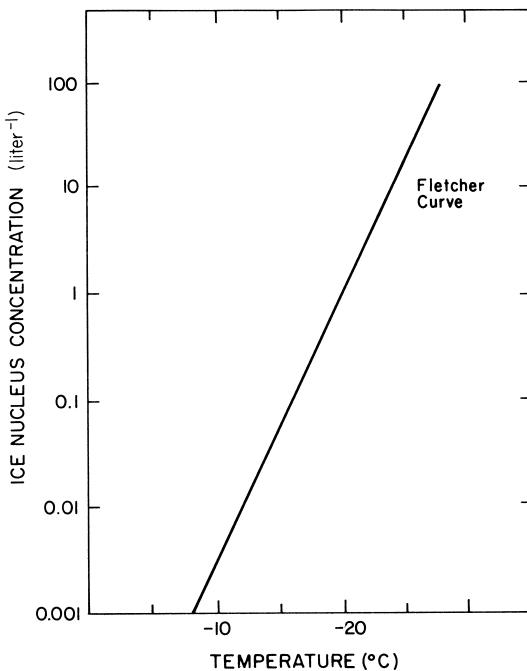


FIGURE 6 Worldwide “average” of measured ice nucleus concentrations versus temperature (the Fletcher curve).

(This is the so-called Fletcher worldwide average ice nucleus curve. It must be emphasized, however, that there is a great deal of scatter of the data about the Fletcher curve, and it may not provide a useful representation of the concentrations in the atmosphere.)

When the temperatures in-cloud are near or below -40°C and the ambient relative humidity is close to 100% with respect to water saturation—conducive to cirrus cloud formation—homogeneous ice nucleation is likely to be an important mechanism for ice production. The theory of homogeneous nucleation of ice has shown that when molecules of liquid water comprising a pure droplet a few microns in diameter bind to form a stable ice structure at an observable rate only when temperatures near -40°C is reached. This theory has been confirmed in the laboratory. What has been confirmed in studies in orographic wave clouds, and in cirrus cloud as well, is that when water condenses on a cloud condensation nucleus at temperatures near and below -40°C and grow to a sizes of $1\ \mu\text{m}$, the droplets freeze homogeneously. This mechanism is thought to be important in the formation of cirrus clouds.

B. Secondary Ice Crystals

Discrepancies between concentrations of ice crystals and ice nuclei have been clearly demonstrated in clouds containing air of maritime origin. These clouds, often with tops no colder than -10°C , contain concentrations of ice

crystals that can exceed that of ice nuclei by a factor of 10^4 and typically contain droplets larger than $25\ \mu\text{m}$. Whatever process is operating seems to be most effective at a temperature of -6°C .

Various mechanisms have been proposed to explain this important ice nucleation process. In one, drops that are in the process of freezing build up high internal pressures, shatter, and produce secondary ice particles. Laboratory studies indicate that this mechanism applies to droplets with diameters greater than $400\ \mu\text{m}$; it is not likely to be important in most clouds, however. In another, secondary ice particles are thought to be ejected when super-cooled drops freeze onto an ice particle at temperatures between -3° and -8°C . Experiments have demonstrated that this mechanism is operative in the laboratory when the particle collecting the droplets is falling at a velocity of $1\text{--}3\ \text{m sec}^{-1}$, when some of the droplets being collected are larger than $25\ \mu\text{m}$ in diameter and when some are about half this diameter. The nature of this process has been inferred from laboratory experiments, and it appears likely to explain the initiation of many of the secondary crystals observed in the atmosphere. Collections of ice particles at ground level and in clouds often contain fragments of crystals, which indicates that a third mechanism of producing secondary ice particles is the fracturing of ice particles on collisions with other ice particles. Ice multiplication by this process (mechanical fracture) is still not very well understood, nor has its importance been documented. A fourth mechanism involves the aerodynamic breakup of particles in subsaturated layer, where components of crystals can break off from the parent particles.

VII. OBSERVATIONS OF THE GROWTH MECHANISMS OF ICE PARTICLES

Observations of ice particle shapes and sizes from their origin in clouds through their fallout at the ground indicate that growth proceeds by the chain of processes illustrated in Fig. 7. Following nucleation, crystals grow first through vapor diffusion. Observations indicate that they must achieve a diameter of $\sim 200\ \mu\text{m}$ before they begin to collect, or *accrete*, water droplets. Water drops that freeze in the air at a diameter of $200\ \mu\text{m}$ or larger collect water drops immediately. In the absence of liquid water, ice crystals grow into snow crystals and often clump together to form snowflakes. Rimed crystals or rimed snowflakes that continue to rime can grow into spherical or conical low-density ice particles (graupel) and in some instances continue to grow to diameters larger than $0.5\ \text{cm}$ (hail) and are then nearly solid ice. Frozen drops grow immediately into graupel and can also grow into hail. Ice particles then reach the ground in the form of snow, rain, hail, or graupel.

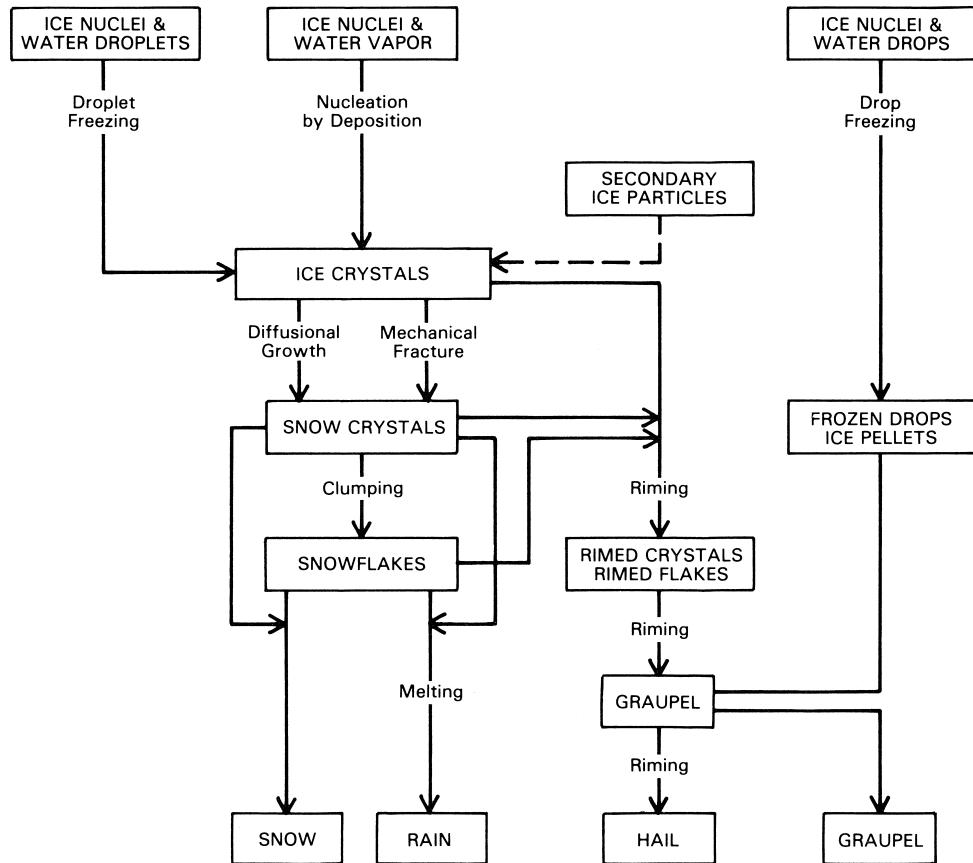


FIGURE 7 Flow chart showing processes of ice particle growth. [Adapted from Braham, R. R., and Squires, P. (1974). *Bull. Amer. Meteorol. Soc.* **55**, 543–556. Courtesy of the American Meteorological Society and the authors.]

Observations also show that ice and snow crystals grow in a large variety of shapes, or *habits*. The major habits illustrated in Fig. 8 are plates, dendrites, columns, needles, bullets, and combinations of bullets. Laboratory and theoretical studies reveal that snow crystals have one common basic shape, that of a sixfold, symmetric (hexagonal) prism composed of two basal planes (e.g., the faces of the crystals; see crystals in Figs. 8a–8c) and six prism planes (the crystal edges, see crystals in Figs. 8d–8f). Laboratory experiments reveal that the rate of propagation of the basal faces (growth along the *c* axis) relative to that of the prism faces (*a* axis) varies with temperature and supersaturation in a characteristic manner. The results of experimental studies of crystal habits are shown as a function of temperature and relative degree of water vapor supply in Fig. 9. A dashed, horizontal line in the figure shows water saturation conditions (100% relative humidity). A solid, horizontal line at the bottom of the figure shows ice saturation conditions. (In terms of relative humidity with respect to water, ice saturation equals 100% at 0°C and decreases nearly linearly with temperature, equaling ~60% at –40°C.) The important points to note from this figure

are twofold. There are changes in habit from columns to plates to columns that occur at precise temperatures, and at a particular temperature, the relative humidity controls some important features of snow crystal growth.

Measurements have illustrated that ice crystal growth rates depend strongly on temperature. Figure 10 shows the results of laboratory measurements of the growth rates along the *a* and *c* axes, designated in the figure as da/dt and dc/dt . These data represent averages over a ~3-min period, which started with crystals ~20 μm in diameter. This figure illustrates that there is a marked maximum in the growth rate along the *a* axis maximum at –15°C and that a secondary broader maximum occurs along the *c* axis at –6°C.

VIII. PHYSICS OF THE GROWTH OF ICE PARTICLES

The ice crystal growth processes presented in the previous section (Fig. 7) are discussed from a physical point of view in this section.

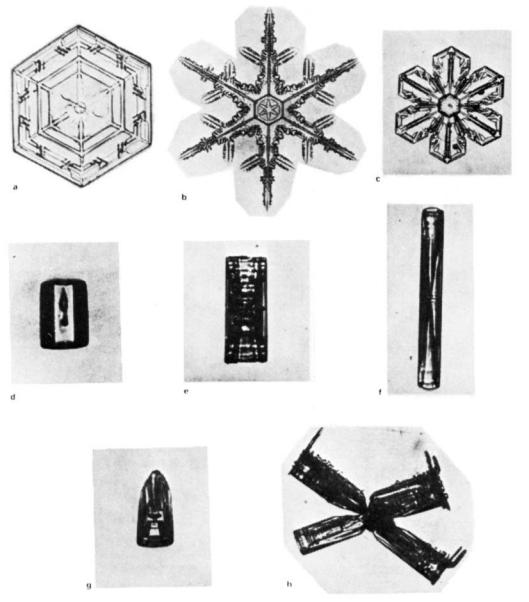


FIGURE 8 Major shapes of snow crystals: (a) simple plate, (b) dendrite, (c) crystal with broad branches, (d) solid column, (e) hollow column, (f) sheath, (g) bullet, and (h) combination of bullets. [From Pruppacher, H. R., and Klett, J. D. (1978). “Microphysics of Clouds,” who used photographs from Nakaya, U. (1954). “Snow Crystals,” Courtesy of D. Reidel Publishing Co., H. R. Pruppacher, and Harvard University Press, copyright 1954 by the President and Fellows of Harvard College.]

A. Growth of Ice Crystals by Diffusion

On formation of ice embryos, whether by sublimation directly from the vapor, by freezing of supercooled droplets, or by a secondary production mechanism, diffusional growth commences because the embryo is in an environment that is probably at or close to water saturation. The growth rate equations can be derived using theory analogous to that presented in Section IV, but since ice crystals are generally not spherical, some alterations must be made.

The approach that has been used to calculate ice crystal growth starts with an analogy between the governing equations and the boundary conditions for electrostatic and diffusion problems. Poisson’s equation in electrostatics and Green’s theorem lead to the following equation for the ice crystal diffusional mass growth rate,

$$dm_d/dt = 4\pi CD(\rho_v - \rho_{iR}) \quad (9)$$

where ρ_{iR} is the vapor density at the ice particle surface, C is the “capacitance” of the ice crystal, and the other terms have been previously defined. In order to apply Eq. (9) to a particular crystal form, C is specified as a function of the crystal geometry. In the simplest case of a spherical crystal of radius R , $C = R$, and Eq. (9) then has a form similar to the mass growth rate of drops [Eq. (3)]. A slightly more complex case is a simple thin hexagonal plate, for which

$C = 2R/\pi$. Laboratory experiments of the capacitances of brass models of snowflakes have confirmed the validity of this approach for calculating the ice crystal growth rate.

The linear rate of growth of a crystal can be derived from Eq. (9), given knowledge of the crystal geometry, thickness-to-diameter ratio (axial ratio, AR), and bulk density ρ_i . For example, a hexagonal plate crystal has a mass given by $m = 5.2R^3 \times AR \times \rho_i$, and so its growth rate along its major axis is

$$\frac{dR}{dt} = \frac{0.51D(\rho_v - \rho_{iR})}{AR \times \rho_i R} \quad (10)$$

For columnar and needle crystals, growth along the crystal length L is given by

$$\frac{dL}{dt} = \frac{3.22D(\rho_v \times \rho_{iR})}{\ln(2AR)\rho_i L} \times AR^2 \quad (11)$$

It is important to point out here that as crystals become increasingly large, their linear growth rate becomes increasingly small: From Eq. (10), dR/dt is proportional to $1/R$, and from Eq. (11), dL/dt is proportional to $1/L$. Data on the crystal axial ratios and bulk density are necessary to solve Eqs. (10) and (11). The data given in Fig. 10 can be used to compute the axial ratio: $AR = (da/dt)/(dc/dt)$. Additional axial ratio information is available from field data. Laboratory and field data indicate that bulk densities have typical values of between 0.5 and 0.8 g cm^{-3} .

Solving for R and L as a function of time in Eqs. (10) and (11) also requires values for D , ρ_v , and ρ_{iR} . Meteorological tables or equations can be used to derive the value of D and ρ_v if the temperature, pressure, and relative humidity of the growth environment are known. Now consider the vapor density at the crystal surface, ρ_{iR} . As the ice crystal grows, its surface is heated by the latent heat of vaporization, and because of this warming, the value of ρ_{iR} is effectively raised above the value that would apply without heating. Under stationary growth conditions, the value of ρ_{iR} is determined from the rate of latent heating and the rate of heat transfer away from the surface. If the crystal has grown to a size of more than a few hundred micrometers, at which time it has an appreciable fall velocity, it is necessary to take into account the effect of “ventilation” on the diffusion of water vapor and heat. Fairly simple expressions have been derived to solve for ρ_{iR} in Eqs. (9) to (11). Values for the term $D(\rho_v - \rho_{iR})$ as a function of temperature are shown in Fig. 11 for water saturation, for $100\text{-}\mu\text{m}$ crystals, and for pressures of 1000 (sea level) and 400 mbar ($\sim 6 \text{ km}$). The important points to note from this figure are that (1) this term is at a maximum at a temperature of about -15°C , suggesting that particles can potentially grow most rapidly at this temperature, and (2) the term is higher at 400 than at 1000 mbar, indicating

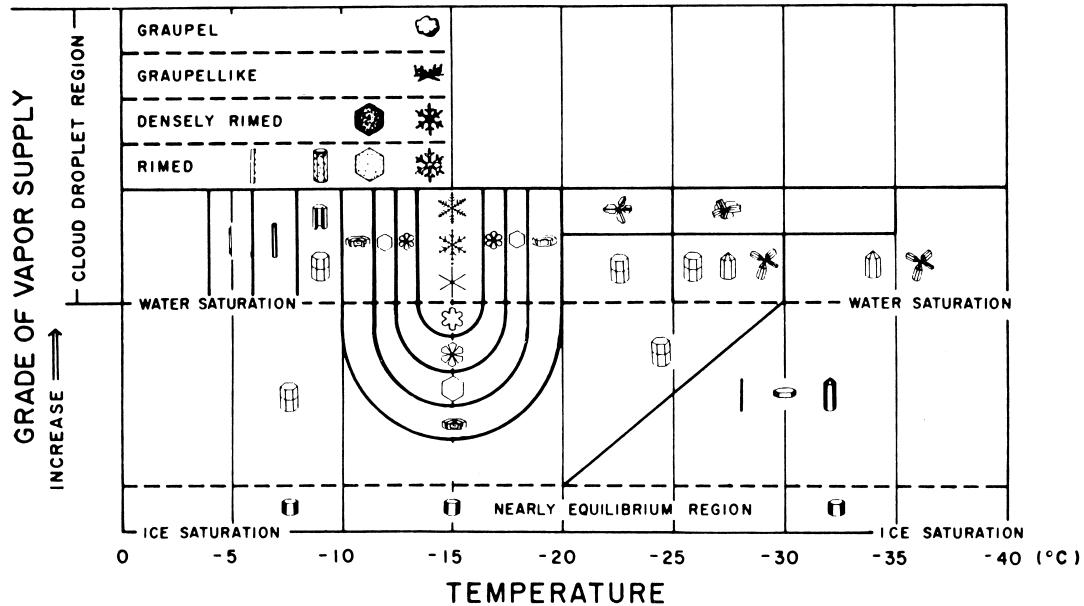


FIGURE 9 Temperature and humidity conditions for the growth of natural snow crystals of various types. [From Magono, C., and Lee, C. W. (1966). *J. Fac. Sci. Hokkaido Univ. Ser. 7*, **2**(4). Courtesy of the Faculty of Sciences, Hokkaido University, Japan.]

that the growth rate increases with altitude, given a constant temperature.

It still remains to be shown that the growth rate equations provide a reasonable representation of the experimentally measured growth rates in Fig. 10. To simulate the laboratory conditions, growth has been calculated over a ~ 3 -min period beginning with crystals $\sim 20 \mu\text{m}$ in diameter. The average values of da/dt and dc/dt are shown in Fig. 12. The growth equations appear to emulate the salient features of the observed growth in Fig. 10, at least during these early stages of crystal growth. Data at later stages are not available for comparison.

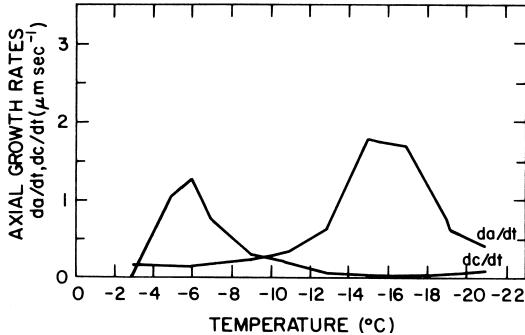


FIGURE 10 Variation of measured crystal axial growth rates with temperature over a period of ~ 3 min following ice crystal nucleation. [Adapted from Ryan, B. F., Wishart, E. R., and Shaw, D. E. (1976). *J. Atm. Sci.* **33**, 842–851. Courtesy of the American Meteorological Society and the lead author.]

Ice crystals are often carried or descend into regions where the relative humidity is below saturation with respect to ice and begin to evaporate; common ice subsaturated regions are near the bases of thunderstorm anvils and at the base of cirrus clouds. The growth rate equations given by Eqs. (9) to (11) can be used to compute the rate

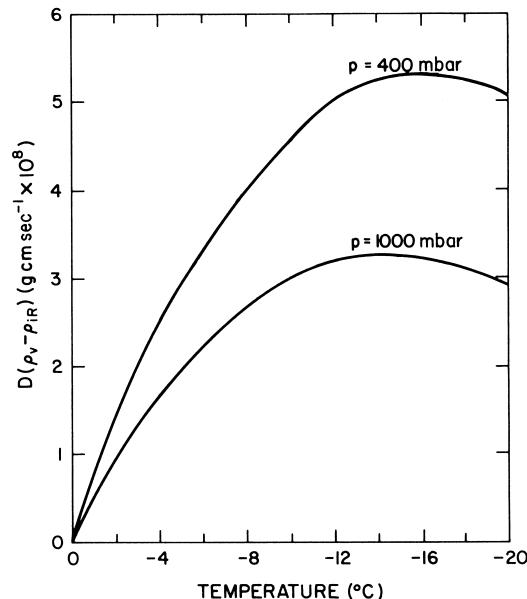


FIGURE 11 Variation of a term in the ice crystal growth rate equation with temperature at two atmospheric pressures.

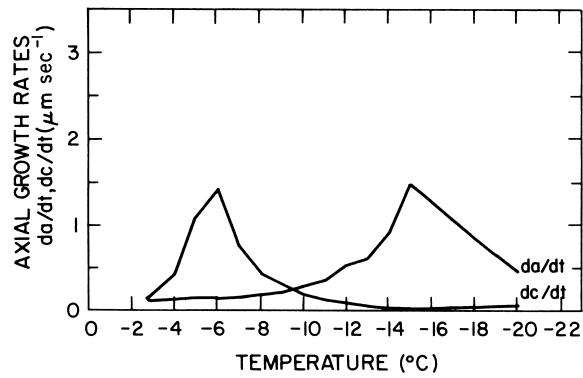


FIGURE 12 Variation of calculated crystal axial growth rates with temperature over a period of ~ 3 min following ice crystal nucleation.

at which crystal mass and linear dimensions decrease in these regions.

B. Growth through Collisions between Ice Particles

Snowflakes rather than individual ice crystals account for most of the precipitation reaching the ground as snow. As crystals become increasingly large, they are more likely to bump into one another, and some stick together or clump to form an aggregate. The physics of aggregation is summarized below.

Theoretical approaches to snowflake growth compute the number of collisions that can occur between the crystals in a given volume of air. Consider particles of radius R_1 and terminal velocity V_1 in concentration N_1 in close proximity to those of radius R_2 , velocity V_2 , and concentration N_2 . The number of the faster falling particles R_1 that collide with R_2 in time dt is given by the product of the following three terms: (1) the volume that they collectively sweep out, $\pi(R_1 + R_2)^2 V_1 N_1 N_2 dt$; (2) their relative terminal velocity, $V_1 - V_2$; and (3) the efficiency at which they collide, E . The evolution of the size spectrum of crystals and aggregates is obtained by calculating the collisions that could occur between all sizes of crystals, aggregates and crystals, and aggregates and aggregates that are present in a given volume of air, taking into account their respective collection efficiencies E .

The theoretical studies have demonstrated that aggregates can grow at a rate of up to $10 \mu\text{m sec}^{-1}$, in contrast to single ice crystals of the same size, which grow at a rate of 1–10% of this value. Small variations in the velocity of particles of the same size can lead to an even more rapid growth rate, and aggregation leads to the development of a size distribution in which the concentration decreases exponentially with increasing size, all in agreement with observations.

C. Growth by Droplet Collection (Accretion)

An ice crystal in a cloud of supercooled water droplets grows by accretion into a rimed crystal, a graupel particle, or a hailstone. Water drops can grow by an analogous process of drop–droplet collection, as previously discussed (Section IV.B). Theoretical treatments of the accretion process are much more complex than those for drops, because they must consider a variety of complex shapes (e.g., Fig. 8) changing with time during the accretion process, the release of latent heat at the crystal surface by droplet freezing, and the density of droplets accreted on the surface of the ice particle since the density depends on the characteristics of the ice particle (these densities vary from 0.1 – 0.91 g cm^{-3}).

The basic equation for the particle mass growth rate during accretion can be written as the sum through diffusion and accretion,

$$dm/dt = (dm_d/dt) + A \times E \times \text{LWC} \times V_T, \quad (12)$$

where dm_d/dt is the diffusional growth rate term, A is the cross-sectional area of the particle normal to the airflow, and the other terms are as previously defined. The diffusional growth rate term must be calculated by considering the effects of latent heating due to sublimation and droplet freezing versus conduction of heat away from the particle. During the early stages of riming, when a crystal is beginning to grow into graupel, the growth rate dm_d/dt is positive, while in later periods, for example, graupel and hail growth stages, dm_d/dt is negative. The particle terminal velocity V_T and cross-sectional area A are the two factors that control the number of droplets a particle sweeps out per unit time, and the terminal velocity is highly dependent on the particle shape (e.g., whether crystal or graupel), the diameter, and the bulk density (Fig. 13).

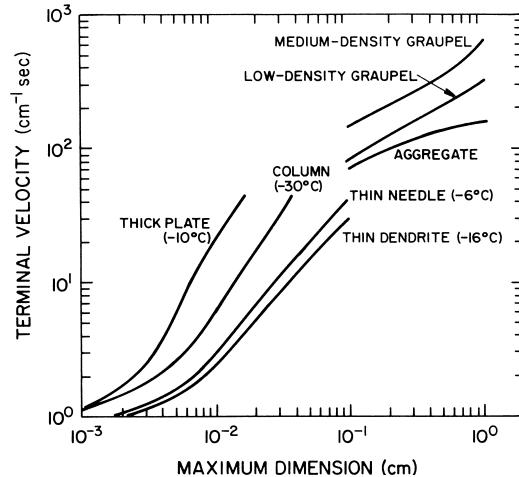


FIGURE 13 Terminal velocities of ice particles as a function of maximum crystal dimension, calculated for particles of different types.

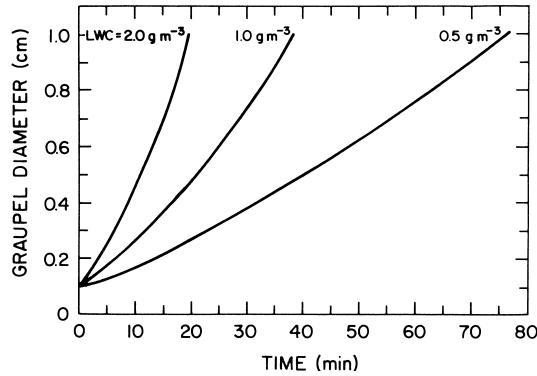


FIGURE 14 Calculated diameters of graupel particles as a function of time for several liquid water contents (LWC). Growth temperature is -15°C .

The collection efficiency is highly dependent on particle terminal velocity and cross-sectional area as well.

Assuming spherical particle growth, the linear accretional growth rate can be written analogously to Eq. (5),

$$dR/dt = \bar{E} \times \text{LWC} \times V_T / (4\rho_A R) \quad (13)$$

such that ρ_A is the accretional density. Calculations of the growth of a spherical (graupel) ice particle from 1 mm to 1 cm in diameter (hail size) at -15°C and liquid water contents commonly observed in clouds (0.5, 1.0, and 2.0 g m^{-3}) are given as a function of time in Fig. 14. Note that hail can grow from very small particles in a period as short as 20 min. The liquid water contents and temperature in Fig. 14 were chosen so that all accreted liquid water was frozen (dry growth). When growth temperatures or liquid water contents are higher, the surface temperature of the particle may rise to 0°C , and not all accreted water is frozen (wet growth). Wet growth can be very important for hailstones growing at temperatures above about -10°C .

IX. IN-CLOUD OBSERVATIONS OF ICE PARTICLES

A. Measurement Devices

Electrooptical devices are currently the primary tools used on aircraft to probe ice particles in clouds. The two probes used most frequently are one-dimensional (1-D) and two-dimensional (2-D) optical array probes, both operating on essentially the same principle. These probes project a laser beam through ~ 10 cm of a cloud onto a linear array of photodiode elements. When a particle intervenes near the object (in-focus) plane of the beam, its imaged shadow momentarily occults a number of optical array elements. For the 1-D probe, the particle size is the measured shadow size divided by the optical magnification. Such probes are widely used in the size ranges 20–300 and 300–4500 μm , obtaining information on concentrations in

15 equally spaced size intervals, which is then recorded on magnetic tape. The 2-D probe uses a photodiode array and electronics similar to the 1-D probe. However, the 2-D probe contains electronics to record many pieces of shadow information for an individual particle as it passes across the photodiode array. As the particle's transit shadows the array, image slices are obtained across the shadow to develop a 2-D image. The 2-D probes are used to size particles in the same ranges as the 1-D probes, but the 2-D probes have twice the size resolution.

Several new probes on aircraft and balloons are revolutionizing the way that cloud physicists observe ice particles. The cloud particle imager (CPI) makes digital images with $2.3 \mu\text{m}$ size resolution. It uses a high-power laser pulsed at a high frequency to freeze images of particle on a CCD camera (Fig. 15). The high-volume particle sampler is a version of a 2-D probe with a very large sampling volume to provide better sampling statistics for precipitation-size particles. The cloudscope is essentially an airborne microscope that obtains images of particles as they impact a window exposed to the airstream. A balloon-borne ice crystal replicator and the hydrometeor-video-sonde (HYVIS) are instruments which capture and either preserve or obtain video images of particles as the balloon ascends through a cloud. The counterflow virtual impactor is an instrument which obtains the water content of the condensate above a size of about $10 \mu\text{m}$.

B. Summary of In-Cloud Measurements

A fairly large number of measurements made in clouds with all droplets smaller than $25 \mu\text{m}$ show that ice crystal concentrations are scattered on either side of the Fletcher ice nucleus curve (Fig. 6), but there is a great deal of variability. Conversely, in clouds where droplets are greater than $25 \mu\text{m}$ (usually in maritime areas), measured concentrations are often $10^4 \mu\text{m}$ times higher than those expected from ice nucleus measurements. A secondary ice crystal production mechanism that operates at about -6°C appears to account for these enhancements (see discussion in Section VI.B).

Measured ice particle size spectra typically exhibit a gamma form, where concentrations first increase with size up to tens to $100 \mu\text{m}$, then decrease exponentially with size thereafter. In fairly quiescent clouds such as cirrus, ice particles at a given location span a size range from 1–1000 μm , while in vigorous clouds such as cumulonimbus, sizes can span a range from 1 μm to as much as $10^5 \mu\text{m}$ (10 cm) when large hail is present. The observations of an exponential “tail” to the size distribution is not surprising, since ice particles found at a particular location could have originated from vastly different locations wherein they experienced different temperatures and humidities during their growth. Processes such as

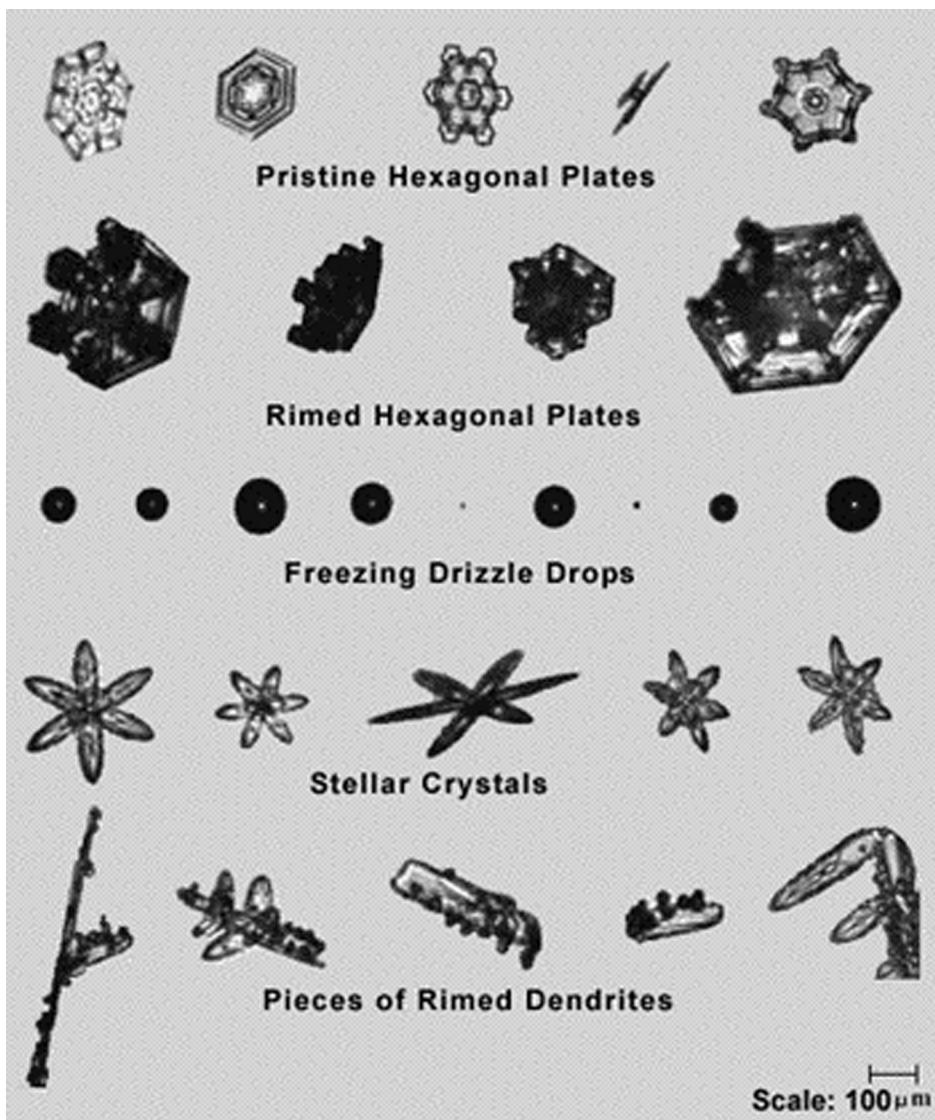


FIGURE 15 Images of ice particles and water droplets from the SPEC, Inc. cloud particle images. Scale is shown in the lower right-hand corner. (Courtesy of R. Paul Jawson.)

aggregation contribute further to the development of exponential spectra.

Ice crystal shapes observed in clouds vary with temperature in a way consistent with the findings from the laboratory experiments (Fig. 9). Exceptions occur when, for example, columnar ice particles fall into a planar crystal growth regime; in such a case, columnar crystals with end plates (capped columns) are observed. There is still a great deal to learn about the dependence of ice crystal shapes on ambient temperature and relative humidity below -25°C . While the probability of ice particle aggregation is a maximum near 0°C , recent observations indicate that it can occur at any temperature, leading to the development of the exponential tail of the size distribution. Rimed ice crystals and graupel form in clouds that

contain both ice crystals and supercooled drops. In such clouds, ice crystals, aggregates, and frozen drops can serve as graupel and hail embryos. In-cloud and ground-based studies indicate that both columnar ice crystals and ice crystal plates have to grow by diffusion to a certain size before they can grow by riming. It turns out that the columnar crystal's minor dimension (width) has to exceed about $50 \mu\text{m}$ by diffusion first, while that of plates must exceed $\sim 300 \mu\text{m}$.

X. EFFECTS OF CLOUDS ON CLIMATE

Over the past decade, cloud physics research has been driven by uncertainties related to the role of clouds on

climate. Clouds absorb the outgoing long-wave radiation emitted by the surface and the troposphere, and re-emit this energy to space at the much colder cloud-top temperatures. Clouds also tend to cool the surface by reflecting solar radiation back to space. The sum of these two effects lead to the net effect of clouds on climate, but since there is such a wide range of cloud sizes, microphysical properties, and height in the atmosphere, it is very difficult to assess the role of clouds on climate.

Uncertainties about the effects of clouds on climate have provided the impetus to conduct focused field campaigns aimed at understanding the effects of cirrus and stratocumulus clouds on incoming short-wave radiation and outgoing long-wave radiation. These experiments involve aircraft *in situ* measurements, ground-based measurements using radar and laser radar (lidar), and overflying aircraft and satellites to provide information on cloud radiative properties. Focused experiments have been conducted in North America, Europe, Japan, Brazil, and in the central and western Pacific. Long-term monitoring stations equipped with a wide range of remote sensing equipment have been placed in Oklahoma, Alaska, and the western Pacific to examine cloud properties over periods of decades.

Two cloud types, stratocumulus and cirrus, cover the greatest portion of the earth's surface and are therefore the most important cloud types from the standpoint of the effect of clouds on climate. The focused observations reveal that stratocumulus have a net cooling effect on the planet. However, the net effect of cirrus on climate is still unclear. The details of the cirrus cloud microphysics, including the mean, maximum, and effective (radiatively important) particle size, crystal shape, temperature, and height in the atmosphere, affect their radiative properties. This is an active area of research.

Satellite observations are now beginning to play a pivotal role in measuring not only the size and spatial distribution and extent of clouds, but their microphysical properties as well. The International Satellite Cloud Climatology Project (ISCCP) has been analyzing combined geostationary and sun-synchronous or polar orbiting satellite datasets since 1983. These datasets have been used to develop a near-global survey of the effective particle size in liquid water clouds and also ice clouds. Satellites are also being used to characterize the size of ice crystals in cirrus clouds and to map out the spatial distributions of aircraft-produced condensation trails (contrails). The Tropical Rainfall Measuring Mission (TRMM) is an important cloud physics platform that carries a radar for measuring precipitation rates and amounts over land and ocean areas in the tropics. The Moderate Resolution Imaging Spectroradiometer on the recently launched Terra spacecraft will enhance the capability for monitoring cirrus cloud microphysical properties. Several satellites to

be launched over the next five years will further improve our ability to measure cloud microphysical properties over global scales.

The modeling of cloud physics processes in climate models has progressed significantly in recent years. Cloud physics processes are now represented in the more advanced climate models, although still rather simplistically. Processes such as rain formation, ice crystal nucleation, growth, advection, and fallout are now being included. A major process has been linking the microphysics to the radiative properties in climate models, although the parameterizations are still rather simplistic.

Over the past five years, there have been a number of studies which have sought to examine the effect of man's activities on clouds and climate. Modification of the microphysical structure of clouds as a result of biomass burning and the potential impact of the affected clouds on rainfall has been a topic of major interest. Field programs in the tropics have shown that biomass burning modifies the cloud condensation nuclei, and therefore there will be a change in the microphysical properties of clouds in the area. Increases in the CCN concentration results in more cloud droplets produced, with associated increases in cloud reflectivity. Assessing the impact of man's activities on cloud properties and reflectivity is an area of intense activity.

A. Aircraft-Induced Condensation Trails (Contrails)

Contrails form when saturation with respect to water is temporarily reached in the plume behind an aircraft, and they persist in ice supersaturated air masses. The ambient temperature necessary for contrail formation can be predicted accurately. Contrail particles consist of ice crystals. Young persistent contrails are composed of more, but smaller, ice crystals than typical cirrus clouds. Persistent contrails develop towards cirrus clouds in the course of time. The average contrail coverage exhibits a value around 0.5% over Europe. The mean global contrail cover is estimated to be of order of 0.1%. The net radiation effect of contrails is believed to enhance warming of the troposphere on average.

XI. CLOUD SEEDING

The main purposes of cloud seeding are (1) to increase precipitation, (2) to dissipate cloud or fog, and (3) to suppress hail. These purposes and their scientific basis are discussed in order below.

1. Experiments that attempt to increase precipitation are based on three main assumptions. (a) The presence of

ice crystals is necessary to produce precipitation in a supercooled cloud, or the presence of fairly large water drops is required to initiate coalescence growth. In earlier sections it was shown that these particles are necessary for producing precipitation. (b) Some clouds precipitate inefficiently or not at all, because the particles listed in (a) are naturally deficient. The Fletcher curve (Fig. 6) shows that ice nuclei are active in concentrations of less than 1 liter⁻¹ down to a temperature of $\sim 20^{\circ}\text{C}$. For most clouds, a concentration of about 1 liter⁻¹ is considered to be necessary to convert water vapor and cloud droplets efficiently to precipitation. Therefore, in clouds whose temperatures at cloud top do not fall below about -20°C , seeding can potentially increase precipitation. Clouds that would obviously not benefit from seeding are those of the maritime type in which large droplets ($>25\ \mu\text{m}$) are present and high concentrations of secondary ice crystals are naturally produced. In clouds that contain small droplets, mechanisms are often not present to initiate coalescence growth. In such cases, addition of salt particles can lead to broader cloud droplet size distributions and initiate precipitation by coalescence growth. (c) Seeding the clouds artificially to produce ice crystals or water drops can alleviate the deficiency of precipitation embryos. Laboratory experiments have shown that two materials are highly effective in producing abundant ice crystals in supercooled clouds: dry ice, in the form of pellets that nucleate a large number of droplets in their path homogeneously and rapidly freeze them, and silver iodide, in the form of small (submicrometer-size) particles, which have a lattice structure very similar to that of ice and are effective ice nuclei. Coalescence growth can be initiated by introducing water drops into clouds or by seeding summertime convective clouds below cloud base with pyrotechnic flares that produce small salt particles in an attempt to broaden the cloud droplet spectrum and accelerate the coalescence process (see Section III). Dry ice is dispersed in clouds by aircraft, while silver iodide is dispersed either by aircraft or by ground-based aerosol generation units.

To produce the desired results, a cloud obviously must be seeded so as to produce an optimal concentration of ice crystals or water drops for the particular cloud conditions. For example, introducing too many particles will result in “overseeding,” which will cause small, nonprecipitating particles to develop. Similarly, introducing too few particles will “underseed” the cloud and will result in too few precipitation particles.

Many problems have been encountered in the precipitation-enhancement seeding experiments conducted to date. In particular, it has been extremely difficult to deliver the optimal amount of seeding material at the correct time and place in a cloud.

2. Fogs and low clouds often pose hazards around airports. The concept for seeding these clouds is to introduce large salt particles or ice nuclei to sweep out the cloud droplets, thus clearing an area temporarily. Some success has been achieved, particularly in supercooled clouds. However, the clearing is often of short duration, and operations can be fairly expensive.

3. Two main concepts are employed in the design of hail suppression experiments. In the first, it is argued that by adding a large amount of seeding material to a thunderstorm so as to freeze most of the cloud liquid water at temperatures below about -15°C , accretional growth will be effectively eliminated, and large hail will not grow. The second approach is to produce artificially many embryos within the regions of a storm where hail growth is occurring, and through competition among particles for the available water, it will be unlikely that any hailstone will grow to a large size.

Many hail suppression experiments are being conducted throughout the world, notably in Russia. Some positive responses have been claimed in hail suppression experiments, but the extent of the result has been difficult to document.

XII. CONCLUDING REMARKS

Technological advances are contributing to rapid increases in the understanding of the key physical processes operative in clouds. For example, sophisticated measurements now obtained by aircraft and radar in clouds are being used in concert to increase the understanding of the symbiotic relationship between cloud dynamics and cloud microphysics. Detailed computer models that account for the dynamics, thermodynamics, and microphysics of clouds are being used to simulate the complex processes that occur in clouds from initial cloud formation through collapse. Laboratory experiments in the carefully controlled environments of wind tunnels and cold rooms are now performed to investigate key hydrometeor growth processes. The role of clouds in the earth’s energy balance and in climate change is the focus of intensive research.

This article has not addressed several subjects in the field of cloud physics. For cloud chemistry and cloud electrification, see Atmospheric Chemistry and Atmospheric Electricity. The topic of cloud radiation and optics is almost as broad as that of cloud physics itself. The reader is referred to the Bibliography for information on this topic.

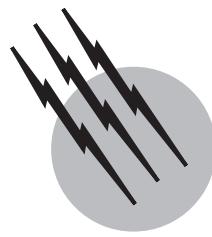
SEE ALSO THE FOLLOWING ARTICLES

CLIMATOLOGY • GREENHOUSE EFFECT AND CLIMATE DATA • IMAGING THROUGH THE ATMOSPHERE • METEOROLOGY, DYNAMIC • THUNDERSTORMS, SEVERE

BIBLIOGRAPHY

Cotton, W. R., and Anthes, R. A. (1989). "Storm and Cloud Dynamics," Academic Press, San Diego.
Gadsen, M., and Schroder, W. (1989). "Noctilucent Clouds," Springer-

- Verlag, New York.
Hobbs, P., and Deepak, A., eds. (1981). "Clouds: Their Formation, Optical Properties and Effects," Academic Press, New York.
Huschke, R. E., ed. (1970). "Glossary of Meteorology," 2nd printing, Am. Meteorol. Soc., Boston, MA.
Knight, C., and Squires, P., eds. (1982). "Hailstorms in the Central High Plains," Vols. 1 and 2, Colorado Associated Univ. Press, Boulder.
Pruppacher, H. R., and Klett, J. D. (1978). "Microphysics of Clouds and Precipitation," Reidel, Dordrecht, Netherlands.
Rogers, R. R. (1976). "A Short Course in Cloud Physics," Pergamon, Oxford.
Scorer, R. (1972). "Clouds of the World," Lothian Publ. Co., Melbourne, Australia.



Coastal Meteorology

S. A. Hsu

Louisiana State University

- I. Coastal Weather Phenomena
- II. Local Winds
- III. Boundary-Layer Phenomena
- IV. Air–Sea Interaction
- V. Hurricanes

GLOSSARY

- Aerodynamic roughness length** Measure of the roughness of a surface over which wind is blowing.
- Ageostrophic wind** Vector difference between measured wind and assumed geostrophic wind.
- Antitriptic wind** Wind in which the pressure gradient force exactly balances the viscous force.
- Baroclinic** State of atmosphere in which surfaces of constant pressure intersect surfaces of constant density or temperature.
- Brunt–Väisälä frequency** Frequency of vertical oscillation of an air parcel released after displacement from its equilibrium position.
- Coriolis force** Apparent force on moving particles caused by the Earth's rotation.
- Density current** Intrusion of colder air beneath warmer air caused mainly by hydrostatic forces arising from gravity and density or temperature differences.
- Frontogenesis** Initial formation of a front as a result of an increase in the horizontal gradient of the air mass property.
- Geostrophic wind** Horizontal wind in which the Coriolis and pressure-gradient forces are balanced.
- Gradient wind** Horizontal wind in which the Coriolis

acceleration and the centripetal acceleration together exactly balance the pressure-gradient force.

Inversion Increase in temperature with height as opposed to the normal condition, when temperature decreases with height.

Mesoscale Scale in which atmospheric motions occur within 2000 km of horizontal length and 24 hr of time.

Millibar (mb) Meteorological unit of pressure that equals 100 N m^{-2} .

Potential temperature Temperature a parcel of dry air would have if it were brought adiabatically (at approximately 1°C per 100 m) from its initial state to 1000 mb.

Rossby radius of deformation Length scale that is equal to $c/|f|$, where c is the wave speed in the absence of rotation effects and f is the Coriolis parameter caused by Earth's rotation.

Significant wave height Average of the highest one-third waves on the sea surface.

Stability parameters Dimensionless ratio of height, Z , and the Monin–Obukhov stability length, L , which relates to the wind shear and temperature (buoyancy) effect. Thus, the term Z/L represents the relative importance of heat convection and mechanical turbulence. The Richardson number is also a dimensionless ratio of wind shear and buoyancy force.

Supergeostrophic wind Type of wind with speed in excess of the local geostrophic value.

Synoptic scale Scale of atmospheric motion with horizontal length scale greater than 2000 km and time scale longer than 24 hr.

Thermal wind Vertical shear of geostrophic wind directed along the isotherms with cold air to the left in the Northern Hemisphere and to the right in the Southern Hemisphere.

Upwelling Slow, upward motion of deeper water.

Vorticity Vector measure of fluid rotation.

Wet-bulb potential temperature Wet-bulb temperature after an air parcel is cooled from its initial state adiabatically and then brought to 1000 mb.

COASTAL METEOROLOGY is an integral part of the total-system approach to understanding coastal environments. Since meteorology is the study dealing with the phenomenon of the atmosphere, coastal meteorology may be defined as that part of meteorology that deals mainly with the study of atmospheric phenomena occurring in the coastal zone. This description includes the influence of atmosphere on coastal waters and the influence of the sea surface on atmospheric phenomena, that is, the air-sea interaction.

The behavior of the atmosphere can be analyzed and understood in terms of basic laws and concepts of physics. The three fields of physics that are most applicable to the atmosphere are radiation, thermodynamics, and hydrodynamics. Owing to limitations of space, only a few topics in coastal meteorology are covered here. More information is available through the bibliography.

I. COASTAL WEATHER PHENOMENA

A. Coastal Fronts

1. Synoptic Scale Phenomena

The coastal front marks the boundary between cold continental air and warm oceanic air. A thermally direct frontal circulation exists, resulting in local enhancement of precipitation on the cold side of the frontal boundary. The coastal front is also a low-level baroclinic zone in which upward vertical motion and temperature advection in a narrow area along the coast exist. Along the U.S. Atlantic coast, both geostrophic and observed wind deformation play a role in coastal frontogenesis. The frontogenetical process involves a weak cyclone, which strengthens the preexisting temperature gradient as it moves northward. A moist baroclinic zone remains in place along the coast in the absence of strong cold advection in the wake of the

weak cyclone. The residual moisture-enhanced baroclinicity and surface vorticity are important factors contributing to a second disturbance, which intensifies as it moves northeastward parallel to the coast along the frontal zone.

2. Mesoscale Phenomena

Frictional convergence at coastlines. The coastline generally represents a marked discontinuity in surface roughness. The resulting mechanical forcing leads to a secondary circulation in the boundary layer and, consequently, to a vertical motion field that may have a strong influence on the weather in the coastal zone. In one example, heavy shower activity along the Belgium and Netherlands coasts was caused by frictional uplifting and frontogenesis occurring when a maritime polar air mass hit the coastline at a critical angle. By utilizing numerical models, scientists from the Netherlands recently found that upward motion is most pronounced when the geostrophic wind makes a small (about 20°) angle with the coastline (in a clockwise direction) and not when the geostrophic wind is perpendicular to the coastline, as is sometimes mentioned. The asymmetry relative to normal to the coastline is caused by Coriolis acceleration and not by a nonlinear effect.

Boundary-layer fronts at sea. These fronts or convergence lines develop in the cold air at sea when there are large bends or kinks in the shape of the upstream land or ice boundary from which the cold air is flowing. The air on one side of the convergence line has had a different over-water trajectory than that on the other side. The existence of such convergence in the cold air may, under the right conditions, very well be a factor in the genesis of polar vortices, which may later be intensified into polar (arctic) lows.

Orographically forced cold fronts. A coastal mountain range, for example, in southeastern Australia, can sometimes block shallow cold fronts with a northwest/southeast orientation. The violent behavior of some cold-front passages, or southerly busters, is found by Australian scientists to be at least orographically initiated. The head of the front has the character of an evolving density current, and its propagation is well predicted by density current theory over more than half of its lifetime. The horizontal roll vortex just behind the front is found to be accelerating relative to the rate of advection of cold air behind the front. This evolution is governed by warm-air entrainment.

Sea-breeze front. A sea-breeze front has the properties of the head of a gravity current. The front head has the shape of a lobe at its nose and a cleft behind the nose

above the denser air from the sea. The cleft can engulf overriding lighter air from the land. The circulation at the head may form a cutoff vortex.

Island-induced cloud bands. The development of cloud bands induced by an island is a complex interaction between the airflow and the geometry of the island. First, the upwind surface flow forms a separation line with an associated stagnation point. Then, a low-level convergence zone develops along this line, resulting in an updraft line. If the updrafts are strong enough, a band cloud forms. To characterize such a flow, the Froude number, Fr , is often employed ($Fr = U/NH$, where U is the upstream windspeed, N is the Brunt–Väisälä frequency, and H is the characteristic height of the island mountain). As an example, $Fr \approx 0.2$ for the island of Hawaii, where a well-defined band cloud was observed offshore of Hilo. A higher Fr tends to induce a stronger band that forms closer to shore. When $Fr \approx 1$, orographic clouds may form. On the other hand, when $Fr \approx 0.1$, the convergence zone moves offshore and the cloud band may be weak or may even disappear.

Cold-air damming. Cold-air damming exists when the cold air over land becomes entrenched along mountain slopes that face a warmer ocean. For example, along the eastern slopes of the Appalachians, the temperature difference can exceed 20°C between the damming region and the coast, a distance of approximately 150 km. This Appalachian cold-air damming was investigated in detail most recently by G. D. Bell and L. F. Bosart, who reveal that damming events occur throughout the year but peak in the winter season, particularly December and March, when three to five events per month might be expected. Cold-air damming is favored in late autumn and early winter when the land is coldest relative to the ocean. The event is critically dependent upon the configuration of the synoptic-scale flow. The presence of a cold dome is indicated by a U-shaped ridge in the sea-level isobar pattern, as well as in the 930-mb (about 700–800 m) height field, which is near the top of the cold dome. The potential temperature contours are also pronouncedly U shaped in the damming region, indicating relatively uniform cold air in the dome bounded by a strong baroclinic zone just to the east.

At the onset of the cold dome, warm air advection was observed over the surface-based layer of cold air advection. This differential vertical thermal advection pattern aided in generating and rapidly strengthening an inversion at the top of the cold dome, resulting in decoupling of the northeasterly flow in the cold dome from the southeasterly flow just above it. Force balance computation indicated that the acceleration of the flow to the speed of the low-level wind maximum at 930 mb was governed by

the mountain-parallel component of the large-scale height (or pressure) gradient force. After the formation of the cold dome, the force balance on the accelerated flow was geostrophic in the cross-mountain direction and antitriptic in the along-mountain direction. Cold dome drainage occurs with the advection of the cold air toward the coast in response to synoptic-scale pressure falls accompanying coastal cyclogenesis.

Cyclogenesis. Cyclogenesis is defined as any development or strengthening of cyclonic circulation in the atmosphere. In certain coastal regions, cyclogenesis is a very important phenomenon, for example, along the mid-Atlantic coast of the United States and in the northwestern Gulf of Mexico. The cyclones that develop over the Yellow Sea and East China Sea often cause strong gales, and the cold air west of the cyclones spreads southward as an outbreak of the winter monsoon. Main features and processes contributing to coastal cyclogenesis along the U.S. East Coast are significant sensible heat transport over the ocean and latent heat release along the East Coast, coastal frontogenesis, and a polar jet streak propagating eastward. In a case of East Asian coastal cyclogenesis, a numerical experiment that included all physical processes simulated the development of a cyclone that developed rapidly in a way similar to that observed. In an experiment without latent heat feedback, only a shallow low appeared when the upper short-wave trough approached the inverted surface trough situated on the coast, but no further development took place. This suggests that the baroclinic forcing was enhanced by the feedback of physical processes. The latent heating had a profound impact on the amplifying jet streak circulation and the vertical coupling within the system, which appeared to prime the rapid cyclogenesis along the coast. Sensible heating contributed nearly 18% to the surface development. It helped to build a potential temperature contrast along the coast below 900 mb. Without sensible heating, the model-latent heat release was reduced. Thus, the impact of sensible heating was partly through the moist processes rather than direct heating.

Cyclogenesis in the western Gulf of Mexico is contributed in most cases by a mountain-induced standing wave developed on a climatological surface baroclinic zone over the Gulf. The effect of surface-layer baroclinicity on cyclogenesis is shown in Fig. 1. A close relationship is found between the frequency of occurrence of frontal overrunning over New Orleans, LA, and the air temperature difference between the shelf (shallow) water and deeper (warm) ocean water. Because the surface-based vorticity is directly and linearly proportional to the temperature difference across the cold shelf water and warmer Gulf water, Fig. 1 indicates that there is definitely a correlation between the surface-based, vorticity-rich air and

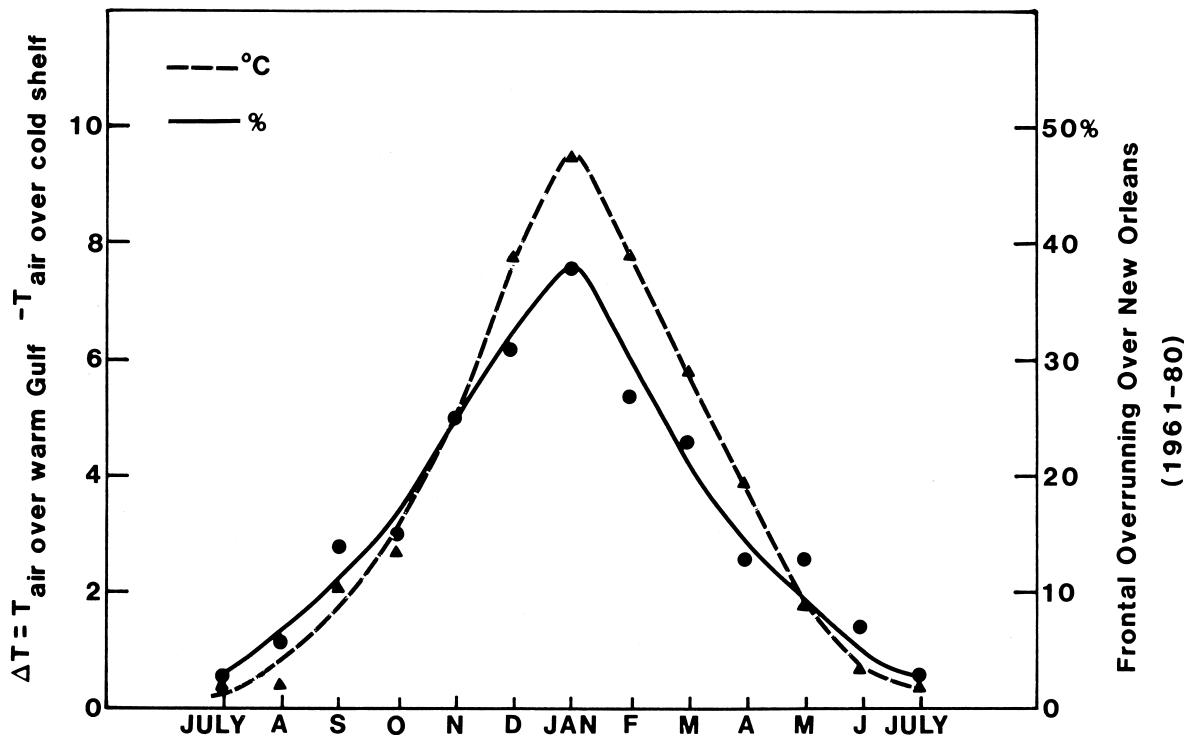


FIGURE 1 Correlation between the frequency of frontal overrunning over New Orleans, LA, and the difference in air temperature over warmer Gulf of Mexico and colder shelf waters.

the temperature difference or baroclinic zone occurring between the colder land/shelf water and the warmer deep-ocean water. A larger scale baroclinic or solenoidal field from Key West, FL, to Del Rio, TX, via Victoria, TX, and the Gulf of Mexico is shown in Fig. 2. The cold pool or cold-air damming, as discussed previously, over the cold shelf water off the coast of southern Texas and the Victoria region is clearly delineated. On the other hand, a warmer region over the Loop Current west of Key West is also illustrated.

An example of the cyclogenesis over the Western Gulf of Mexico including its effect and classification is provided in Figs. 3–6.

B. Fog and Stratiform Clouds

Fog and stratus are clouds, but the base of fog rests on the Earth's surface and stratus clouds are above the surface. Although the substance of fog and cloud is the same, their processes of formation are different. Clouds form mainly because air rises, expands, and cools. Fog results from the cooling of air that remains at the Earth's surface. Scientists at the University of Nevada recently revealed the following characteristics for convection over oceans: when warm air flows over cooler water, the air layer over the sea will usually convect even when the water surface

is 10° or more colder than the initial air temperature. An inversion at stratus cloud tops can be created by the stratus. Such inversions persist after subsidence evaporates the cloud. Radiation heat exchange does not play an essential role in stratus formation or maintenance and can either heat or cool the cloud. Dry air convection does not erode inversions at the top of the convective layer. Fogs are most likely to form at sea, where the water is coldest and needs no radiation effects to initiate cooling, or a boost from patches of warmer water, to begin convection. Both stratus cloud growth and the evaporation of clouds by cloud-top entrainment readjust the vertical structure of the air to leave a constant wet-bulb potential temperature with height.

The stratocumulus cloud deck over the east Pacific has large cloud variability, on 1–5 km scales. The cloud deck slopes upward from 700 to 1000 m in a northeast–southwest direction over a distance of 120 km. In the examples studied, vertical cloud top distributions were negatively skewed, indicating flat-topped clouds. The dominant spectral peak of the cloud-top variations was found at 4.5 km, which is 5–7 times the depth of the local boundary layer. The cloud layer was stable with respect to cloud-top entrainment instability. Structural properties of stratocumulus clouds observed off the coast of southern California, near San Francisco, and in the Gulf of Mexico are

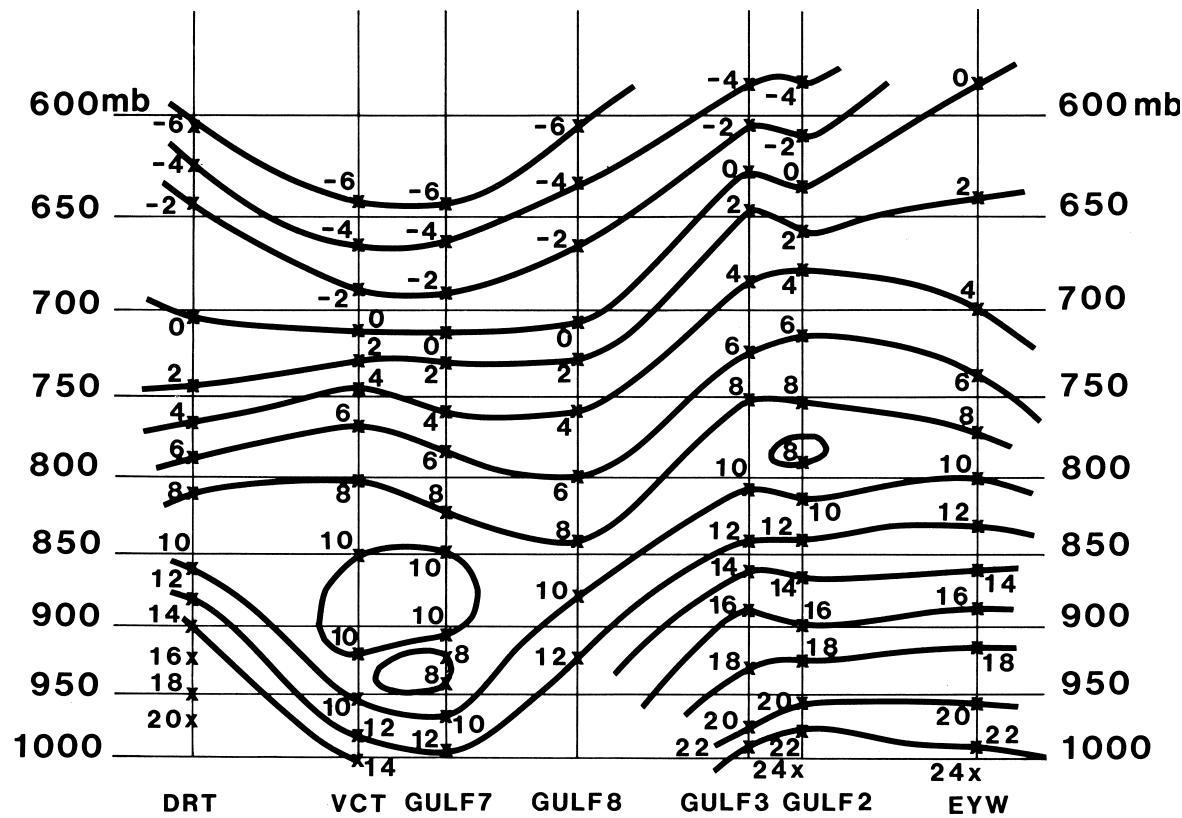


FIGURE 2 An example of a baroclinic (or solenoidal) field from Key West, FL, to Del Rio, TX, via the Gulf of Mexico and Victoria, TX, on February 22, 1986, during a special experiment. Based on radiosonde ascents from weather stations and radiosonde drops from airplanes.

similar. Marine stratocumulus cloud fields composed primarily of small cells have very steep slopes and reach their asymptotic values at short distances from the origin. As the cells composing the cloud field grow larger, the slope becomes more gradual and the asymptotic distance increases accordingly.

II. LOCAL WINDS

A. Land and Sea Breezes

The best example of local winds in the coastal zone is perhaps the land-sea breeze system. This coastal air-circulation system brings fresh air from the sea in the afternoon to cool coastal residents, whereas farther inland hot and still air is the general rule. On coasts and shores of relatively large lakes, because of the large diurnal temperature variations over land as compared to that over water, a diurnal reversal of onshore (sea breeze) and offshore (land breeze) wind occurs.

A sea breeze develops a few hours after sunrise, continues during the daylight hours, and dies down after sunset. Later, a seaward-blowing land breeze appears and contin-

ues until after sunrise. The sea breeze may extend up to 50 or 100 km inland, but the seaward range of the land breeze is much smaller. In the vertical, the sea breeze reaches altitudes of 1300–1400 m in tropical coastal areas, with a maximum speed at a few hundred meters above the ground. In contrast, the nocturnal land breeze is usually rather shallow, being only a few hundred meters deep. Typical horizontal speeds of the sea breeze are of the order of meters per second, while the vertical components are only a few centimeters per second. At specific locations, large and abrupt temperature and relative humidity changes can occur with the passage of the sea-breeze front. An example of the land- and sea-breeze system is shown in Fig. 7. The onshore and offshore wind components are shown at 3-hr time intervals during the day. The lower portion of the onshore flow is the sea breeze and that of the offshore flow is the land breeze. The maximum wind speed and its approximate height in each current are depicted by arrows. The elliptical shapes in the figure illustrate the horizontal and vertical extent of the land-sea breeze circulation. The dashed horizontal line represents the 900-mb pressure surface (approximately the convective condensation level). At 0900 LST (local standard time), the air



FIGURE 3 An example of cyclogenesis which took place over the Gulf of Mexico on February 16, 1983. This shot was taken from the GOES satellite. Notice the comma-shaped whirlpool cloud pattern and also the fact that this system was not linked to other larger scale systems. This was one of the top five cyclones generated over the Gulf of Mexico during the 1982–1983 El Niño period. Not only do surface conditions, such as sea surface temperatures, play an important part in the development and intensification of these storms, but the upper atmospheric conditions are critical as well.

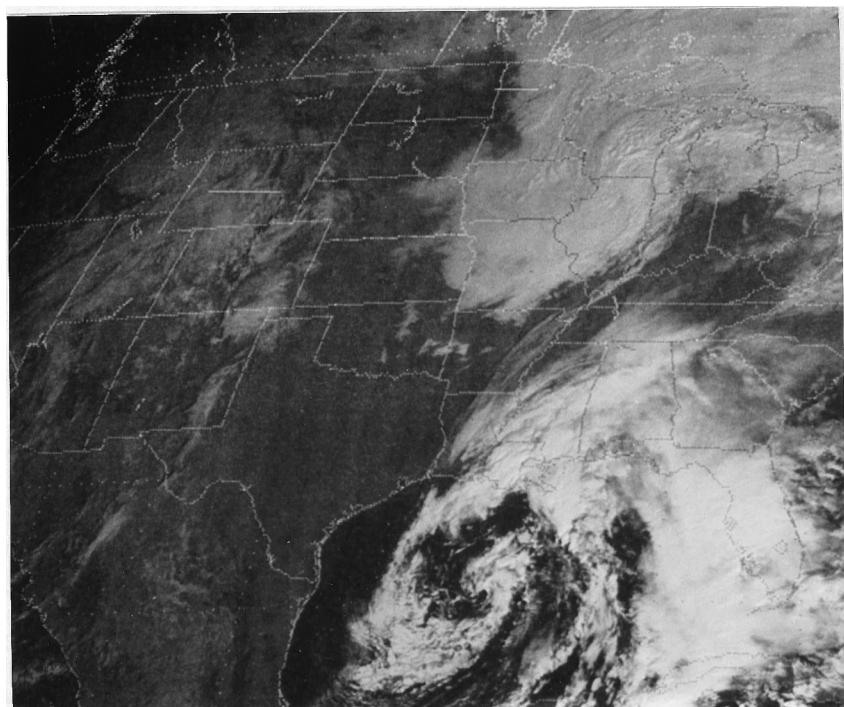


FIGURE 4 An enlargement of Fig. 3 over the western Gulf of Mexico.

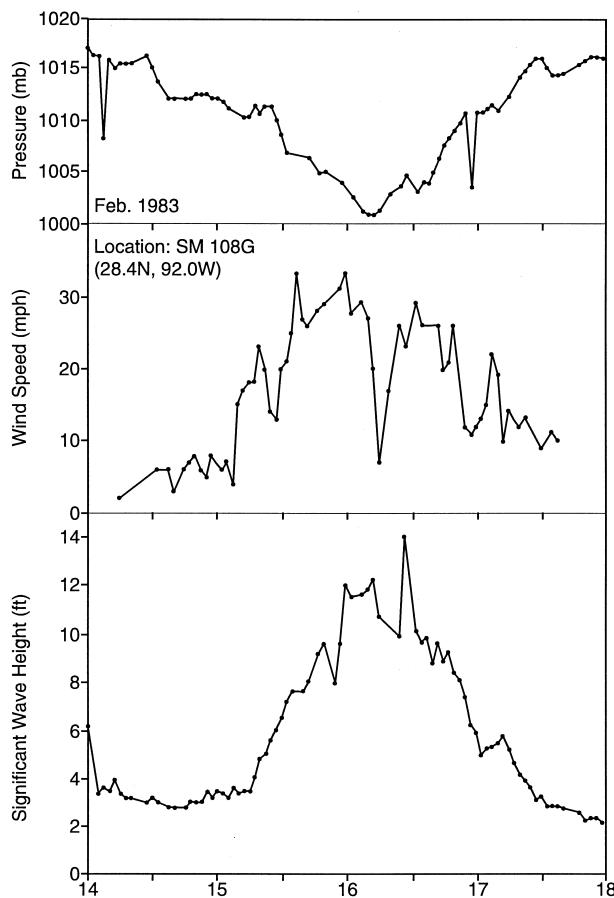


FIGURE 5 The time series analysis for this storm (Fig. 4) was made from a data buoy for atmospheric pressure, wind speed, and significant wave height during the period of cyclogenesis. Note the relationship between pressure and winds. The maximum wind speed does not usually occur at the time of the lowest pressure, but in general the lower the pressure the stronger the wind will be. This particular time series very much resembles a typical tropical cyclone plot as the wind speed would be expected to drop off dramatically in the eye or center of lowest pressure.

temperature over land is still cooler than over the sea and the land breeze is still blowing. By 1200 LST, the land has become warmer than the water, and the circulation has reversed. At this time, a line of small cumulus may mark the sea-breeze front. At 1500 LST, the sea breeze is fully developed, and rain showers may be observed at the convergence zone, 30 to 40 km inland. Because of a low-level velocity divergence, there is a pronounced subsidence and thus a clear sky near the coastal area at this time. At 1800 and 2100 LST, the sea breeze is still clearly present but is gradually weakening in intensity. By midnight or 0000 LST, the sea breeze is barely evident aloft, and the surface wind is nearly calm over land. At this time, a temperature inversion and occasionally fog appear over land. After land again becomes cooler than the water, a land

breeze becomes well developed by 0300 LST and reaches its maximum intensity near 0600 LST. A weak land-breeze convergence line and associated line of cumulus clouds develop offshore near sunrise. The land breeze continues until midmorning, when the sea-breeze cycle starts over again. It is interesting to note that in this model the maximum strength of the land breeze in the near-surface layer is comparable to that of the sea breeze. Because of day-night differences in stability and frictional effects, however, the observed strength of the daytime sea breeze at the surface is considerably greater than the nighttime land breeze.

The importance of the effect of latitude on the sea-breeze circulation has been investigated numerically by scientists at the U.S. National Center for Atmospheric Research. They show that at the equator the absence of the Coriolis force results in a sea breeze at all times. At the other latitudes, the Coriolis force is responsible for producing the large-scale land breeze. At 20°N, the slower rotation of the horizontal wind after sunset produces a large-scale land breeze that persists until several hours after sunrise. At 30°N, the inertial effects produce a maximum land breeze at about sunrise, and the land breeze is strongest at this latitude. At 45°, the rotational rate of the horizontal wind after sunset is faster, so that the maximum land breeze occurs several hours before sunrise. These results indicate that the Coriolis force may be more important than the reversal of the horizontal temperature gradient from day to night in producing large-scale land breezes away from the equator.

Onshore penetration of the sea breeze varies with latitude also. In midlatitude regions (generally above 40°N), even under favorable synoptic conditions, the sea breeze may extend to 100 km. At latitudes equatorward of about 35°, much greater penetrations have been reported, for example, 250 km inland of the Pakistan coastline. In tropical regions of Australia, the sea breeze can penetrate to 500 km. In such cases, the sea breeze traveled throughout the night before dissipation occurred shortly after sunrise on the second day.

The effect of the sea breeze on the long-range transport of air pollutants to cause inland nighttime high oxidants has recently been investigated by Japanese scientists. On clear nights with weak gradient winds, a surface-based inversion layer often forms between sunset and sunrise. A strong inversion forms mainly in basin bottoms in the inland mountainous region between Tokyo on the Pacific Ocean and Suzaka near the Japan Sea. An air mass that passes over the large emission sources along the coastline can be transported inland by the sea breeze in the form of a gravity current. In the case studied, a high-concentration layer of oxidants was created in the upper part of the gravity current. It descended at the rear edge of a gravity-current head because of the internal circulation within the

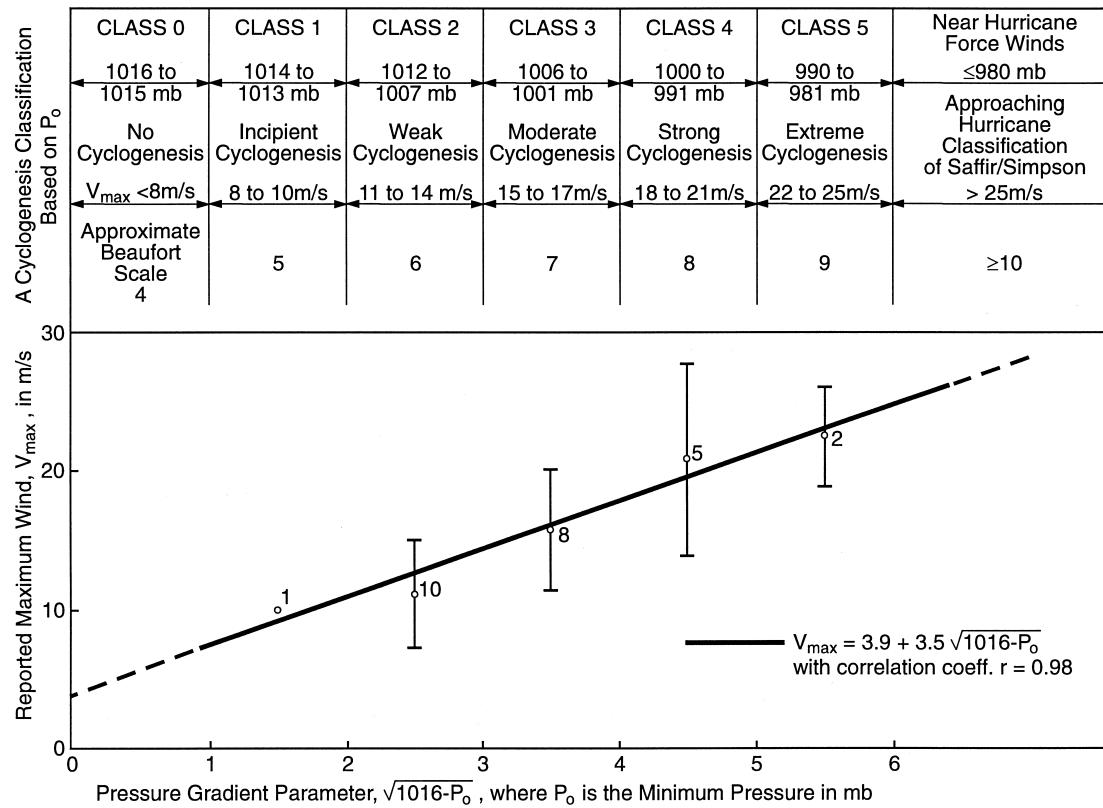


FIGURE 6 The cyclogenesis classification (top panel) is based on the minimum pressure of winter storms in the Gulf of Mexico. The bottom panel shows the number of storms studied, the relationship between the pressure gradient parameter and the reported maximum winds, while the vertical bars are the standard deviations. [Reprinted with permission from Hsu, S. A. (1993). *Mariners Weather Log* 37(2), 4.]

head, thus yielding the highest concentration of oxidants near the ground.

An example of the sea-breeze system along the coasts of Texas and Louisiana is presented in Figs. 8 and 9.

B. Low-Level Jets

Low-level jets have long been known to meteorologists. There are many manifestations of low-level jets around the world and many mechanisms for their formation. For instance, in the Northern Territory of Australia the mechanism for a jet would seem to be an inertial oscillation set up when the turbulent shearing stress falls dramatically with the formation of the nocturnal inversion. Observations from that region indicated that for geostrophic winds in the range of 10–20 m/sec ageostrophic wind magnitudes of 5–10 m/sec were common above the surface layer near sunset, with cross-isobar flow angles of above 40°. The jet that then developed by midnight was probably the result of these large ageostrophic winds, strong surface cooling, and favorable baroclinicity and sloping terrain.

Low-level jets usually have a well-marked super-geostrophic maximum in the boundary-layer wind speed

profile within a few hundred meters above the ground. They are modified by thermal stratifications, baroclinicity of the lower atmosphere, advective accelerations, and nonstationarity of the boundary layer. Some low-level jets have a diurnal life cycle with pronounced maxima during the night. An improved numerical model developed recently by German scientists is able to simulate the low-level jet. The improvements stem from the incorporation of a diurnally varying drag coefficient rather than a constant value for the entire 24 hr.

In certain regions, heavy rainfall is closely related to the low-level jet. For example, during the early summer rainy season in subtropical China and Japan, extremely heavy rainfall (at least 100 mm/day) is one of the most disastrous weather phenomena. As in other parts of East Asia, extremely heavy rainfall is also found by meteorologists in Taiwan to be closely associated with a low-level jet. For example, there was an 84% likelihood that a low-level jet of at least 12.5 m/sec would be present at 700 mb (about 3 km) 12 hr before the start of the rainfall event. They concluded that the low-level jet may form to the south of heavy rainfall as part of the secondary circulation driven by convective latent heating.

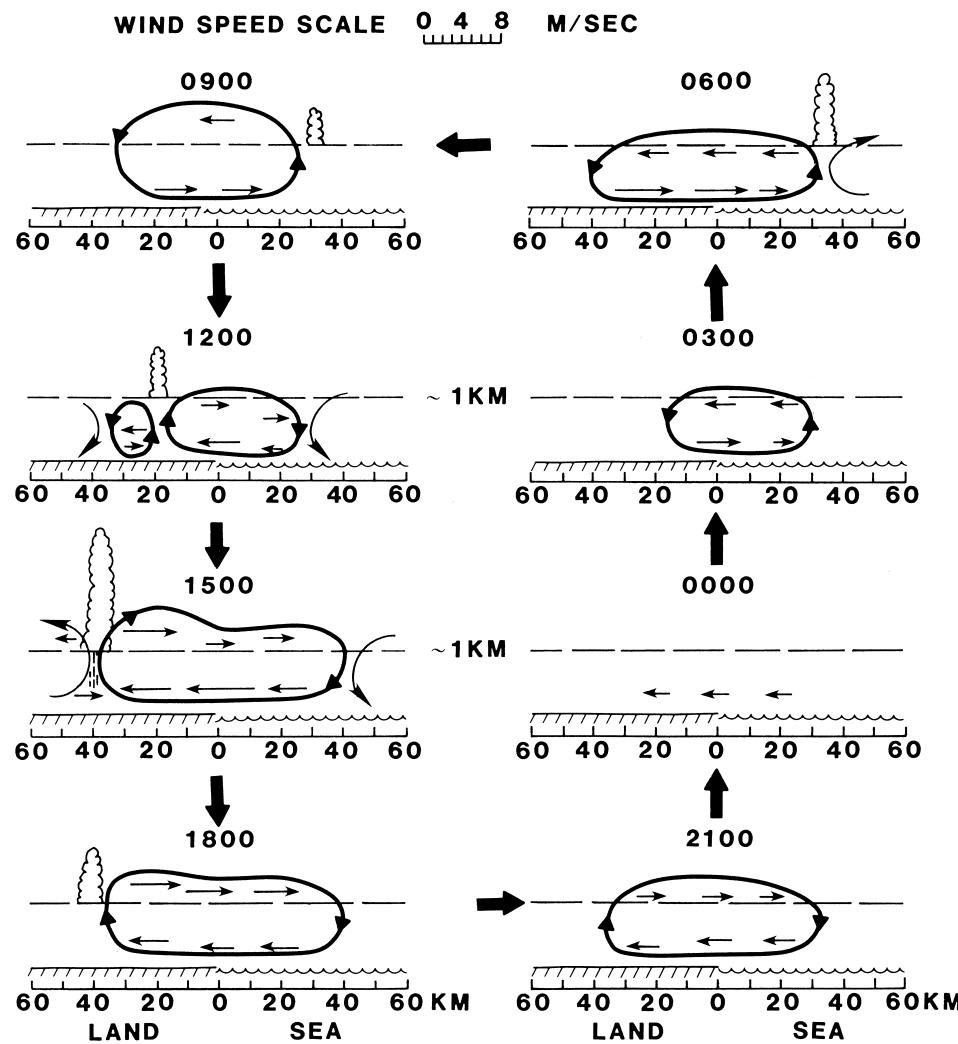


FIGURE 7 A simplified synthesized observed life cycle of the land-sea breeze system along the Texas Gulf Coast. Arrow lengths are proportional to wind speed. See text for explanation.

Along the coast of California during north-westerly upwelling favorable winds, the marine atmospheric boundary layer is characterized by a low-level jet, with peak wind speeds of as much as 30 m/sec at elevations of a few hundred meters. The vertical structure is marked by an inversion, usually at or near the elevation of the wind speed maximum. Above the inversion, the stratification is stable, and the wind shear is caused primarily by baroclinicity (thermal wind) generated by the horizontal temperature gradient between the ocean and land. Below the inversion, the flow is turbulent.

C. Other Coastal Winds

1. Wind Reversals along the California Coast

North winds along the northern California coast in summer may be interrupted by southerly winds. At the start of the particular event studied, the marine layer thickened in the

southern California bight. A couple of days later the marine layer thickened from Point Conception to Monterey. Then, the marine layer thickness increase surged to the north along the coast to Point Arena, where progression stopped and an eddy formed. In this surging stage, winds switched to southerlies as the leading edge of the event passed. A day later, the leading edge surged farther to the north. Inshore winds were southerly, and the lifted marine layer extended to Cape Blanco in Oregon. This event is interpreted by C. Dorman as a gravity current surging up the coast.

2. Offshore-Directed Winds over the Gulf of Alaska

The strong thermal contrast between relatively warm offshore waters and frigid air over the interior plateau of Alaska creates a region of hydrostatic pressure contrast

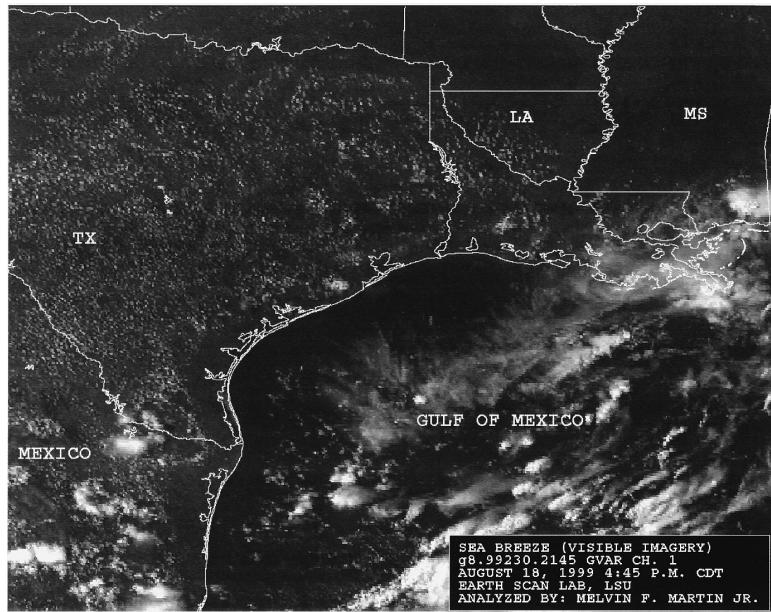


FIGURE 8 The sea-breeze system along the Gulf coast of Mexico, Texas, and Louisiana. This visible imagery from the GOES satellite shows the sinking air (or subsiding and thus clearing) on both sides of the shoreline. Notice the existence of the sea-breeze front or the convergence line displaced inland from the shore.

along the southern gulf coast of Alaska during the cold season. As a result, there is frequent regional offshore flow and nearly continuous drainage flow through mountain gaps in this region during the winter months. Scientists from the U.S. Pacific Marine Environmental Labora-

tory found that coastal mountains around Prince William Sound contribute to offshore winds in three ways: (1) by forming a physical barrier to low-level coastal mixing of cold continental and warm marine air, (2) by providing gaps through which dense continental air may be

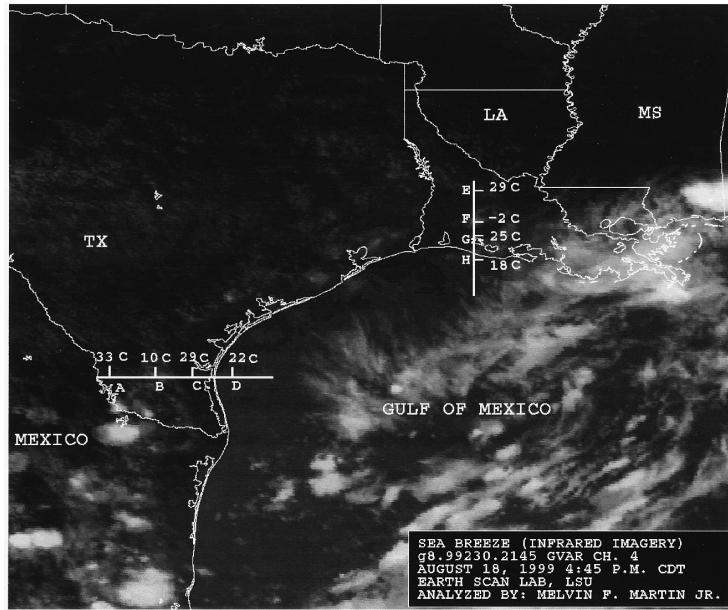


FIGURE 9 Infrared imagery from the GOES satellite for Fig. 8. Two lines across the sea-breeze front are delineated, one in south Texas and the other in west Louisiana. They provide the horizontal temperature distribution of the cloud top at B (south Texas) and F (west Louisiana), sea surface (at D and H), and ground (at A, C, E, and G). Note that both B and F are located on the sea-breeze front. Also, there is 4°C difference between A and C as well as between E and G, indicating the advancing of cooler air onshore associated with the sea-breeze system.

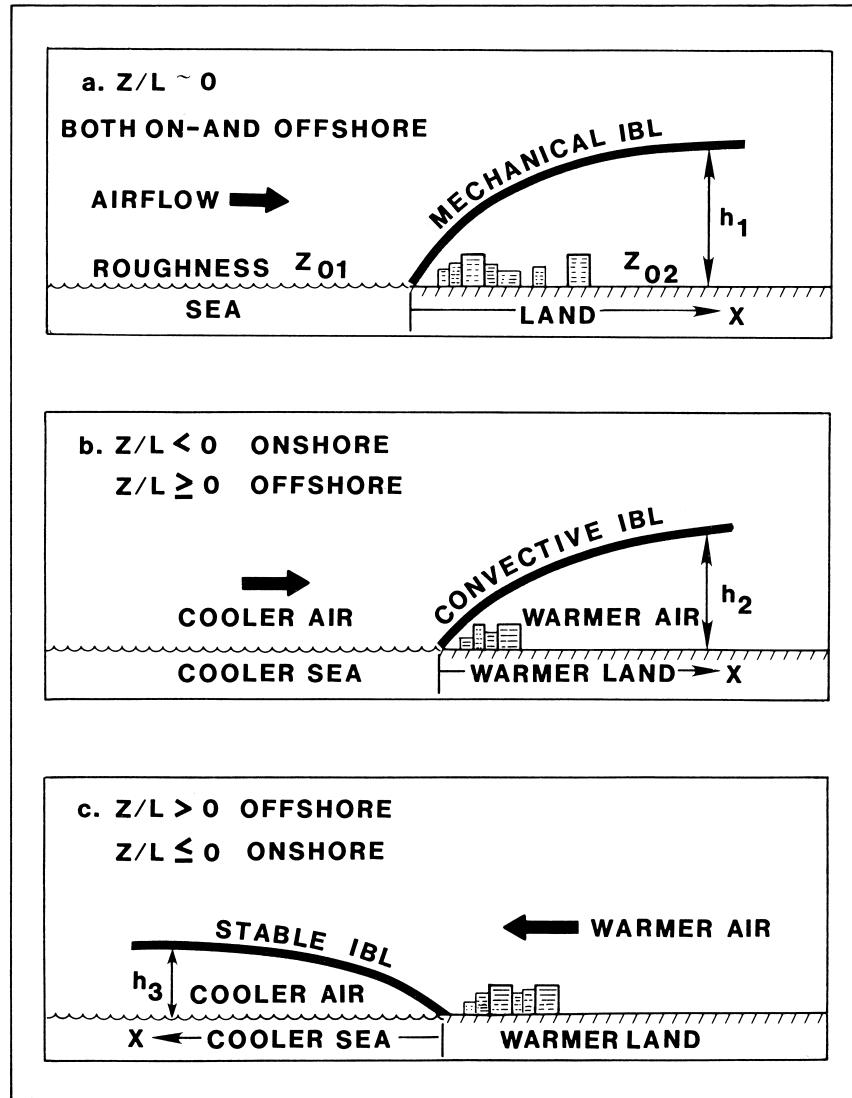


FIGURE 10 Schematics of the development of internal boundary layers across a shoreline.

channeled, and (3) by exciting mountain-lee waves. Nearshore winds are perturbed and highly localized. Cold drainage winds are eroded by heat and momentum transfer within a few tens of kilometers of the coast. Geotriptic adjustment of the regional surface-wind field is nearly achieved within a distance from the barrier corresponding to the Rossby radius of deformation. Note that in a geotriptic flow the acceleration term in the horizontal equations of motion is balanced by the pressure gradient force, Coriolis force, and frictional force.

III. BOUNDARY-LAYER PHENOMENA

A. Internal Boundary Layer

When air flows across a shoreline, its boundary layer undergoes significant modification in aerodynamic rough-

ness, potential temperature, mixing ratio, and aerosol concentration. Because these changes in atmospheric layering occur within the planetary or atmospheric boundary layer, the modified layer is dubbed the internal boundary layer (IBL). There are two major IBLs, one thermal and the other mechanical. The thermal IBL can be further classified as convective or stable, depending upon the wind direction with respect to the temperature contrast between the land and sea. Figure 10 shows these IBLs. A review of the IBL is given in Garratt (1990).

B. Mechanical Internal Boundary Layer

The mechanical IBL develops across the shoreline because of roughness changes. Mechanical turbulence overpowers thermal contrast to make the stability parameter, Z/L , close to zero, or neutral. As shown in Fig. 10a, the

height of the mechanical IBL, h_1 , is given by the general form

$$h_1 = a \times Z_{02} \times (x/Z_{02})^b. \quad (1)$$

Experimental results have shown that

$$a = 0.75 + 0.03 \ln(Z_{01}/Z_{02}),$$

where Z_{01} and Z_{02} are the aerodynamic roughness lengths over the water (upwind) and land (downwind), respectively; X is the fetch downwind from the shoreline; and the power, b , is equal to approximately 0.8. Typical values of roughness length are about 0.01 cm for the water surface and about 10 cm for a relatively flat coast.

C. Convective Internal Boundary Layer

A convective IBL develops when cooler air flows from the sea to warmer land. It is modified by the temperature contrast, as shown in Fig. 10b. The height of the convective IBL, h_2 , as derived by A. Venkatram, is

$$h_2 = \left[\frac{2C_d(\theta_{\text{land}} - \theta_{\text{sea}})X}{\gamma(1 - 2F)} \right]^{1/2}, \quad (2)$$

where C_d is the drag coefficient inside the convective IBL; γ is the lapse rate above the boundary layer or upwind conditions; F is an entrainment coefficient, which ranges from 0 to 0.22; θ_{land} and θ_{sea} are the potential air temperatures over land and water, respectively; and X is the distance or fetch downwind from the shoreline. The dependency of h on $X^{1/2}$ has been predicted by dimensional analysis and by the moddynamic approaches.

Equation (2) can be rewritten as

$$h_2 = AX^{1/2}. \quad (3)$$

In flat coastal regions, $A \approx 60$ if h_2 is in meters and X is in kilometers.

Over urban areas in the coastal zone, however, the following formulas for the height of a daytime convective IBL have been recommended by S. R. Hanna for operational application:

$$h_2 = 0.1X \quad \text{for } X \leq 200 \text{ m} \quad (4a)$$

$$h_2 = 200 \text{ m} + 0.03(X - 2000 \text{ m}) \quad \text{for } X > 2000 \text{ m}, \quad (4b)$$

where X is the distance from the shoreline.

Under convective IBL conditions, the phenomenon of fumigation near the shoreline is a common occurrence. Recent studies have shown that for maximum ground-level concentrations

$$X_{\text{max}} \simeq 0.3Q/(uh^2)$$

which occur at a downwind distance of

$$X_{\text{max}} \simeq 10(u/W_*)h,$$

where W_* is the convective velocity scale,

$$W_* = (gHh/C_p\rho T)^{1/3},$$

where H is the sensible heat flux, T is the air temperature near the surface, u is the wind speed, h is the IBL height, and C_p is the specific heat at constant pressure, all inside the convective IBL.

A similar phenomenon exists when cooler air flows from a colder sea toward an oceanic warm front. For example, over the Gulf Stream when the wind direction is approximately perpendicular to the edge of the stream, a convective IBL may develop. At approximately 70 km downwind from the edge of the Gulf Stream, the height of the convective IBL, h_2 , was observed at 300 m. This says that over the Gulf Stream when conditions are right

$$h_2 \approx 36X^{1/2}. \quad (5)$$

Comparison of Eqs. (3) and (5) indicates that the height of the convective IBL over a coast is higher than that over an ocean for a given fetch and temperature difference. This is mainly due to the higher drag coefficient ($\approx 10 \times 10^{-3}$) across the shoreline on land than that ($\approx 1.5 \times 10^{-3}$) across an oceanic front at sea.

D. Stable Internal Boundary Layer

Contrary to the convective internal boundary layer, a stable IBL develops when warmer air is advected from an upstream warmer land (or sea) surface to a cooler sea downstream. This situation is shown in Fig. 10c. The main difference between convective and stable IBLs is that the heat flux is directed upward (from a warm sea to cooler air) for a convective IBL and downward (from warm air to a cooler sea) for a stable IBL. A two-dimensional numerical mesoscale model was used by J. R. Garratt to investigate the internal structure and growth of a stably stratified IBL beneath warm continental air flowing over a cooler sea. An analytical model was also used by Garratt to study a stable IBL, and excellent agreement with the numerical results was found. This analytical model states that

$$h_3 = BX^{1/2}U(g\Delta\theta/\theta)^{-1/2}, \quad (6)$$

where h_3 is the depth of the stable IBL, which relates to X , the distance from the coast; U , the large-scale wind (both normal to the coastline); $g\Delta\theta/\theta$, in which $\Delta\theta$ is the temperature difference between continental mixed-layer air and sea surface; the mean potential temperature, θ ; the acceleration caused by gravity, g ; and other parameters combined as B .

From numerical results, Garratt suggests that B is a constant with a value of 0.014. Actually, B relates several other parameters, such as the flux Richardson number, the geostrophic drag coefficient, and the angle of the geostrophic wind measured counterclockwise from the positive X direction.

Comparing Eqs. (2) and (6) indicates that, for a convective IBL, $h \propto |\Delta\theta|^{1/2}$ and, for a stable IBL, $h \propto |\Delta\theta|^{-1/2}$. This difference is important, since both equations are consistent with energetic considerations.

For operational purposes, Eq. (6) may be simplified. Appropriate experiments in a flat coastal zone show that (see Fig. 10c)

$$h_3 \approx 16X^{1/2}, \quad (7)$$

where h_3 is in meters and X is in kilometers.

IV. AIR-SEA INTERACTION

A. Meteorological Fluxes

Meteorological transport processes in coastal marine environments are important from several points of view. For example, the wind stress or momentum flux is one of the most essential driving forces in water circulation. Heat and convection are the origin of some localized coastal weather systems. Sensible heat and water vapor fluxes are necessary elements in radiation and heat budget considerations, including computation of evaporation and salt flux for a given estuarine system.

In the atmospheric surface boundary layer, the vertical turbulent transports are customarily defined as, for practical applications,

$$\text{Momentum flux} = \rho u_*^2 = \rho C_d U_{10}^2 = \tau$$

$$\text{Sensible heat flux} = \rho C_p C_T (T_{\text{sea}} - T_{\text{air}}) U_{10} = H_s$$

$$\text{Latent heat flux} = L_T \rho C_E (q_{\text{sea}} - q_{\text{air}}) U_{10} = H_l$$

$$\text{Moisture flux} = \rho C_E (q_{\text{sea}} - q_{\text{air}}) U_{10} = E$$

$$\text{Buoyancy flux} = C_T U_{10} (T_{\text{sea}} - T_{\text{air}}) \left(1 + \frac{0.07}{B} \right)$$

$$\text{Bowen ratio} = B = \frac{H_s}{H_l}$$

where

ρ = air density

u_* = friction velocity

C_d = the drag coefficient

U_{10} = wind speed at 10 m above the sea surface

C_p = the specific heat capacity at constant air pressure

C_T = the sensible heat coefficient

T_{sea} = the “bucket” seawater temperature in the wave-mixed layer (in °C)

T_{air} = the mean air temperature at the 10 m reference height (in °C)

L_T = the latent heat of vaporization

C_E = the latent heat flux coefficient

q_{sea} = the specific humidity for the sea

q_{air} = the specific humidity for the air

E = evaporation

Operationally, according to the [WAMDI Group \(1988\)](#), $u_* = U_{10} \sqrt{C_d}$, where

$$C_d = \begin{cases} 1.2875 * 10^{-3}, & U_{10} < 7.5 \text{ m sec}^{-1} \\ (0.8 + 0.065U_{10}) * 10^{-3}, & U_{10} \geq 7.5 \text{ m sec}^{-1} \end{cases}$$

According to [Garratt \(1992\)](#), $C_T = C_E \simeq 1.1 * 10^{-3}$ ($\pm 15\%$); and according to [Hsu \(1999\)](#),

$$B = 0.146(T_{\text{sea}} - T_{\text{air}})^{0.49}, \quad T_{\text{sea}} > T_{\text{air}}.$$

The atmospheric stability parameter, Z/L , at height Z over the sea surface is defined as

$$\frac{Z}{L} = -\frac{\kappa g Z C_T (T_{\text{sea}} - T_{\text{air}}) \left(1 + \frac{0.07}{B} \right)}{(T_{\text{air}} + 273) U_Z^2 C_d^{3/2}},$$

where L = Monin-Obukhov stability length

κ = Von Karman constant = 0.4

g = gravitational acceleration = 9.8 m sec⁻²

U_Z = wind speed at height Z , normally set to 10 m

Thus, Z/L represents the relative importance between (or simply the ratio of) the buoyancy effect (or thermal turbulence) and the wind-shear or mechanical turbulence. Note that if $T_{\text{sea}} > T_{\text{air}}$, Z/L is negative, this stands for an unstable condition. On the other hand, if $T_{\text{air}} > T_{\text{sea}}$, Z/L is positive, the stability is said to be stable. When $T_{\text{air}} \simeq T_{\text{sea}}$, $Z/L \simeq 0$, the stability is near neutral.

B. Wind-Wave Interaction

Ocean surface waves are primarily generated by the wind. Because the water surface is composed randomly of various kinds of waves with different amplitude, frequency, and direction of propagation, their participation is usually decomposed into many different harmonic components by Fourier analysis so that the wave spectrum can be obtained from a wave record. Various statistical wave parameters can then be calculated. The most widely used parameter is the so-called “significant wave height,” $H_{1/3}$, which is defined as the average height of the highest one-third of the waves observed at a specific point. Significant wave height is a particularly useful parameter because it is approximately equal to the wave height that a trained observer would visually estimate for a given sea state.

In the fetch-limited case (i.e., when winds have blown constantly long enough for wave heights at the end of the fetch to reach equilibrium), the parameters required for wave estimates are the fetch F and U_{10} . The interaction among wind, wave, and fetch are formulated for

operational use as simplified from the U.S. Army Corp of Engineers (1984):

$$\frac{gH_{1/3}}{U_{10}^2} = 1.6 * 10^{-3} \left(\frac{gF}{U_{10}^2} \right)^{1/2}$$

$$\frac{gT_p}{U_{10}} = 2.857 * 10^{-1} \left(\frac{gF}{U_{10}^2} \right)^{1/3}.$$

The preceding equations are valid up to the fully developed wave conditions given by

$$\frac{gH_{1/3}}{U_{10}^2} = 2.433 * 10^{-1}$$

$$\frac{gT_p}{U_{10}} = 8.134,$$

where g is the gravitational acceleration, $H_{1/3}$ is the spectrally based significant wave height, T_p is the period of the peak of the wave spectrum, F is the fetch, and U_{10} is the adjusted wind speed.

In order to estimate hurricane-generated waves and storm surge, the following formulas are useful operationally, provided the hurricane's minimum (or central) pressure near the surface, P_0 , in millibars is known (see Hsu, 1988, 1991, and 1994).

$$H_{\max} = 0.20(1013 - P_0)$$

$$\frac{H_r}{H_{\max}} = 1 - 0.1 \frac{r}{R}$$

$$\Delta S = 0.069(1013 - P_0) = 0.35H_{\max},$$

where H_{\max} (in meters) is the maximum significant wave height at the radius of maximum wind, R (in kilometers); H_r (in meters) is the significant wave height at the distance r away from R ; and ΔS (in meters) is the maximum open-coast storm surge (i.e., above the astronomical tide) before shoaling. The mean R value for hurricanes is approximately 50 km.

V. HURRICANES

Hurricanes are tropical cyclones that attain and exceed a wind speed of 74 mph (64 knots or 33 m sec⁻¹). A hurricane is one of the most intense and feared storms in the world; winds exceeding 90 m sec⁻¹ (175 knots or 200 mph) have been measured, and rains are torrential. The Saffir–Simpson damage-potential scale is used by the U.S. National Weather Service to give public safety officials a continuing assessment of the potential for wind and storm-surge damage from a hurricane in progress. Scale

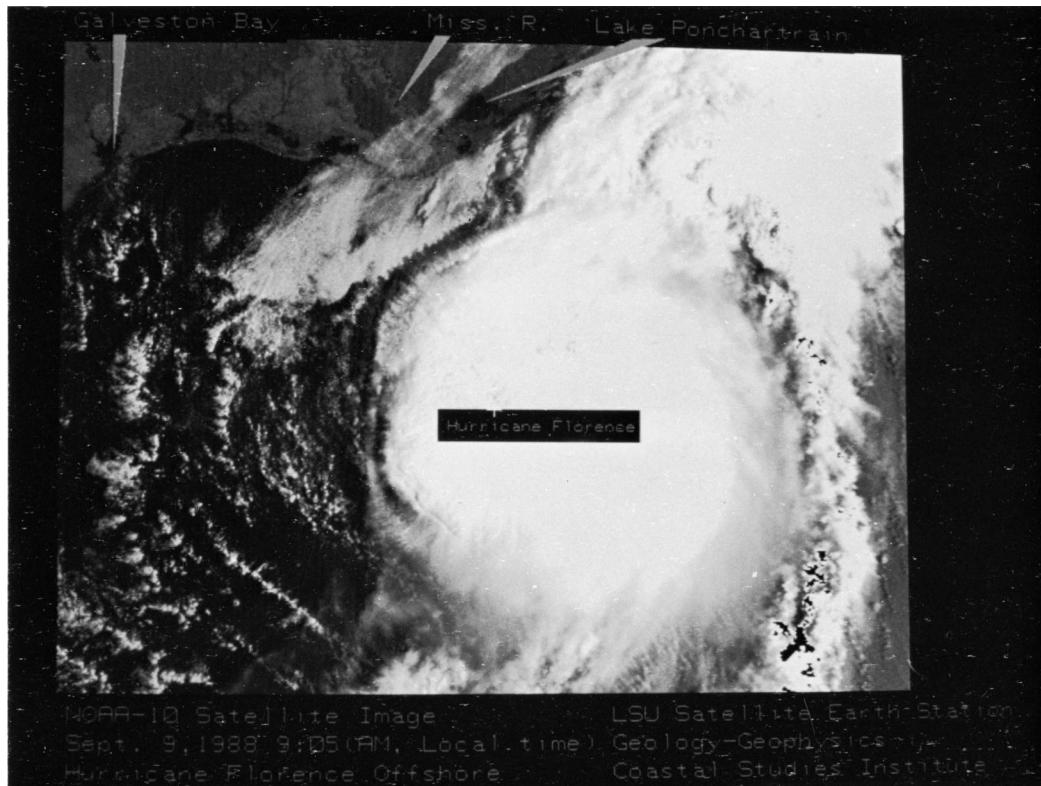


FIGURE 11 Hurricane Florence over the north-central Gulf of Mexico at 9:05 AM (CDT) on September 9, 1988. (Courtesy of Oscar Huh and David Wilensky, Louisiana State University.)



FIGURE 12 This pre-Florence photo was taken on July 16, 1988, along the north-central section of Curlew Island, which is an undeveloped barrier island in St. Bernard Parish, LA. The photo shows a partially vegetated barrier island with a large area of washover flat deposit to the right. However, no major washover channels exist at this time period. (Courtesy of Louisiana Geological Survey, Coastal Geology Section.)



FIGURE 13 Significant hurricane impact features associated with Hurricane Florence can be seen in this photo, which was taken on October 13, 1988, at the same locality. Waves in combination with the hurricane storm surge cut several large washover channels across Curlew Island, depositing washover fans into the Chandeleur Sound. These washover channels break the barrier into smaller pieces and limit the ability of the barrier island to act as a buffer against subsequent hurricane impacts. (Courtesy of Louisiana Geological Survey, Coastal Geology Section.)

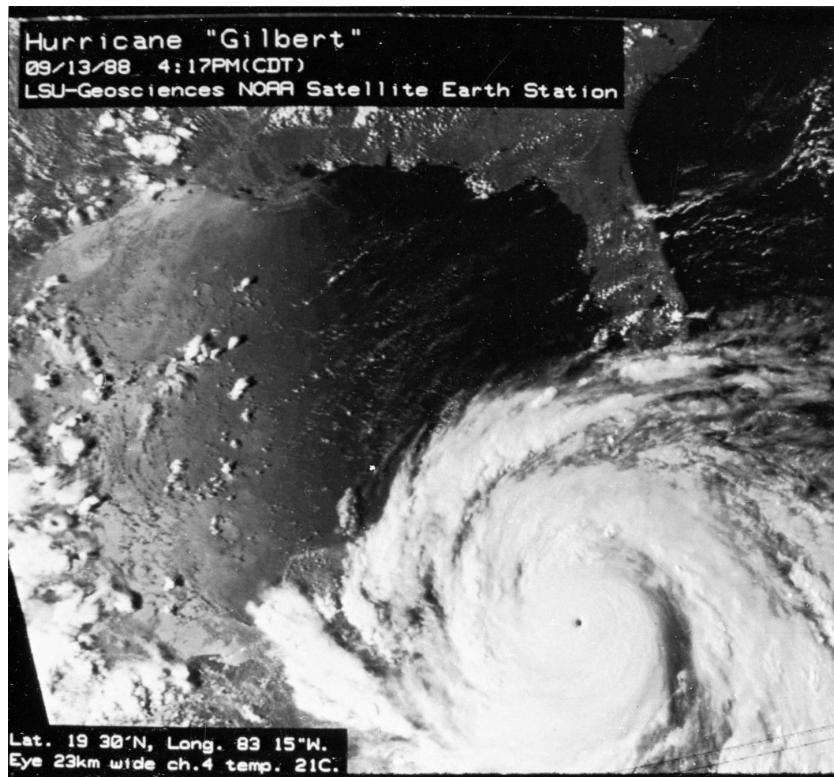


FIGURE 14 Hurricane Gilbert near Yucatan, Mexico at 4:17 PM (CDT) on September 13, 1988. (Courtesy of Oscar Huh and David Wilensky, Louisiana State University.)



FIGURE 15 This pre-Gilbert photo was taken on July 15, 1988, along the central part of Grand Isle, which is the only commercially developed barrier island in Louisiana. In 1984, the U.S. Army Corps of Engineers constructed an 11-ft-high artificial dune along the entire length of Grand Isle with wooden beach-access structures placed over the dune at regular intervals. The artificial dune provides protection against hurricane storm surge and flooding. Note the position of the shoreline in relation to the wooden beach-access structure. (Courtesy of Louisiana Geological Survey, Coastal Geology Section.)



FIGURE 16 Although Hurricane Gilbert made landfall over 400 miles away at the Mexico–Texas border, Grand Isle still experienced significant coastal erosion. This post-Gilbert photo was taken on October 12, 1988. The artificial dune was totally eroded away, and the shoreline migrated approximately 10–20 m to the landward side of the wooden beach-access structure. (Courtesy of Louisiana Geological Survey, Coastal Geology Section.)

numbers are made available to public safety officials when a hurricane is within 72 hr of landfall.

In September 1988, two hurricanes affected the coastal regions of the Gulf of Mexico. One was Florence, a minimal hurricane (scale 1), and the other was Gilbert, a catastrophic storm (scale 5). The effect of these two hurricanes on the barrier islands of Louisiana is shown in Figs. 11–16. Figure 11 is an NOAA-10 satellite image made at 9:05 AM (CDT) on September 9, 1988, as received at Louisiana State University, Baton Rouge. Wind

speeds at the eye of Florence, about 30 miles wide, were 80 mph. The eye passed 30 miles east of New Orleans around 3 AM then headed across Lake Pontchartrain, slightly west of the predicted path, over the city of Slidell, LA. The effect of Florence on a barrier island is shown in Figs. 12 (prestorm) and 13 (poststorm). Waves and surges produced by Florence physically cut several large washover channels, which broke the barrier island into smaller pieces and limited its ability to act as a buffer against subsequent hurricane impacts.

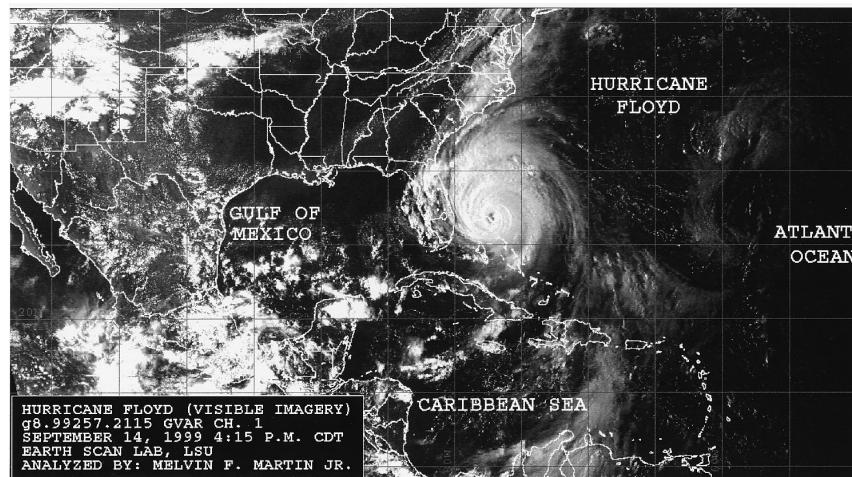


FIGURE 17 Visible imagery from the GOES satellite while Hurricane Floyd was still over the Bahama region.

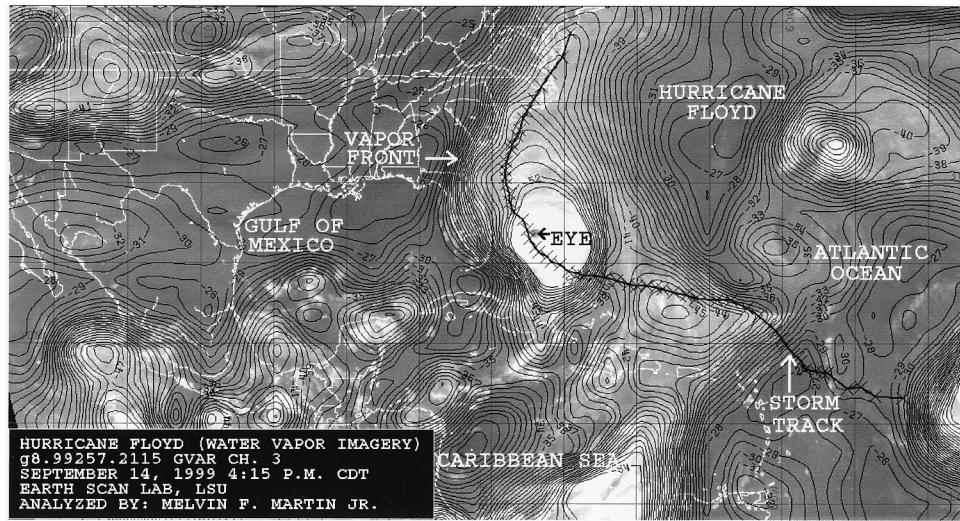


FIGURE 18 Water vapor imagery at the same time as the visible shown in Fig. 17. Note that the tongue of moisture (or the water vapor plume) extended northeastward over North Carolina. The hurricane track appears to follow this vapor plume. Note also that the vapor front which was located west of the vapor plume acts as a blocker to prohibit possible landfall of Floyd over Florida, Georgia, and South Carolina.

Hurricane Gilbert (Fig. 14), one of the most powerful (category 5) storms on record, devastated the Yucatan Peninsula, Mexico, on September 14, 1988. Two days earlier, it had destroyed an estimated 100,000 of Jamaica's 500,000 houses. Two days later, it again made landfall, striking northeastern Mexico and causing more than 200 people to perish.

The damage-potential scale categories of the Saffir-Simpson scale of hurricanes range from category 1, a minimal-size hurricane with central pressure equal to or greater than 980 mbar, up to category 5, a catastrophic storm with central pressure lower than 920 mbar. Two other category 5 hurricanes have affected the Gulf of Mexico region in this century: one in September 1935, which devastated Key West, FL, and Hurricane Camille in August 1969, which caused extensive damage along the Mississippi and Louisiana coasts.

The NOAA satellite advanced very high resolution radiometer imagery of Hurricane Gilbert was acquired with a 1.2-m program track antenna of the SeaSpace Co., Terascan System, NOAA satellite Earth station, established in the summer of 1988 at Louisiana State University, Baton Rouge. The hurricane photographs are of the 1.1-km resolution channel 2 (3.55–3.93 μm NIR) imagery. At the time of this NOAA-9 overpass, on September 13 at 4:17 PM (CDT), Gilbert was a category 5 hurricane nearly the size of the Gulf of Mexico, with the eye located at latitude 19°25.51'N, longitude 83°15.96'W. The eye, as measured with the channel 4 data, was 21 km in diameter. Channel 4 radiation temperatures ranged from 23.5°C in the center of the eye to below –83°C on the surrounding cloud tops.

Photographs taken before (Fig. 15) and after (Fig. 16) Gilbert at Grand Isle, LA, show that, even though Gilbert made landfall over 400 miles away, at the Mexico–Texas border, Grand Isle experienced significant coastal erosion.

Hurricane Floyd (Figs. 17 and 18) in 1999 caused extensive damage in North Carolina. Because its track was along the southeast coast of the United States, the exact landfall position was a challenge to forecast. With the aid of hourly water vapor imagery from the GOES satellite, improvements in earlier warnings can be made.

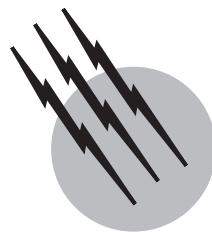
SEE ALSO THE FOLLOWING ARTICLES

ATMOSPHERIC TURBULENCE • CLOUD PHYSICS • COASTAL GEOLOGY • METEOROLOGY, DYNAMIC • OCEAN-ATMOSPHERIC EXCHANGE

BIBLIOGRAPHY

- Bell, G. D., and Bosart, L. F. (1988). *Mon. Weather Rev.* **116**, 137.
- Beyrich, F., and Klose, B. (1988). *Boundary-Layer Meteorol.* **43**, 1.
- Boers, R., Spinhirne, J. D., and Hart, W. D. (1988). *J. Appl. Meteorol.* **27**, 797.
- Brook, R. R. (1985). *Boundary-Layer Meteorol.* **32**, 133.
- Chen, G. T. J., and Yu, C.-C. (1988). *Mon. Weather Rev.* **116**, 884.
- Chen, S.-J., and Dell'Osso, I. (1987). *Mon. Weather Rev.* **115**, 447.
- Coulman, C. P., Colquhoun, J. R., Smith, R. K., and Melnnes, K. (1985). *Boundary-Layer Meteorol.* **32**, 57.
- Dorman, C. (1987). *J. Geophys. Res.* **92**(C2), 1497.
- Garratt, J. R. (1985). *Boundary-Layer Meteorol.* **32**, 307.
- Garratt, J. R. (1987). *Boundary-Layer Meteorol.* **38**, 369.
- Garratt, J. R. (1990). *Boundary-Layer Meteorol.* **50**, 171.

- Garratt, J. R. (1992). "The Atmospheric Boundary Layer," Cambridge Univ. Press, Cambridge, U.K.
- Hanna, S. R. (1987). *Boundary-Layer Meteorol.* **40**, 205.
- Hsu, S. A. (1988). "Coastal Meteorology," Academic Press, San Diego, CA.
- Hsu, S. A. (1991). *Mariners Weather Log* **35**(2), 57.
- Hsu, S. A. (1993). *Mariners Weather Log* **37**(2), 4.
- Hsu, S. A. (1994). *Mariners Weather Log* **35**(1), 68.
- Hsu, S. A. (1999). *J. Phys. Oceanogr.* **29**, 1372.
- Kotsch, W. J. (1983). "Weather for the Mariner," 3rd ed., U.S. Naval Institute, Annapolis, MD.
- Macklin, S. A., Lackmann, G. M., and Gray, J. (1988). *Mon. Weather Rev.* **116**, 1289.
- McIlveen, J. F. R. (1986). "Basic Meteorology, A Physical Outline," Van Nostrand-Reinhold, Berkshire, England.
- Roeloffzen, J. C., Van den Berg, W. D., and Oerlemans, J. (1986). *Tellus* **38A**, 397.
- Smolarkiewicz, P. K., Rasmussen, R. M., and Clark, T. L. (1988). *J. Atmos. Sci.* **45**, 1872.
- Stull, R. B. (1988). "An Introduction to Boundary Layer Meteorology," Kluwer, Dordrecht, The Netherlands.
- Telford, J. W., and Chai, S. K. (1984). *Boundary-Layer Meteorol.* **29**, 109.
- The WAMDI Group (1988). *J. Phys. Oceanogr.* **18**, 1775.
- Toba, Y., Okada, K., and Jones, I. S. F. (1988). *J. Phys. Oceanogr.* **18**, 1231.
- U.S. Army Corps of Engineers (1984). "Shore Protection Manual," Vicksburg, MS.
- Ueda, H., and Mitsumoto, S. (1988). *J. Appl. Meteorol.* **27**, 182.
- Venkatram, A. (1986). *Boundary-Layer Meteorol.* **36**, 149.
- Welch, R. M., Kuo, K. S., Wielick, B. A., Sengupta, S. K., and Parker, L. (1988). *J. Appl. Meteorol.* **27**, 341.
- Yan, H., and Anthes, R. A. (1987). *Mon. Weather Rev.* **115**, 936.
- Zemba, J., and Friehe, C. A. (1987). *J. Geophys. Res.* **92**(C2), 1489.



Greenhouse Effect and Climate Data

Philip D. Jones

University of East Anglia

- I. Greenhouse Effect and Climate Change
- II. Surface Temperature Data
- III. Precipitation Records
- IV. Climate Change Detection and Attribution

GLOSSARY

Attribution of climate change Relating the climate change to known forcing factors (including the effects of human influences).

Detection of climate change The detection of a change in climate (either in the mean or variability) between one period and another.

Forcing factors Influences that force the climate system, such as natural (solar output changes and volcanic eruptions) and human factors (greenhouse gases, land-use changes).

Global climate models Computer models (general circulation models) which simulate past, present, and future climates from changes in greenhouse gas concentrations and natural forcing factors.

Greenhouse effect The change in surface temperature caused by the radiative properties of some atmospheric constituents (water vapor and greenhouse gases). The natural greenhouse effect is being enhanced by increases in the major greenhouse gases.

Homogeneous Property of a climate series whereby all

variations of the series are caused solely by the vagaries of weather and climate.

Major greenhouse gases Radiatively active gases in the atmosphere such as carbon dioxide, methane, nitrous oxide, and chlorofluorocarbons. They are increasing as a result of fossil fuel burning, land-use change, and some agricultural activities.

Proxy (paleo) climatic variables Noninstrumental indicators of climate variations (e.g., historical documents, tree growth, ice cores, corals, lake and marine varves), used as proxy climate evidence for preinstrumental periods.

Sulfate aerosols Small particles in the atmosphere, due to fossil fuel burning, which reflect incoming radiation and offset the effects of greenhouse gases.

INCREASES IN THE ATMOSPHERIC COMPOSITION of natural and man-made greenhouse gases are expected, from both theoretical and modeling studies, to raise surface temperatures. This article considers how well we measure surface temperature and precipitation and

what has happened over the past 150 years, the period of instrumental records. Over this time, the average global temperature has risen by 0.6°C to levels that are unprecedented, based on proxy climatic information, for at least a thousand years. The final section considers the climate change detection and attribution issues: greenhouse gas concentrations have risen, temperatures are higher now than for a millennium, but are the two related? Evidence from many studies during the 1990s indicates that the observed patterns of temperature change show strong similarities to patterns generated by global climate models.

I. GREENHOUSE EFFECT AND CLIMATE CHANGE

Climate is controlled by the long-term balance between incoming radiation from the sun, which is absorbed by the earth's surface and the atmosphere above, and energy returned to space in the form of infrared radiation. Changes in climate over the last 1000 years that have occurred have been due to modifications of this energy balance caused by natural factors, external to the climate system. More recently, anthropogenic (human) activities (affecting the composition of the atmosphere) have begun to be important. Over the last 10,000 years (the period termed the Holocene), solar output received at the surface has also varied due to differences in the position of the earth's orbit relative to the sun. This has caused slight variations in summer insolation between different latitudes (the Milankovitch effect).

Natural external influences take two forms: changes in solar output on decade-to-century and longer time scales and reductions in incoming radiation caused by explosive volcanic eruptions. The latter put significant amounts of dust and aerosols into the stratosphere, where it might reside for up to a year or two, lowering radiation receipts at the surface. A useful proxy for solar output is the number of sunspots, lower numbers reducing radiation received at the surface by a few tenths of 1%. Impacts on the 11-year time scale are negligible, as the atmosphere/ocean system has little time to respond to the slight changes. Gradual impacts, occurring on near-century time scales, are more likely, but effects are very difficult to detect. Stratospheric dust veils impact radiation balances much more quickly and reduce surface receipts by up to 1%, but only for a year or two, while the dust/aerosols slowly settle back to earth. The effects on surface temperature will tend to be rapid, leading to cooling, particularly over Northern Hemisphere land areas during the summer season, when there are greater amounts of radiation to perturb. Effects are less noticeable in the Southern Hemisphere because of the greater area of ocean which moderates influences.

Unless explosive eruptions occur closely together in time, the effects of one eruption will have dissipated by the time the next occurs. Volcanic effects therefore occur on high-frequency (interannual) time scales, while solar output changes occur on decade-to-century time scales. Climate is also influenced internally by changes within the ocean (strengths and directions of currents, rates of upwelling, and deepwater formation).

The earth's atmosphere contains relatively small quantities of greenhouse gases (principally water vapor, carbon dioxide, and methane), which trap some of the infrared radiation, causing average surface temperatures to be much warmer (34°C) than if the content was just nitrogen and oxygen. CO_2 , CH_4 , and N_2O levels are, however, increasing as a result of industrial (fossil fuel burning) and agricultural activities, and deforestation, causing the natural greenhouse effect to be enhanced. If current trends in emissions continue, the amount of CO_2 in the atmosphere will double during the twenty-first century. The amounts of several other human-made greenhouse gases (e.g., CFCs) are also increasing substantially as well.

This enhancement of greenhouse gases in the atmosphere, due almost entirely to human activities, will change the climate, leading to an increase in the average surface temperature of the earth. In some regions the enhancement is being countered by related air pollution through emissions of sulfate aerosols and soot (also from fossil fuel burning), but cleaner energy systems (principally in North America and Europe) are reducing this influence. Current best estimates from climate models are that, relative to 1990, surface temperatures will rise by $1\text{--}3.5^{\circ}\text{C}$ by 2100. Because some greenhouse gases have long lifetimes (~ 100 years) in the atmosphere, even if emissions were to cease immediately, the effects of past emissions would continue for centuries.

This article considers principally the past surface climate record, showing why we are confident that surface temperatures are rising. A shorter section considers precipitation changes, as these are generally more important for many factors of human life. Changes in global climate, though, are nearly always considered in terms of temperature. Relating the observed rises in temperature to the greenhouse gas increases is more than one of simple cause and effect, however, as changes in temperature have occurred in the past. The rise in surface temperature and the increases in greenhouse gas concentrations may be unrelated. Although this is unlikely, given present knowledge and advances in climate modeling, it is vital to relate the two if nations are to be persuaded to change their energy policies. The linking of cause and effect is the climate change detection issue: is the rise in temperature unequivocally due to the greenhouse gas changes? Related to this, modeling is used to estimate future rates

of temperature increase, enabling either adaptation measures to be taken or to determine the levels of mitigation that will be required to minimise rapid rates of change in the near future. After considering the evidence for change over the twentieth century, attempts to explain the changes in temperature are discussed in the detection context.

II. SURFACE TEMPERATURE DATA

A. Quality of Temperature Data

Any assessment of trends or changes in temperature requires that all the observations have been taken in a consistent manner. Climatologists refer to this property as homogeneity. Time series of temperature are homogeneous if the variations exhibited are due solely to the vagaries of the weather and climate. Numerous nonclimatic factors influence the basic data, causing erroneous conclusions to be drawn regarding the course of temperature change. The factors vary depending on the data source and are briefly considered in the next two subsections for the terrestrial and marine components of the earth's surface.

1. Land

It is extremely rare for observational protocols and the environment around the observing location to have remained exactly the same during the station's history. Changes are likely to have occurred with the instruments, their exposure and measurement techniques, in the location of station and the height of the instruments, in the times of observations per day, and the methods used to calculate daily and monthly averages.

The commonly used louvered screen developed by Stevenson in the 1860s/1870s is now the standard around the world, although different countries use variants of a similar design. Prior to this, most thermometers were positioned on poleward-facing walls (i.e., out of direct sunlight), but this poses problems in high-latitude regions in the summer. Most stations have been moved at least once during their lifetime. Also of importance is the time when observations are made each day. Even today there is no accepted standard, countries being allowed to choose whatever times suit them. English-speaking countries have tended to use the average of the daily maximum and minimum readings each day to measure daily and monthly averages. Some countries have switched to this method, mainly because of its ease, while others retain their national standards (averages of measurements made at fixed hours, between 3 and 24 times per day).

All these problems influence series, often in an abrupt manner (temperatures jumping to a new level by up to 2°C in extreme cases). Ideally, when new sites or ob-

servation protocols are adopted, parallel measurements are recommended, enabling corrections to be calculated. Sadly, although clearly recognized as being necessary, few countries carry out sufficient overlapping measurements. The most common problems relate to location moves, particularly to airports in the 1940s and 1950s. Recently, many countries have switched from mercury-in-glass thermometers to electrical resistance thermisters, to reduce manpower, automating measurements. The sum total of all these problems can be disentangled if adequate station history information is available, but it is a generally a tedious process to locate all the necessary information. In some countries, it is just not available in sufficient detail.

Potentially the most important factor with respect to homogeneity is changes in the environment around the station. The location may have been a small town in the nineteenth century, but now it could be a city of several million. Development around the site (urbanization) leads to relative warming of city sites compared to, still rural, neighbors. On certain days, particularly calm sunny days, cities can be warmer than rural surroundings by up to 10°C. For monthly averages this reduces to up to 2°C, larger for inland continental, compared to coastal, locations. Cities which have grown rapidly over the twentieth century tend to be more affected, compared particularly to European locations where development has taken place over many centuries.

The sum total of these problems can lead to gradual warming due to environmental changes and abrupt changes (both to warmer or colder absolute temperatures) for all other problems. Several groups in the United Kingdom and the United States have extensively analyzed the basic surface temperature data (between one and 7000 stations), adjusting the data for the abrupt changes and removing urban-affected stations, and have reached similar conclusions about the course of temperature change over the instrumental period since 1850 (Sections II.C and D). It is highly unlikely that every problem has been corrected for, but the different techniques used give confidence that large-scale changes over the last 150 years are both real and well documented.

2. Marine

Terrestrial parts of the world constitute only 30% of the earth's surface, so it is vital that we also monitor the oceans if we are to gain more of a global picture. Historical temperature data over marine regions are derived largely from *in situ* measurements of sea surface temperature (SST) and marine air temperature (MAT) taken by ships and buoys. To be of use, each measurement must be associated with a location. Up to 15% of marine data

is thought to be mislocated (ships located on the land!), and these values must be discarded. It is obviously harder to reject data still located over the ocean, but all analyses of the raw data also attempt to remove or correct these problems.

Marine data are also beset with homogeneity problems, but they are distinctly different from the terrestrial realm. For MAT data the average height of ship's decks above the ocean has increased during the twentieth century, but more important, daytime measurements are influenced by the solar heating of the ship, rendering only the nighttime MAT (NMAT) data of any value. For SST data, the changes in sampling method from uninsulated canvas buckets (generally prior to the early 1940s) to engine intake measurements (early 1940s onwards) causes an artificial rise in SST values of 0.3–0.7°C.

In combining marine data with land-based surface temperatures, SST data is preferred to NMAT, because they are generally more reliable, principally as there are at least twice as many observations, daytime MAT values having been contaminated by the ships' infrastructure. Absolute values of SST and land air temperatures may differ by up to 10°C near some coastlines, so we cannot combine the two directly. Instead, we use anomalies (departures or differences from average), assuming that anomalies of SST and MAT agree on climatological (monthly and greater) time scales. Correction of the SST data for the change from canvas buckets is achieved using a physical-empirical model to estimate the degree of sea-water cooling that occurs in buckets of varying design. The cooling depends on the ambient weather conditions, but this can be approximated by climatological averages. Corrections are greatest in regions with the largest air-sea temperature differences (i.e., winters compared to summers), and the technique minimizes residual seasonal cycles in pre-World War II SST values compared to post-1945 values.

Since the marine and land components are independent, the two records can be used to assess each other after they have been separately corrected. The components have been shown to agree by several groups on both hemispheric scales, but also using island and coastal data.

B. Aggregation of the Basic Data

Both the land and marine data are irregularly located over the earth's surface. To overcome the greater density of data on land, it is necessary to interpolate the data, generally to some form of regular latitude/longitude grid.

1. Land

Differing station elevations and national practices with regard to the calculation of monthly mean temperatures

means that interpolation to a regular grid is much more easily achieved by converting all the monthly data to anomalies from a common reference period. The period with best available data is 1961–1990. The simplest interpolation scheme is the average of all stations that are located within each $5^\circ \times 5^\circ$ grid box. More complex interpolation methods yield essentially the same results on all spatial scales. A potential drawback of gridding schemes is that the variance of grid-box time series is affected by changing numbers of stations within each grid box through time, although it is possible to correct for this.

2. Marine

For SST the aggregation is approached in a somewhat different manner. The random location of each observation means that it is necessary, by interpolation, to derive a climatology for each $1^\circ \times 1^\circ$ square of the world's oceans for each 5-day period (pentad). SST anomaly values with respect to this climatology are then averaged together for each month for each $5^\circ \times 5^\circ$ grid box, the same as used for the land component.

3. Combination into One Dataset

Combination of the two components occurs in the simplest manner. Anomaly values are taken from each component. When both are available the two are weighted by the fraction of land/ocean within each grid box. Because island and coastal data in some regions are likely to be considerably more reliable than a few SST observations, no land or marine component can be less than 25%.

C. Hemispheric and Global Time Series

With the basic data now in 5° latitude/longitude grid boxes, calculation of large-scale averages is relatively simple but must take into account the different size of grid boxes in tropical, compared to polar, latitudes. This is simply achieved by weighting each grid box by the cosine of its central latitude value.

[Figure 1](#) shows annual hemispheric and global time series for the 1861–1999 period. [Table I](#) gives monthly linear trend values, estimated by least squares, for the three domains calculated over the 139-year period and for some other subperiods. For the global average, surface temperatures have risen by 0.60°C, a value that is statistically significant at the 99.9% level. All the monthly values also exhibit a significant warming. Warming is marginally greater over the Southern Hemisphere (SH) compared to the Northern Hemisphere (NH). Warming is least in magnitude during the NH summer months and may be influenced by the exposure of thermometers (Section II.A)

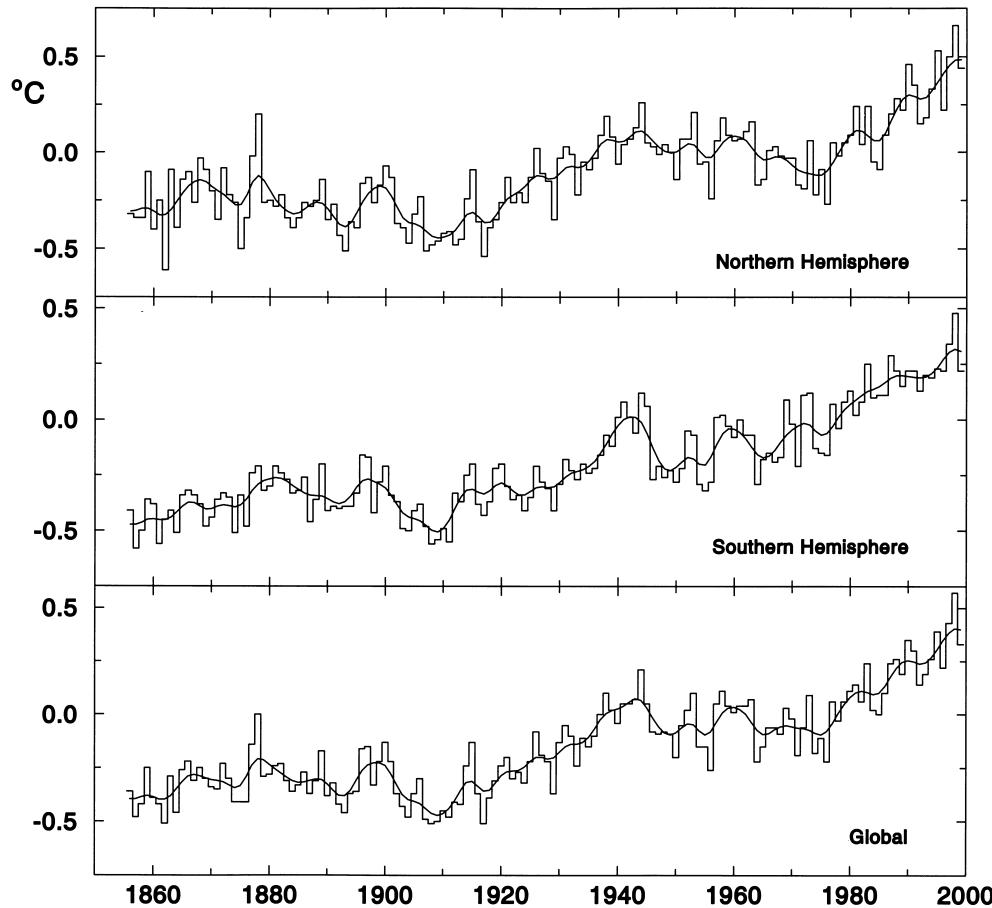


FIGURE 1 Hemispheric and global temperature averages on the annual time scale (1856–1999 relative to 1961–1990). The smooth curves highlight variations on decade time scales.

during the 1860s and 1870s. Seasonal differences in temperature are much more marked in the NH (with its greater landmass) than over the SH.

While both hemispheres show similar degrees of warming, it is also apparent that many warm and cool years, relative to the underlying trend, are in common. Many anomalous warm years are coincident because they relate to El Niño years in the eastern equatorial Pacific. El Niño events cause somewhat predictable patterns of temperature and precipitation patterns over the world, with more regions experiencing warmer than cooler conditions. Polar regions and much of northern Eurasia are largely unaffected by such an influence. Cooler-than-normal years generally relate to the counterpart of El Niño, La Niña, when anomalous patterns which, to the first order are opposite, occur. A few cool years can be related to the climatic effects of explosive volcanic eruptions which are large enough to put considerable amounts of dust into the stratosphere. Once there, the dust forms a veil over the earth, reducing solar radiation and cooling the surface, particularly land areas. Surface cooling of about 0.2–0.3°C

followed the eruption of Pinatubo in the Philippines in June 1991, mainly in the northern summer months of 1992 and 1993. Volcanic eruptions which affect only the troposphere (e.g., Mt. St. Helens in 1980) have little climatic effect, as their ejecta are quickly dispersed by rain-making processes.

Both hemispheres show long-term warming, but it clearly has occurred in two phases (1920–1944 and since about 1975, see also Table I). Spatial patterns of the changes will be considered later. For the global average, 12 of the 14 warmest years occurred between 1986 and 1999. The warmest year of the entire global series was 1998, 0.57°C above the 1961–1990 average. The next four warmest years were 1997 (0.43), 1995 (0.39), 1990 (0.35), and 1999 (0.33).

1. Accuracy of the Hemispheric and Global Series

The series in Fig. 1 is subject to three sources of error: bad measurements, residual effects of the homogeneity checks due to urbanization and the bucket corrections over the

TABLE I Temperature Change ($^{\circ}\text{C}$) Explained by the Linear Trend over Four Periods

	1861–1999			1901–1999			1920–1944			1975–1999		
	NH	SH	Globe									
Jan.	0.68	0.60	0.64	0.65	0.63	0.64	0.26	0.44	0.35	0.65	0.33	0.49
Feb.	0.78	0.59	0.68	0.85	0.64	0.75	0.67	0.28	0.47	0.98	0.37	0.68
Mar.	0.69	0.60	0.65	0.78	0.69	0.74	0.20	0.31	0.25	0.56	0.48	0.52
Apr.	0.56	0.62	0.59	0.71	0.66	0.68	0.40	0.41	0.40	0.56	0.39	0.48
May	0.53	0.70	0.61	0.66	0.76	0.71	0.22	0.56	0.39	0.55	0.41	0.48
Jun.	0.35	0.76	0.56	0.57	0.72	0.64	0.19	0.67	0.43	0.62	0.42	0.52
July	0.33	0.63	0.48	0.53	0.72	0.62	0.26	0.71	0.48	0.63	0.43	0.53
Aug.	0.44	0.61	0.53	0.53	0.70	0.62	0.35	0.45	0.40	0.63	0.41	0.52
Sept.	0.44	0.58	0.51	0.52	0.69	0.60	0.52	0.28	0.40	0.43	0.37	0.40
Oct.	0.66	0.64	0.65	0.55	0.69	0.62	0.64	0.36	0.50	0.69	0.42	0.56
Nov.	0.77	0.62	0.69	0.62	0.70	0.66	0.49	0.18	0.33	0.43	0.33	0.38
Dec.	0.74	0.62	0.68	0.81	0.61	0.71	0.56	0.32	0.44	0.71	0.25	0.48
Year	0.58	0.63	0.60	0.65	0.68	0.67	0.40	0.41	0.40	0.62	0.39	0.50

ocean, and most important, changes in the availability and density of the raw data through time. The latter are referred to in statistical terms as sampling errors. Taking them into account and allowing for the areas always missing from the grid-box analyses (see later gaps in coverage in Figs. 2 and 3) indicates that annual hemispheric averages are now accurate to within $\pm 0.05^{\circ}\text{C}$ (one standard error). Errors in the mid-nineteenth century were roughly twice modern values.

D. Analyses of the Temperature Record

The surface record has been extensively analyzed, principally over the past 25 years. The series in Fig. 1 has become one of the foremost series in major international reviews of the climate change issue, most recently by the Intergovernmental Panel on Climate Change (IPCC). Here several diverse aspects of the record are analyzed:

Patterns of recent change

Trends in areas affected by monthly extremes

Trends in maximum and minimum temperature

Daily extremes of temperature in some long European series

The last 150 years in the context of the last thousand years

Figure 1 clearly shows recent warming since the late 1970s. This warming has been called into question by satellite data, which measures the average temperature of the lower troposphere. The satellite record shows only a very small, statistically insignificant warming, whereas the surface temperature has risen by about 0.15°C per

decade since 1979. The warming at the surface is, however, not doubted. The agreement between the marine and terrestrial components and the widespread retreat of Alpine glaciers around the world are pervasive arguments. We would expect, however, that the lower troposphere should also be warming in as strong a fashion as the surface, and it is casting doubt on our ability to understand and model atmospheric processes and their response.

1. Patterns of Recent Change

Spatial patterns of change are shown in Figs. 2 and 3 for the two 25-year periods that show the strongest warming of the twentieth century (1920–1944 and 1975–1999). Patterns are shown seasonally and for the annual average. Even for the most recent period, coverage is not complete, with missing areas over most of the Southern mid to high-latitude oceans, parts of the Antarctic and central Arctic, and some continental interiors. Available data for the 1920–1944 period, however, only enables patterns to be shown for about half the earth's surface, compared with about three-quarters now. Both periods exhibit strong and highly significant warming in the hemispheric averages, but statistical significance is achieved in only relatively few areas on a local basis. More areas achieve significance in the recent period, but this is no more than would be expected, given the greater areas with data and the stronger warming (see Table I). Warming is not spread evenly across the seasons, although annually most regions indicate warming. The warming patterns of the two periods show different patterns, suggesting that they might be related to different combinations of causes.

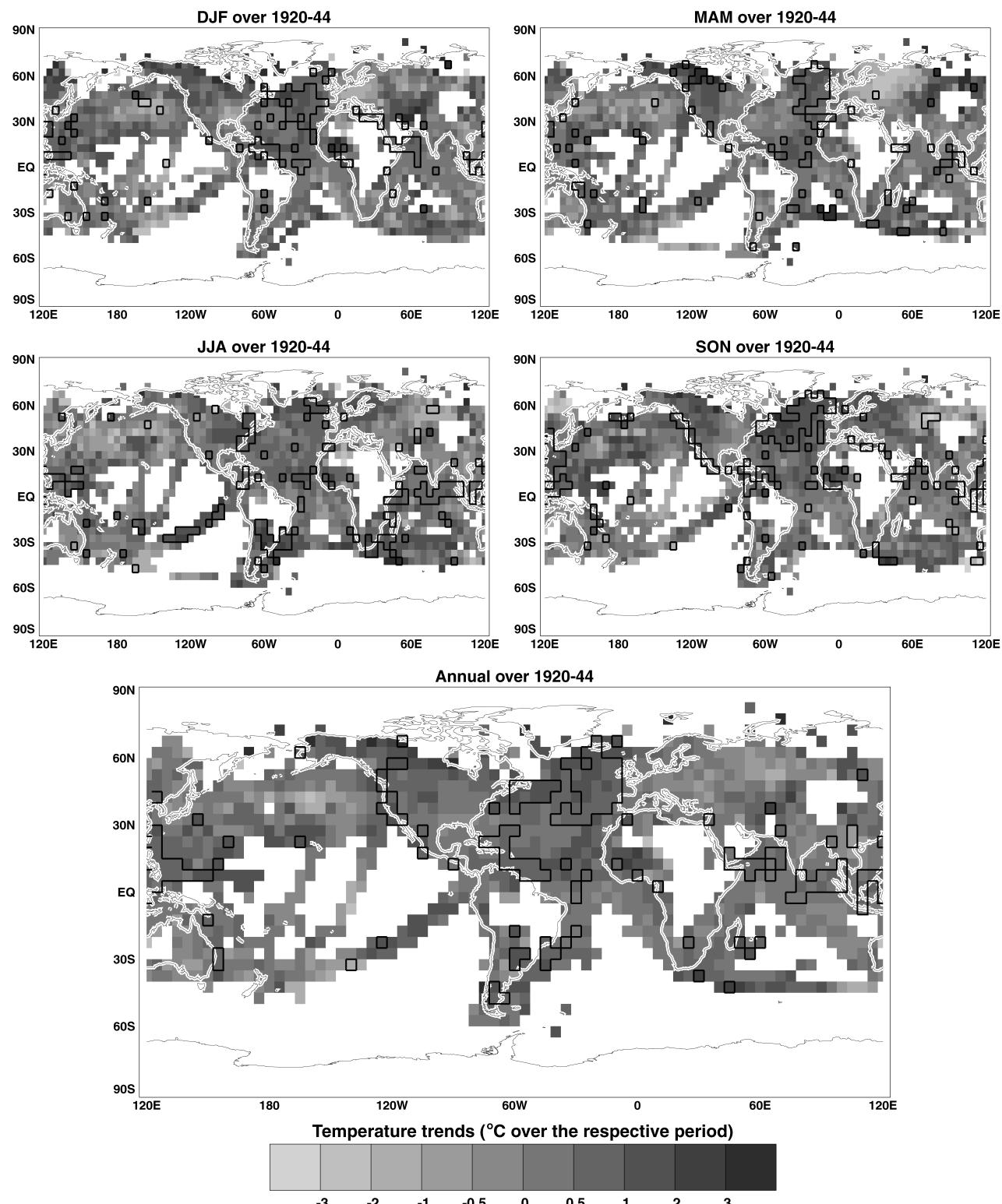


FIGURE 2 Trend of temperature on a seasonal and annual basis for the 25-year period 1920–1944. Boxes with significant linear trends at the 95% level (allowing for autocorrelation) are outlined by heavy black lines. At least 2 (8) months' data were required to define a season (year), and at least half the seasons or years were required to calculate a trend.

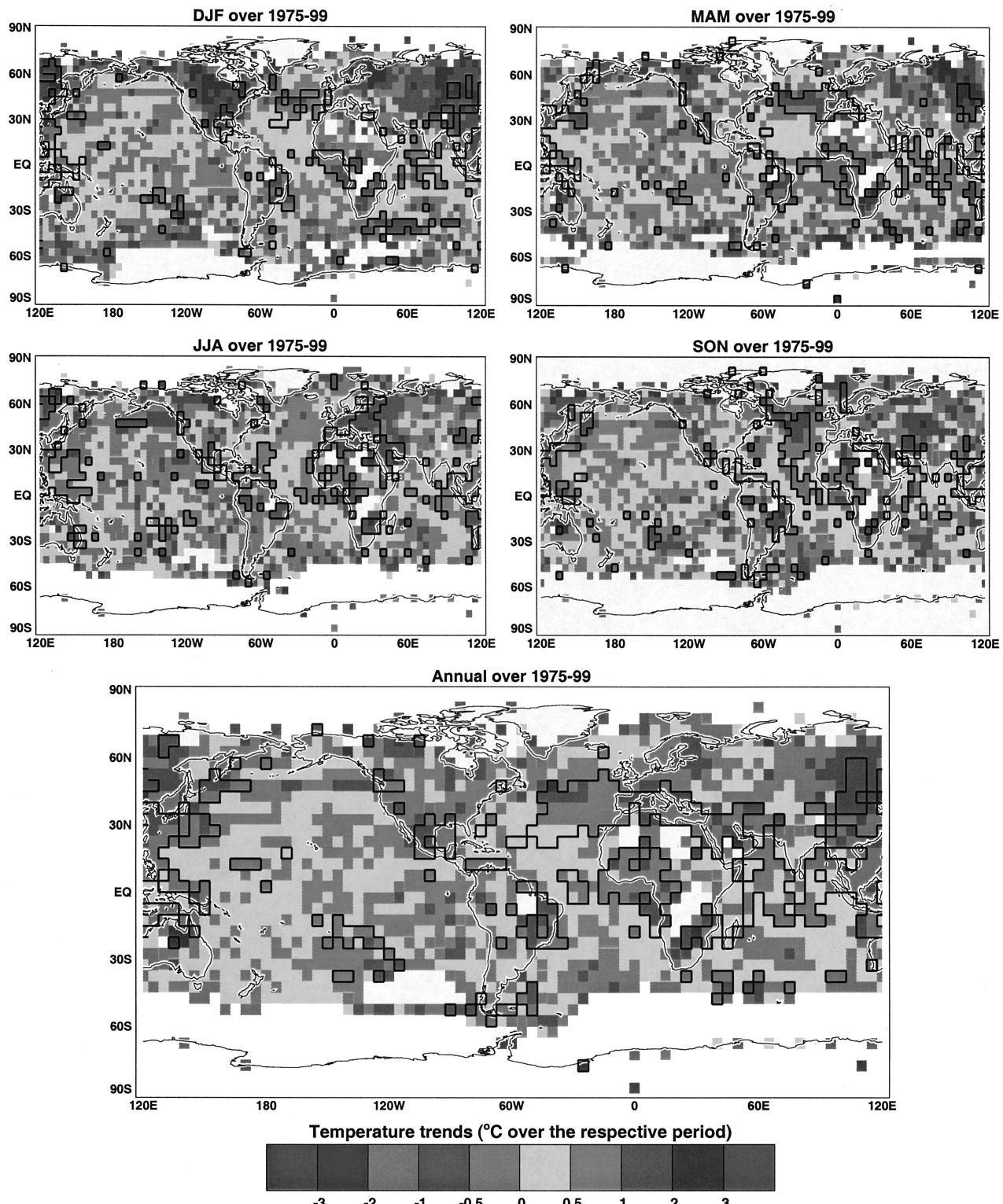


FIGURE 3 Trend of temperature on a seasonal and annual basis for the 1975–1999 period. Boxes with significant linear trends at the 95% level (allowing for autocorrelation) are outlined by heavy black lines. At least 2 (8) months' data were required to define a season (year), and at least half the seasons or years were required to calculate a trend.

2. Trends in Areas Affected by Monthly Extremes

The previous section has clearly shown the effects that differences in temporal variability have on the estimation of trends. Warming in Siberia may be large, but it is barely significant because of the large year-to-year variability. Trends in some tropical oceanic areas indicate highly statistically significant warming, but it may only be about 0.3–0.5°C. To enable easier intercomparison of trends and extremes, from a local impact point of view, removing the effects of year-to-year variability highlights where significant changes are occurring. An appropriate transformation is the gamma distribution. Each grid-box time series, on a monthly basis, is transformed from one in anomalies with respect to 1961–1990 to percentiles based on the same period. Percentiles can be easily related to return periods (e.g., the 5/95th percentile is equivalent to the 1-in-20-year return period). Using a normal distribution (i.e., simply dividing the grid-box anomaly series by the standard deviation calculated over the 1961–1990 period) works almost as well as the gamma distribution, but the latter is better in many regions of the world, as monthly temperatures are often significantly negatively skewed.

[Figure 4](#) compares the anomaly and percentile method for displaying annual temperatures for 1999. The zero anomaly and the 50th percentile contour are essentially the same in both plots. The percentile map, however, indicates extremely warm annual temperatures over many tropical and oceanic regions that might not warrant a second glance in anomaly form. The year 1999 shows 36% of the world's surface with data above the 90th percentile and 5% below the 10th percentile. How unusual is this, compared to other years? [Figure 5](#) shows the percentage of the world's surface with data with temperatures greater than the 90th and less than the 10th percentile since 1900. An increase in the percentage of the analyzed area with warm extremes is evident [the largest area being 56% in the warmest year (1998)], but by far the greatest change is a reduction in the percentage of the analyzed area with cold extremes. Some caution should be exercised when interpreting these results because of the large changes in coverage, particularly before 1951 (as seen in [Fig. 2](#)). The implicit assumption being made is that the average of the unsampled regions is the same as the average of the sampled regions. Coverage changes since 1951 are minimal, though, and even analyzing only those regions with data for the 1900–1920 period produces similar series to those seen in [Fig. 5](#). The implications of these series are that before the mid-1970s, most of the warming this century was more apparent through less cold annual averages than excessively warm ones. Over the last 25 years, regions experiencing very warm annual anomalies have begun to increase dramatically.

3. Trends in Maximum and Minimum Temperatures

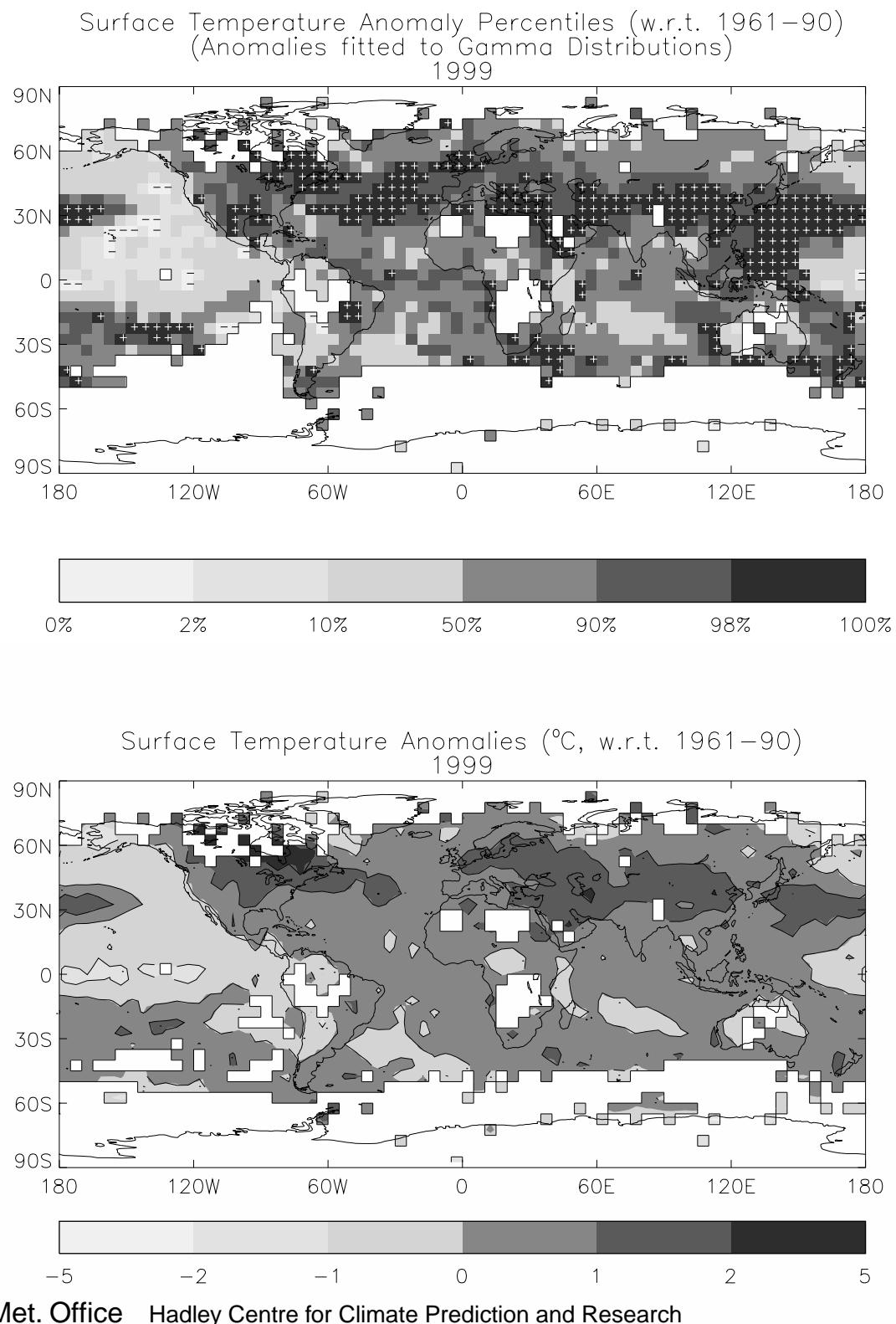
Up to now, all the surface temperature analyses have been based on monthly mean temperatures. This situation has arisen due to the widespread availability of this variable. As mentioned earlier, English-speaking countries have tended to measure daily and monthly means using maximum and minimum temperatures. Recently, extensive datasets of monthly mean maximum and minimum temperature have become available for periods since the 1950s. These enable recent warming patterns to be assessed for both day (maximum) and night (minimum) temperature. The difference between day and night (the diurnal temperature range, DTR) should prove a useful variable when considering what the causes of changes might be due to.

Homogeneity of the series poses more severe problems than for mean temperatures, as the various factors discussed earlier generally cause differential effects in the maximum and minimum series, and station history information is even more important to decide upon adjustments. Analyses are restricted to the period 1950–1993 because of data availability issues in many regions of the world. [Figure 6](#) shows trends over these 44 years for maximum and minimum temperatures and the DTR. Minimum temperatures decrease in only a few areas (which represent 54% of the world's land area), while maximums show decreases over large areas, notably eastern Canada, southern United States, parts of Europe, and much of China and Japan. The DTR shows decreases almost everywhere except in Arctic Canada and some Pacific islands.

Combining all regions, “global” minimum averages warmed by 0.79°C over 1950–1993, while maximums warmed by only 0.36°C. The DTR decreased by 0.35°C. Urbanization influences have been shown to have the same signature (warmer nights compared to days), so these studies have restricted analyses to nonurban stations. In most regions, these differential trends can be clearly related to increases in cloudiness, which will raise nighttime, compared to daytime, temperatures. Longer records, back to the turn of the twentieth century, are available in a few limited regions. Analyses over the United States and southern Canada, for example, show little change over the first half of the twentieth century.

4. Daily Temperature Extremes in Long European Series

The last two sections have considered extremes on a monthly basis, but public perception of climate change is often considered by daily extremes or runs of warm/cold days. In the context of the global warming issue, daily data



The Met. Office Hadley Centre for Climate Prediction and Research

FIGURE 4 Surface temperatures for 1999, relative to the 1961–1990 average (a) as percentiles and (b) as anomalies. The percentiles were defined by fitting gamma distributions to the 1961–1990 annual deviations relative to the 1961–1990 base period, for all $5^{\circ} \times 5^{\circ}$ grid boxes with at least 21 years of annual data in this period.

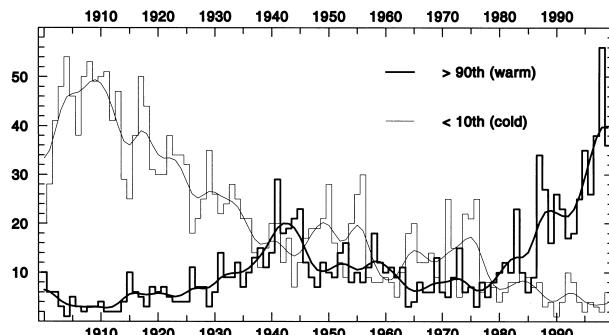


FIGURE 5 Percentage of the monitored area of the globe, for each year from 1900 to 1999, with annual surface temperatures above the 90th percentile and below the 10th percentile. The percentiles were defined by fitting gamma distributions based on the 1961–1990 period, using all $5^\circ \times 5^\circ$ boxes with at least 21 years of data. The smooth curves highlight variations on decadal timescales.

are relatively unimportant, as detection and attribution of human influences is concerned primarily with underlying trends on decadal time scales. In the public and political worlds, though, changing frequencies of daily extremes are how much of the global warming debate is perceived.

Daily temperature series present even greater problems to climatologists with respect to homogeneity than monthly data. Site and observation time changes are particularly important, and in some cases it may not be possible to fully correct for all the problems. Few long daily temperature series, therefore, are totally homogeneous. Furthermore, the availability of long series is often restricted to the last 50 years. Changes in the frequency of extremes may be occurring, but without long series it is difficult to judge whether recent changes are really unprecedented. In Europe, however, several series of 200+ years have recently been developed, which will be ideal for analysis.

The public perception of extremes is clearly cold winter and hot summer days, but in different regions it is necessary to define somewhat arbitrarily what is meant by cold and hot. A cold-day threshold of 0°C clearly has important consequences, but what is hot in northern Europe clearly differs from what would be regarded as hot in southern Europe. Also, considering only absolute extremes ignores changes that might be taking place in the transition seasons. A better and universally applicable means of defining extremes is to let the data define the thresholds and to allow these to change during the year.

The first step is an analysis to define the annual cycle of temperature on a daily basis, based on a common period such as 1961–1990. Some smoothing of this cycle is necessary, as 30 years is a relatively short period for definition. The 1961–1990 period is chosen for compatibility with

the other analyses in this section. Variability of a single day's temperatures from the annual cycle shows greater variability in Europe during winter compared to summer. Also, most station data series throughout the year, but particularly in winter, tend to be negatively skewed, so a normal distribution would be inappropriate as this would give a bias to the cold-day count. Instead, it is necessary to fit a gamma distribution to the daily anomalies for each day of the year, again using the 30 1961–1990 days for each day. Now it is a simple matter to count the number of days above the 90/95th (warm/very warm) and below the 10th/5th (cold/very cold) percentiles in a calendar year or in a season.

Figure 7 shows counts of warm/cold days for some of the long European series (central England, Stockholm, Uppsala, St. Petersburg, Brussels, Milan, Padua, and Cadiz). Although there are differences between the stations in the timings of change, the overall picture is of an increase in warm days in the second half of the twentieth century, but the largest trend is a reduction in the number of cold days. Recent increases in warm days at the sites with longer records have only just exceeded similar counts in some decades of the eighteenth century. Cold-day counts, in contrast, are clearly lower than at any period in the long records. The analysis method is insensitive to the choice of base period, another choice producing similar trends but centered around a different base.

All the last three sections considered extremes in different ways, but all show similar conclusions. Until the recent 25 years, the warming of the twentieth century is mostly manifest, not by increases in warm extremes, but by reduction in cold extremes. Cold extremes often pass by unnoticed by the majority, except in sectors where they are important. Pesticide use is much greater in some regions because insect-killing frosts are less severe, skiing seasons are often shorter, and expected freeze-up periods of some major rivers are reduced.

5. The Last 150 Years in the Context of the Last 1000 Years

Global average surface temperature has clearly risen over the last 150 years (Fig. 1), but what significance does this have when compared to changes over longer periods of time? The last millennium is the period for which we know most about the preinstrumental past, particularly spatially, but it must be remembered that such knowledge is considerably poorer than since 1850. The millennium, particularly the last 500 years, is also the most important when considering attribution of recent changes to human influences. Earlier millennia are also important, but they are known to have experienced longer-time-scale changes in solar irradiance caused by orbital changes

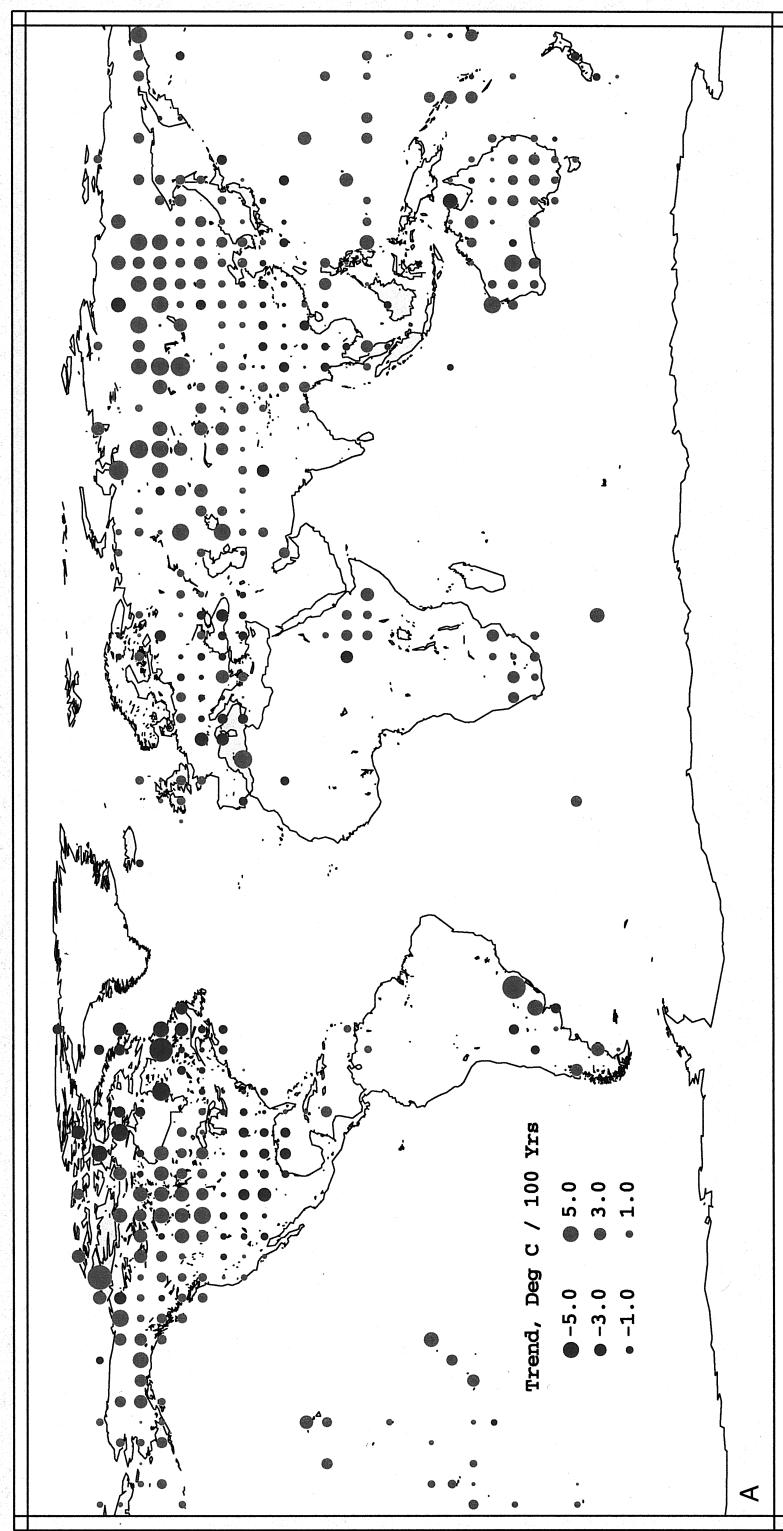


FIGURE 6 Trends (in degrees Celsius per 100 years) for each $5^\circ \times 5^\circ$ grid box for (a) maximum, (b) minimum, and (c) diurnal temperature range for nonurban stations. [Redrawn with permission from Easterling et al. (1997). Copyright 1997 American Association for the Advancement of Science.]

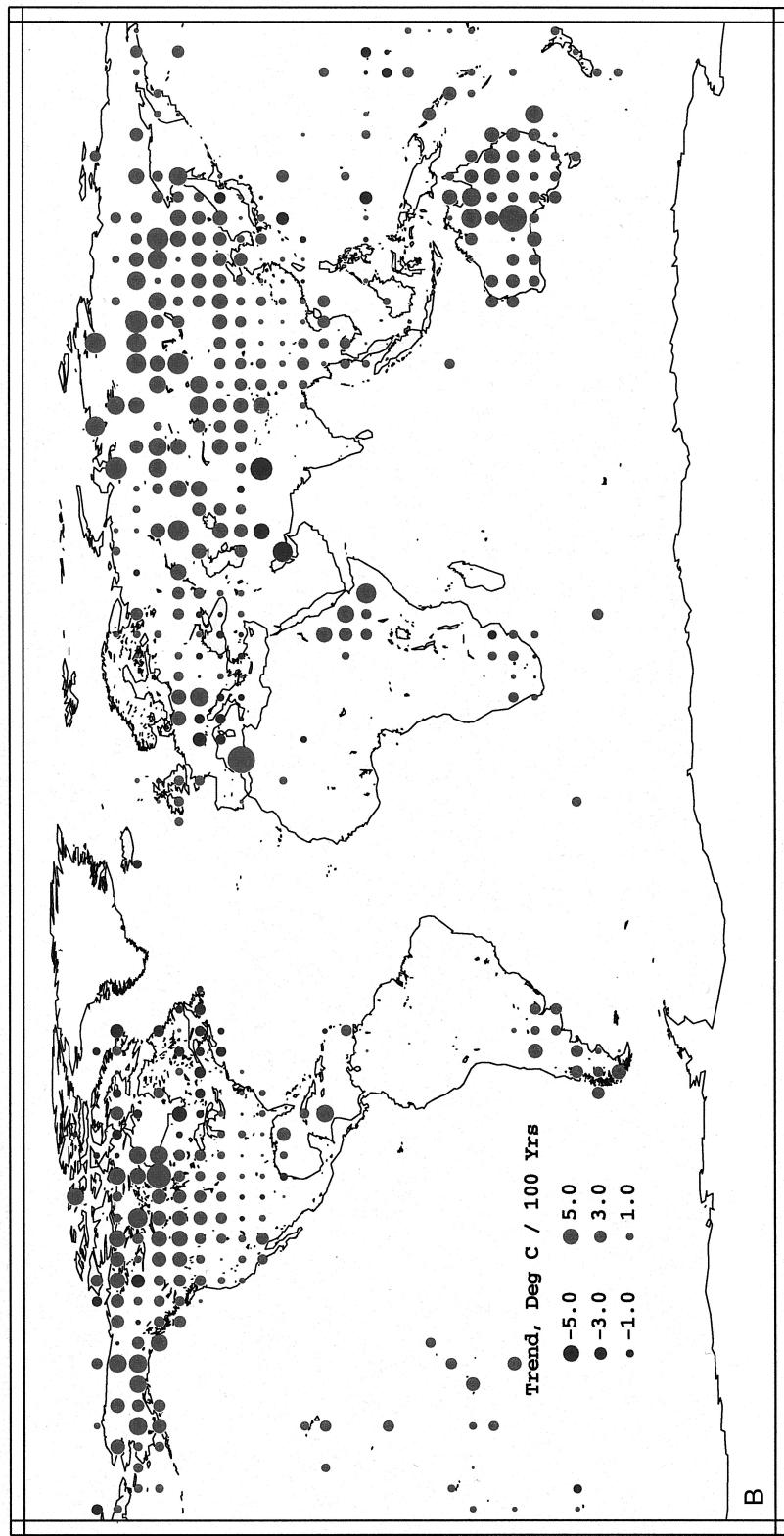


FIGURE 6 (Continued)

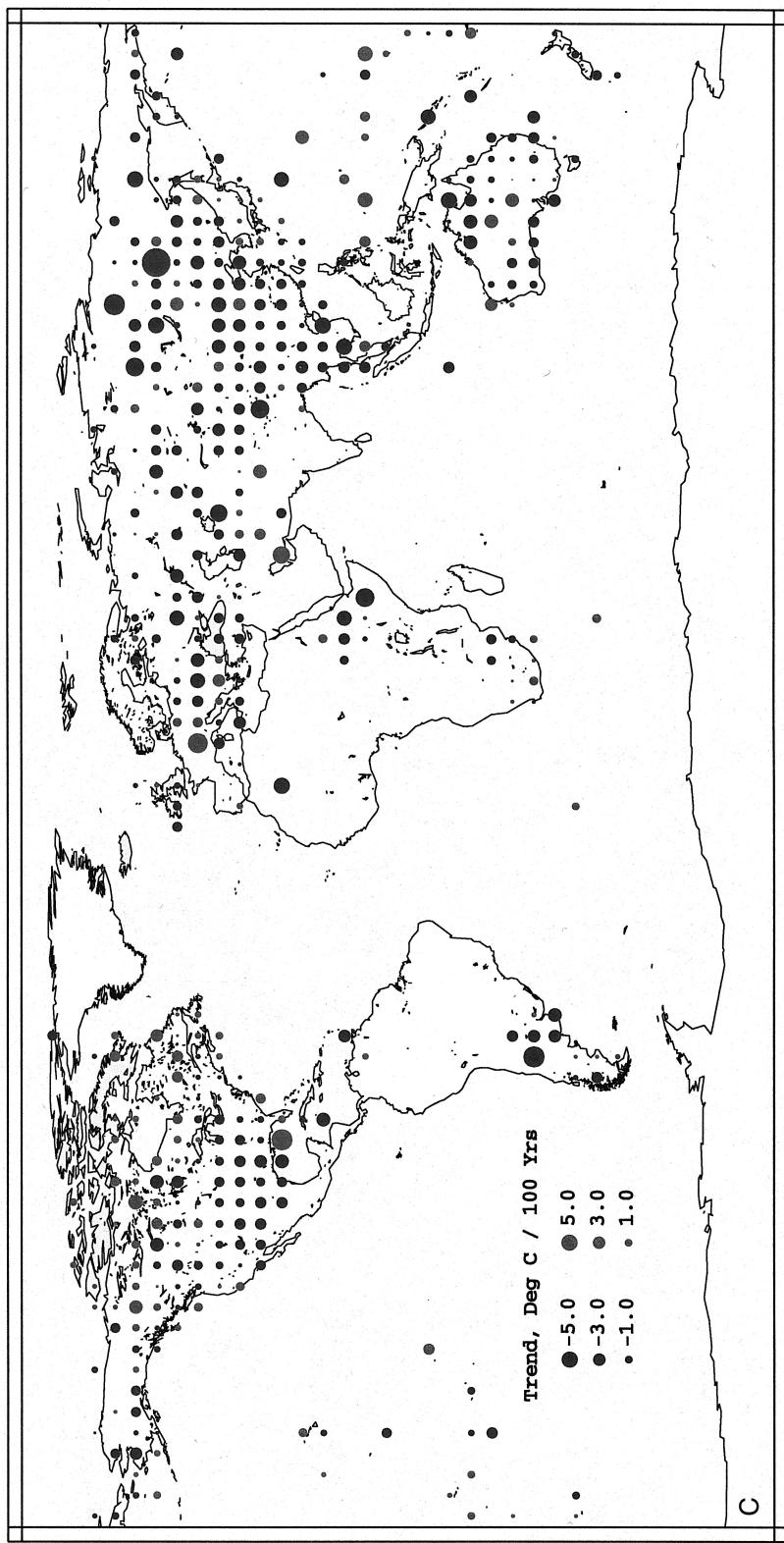


FIGURE 6 (Continued)

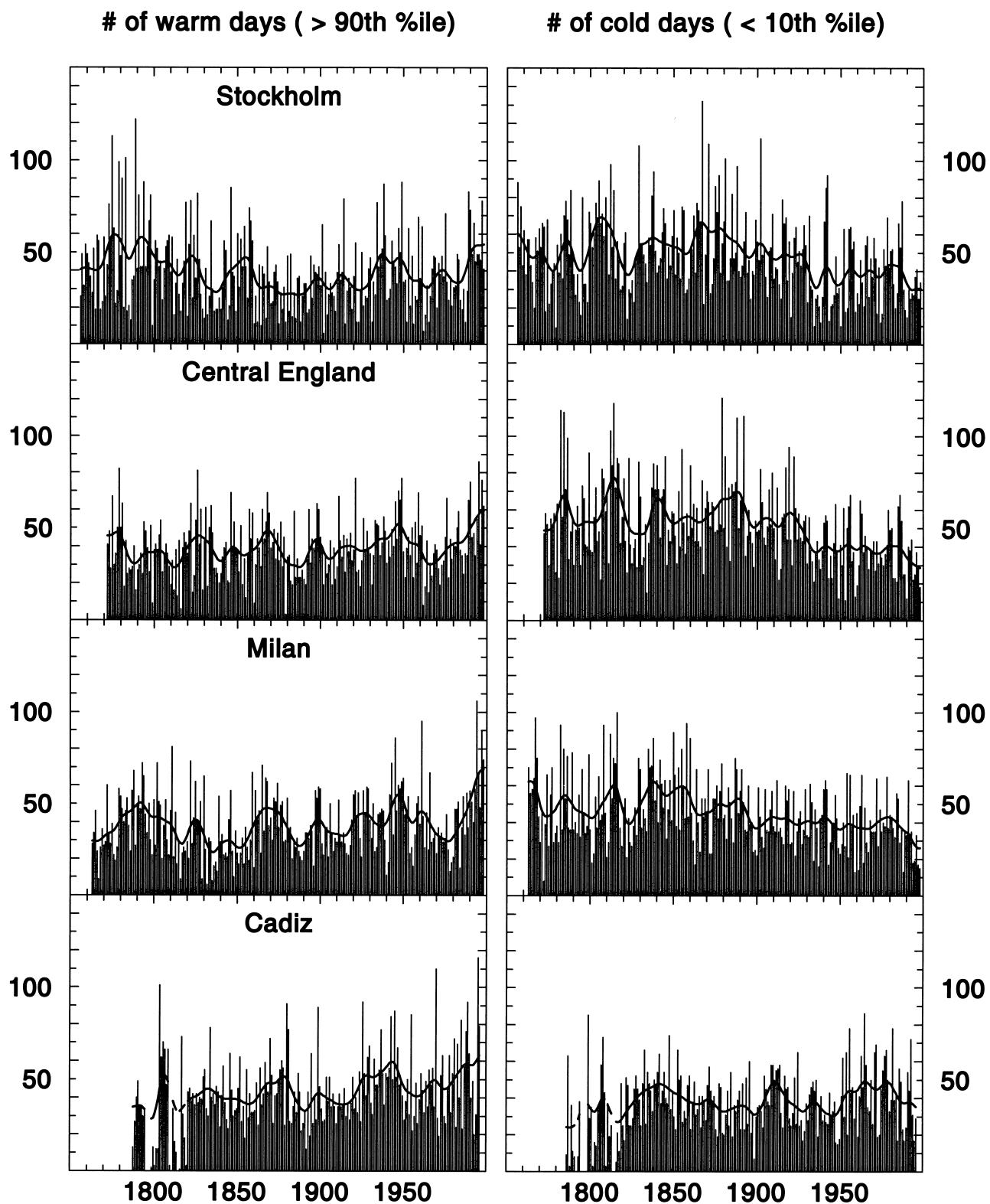


FIGURE 7 Numbers of cold days (<10th percentile) and warm days (>90th percentile) for four European locations with long daily records (Stockholm, central England, Milan, and Cadiz). The smooth curves highlight variations on decade time scales.

(the Milankovitch effect), bringing, for example, higher irradiance in summer to northern high latitudes around 9000 years ago.

Information about the past millennium comes from a variety of high-frequency and low-frequency proxy sources. High-frequency sources, giving information on the annual time scale, include early instrumental records (back to the late seventeenth century in Europe), written historical documents (mainly Europe and the Far East), tree-ring densities and widths (mid to high latitudes of both hemispheres), ice cores (both polar ice caps and also high-elevation tropical and smaller polar-latitude ice caps), corals (tropical), and some highly resolved lake and marine sediments. Low-frequency (decade-to-century time scale change) evidence comes from boreholes, glacial advances/retreats, and peat, lake, and marine cores. Uncertainties in all proxy information are considerable, both because evidence is restricted to where these written and natural archives survive, and more important, all proxy records are only imperfect records of past temperature change.

The last decade has seen a dramatic improvement in both availability of past evidence and also in information from diverse regions and sources. Figure 8 compares several different reconstructions of Northern Hemisphere temperature change for most of the last millennium. The reconstructions are of different seasons so, based on the instrumental record, would be expected to differ somewhat. None of the series is strictly independent of the others, as they contain some common sources, but each has made different assumptions in averaging.

The most striking feature of the multiproxy averages is the warming over the twentieth century, both for its magnitude and duration. Agreement with the instrumental record should be taken as read, as all the components of the series have, to some extent, been calibrated against instrumental data, either locally or as a whole. The twentieth century was the warmest of the millennium, and the warming during it was unprecedented. It is clear that the 1990s was the warmest decade and 1998 the warmest year. The coldest year is likely to have been 1601, about 1°C below the 1961–1990 average, although this is considerably more uncertain than the warmth of 1998. If standard errors are assigned to these series, as was the case for the instrumental period, errors would be considerably larger and it is probably prudent to consider only changes on decade-to-century time scales.

Earlier studies, using considerably fewer proxy datasets, have considered the past millennium and two periods, the Little Ice Age (variously defined as 1450–1850) and the Medieval Warm Epoch (less well defined in the literature, but 900–1200 encompasses most earlier work) are often discussed. To some extent, these two periods have become accepted wisdom, but the various curves in Fig. 8 indicate only partial support. Spatial analysis of the proxy data shows that no periods in the millennium were universally colder or warmer everywhere, considerable variability being present. The latter is to be expected even by studying the instrumental period since 1850. Just as the early 1940s were warm in many parts of the world but Europe was cold, the early seventeenth century was

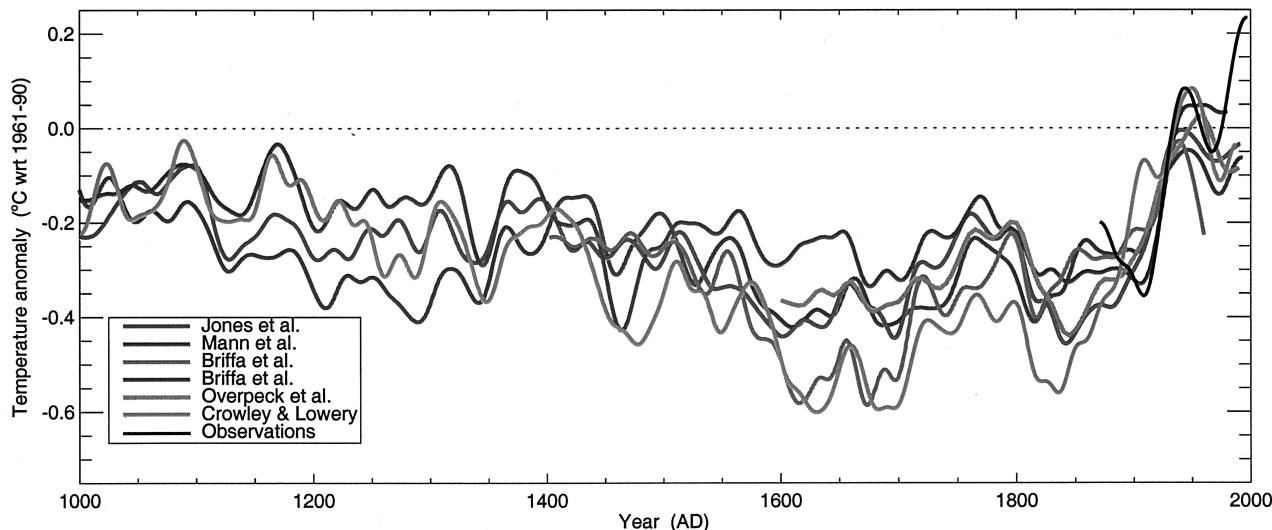


FIGURE 8 Reconstructions of Northern Hemisphere temperatures from several different combinations of proxy data (multiproxy averages). All the series have been smoothed with a 50-year Gaussian filter and all are plotted as departures from the 1961–1990 average. The different reconstructions are shown by the colored lines, the black being the instrumental record for April to September for land areas N of 20°N. All the series have been assembled recently and represent cutting-edge research in paleoclimatology.

cool in many regions but was relatively mild in Iceland. In many respects, therefore, paleoclimatology is in the process of reassessing the evidence for these past periods, and further changes are in prospect as more evidence becomes available.

The various series in [Fig. 8](#) differ in some respects with regard to the coldest and warmest periods of the millennium, but they have all analyzed orders of magnitude more data than was available in the early 1970s. The cooler centuries of the millennium were the sixteenth to the nineteenth, the seventeenth being the coldest in Europe and the nineteenth the coldest in North America. These regions are still the best studied, and it will be vital in the future to extend our knowledge to other areas, particularly in the Southern Hemisphere. At present, for every one long SH record there are probably 20 in the NH. Just as with the instrumental record, it is important to gain as much evidence from as many regions as possible, if we are to understand fully how global and hemispheric temperatures have varied over this long time. Contrasts in the timing of changes between regions and particularly between the hemispheres must be recognized if we are to fully understand the causes of the changes. A more complete understanding of the causes of the changes will allow us to determine how much climate can change naturally, enabling us to better distinguish the degree of human influence on surface temperature during the twentieth century.

III. PRECIPITATION RECORDS

A. Homogeneity and Aggregation

As with temperature data, precipitation measurements are subject to problems with homogeneity. The greater spatial variability of monthly precipitation totals means that slight changes in location, rain gauge design, and elevation can seriously impair the reliability of long-term records. Furthermore, although the network of available data is slightly better (in terms of gauges per area) than temperature, grid-box or regional averages are less reliable because of the higher levels of spatial variability. Time series of neighboring (<100 km) temperature sites are likely to be highly correlated with one another ($r > 0.95$), but for precipitation the correlation is likely to be only 0.6–0.8, lower if the precipitation is mainly convective, higher if it is principally of frontal origin.

Despite the issues of accuracy, several groups have produced grid-box datasets at various resolutions from $1^\circ \times 1^\circ$ to $5^\circ \times 5^\circ$. Unlike their temperature counterparts, datasets are confined to land areas. Assessments over marine regions are beginning to be made using various satellite sensors, but they are confounded by the inevitable

changes in sensors and records only extend back to the mid-1970s. Attempts to improve the records are important, though, as it is necessary to provide global-scale observational databases with which to test simulations of general circulation models.

B. Analyses of Grid-Box Datasets

[Figure 9](#) shows trends in precipitation on an annual basis for two different periods in the twentieth century (1931–1960 and 1961–1990). Patterns of changes are seen to be considerably more spatially variable than for temperature. Also, because of different rainy seasons in many tropical and subtropical regions, annual trends are not that meaningful, but they give insights into the available data and the sorts of analyses that can be achieved. Analysis of precipitation trends and understanding the reasons for the changes is important to many facets of society, such as agriculture, water resources, and forestry. Changes are generally more important than for temperature, but it is much harder to document and diagnose the trends and relate them to human influences.

C. Longer-Term Series

As with temperature, there are precipitation records extending back to the eighteenth century in Europe. Proxy reconstructions are possible in some regions, as tree growth is crucially dependent on precipitation amounts in some regions. Also, documentary records from Europe and the Far East contain more references to precipitation than temperature, generally in the form of diary information allowing monthly rain-day counts to be made. This is particularly evident in China and Japan. In Japan, official records enable wet and dry days to be mapped each day for the whole of the Edo period from 1600 to 1850.

[Figure 10](#) shows seasonal and annual precipitation series for England and Wales (EWP) back to 1766. This region has always had a high density of gauges, official records run by the Meteorological Office since about 1860, but always many amateur meteorologists recording amounts, sometimes through many generations of the same family. EWP is comprised of seven records in each of five regions covering the area. All the seasons show considerable variability from year to year, with little evidence of longer-time-scale changes. Winters have, however, become wetter in the second half of the record since about 1860 and summers recently have become slightly drier. This region is generally in the path of the main westerly wind belts of the Northern Hemisphere. Averages for more tropical regions will exhibit even greater year-to-year variability, particularly in regions that are strongly influenced by the El Niño/Southern Oscillation phenomenon. Despite

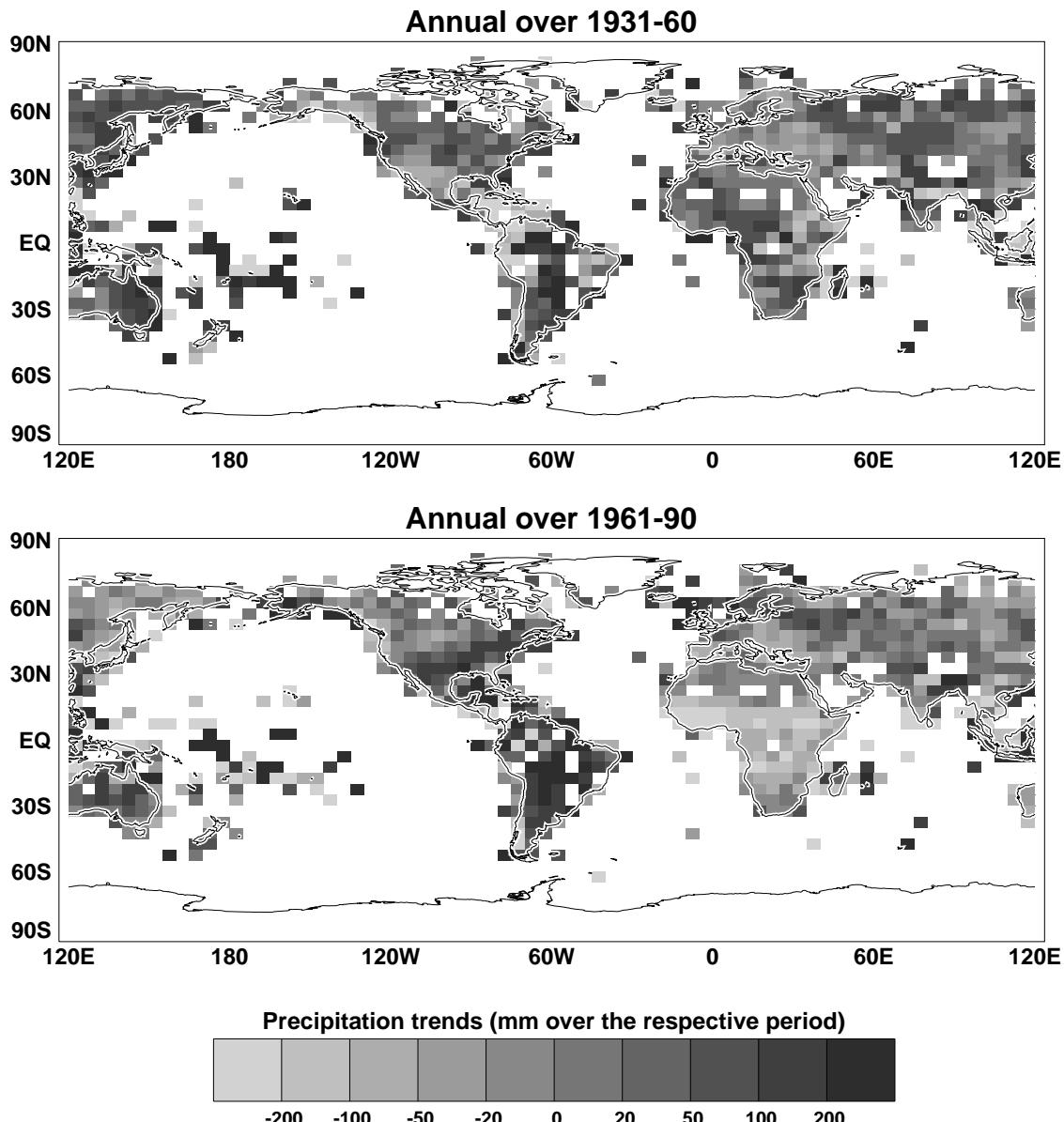


FIGURE 9 Trends in annual precipitation totals for two periods in the twentieth century (1931–1960 and 1961–1990).

this, the slight variations on the 5- to 10-year time scale often severely impact water resources and agriculture over England and Wales, leading to regular droughts and floods.

D. Changes in Daily Extremes

The database of precipitation data analyzed in Fig. 9 is only available monthly. Although the measurements have been taken daily, few long and dense networks of data are digitally available at the daily time scale. Data have recently been analyzed for a few regions (United

States, southern Canada, the former Soviet Union, China, Australia, and parts of Europe). In some of these regions, increases in heavy rainfall have been noted, but the signal is far from uniform. The result is not unexpected, as a warmer atmosphere can hold more moisture and potentially release more in each event. The result is therefore consistent with modeling scenarios, but attributing it to human influences requires a stronger response or considerably more comprehensive and longer datasets so that greater understanding of the natural variability of daily rainfall amounts can be produced.

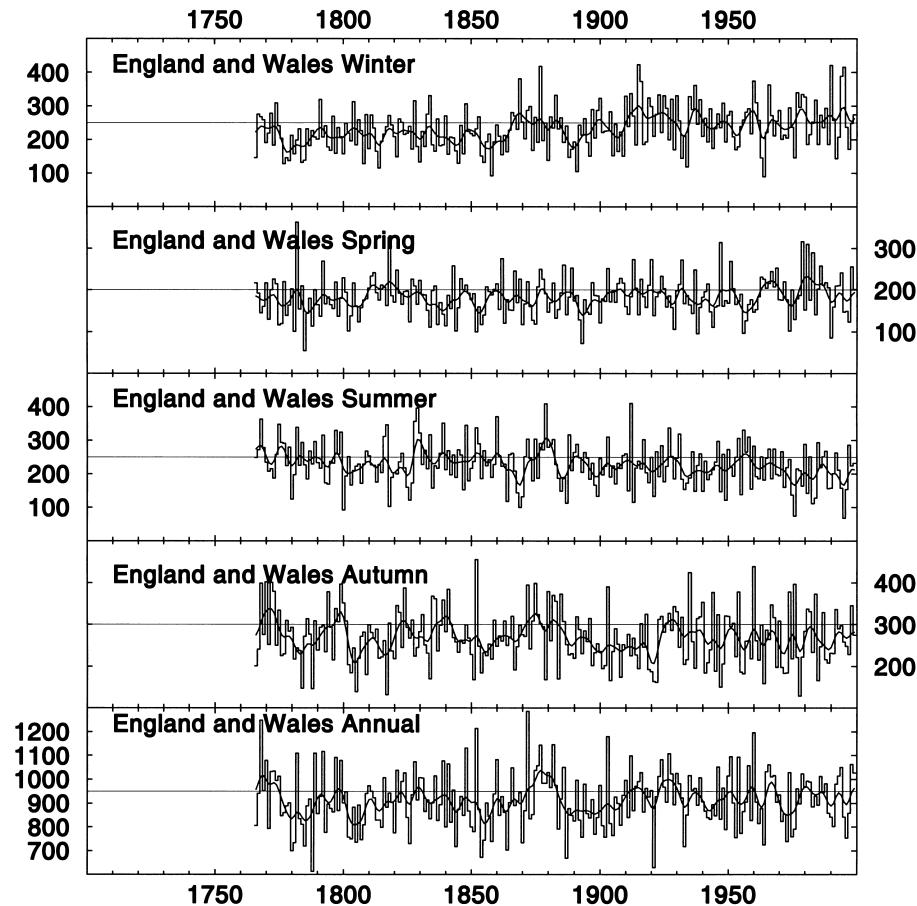


FIGURE 10 Seasonal and annual precipitation totals for the England and Wales precipitation average, 1766–1999. The smooth curves highlight variations on decade time scales.

Precipitation is vital to measure, and to attempt to predict, as it has a strong influence on a number of sectors, but its strongly variable nature both in space and time, and the fact that it is much more poorly modeled than temperature means that it is a poor choice with regard to climate change detection.

IV. CLIMATE CHANGE DETECTION AND ATTRIBUTION

Climate modeling work indicates that increases in greenhouse gases as a result of anthropogenic activities will lead to warming. The previous section has clearly shown that the world has warmed, the increase in global average surface temperatures being highly significant both compared to year-to-year variability over the instrumental period and to longer-time-scale changes over the last millennium. Although we have detected the change, temperatures have varied in the distant and recent past, so how can we attribute the changes to human activities?

The increase might have occurred without any human interference.

The second step in detection studies, attributing the changes to human causes, makes use of climate models, which allow us to estimate the climatic effects of a range of human-induced and natural factors. Attribution requires that the changes agree, with statistical confidence, with those modeled. The human factors considered by models include both the increasing atmospheric concentrations of greenhouse gases and the effects of sulfate aerosols. Natural factors included are solar output changes, the effects of volcanic eruptions, internal variability of the climate system (mainly with respect to the oceans), and interactions among all the external and internal forcing factors.

The 0.6°C rise in globally averaged temperatures is compatible with, in terms of both magnitude and timing, that predicted by models which take the combined influences of human factors and solar variability into account. More recent studies of attribution have considered the patterns of temperature change, both spatially and in the vertical column of the atmosphere. Climate models indicate

that the warming will not take place evenly, with cooling possible in a few regions. Warming will occur throughout most of the troposphere, but with cooling evident in the stratosphere. Looking at these detailed changes in patterns, therefore, provides more powerful techniques than dealing only with large-scale averages. It is likely that different forcing factors have different signatures in their temperature response, so with adequate data and a large enough response we should be able to distinguish the different influences. There is the tacit assumption in such studies that our modeling is a sufficiently good simulation of reality, i.e., we have considered everything. While this assumption is unlikely to be strictly true, we should be confident that we have considered and correctly parameterized all important factors.

Comparisons between observed and modeled patterns at the surface and in the vertical column of the atmosphere have now been made. Model predictions show increasing agreement with the changes in temperature that have occurred over the last 50 years. Agreement is not perfect, but we would not expect it to be, as there are known inadequacies in both the models and our past history of known forcing factors. The best agreements come from the models which account for both greenhouse gases and the effects of sulfate aerosols. In statistical terms, the correspondence between the two is unlikely to have occurred by chance. Studies of this kind are at the forefront of climate change science and are likely to be revised after each new generation of climate models becomes available. Improvements in data are also a factor, but of less importance compared to computer developments enabling us to improve the realism of the models by going to finer and finer resolution.

The agreements between observations and models are seen through similarities at the largest of spatial scales, contrasts between the hemispheres, between the land and ocean parts of the world, and at different levels in the vertical. It is not possible, nor will it be for several decades to come, to say that particular changes locally can be attributed to human influences. Although attribution studies are heavily dependent on statistics, the modeling results accord with our basic physical understanding of the climate system. They should not be viewed as a “black box” only understood by a few statisticians and climate modelers. The large-scale nature of current detection and attribution studies and the decade time scales must always be remembered. Media questions relating to a run of warm extremes in a region can only be answered by saying that the changes are not inconsistent with greenhouse-enhanced warming. Even a run of a few cold years is compatible because of the time scales of detection/attribution studies.

Despite the climate community concluding in its second (1995) IPCC report that “the balance of evidence suggests that human activities have led to a discernible influence

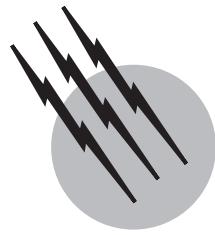
on global climate and that these activities will have an increasing influence in the future,” we need to continue efforts to improve both our modeling and our data collection and analysis, particularly for the recent preinstrumental past. Our knowledge at this time, while being highly suggestive, is still incomplete in many aspects. For example, although the majority of climate scientists are happy with detection and attribution claims, we cannot say with much confidence how much of the warming is human-induced. Ideally, if our models were totally adequate and our knowledge of the changes in forcing over the last millennium much better than it is, we would expect to be able to retrodict the course of regional and global temperature change over this time. The fact that we are in the same ballpark shows the advances that have been made over the last 20–30 years. There should, however, be neither room for complacency by the scientists, nor reluctance to continue funding by agencies, as there is still much work to be done and much more to understand.

SEE ALSO THE FOLLOWING ARTICLES

CARBON CYCLE • CLIMATOLOGY • GREENHOUSE WARMING RESEARCH • HYDROLOGIC FORECASTING • MESOSCALE ATMOSPHERIC MODELING • METEOROLOGY, DYNAMIC • OCEAN-ATMOSPHERIC EXCHANGE • OZONE MEASUREMENTS AND TRENDS • RADIATION, ATMOSPHERIC • THERMOMETRY

BIBLIOGRAPHY

- Bradley, R. S. (1999). “Paleoclimatology: Reconstructing Climates of the Quaternary,” Academic Press, San Diego, CA.
- Easterling, D. R., et al. (1997). “Maximum and minimum temperature trends for the globe,” *Science* **277**, 364–367.
- Folland, C. K., and Parker, D. E. (1995). “Corrections of instrumental biases in historical sea-surface temperatures,” *Q. J. Roy. Met. Soc.* **121**, 319–367.
- IPCC reports (1990, 1995, 2001), particularly the chapters on “Observed climate variability and change,” Cambridge University Press, Cambridge, U.K.
- Jones, P. D., Briffa, K. R., Barnett, T. P., and Tett, S. F. B. (1998). “High-resolution palaeoclimatic records for the last millennium: interpretation, integration and comparison with General Circulation Model control run temperatures,” *The Holocene* **8**, 455–471.
- Jones, P. D., New, M., Parker, D. E., Martin, S., and Rigor, I. G. (1999). “Surface air temperature and its changes over the past 150 years,” *Rev. Geophys.* **37**, 173–199.
- Mann, M. E., Bradley, R. S., and Hughes, M. K. (1998). “Global-scale temperature patterns and climate forcing over the past six centuries,” *Nature* **392**, 779–787.
- National Research Council (NRC) (2000). “Reconciling Observations of Global Temperature Change,” National Academy Press, Washington, DC.
- Peterson, T. C. et al. (1999). “Global rural temperature trends,” *Geophys. Res. Lett.* **26**, 329–332.



Greenhouse Warming Research

Richard T. Wetherald

National Oceanic and Atmospheric Administration

- I. Historical Background and Simple Models
- II. General Circulation Model Description and Experimental Procedure
- III. Experiments with Idealized Geography
- IV. Experiments with Realistic Geography
- V. Comparison with Other Modeling Groups
- VI. Coupled Air-Sea Models
- VII. Committed Warming
- VIII. Summary
- IX. Appendix: Illustration of GCM Procedure

GLOSSARY

- Albedo** Degree of reflection of solar radiation.
- Black body** A substance that absorbs all of the radiation which falls upon it regardless of wavelength.
- Conduction** Transfer of heat through a substance.
- Convection** Transfer of heat by upward atmospheric motions.
- Dynamical** Pertaining to energy, force, or motion in relation to force.
- Feedback** The return of a portion of the output of a process or system to the input.
- Gradient** Rate of change of a quantity from one place to another.
- Greenhouse effect** Absorption of solar radiation by the earth, its conversion and reemission in the infrared, and the absorption of this radiation by the atmosphere resulting in a gradual rise of atmospheric temperature.
- Heat capacity** Ability of a substance to store heat.
- Heat flux** Transfer of heat from one place to another.

Hydrologic cycle Transfer of water vapor/liquid water from the ground to the atmosphere and back again.

Lapse rate The rate of decrease of tropospheric temperature with height.

Latent heat Transfer of heat from the ground upward by the vertical gradient of moisture carrying heat of vaporization.

Opacity Degree of impenetrability.

Radiation forcing Change in the heat balance of the atmosphere–earth system.

Rainbelt Region of maximum precipitation generally located in the upward branch of the Indirect Cell in middle latitudes.

Sensible heat Transfer of heat from the ground upward by the vertical temperature gradient.

DURING THE SUMMER of 1988, one of the worst droughts in history occurred across most of the North American continent. During the subsequent winter, in

the eastern United States, particularly in the mountainous watershed regions along the Appalachian range, very little snow fell. Since then, other anomalous weather events have occurred; severe flooding of the Mississippi River basin in the summer of 1993 and a series of abnormally dry summers and warm winters in the eastern United States with little or no snow in the late 1990s. Regardless of what caused these phenomena, they serve as graphic examples of what can happen if our climate changes significantly from that to which we have become accustomed. In particular, the summer of 1988, as well as an overall tendency for global warming since then has sparked a great deal of discussion on the topic of greenhouse warming and whether or not it has actually begun.

The Climate Dynamics Group of the Geophysical Fluid Dynamics Laboratory of NOAA, formally headed by Dr. Syukuro Manabe, began researching the greenhouse effect in the late 1960s and early 1970s. During this period, data on atmospheric carbon dioxide (CO_2) obtained by Dr. Keeling and his colleagues working at the Mauna Loa Observatory in Hawaii and Antarctica became available and indicated that atmospheric concentrations of CO_2 were, indeed, increasing and increasing at a fairly consistent rate. These observations coupled with the theoretical research work being done at the Geophysical Fluid Dynamics Laboratory (GFDL) laid the foundation for a transition of greenhouse theory from science fiction to science.

I. HISTORICAL BACKGROUND AND SIMPLE MODELS

In the latter half of the nineteenth century, J. Tyndal and S. Arrhenius suggested that a climate change may be induced by a change in CO_2 concentration in the earth's atmosphere. This research work was followed with other studies by G. Callendar, G. Plass, K. Kondratiev, H. Niilisk, L. Kaplan and F. Möller. All of these earlier studies employed relatively simple radiative models based upon the radiative balance at the earth surface alone and, because of this formulation, obtained a variety of sensitivities of surface air temperature to similar increases of atmospheric CO_2 . To overcome this difficulty, S. Manabe and R. T. Wetherald in 1967, employed a one-dimensional model which included both radiative processes and a temperature lapse-rate adjustment or a "radiative, convective model" which was based upon the radiation balance of the entire atmosphere as a whole, not just for the earth's surface. In doing so, they used a model which conserved heat and radiative energy throughout the entire model atmosphere and moved greenhouse research from a "back of the envelope" calculation to a computer.

The above model was used to evaluate, among many other things, the CO_2 -induced change of atmospheric temperature throughout the model atmosphere. This was a one-dimensional model and consisted of a system of equations which represented the effects of radiative transfer and vertical convective mixing upon the heat balance of the atmosphere. The mathematical formulation was based upon the following conditions: (1) the net radiative flux at the top of the model atmosphere is zero, (2) the lapse rate (temperature decrease with height) throughout the model atmosphere cannot exceed a certain critical value (taken to be $6.5^\circ\text{C}/\text{km}$) due to convective and other dynamical processes, (3) in a convective stable layer, the net radiative flux is zero, (4) the surface has zero heat capacity which implies that the net downward flux of radiation is equal to the upward convective heat flux, (5) the vertical distribution of relative humidity rather than absolute humidity was prescribed. Three concentrations of atmospheric CO_2 were considered: 150, 300 and 600 parts per million (ppm). The model was, then, integrated until steady state or "equilibrium" temperature and moisture profiles were obtained in each case.

Figure 1 illustrates the vertical distributions of the equilibrium temperature of the model atmosphere with the normal (300 ppm), half the normal (150 ppm) and twice the normal (600 ppm) concentration of CO_2 . This figure indicates that, in response to the doubling of CO_2 , the temperature of the model troposphere increases by approximately 2.3°C whereas that of the middle stratosphere decreases by several degrees. In addition, it reveals that the magnitude of the warming resulting from the doubling of CO_2 concentration is approximately equal in magnitude to the cooling from the halving of the CO_2 concentration. This result suggests that the CO_2 -induced temperature change is linearly proportional to the logarithm of CO_2 concentration rather than to the CO_2 concentration itself.

The physical processes of the warming due to an increase in CO_2 concentration have traditionally been ascribed, in an analogous manner, to those operating in a "greenhouse." However, this analogy is not quite correct and the actual processes operating in the radiative, convective model may be explained as follows. It is well known that tropospheric temperature generally decreases with increasing altitude. From the physics of "black-body" radiation, it is also known that the amount of infrared (long-wave) radiation is proportional to the fourth power of temperature in degrees Kelvin [degrees Kelvin (K) = degrees centigrade (C) + 273.2] according to a mathematical expression called the Stefan-Boltzmann law. As the concentration of CO_2 is increased throughout the model atmosphere, the height of the emitting source of the infrared (or cooling) radiation to space also increases due to the increased opacity of the model atmosphere. Since

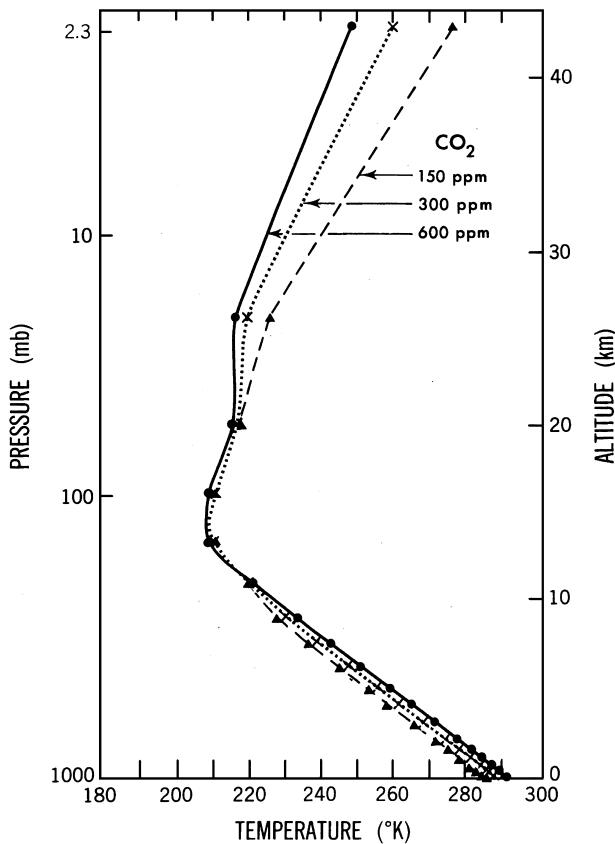


FIGURE 1 Vertical distribution of temperature in radiative, convective equilibrium for various values of atmospheric carbon dioxide concentrations, i.e., 150, 300, 600 parts per million (ppm) by volume [From Manabe, S., and Wetherald, R. T. (1967). *J. Atmos. Sci.* **24**, 241–259.]

tropospheric temperature decreases with increasing altitude, this results in the reduction of the effective emission temperature for the outgoing radiation and hence the outgoing radiation itself at the top of the model atmosphere due to the Stefan-Boltzmann relationship. In order to satisfy the condition that the outgoing infrared or terrestrial radiation be equal to the incoming solar radiation (which has not changed), it is necessary to raise the temperature of the entire model troposphere underneath to prevent the reduction of the outgoing terrestrial radiation. In other words, the troposphere must become warmer to compensate for the above reduction of longwave radiation to space in order to maintain an energy balance at the top of the model atmosphere.

In response to this warming, the absolute humidity (moisture content) of the model atmosphere also increases due to the condition of prescribed relative humidity. This causes further increases in both the opacity and height of the effective source of emission of outgoing infrared radiation just as it did for the additional CO₂ resulting in an

additional increase of tropospheric temperature to compensate. This further enhancement is called “water-vapor feedback” and is responsible for doubling the sensitivity as compared with the same model integrated under the condition of prescribed absolute humidity. As it turns out, this water-vapor feedback is extremely important in the three-dimensional general circulation model studies which will be described in the following sections.

II. GENERAL CIRCULATION MODEL DESCRIPTION AND EXPERIMENTAL PROCEDURE

Atmospheric scientists have developed general circulation models (GCMs) to study a variety of problems ranging from daily forecasting to long-term climate change including the climatic consequences of increased CO₂. In simple terms, a GCM is a complex mathematical model composed of a system of partial differential equations derived from the basic laws of physics and fluid motion and “retooled” to consider air as a fluid. These equations describe the dynamic, thermodynamic, radiational and hydrologic processes of the atmosphere. A block diagram of a GCM is illustrated by Fig. 2 where the large rectangles denote the main components of the model and the arrows joining one rectangle to another denote the interactions that take place between each component. The equations represented by this block diagram are too complex to be solved analytically, so they must be converted into an arithmetic form that is suitable for computation by a large digital computer. This alternate mathematical form is generally referred to as “finite differences.” The

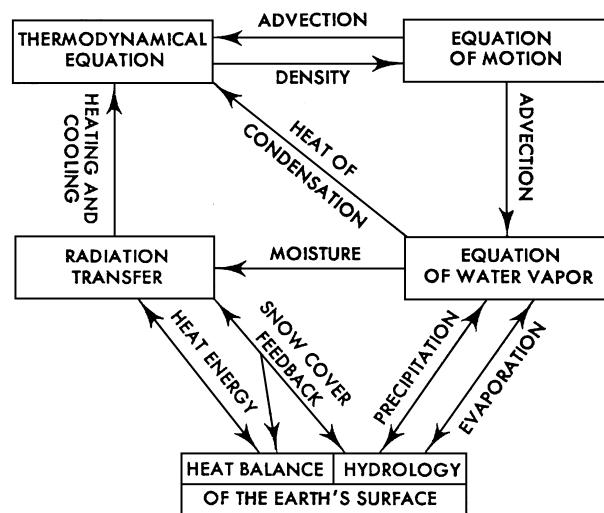


FIGURE 2 Block diagram depicting the structure and major components of a general circulation model.

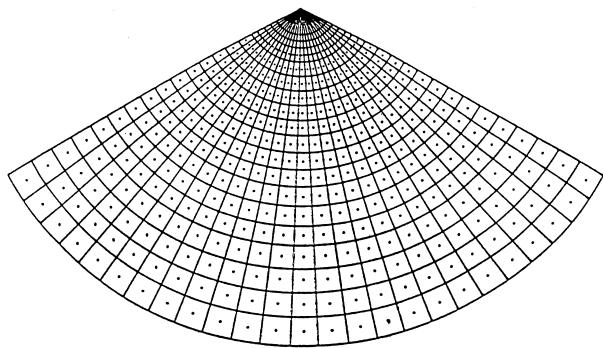


FIGURE 3 Diagram of a sample grid system typically used in performing general circulation experiments.

above procedure is necessary since digital computers can only add, subtract, multiply, and divide. The next step is to divide the entire three-dimensional atmosphere into a systematic series of “boxes” to which the basic equations must be applied and evaluated. In each box, the motion of the air (winds), heat transfer (thermodynamics), radiation (long- and shortwave), moisture content (relative humidity and precipitation), and surface hydrology (evaporation, snowmelt, and runoff) are calculated as well as the interactions of these processes among each of the boxes. The model is, then, programmed to run on this finite differ-

ence network in a series of “time steps” until the particular forecast period is completed. A sample grid system for this purpose, viewed in two dimensions, is illustrated by **Fig. 3**. For the sake of clarity, a simple example of how this entire procedure is accomplished is given in the Appendix (Section IX of this article).

The model is integrated with current CO₂ concentrations until it reaches a steady state or a state where the global mean temperature does not increase or decrease over a long period of time. This run is called the “control” or present-day climate experiment. The above procedure is, then, repeated by assuming twice as much CO₂ with no other change. After both experiments have been completed, the two computed “climates” obtained are averaged over a sufficiently long time period to remove the natural variation present in each climate simulation (usually 10 or more model years). Finally, we compare the two climates to determine the changes caused by the doubling of CO₂. A diagram illustrating this procedure is given in **Fig. 4**.

III. EXPERIMENTS WITH IDEALIZED GEOGRAPHY

To initiate the research work through the use of three-dimensional models, Manabe and Wetherald used a GCM

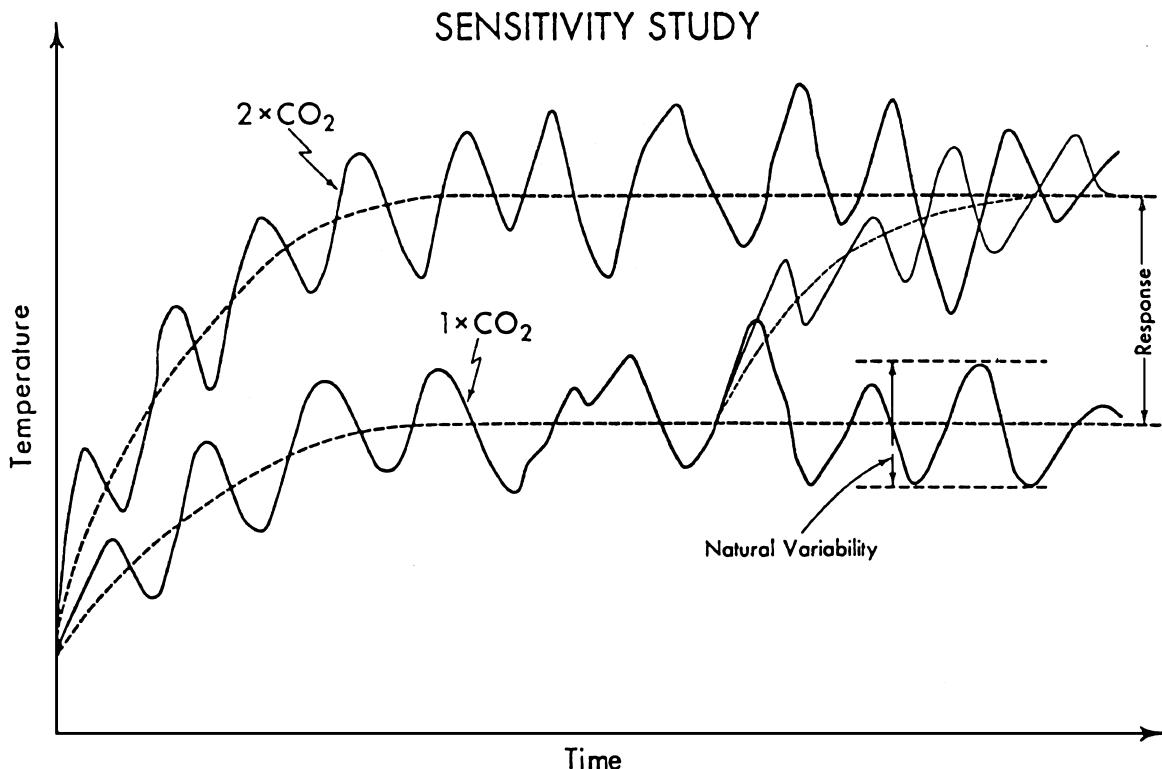


FIGURE 4 A schematic diagram depicting the method used in conducting general circulation experiments. In general, two computer runs (integrations) are performed, one for the normal CO₂ (control) experiment and the other for a higher concentration of CO₂ (usually a doubling of CO₂).

with a simplified land-sea distribution to conserve computer time and to better understand the feedback mechanisms caused by the increase of atmospheric CO₂. A diagram depicting this land-sea distribution is shown in Fig. 5. The “ocean” was considered to be a “swamp ocean” which was, basically, a perpetually wet land surface. Also, surface hydrology was incorporated into the model in a simplistic way namely, a 15-cm-deep soil moisture “bucket” which may be filled by rainfall or snowmelt and depleted by evaporation. Runoff occurs as any excess water in the bucket over the 15-cm capacity. The surface air temperature over both the ocean and land surfaces is computed under the assumption that these surfaces have no capacity to store heat. To incorporate the effect of albedo (reflectivity) change into the model, the depth of snow cover was predicted by an equation of snow budget whereas the extent of “sea ice” was determined by the temperature of the swamp surface. The albedos of snow cover and sea ice were assumed to be much larger than those of bare soil or open sea. Annual mean insolation was assumed in this model.

Figure 6 shows the latitude-height difference of zonal mean temperature caused by a doubling of CO₂ obtained from the model described above. This distribution of temperature change is in qualitative agreement with that obtained from the one-dimensional radiative, convective model, namely, it increases in the model troposphere and decreases in the model stratosphere. For this model, the area mean change of surface air temperature is about 3°C and is somewhat larger than that obtained from the one-dimensional model. This is due to the fact that the three-dimensional GCM included the effects of the recession of snow cover in higher latitudes which was not present in the one-dimensional model. This mechanism can be described as follows. As the climate warms due to increasing greenhouse gases, snow cover at or near the average snow

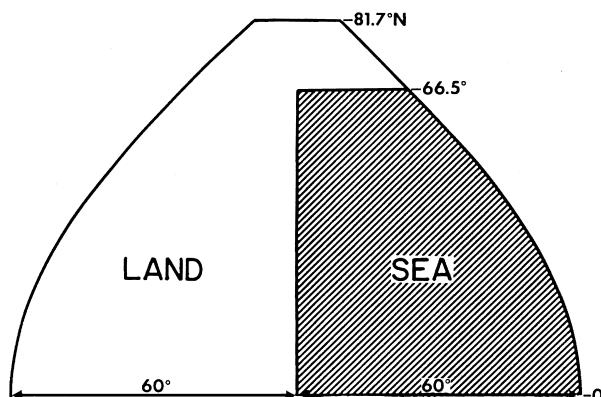


FIGURE 5 Diagram illustrating the distribution of continent and ocean in a model with idealized geography. [From Manabe, S., and Wetherald, R. T. (1975). *J. Atmos. Sci.* **32**, 3–15.]

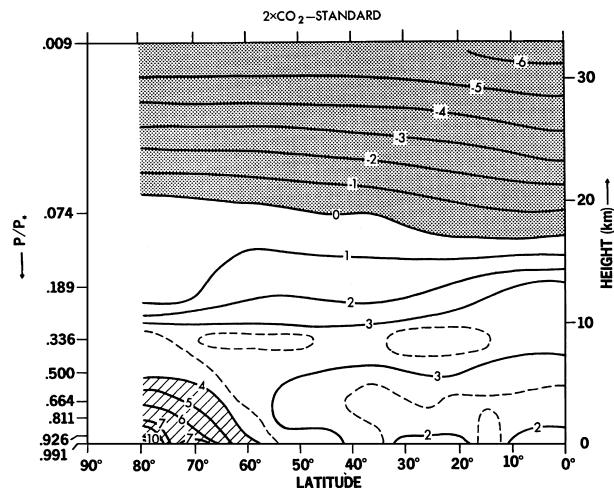


FIGURE 6 Latitude-height distribution of the zonal mean temperature difference (K) between a doubling of CO₂ and control experiments. Shaded region denotes negative values. [From Manabe, S., and Wetherald, R. T. (1975). *J. Atmos. Sci.* **32**, 3–15.]

boundary begins to melt, uncovering bare soil with a much lower reflectivity. This enables greater absorption of direct solar radiation, thereby further heating the ground there. This, in turn, results in further removal and recession of highly reflective snow cover. This cycle repeats until a new equilibrium snow boundary is reached at a higher latitude. This process is referred to as “snow cover-albedo feedback” and is responsible for amplifying the warming of surface air temperature as shown in high latitudes of Fig. 6. Since the model atmosphere is very stable in higher latitudes, the heating due to the snow cover-albedo feedback is limited to just the surface layer.

On the other hand, the increase of surface air temperature in the tropics is considerably less because convective processes spread the CO₂-induced surface heating throughout the model troposphere. Therefore, much less heating is available for increasing the surface air temperature there. In fact, Fig. 6 indicates that, at low latitudes, temperature increases more in the upper troposphere than at the surface. This is due to increased heat release through condensation by more vigorous convection which occurs in the upper tropospheric levels of the tropics.

The increase of CO₂ concentration not only affects the thermal structure but also the hydrologic cycle of the model atmosphere. One of these basic changes is the overall intensification of the hydrologic cycle as shown in Fig. 7. In response to the increase of CO₂ and accompanying warming of the model atmosphere, evaporation also increases significantly, particularly over the idealized “ocean.” To maintain an overall balance of water vapor in the model atmosphere, the increase of the global rate of evaporation stated above must be matched by a similar increase in the global rate of precipitation. These increases of both evaporation and precipitation result in an overall

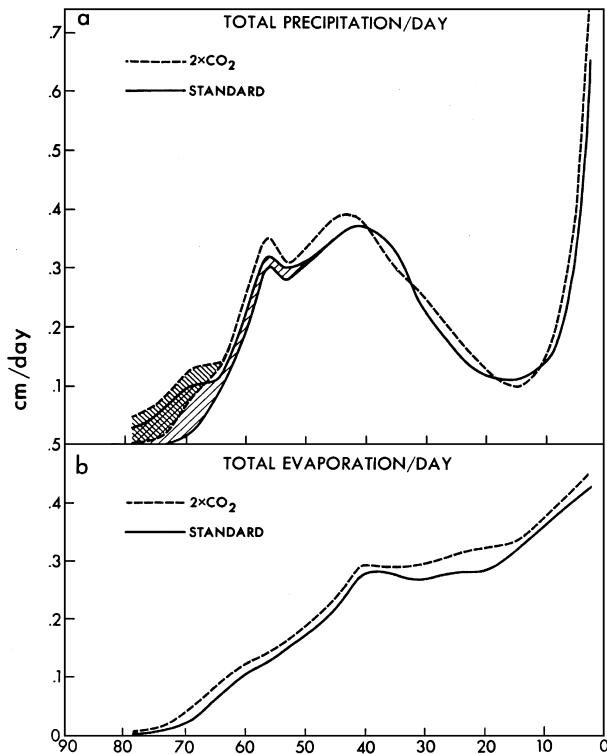


FIGURE 7 Zonal mean rates of (a) total precipitation and (b) evaporation for both a doubling of CO₂ and control experiments (cm/day). Shaded regions denote portion of total precipitation attributed to snowfall. [From Manabe, S., and Wetherald, R. T. (1975). *J. Atmos. Sci.* **32**, 3–15.]

intensification of the entire hydrologic cycle. In this particular study, an increase of 3.7% was obtained due to the doubling of CO₂.

One of the most important factors responsible for this intensification is the change in the surface radiation budget. There are five components of this balance, namely, downward solar radiation, downward longwave radiation, upward longwave radiation, upward latent heat flux (upward motion of moisture carrying heat of vaporization), and upward sensible heat flux (upward motion of heat). For example, the increase in atmospheric CO₂ enhances the downward flux of longwave radiation reaching the earth's surface. In addition, the CO₂-induced warming of the troposphere results in the increase of absolute humidity (moisture content) as discussed in the preceding paragraphs on the radiative, convective model and also contributes to the increase of downward longwave radiation. Therefore, a larger amount of radiative energy is received by the earth's surface which must be removed by turbulent fluxes of latent and sensible heat. Due to the Clausius-Clapeyron relationship between the saturation vapor pressure and temperature, the saturation vapor pressure increases almost exponentially with a linear in-

crease of temperature. This accounts for the increase of global mean rate of evaporation stated above. Therefore, the latent heat flux term becomes a more efficient means of removing heat from the surface than sensible heat.

In a later study, an effort was made to examine the distribution of hydrologic change induced by increases of CO₂ over the idealized land surface (which has now been modified to a half-land-half-sea configuration). Sample results of this analysis are displayed in Fig. 8 which show the zonal mean rates of precipitation (a), evaporation, (b)

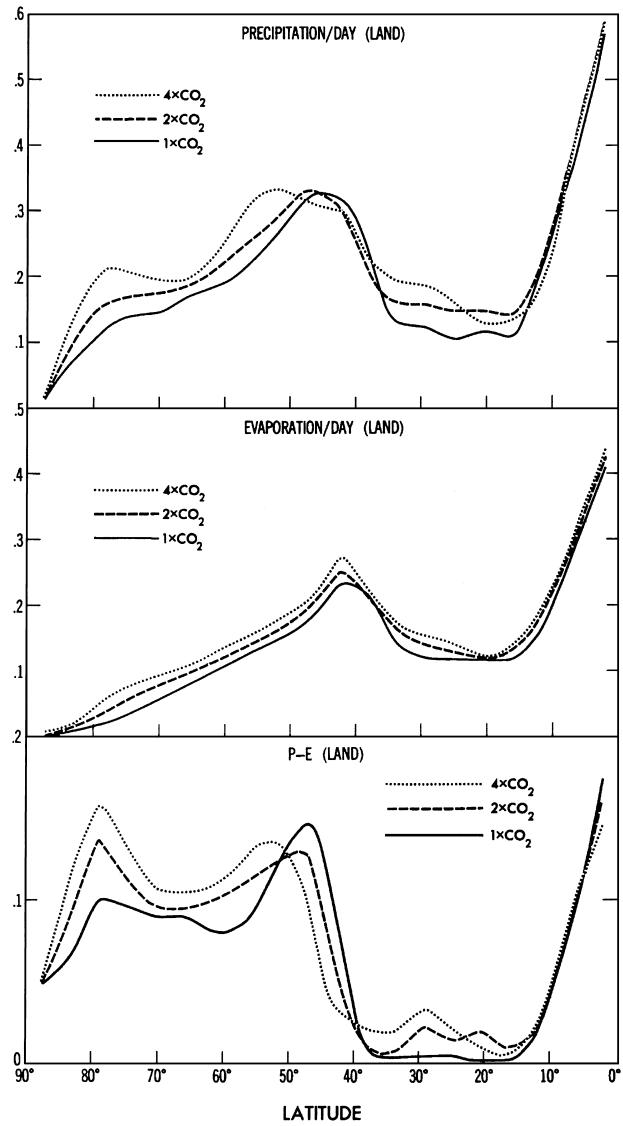


FIGURE 8 Zonal mean rates over the idealized continent of total precipitation (upper), evaporation (middle), and precipitation minus evaporation (lower) for the control experiment, a doubling of CO₂, and a quadrupling of CO₂, respectively (cm/day). [From Manabe, S., and Wetherald, R. T. (1980). *J. Atmos. Sci.* **39**, 99–118.]

and the difference between the two ($P - E$) over the land surface (c). According to these curves, precipitation minus evaporation ($P - E$) increases poleward of 50° latitude and decreases in the latitudinal zone of $40\text{--}50^\circ$ in response to a doubling and quadrupling of CO_2 . These results imply that wetter conditions should prevail poleward of 50° whereas dryer conditions may be expected in the latitudinal zone of 40 to 50° . This is shown in Fig. 9 where there is a considerable reduction of soil moisture in a zonal belt located approximately between 35 and 50° latitude for both concentrations of CO_2 whereas there is a small increase poleward of 50° . The decrease of $P - E$ in middle latitudes may be explained by both a decrease of precipitation and an increase of evaporation there. The decrease in precipitation in middle latitudes may be explained by a poleward shift of the middle latitude rainbelt which moves the maximum of precipitation from middle to higher lat-

itudes (Fig. 9a). The increase of precipitation (and hence $P - E$) poleward of 50° latitude was found to be caused by an increased poleward transport of latent heat and moisture due to the general warming of the model atmosphere.

These studies were followed by versions of the GCMs in which seasonal variation of solar radiation was incorporated into the model; one with idealized geography (reconfigured to two identical hemispheres of half land-half sea) and one with realistic geography. Both of these models necessitated the use of a thermal conducting or “mixed-layer” 50-m deep to simulate the “ocean” surface and its capacity to store heat energy. In the version with idealized geography, the corresponding distributions of $P - E$ and soil moisture, shown previously, now vary with season as Fig. 10 indicates. Specifically, the bottom portion of Fig. 10 shows a zone of enhanced continental dryness which is centered around 35° latitude during

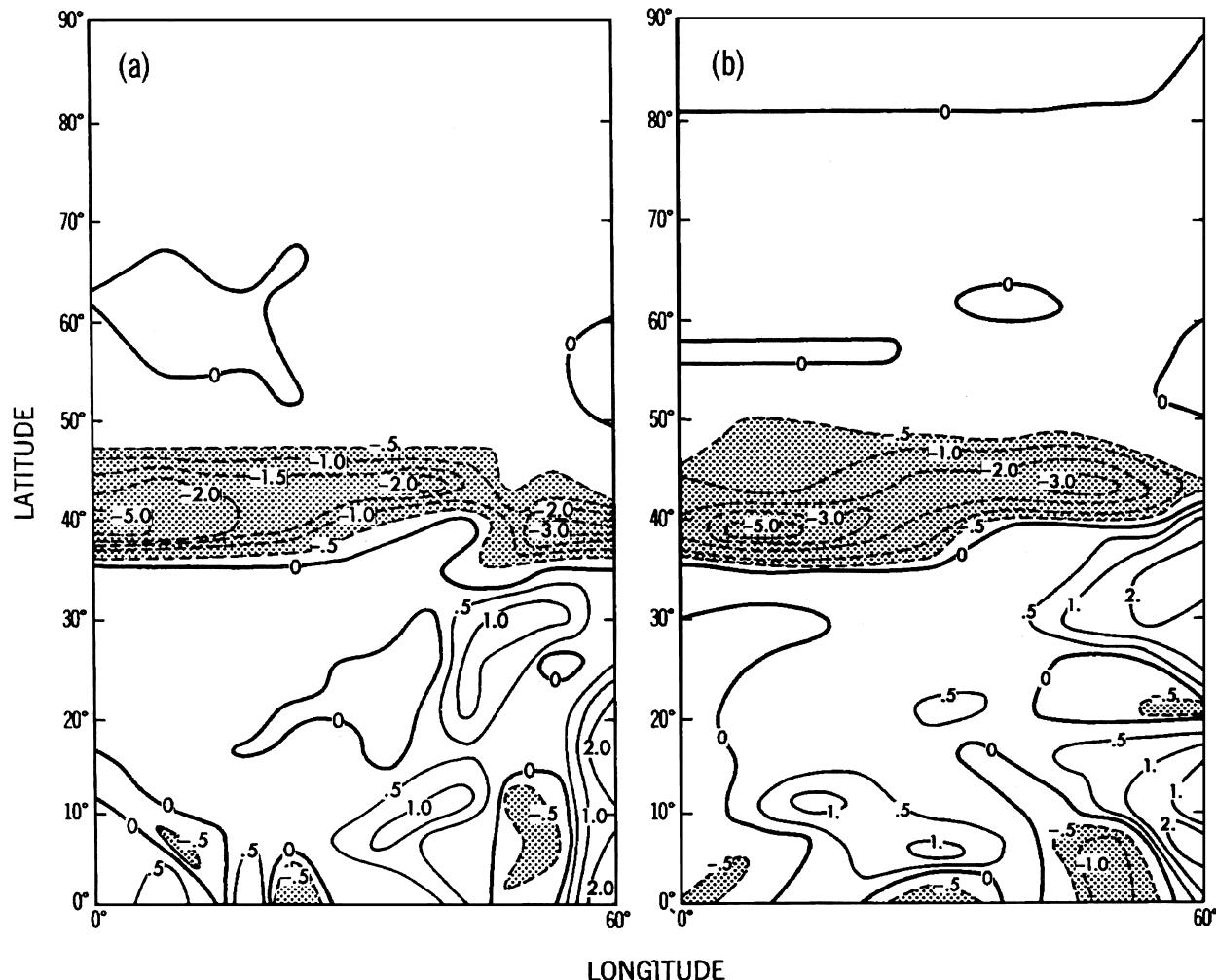


FIGURE 9 Geographical distribution of soil moisture response over the idealized continent to the (a) doubling of CO_2 and (b) quadrupling of CO_2 (cm/day). [From Manabe, S., and Wetherald, R. T. (1980). *J. Atmos. Sci.* **39**, 99–118.]

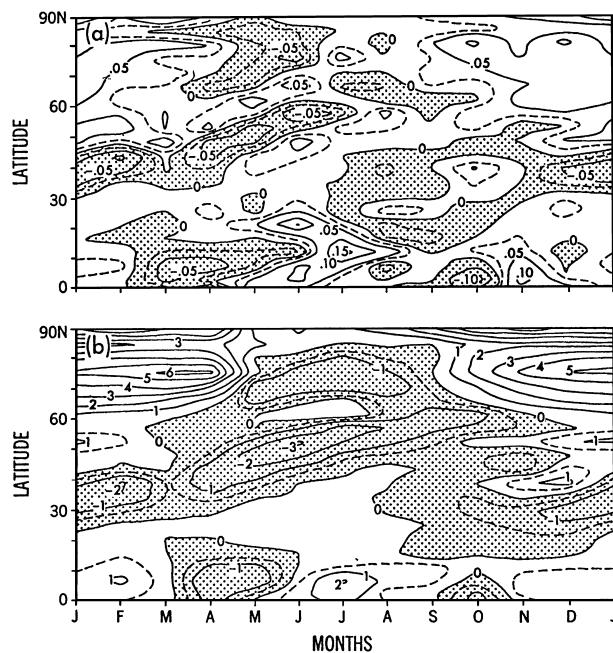


FIGURE 10 Latitude–time difference distribution over the idealized continent of zonal mean precipitation minus evaporation (a, cm/day) and soil moisture (b, cm) in response to a quadrupling of CO₂. [From Wetherald, R. T., and Manabe, S. (1981). *J. Geophys. Res.* **86**, 1194–1204. Reproduced by permission of American Geophysical Union.]

the winter but shifts poleward until it becomes centered at approximately 45° latitude during the summer season. In general, the magnitude of this increased dryness over the continent is greatest in middle latitudes during summer. This summer dryness pattern was found to be caused by two factors.

The first of these is an earlier disappearance of snow cover during the late winter, which causes an earlier beginning of relatively large evaporation from the soil. Because snow cover reflects a large fraction of solar radiation, its disappearance increases the absorption of solar energy by the land surface that is used as latent heat for evaporation. Thus, the end of the spring snowmelt period marks the beginning of the summer drying out of the soil. In the high-CO₂ experiment, the period of snowmelt ends earlier, bringing an earlier start of the spring to summer reduction of soil moisture.

The second factor involves changes in the middle-latitude precipitation pattern caused by a poleward shift of the middle-latitude rainbelt, a region associated with large-scale cyclonic disturbances. In the high-CO₂ atmosphere, warm, moisture-rich air penetrates further north than in the normal-CO₂ atmosphere. This is caused by a greater transport of moisture from lower to higher latitudes. Thus, precipitation increases significantly in the northern half of the rainbelt, whereas it decreases in the

southern half. Because the rainbelt moves northward from winter to summer, a middle-latitude location lies in the northern half of the rainbelt in winter and in its southern half in summer. Therefore, at middle latitudes, the CO₂-induced change of precipitation becomes negative in early summer, contributing to a reduction of soil moisture.

These two mechanisms are illustrated by Fig. 11, which shows the latitude–time distribution of the continental snow cover for both the normal- and high-CO₂ experiments and the latitude–time distribution of total precipitation amount for the normal-CO₂ experiment. The upper and middle portions of Fig. 11 indicate that, not only is there less snow depth, but snow cover is less extensive in the high-CO₂ case as compared with the normal-CO₂ case in middle latitudes. This implies that there is less snowmelt runoff during the spring season there. The snow cover also appears later in fall and disappears earlier in spring.

In the lower portion of Fig. 11, the mean position of the middle-latitude rainbelt for the high-CO₂ experiment (dashed line) is located poleward of its mean position in the normal-CO₂ experiment (solid line). Such a redistribution of the precipitation pattern results in wetter conditions to the north and dryer conditions to the south of the rainbelt in middle latitudes during the summer season.

The summer reduction of soil moisture does not continue throughout the winter season. In response to the increase of atmospheric CO₂, soil wetness increases during the winter season over extensive continental regions of middle and higher latitudes (the lower portion of Fig. 10). In middle latitudes, this increase of soil moisture is mainly due to the increase of precipitation in the northern half of the middle-latitude rainbelt. In general, total precipitation increases in both middle and higher latitudes during the winter season. Also, a larger fraction of the total precipitation occurs as rainfall rather than snowfall due to the warmer atmosphere. Both of these factors combine to cause the soil to become wetter. The lower portion of Fig. 10 also indicates that soil wetness is reduced during winter at 25 to 40° latitude. The reduced rainfall in the southern half of the middle-latitude rainbelt is, again, responsible for this region of enhanced dryness during the winter season in these lower latitudes.

Figure 12 shows the CO₂-induced seasonal variation of the various water-budget components centered approximately at 45° latitude. According to this illustration, both rainfall and evaporation increase during the winter months. However, as the spring season approaches, the CO₂-induced enhancement of rainfall decreases rapidly and actually changes sign (Fig. 12c). A similar tendency occurs for evaporation but it does not occur as quickly and never changes sign. During summer, there is a decrease of rainfall which continues until early fall. Evaporation

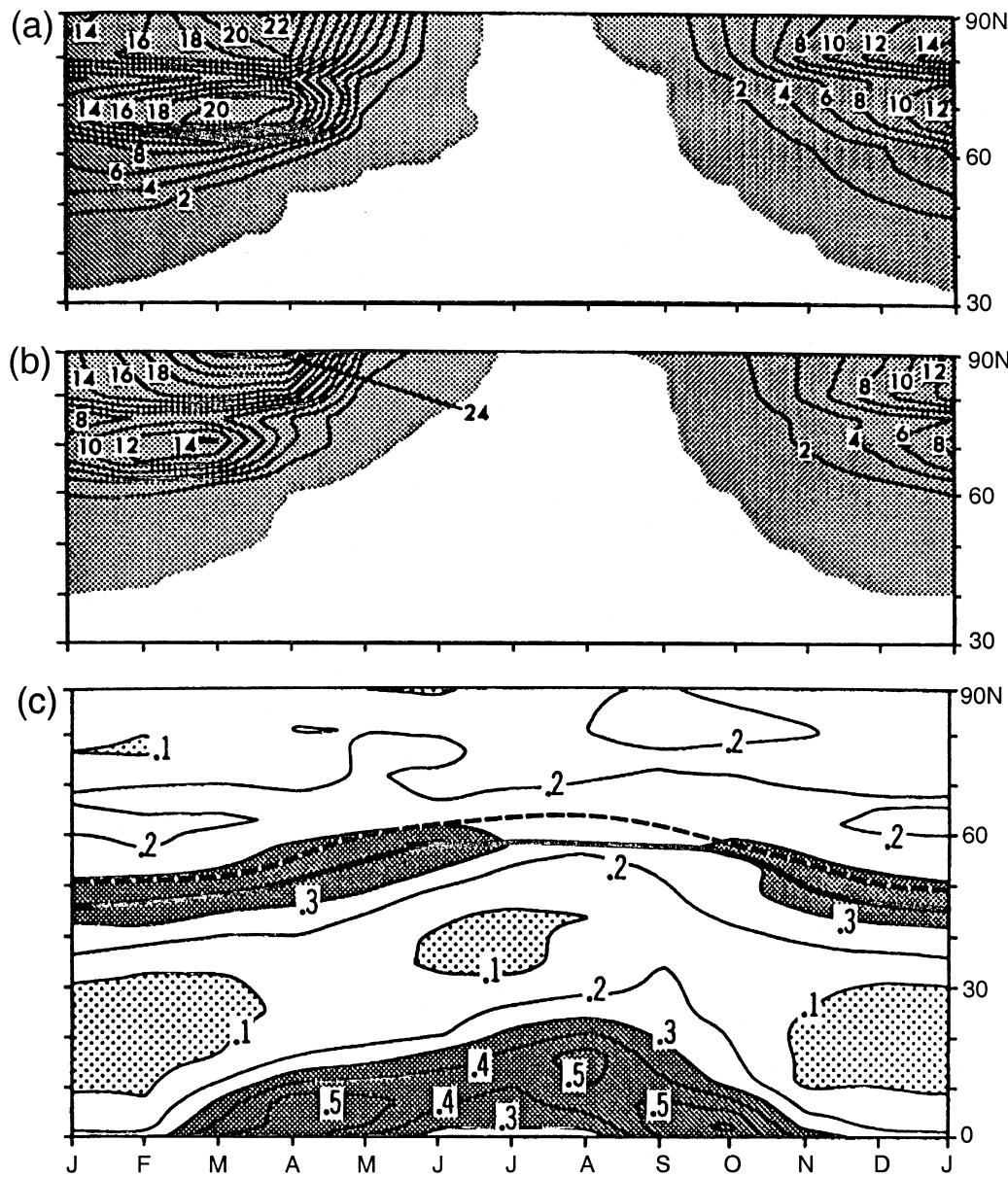


FIGURE 11 Latitude–time distributions of zonal mean snow depth for (a) the control experiment, (b) a quadrupling of CO₂, and (c) zonal mean precipitation rate (cm/day) for the control experiment over the idealized continent. The positions of the middle-latitude rainbelts are depicted by a solid line for the control experiment and a dashed line for the higher CO₂ experiment. [From Wetherald, R. T., and Manabe, S. (1981). *J. Geophys. Res.* **86**, 1194–1204; Manabe, S., Wetherald, R. T., and Stouffer, R. J. (1981). *Clim. Change* **3**, 347–386. Reproduced by permission of American Geophysical Union.]

during summer also finally begins to decrease because there is no longer enough soil moisture to evaporate at the higher rate. These seasonal changes in rainfall and evaporation are consistent with the summertime soil dryness (Fig. 13) due to the earlier removal of snow cover and the poleward shift of the middle-latitude rainbelt. Changes of snowmelt and runoff indicate earlier melting of snow cover and an earlier runoff period during spring.

IV. EXPERIMENTS WITH REALISTIC GEOGRAPHY

Parallel to the studies noted above, S. Manabe and R. J. Stouffer, in 1980, conducted an investigation with a GCM based upon the “spectral” method which incorporated realistic geography rather than the idealized geography used previously. Basically, the model is a variant of the earlier

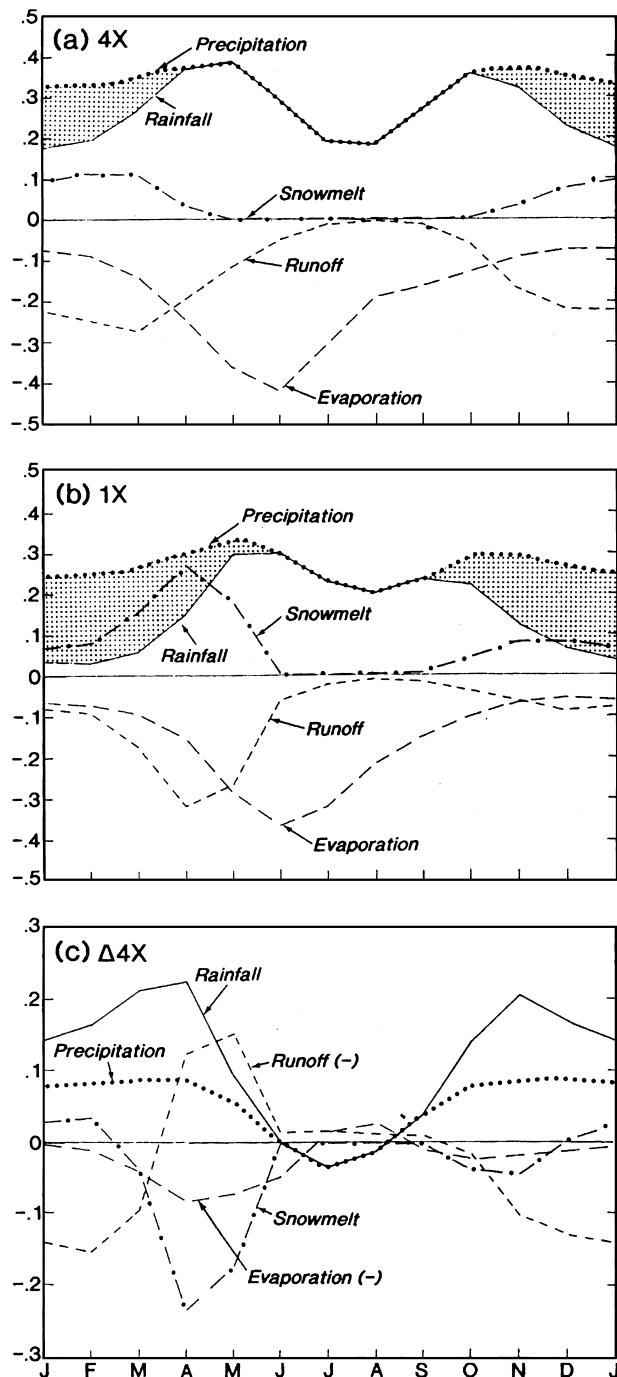


FIGURE 12 Seasonal variation of the components of the monthly mean surface water budget (cm/day) averaged over the continental region of 45–60° latitude for (a) a quadrupling of CO₂, (b) the control experiment, and (c) the difference between the two experiments. Here, both evaporation and runoff are plotted in the negative (−) because they represent losses of water from the model surface. Shaded regions in both (a) and (b) denote the portion of total precipitation (dotted lines) attributable to snowfall. [From Wetherald, R. T., and Manabe, S. (1995). *J. Clim.* **8**, 3096–3108.]

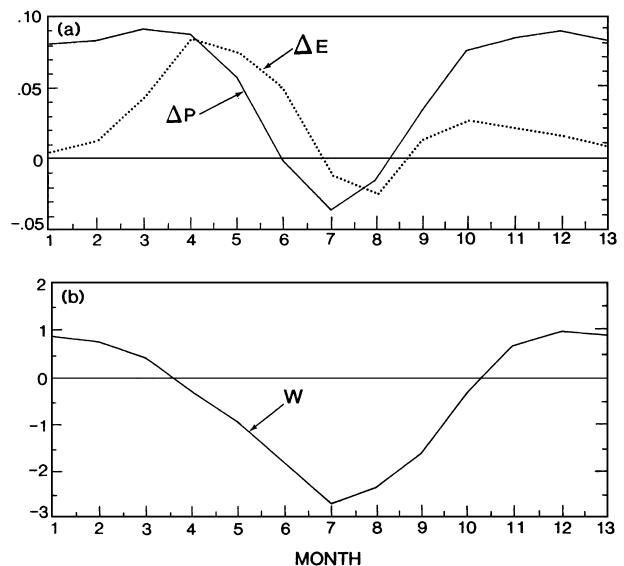


FIGURE 13 Seasonal variation of the monthly mean difference between the quadrupling of CO₂ and control experiments of (a) precipitation and evaporation (cm/day) and (b) soil moisture (cm) averaged over the continental region of 45 to 60° latitude. [From Wetherald, R. T., and Manabe, S. (1995). *J. Clim.* **8**, 3096–3108.]

ones except that it is a spectral model rhomboidally truncated at wave number 15 (R15) with a transform grid spacing of 7.5° longitude by 4.5° latitude. The “spectral” method is based upon the concept of performing linear operations in spectral space (expressing the horizontal distributions and their derivatives of the dynamic variables in terms of trigonometric and Associated Legendre functions) and performing nonlinear operations in “grid point” space (normal finite differences). All other features that were described previously are incorporated into the model such as a seasonal cycle of insolation, heat and water budgets over the land surface, “bucket” hydrology, mixed-layer ocean, etc. This model was, then, integrated in the same manner as described previously, with a control experiment and an experiment in which the CO₂ concentration was quadrupled.

The latitude–height distribution of the CO₂-induced change of zonal mean annually averaged temperature obtained from this model is shown in Fig. 14. In qualitative agreement with the results from both the radiative, convective model and the GCM with idealized geography (Fig. 1 and Fig. 6), the temperature of the model troposphere increases whereas that of the model stratosphere decreases in response to the quadrupling of CO₂. As was the case in Fig. 6, the increase of surface air temperature is particularly pronounced in the polar regions of higher latitudes where the poleward retreat of snow and ice cover enhances the warming due to the snow-, ice-albedo feedback

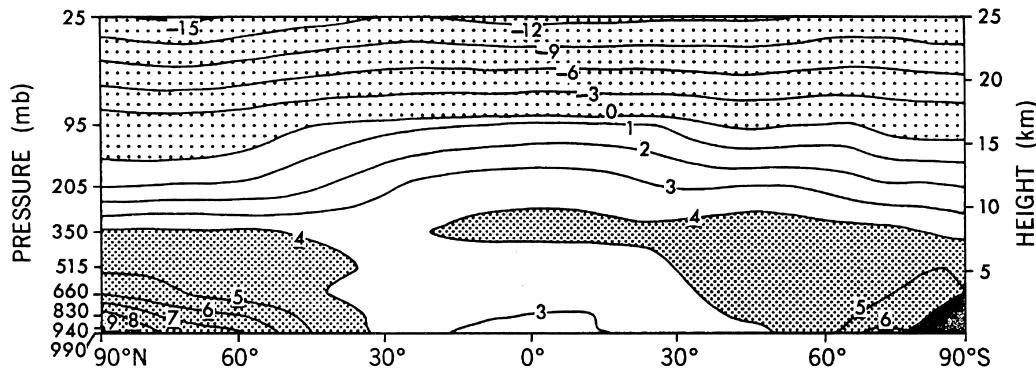


FIGURE 14 Latitude–height distribution of the zonal mean difference in annual mean temperature (K) in response to the quadrupling of CO₂. The Manabe–Stouffer model utilized realistic geography as compared with the earlier models with idealized geography. [From Manabe, S., and Stouffer, R. J. (1980). *J. Geophys. Res.* **85**, 5529–5554. Reproduced by permission of American Geophysical Union.]

process described previously. In low latitudes, the CO₂-induced warming is spread over the entire model troposphere due to the effect of moist convection which causes a greater warming in the model's upper troposphere than at the surface. This feature was also identified in the earlier GCM.

The corresponding latitude–time distribution of surface temperature change is shown in Fig. 15. This figure is

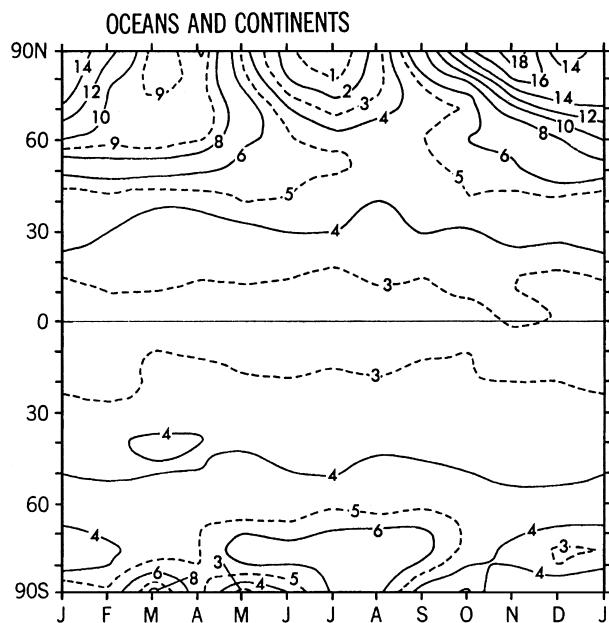


FIGURE 15 Latitude–time distribution of the zonal mean difference in surface air temperature (K) in response to the quadrupling of CO₂. Zonal averaging is taken over both oceans and continents. [From Manabe, S., and Stouffer, R. J. (1980). *J. Geophys. Res.* **85**, 5529–5554. Reproduced by permission of American Geophysical Union.]

constructed in much the same way as Figs. 10 and 11 except now the zonal averages are taken over the entire domain rather than just the continental regions. In low latitudes, the warming due to the quadrupling of CO₂ is quite small and is almost invariant with season whereas at high latitudes, it is much larger and varies markedly with season particularly in the Northern Hemisphere. In the vicinity of the Arctic Ocean, the warming is at a maximum in early winter and is at a minimum during summer. A similar pattern of response is also evident in the vicinity of the Antarctic Ocean but the amplitude of the variation is considerably smaller than it is in the north polar regions.

An analysis of the surface heat fluxes over both oceans indicates that the CO₂-induced reduction of sea ice is mainly responsible for the relatively large wintertime warming of surface temperature. In early winter, the upward conductive heat flux through the ice sheet in the warmer CO₂ experiment is considerably larger than the corresponding flux in the control experiment due to the reduction of sea-ice thickness. Because of this the corresponding flux of sensible heat from the ice surface to the model atmosphere is also larger in the higher CO₂ experiment than in the control experiment. The winter warming is further enhanced by a stable stratification which limits the heating to the lowest layer of the model atmosphere at higher latitudes. This feature was also discussed previously.

As Fig. 15 indicates, the magnitude of the warming during summer is much less than the corresponding warming seen during the winter. Because of the reduction of sea ice, the surface albedo decreases significantly from the control to the high-CO₂ experiment. However, the additional energy of solar and terrestrial radiation absorbed by the ocean surface is used mainly for heating the ice-free, mixed-layer ocean which has a relatively high heat

capacity. Therefore, the warming of surface air turns out to be fairly small during this season. It should be noted, however, that the additional energy absorbed by the ocean either delays the appearance of sea ice or reduces its thickness, thereby increasing the conductive heat flux during early winter when the difference between the water and air temperature becomes relatively large.

Over the continents, the situation is considerably different with regard to both mechanism and timing. As was noted in the previous study, the snow-albedo feedback process was quite important in creating a relatively large warming in higher latitudes over continental regions. However, because of the incorporation of seasonal variation into the current study, the recession of snow cover results in a substantial surface warming only during the spring snowmelt season when the incident solar flux is relatively high rather than in fall or early winter when it is near a minimum. Furthermore, because of the particular distribution of the continents in northern higher latitudes, the increase of conductive heat flux through thinner ice during early winter also affects the continental regions there resulting in two maxima of surface temperature increase, one in higher latitudes during early winter and the other at more middle latitudes due to the recession of highly reflective snow cover during early spring.

The discussion of the Manabe–Stouffer investigation indicates that the mechanism involved with CO₂-induced warming over the continents is significantly different from that causing the corresponding warming over the oceans in high latitudes. Over the continents, the snow cover-albedo feedback mechanism dominates whereas over ice-covered oceans, changes of conductive heat fluxes are the most important. This is because sea ice not only reflects a large fraction of solar radiation, but also inhibits the heat exchange between the atmosphere and the underlying ocean surface. Therefore, the CO₂-induced warming over the north polar ocean is maximum during early winter whereas the maximum warming takes place over the continents during the spring snowmelt season.

The corresponding changes of hydrology are qualitatively similar to those described in the model with idealized geography and seasonal variation of solar radiation. As noted previously, the CO₂-induced change in the hydrologic cycle has a significant seasonal variation. Figures 16 and 17 illustrate the time–latitude changes of both precipitation minus evaporation and soil moisture, respectively. According to Fig. 17, the difference of zonal mean soil moisture in high latitudes of the model has a large positive value throughout most of the year with the exception of the summer season. As discussed previously, this increase in high latitudes is caused by the penetration of warm, moisture-rich air into high latitudes of the model. Figure 17 also indicates two zones of reduced soil

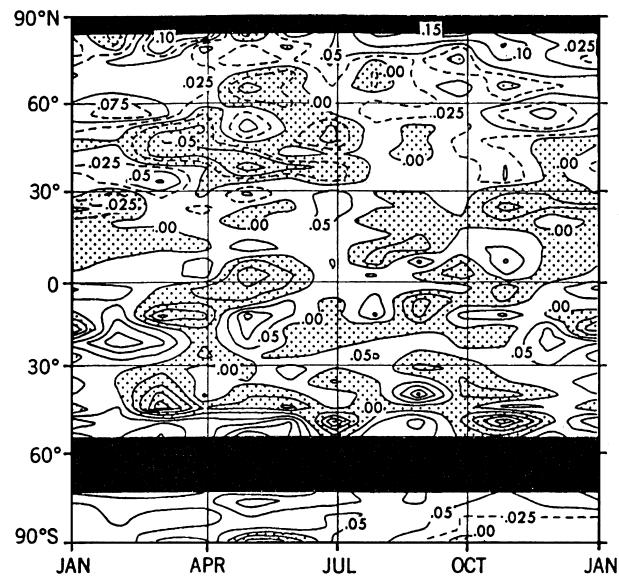


FIGURE 16 Latitude–time distribution of the zonal mean difference in precipitation minus evaporation ($P - E$) (cm/day) in response to the quadrupling of CO₂. Zonal averages are taken over the continents only. [From Manabe, S., and Stouffer, R. J. (1980). *J. Geophys. Res.* **85**, 5529–5554. Reproduced by permission of American Geophysical Union.]

wetness at middle and high latitudes during the summer season. Qualitatively similar results were obtained by the earlier study with idealized geography. The mechanisms responsible for these changes were analyzed and found to be similar to those already identified; namely, an earlier

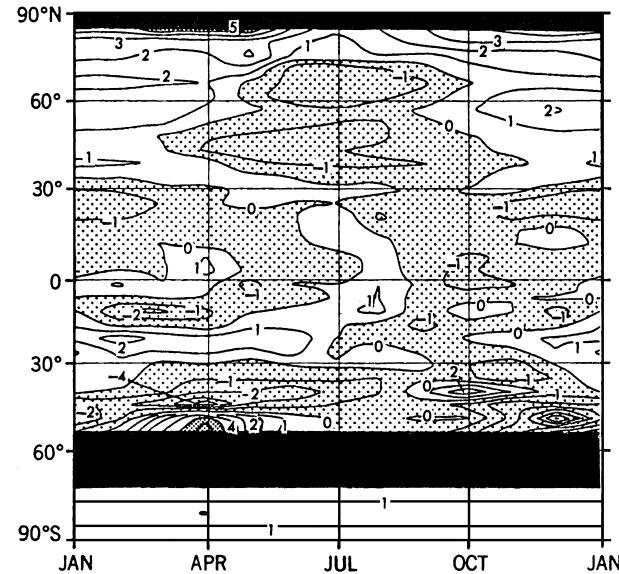


FIGURE 17 Latitude–time distribution of the zonal mean difference in soil moisture (cm) in response to the quadrupling of CO₂. Zonal averages are taken over the continents only. [From Manabe, S., and Stouffer, R. J. (1980). *J. Geophys. Res.* **85**, 5529–5554. Reproduced by permission of American Geophysical Union.]

ending of the snowmelt season which initiates the period of increased evaporation from the soil surface, an earlier beginning of the spring-to-summer reduction of the precipitation rate, an earlier reduction of baroclinicity (storminess), and a poleward shift of the middle-latitude rainbelt. Although the middle-latitude rainbelt is more difficult to identify in the model with realistic geography, the maximum belt of precipitation in middle latitudes was found to shift poleward in the high-CO₂ experiment thus contributing to a reduction of soil wetness both in middle latitudes during summer and in the subtropics during winter. In summary, evaporation is found to increase while precipitation decreases from late spring to summer (Fig. 16), thereby creating dryer conditions in midlatitude continental regions throughout the entire period from late spring to early fall.

Up until this point, the clouds were assumed to be prescribed and invariant with time. The next stage in our GCM development was to incorporate a simple cloud prediction scheme into the Manabe–Stouffer model. The CO₂ forcing in this case was taken to be a doubling rather than a quadrupling of CO₂. This scheme basically consisted of placing cloud cover wherever the relative humidity exceeded a certain critical value (taken to be 99% in this model). In general, the patterns of zonal mean temperature and soil moisture differences obtained from this later version of the model (Fig. 18) are qualitatively similar to those already shown from the earlier model (Figs. 15, 17). The mechanisms responsible for the temperature and soil moisture changes for this study are identical to the processes described previously. However, it is seen that the amplitudes of the changes in response to a doubling of CO₂ are comparable to those noted for the earlier study with prescribed clouds in response to a quadrupling of CO₂. This implies that the model utilizing predicted clouds is more sensitive than the model with fixed or prescribed clouds. An analysis of the CO₂-induced change of cloud cover and relative humidity (Fig. 19) revealed that the following two features of the CO₂-induced change of cloud cover were responsible for this increase in sensitivity: (1) low and upper tropospheric relative humidity and cloud cover reduce in low and middle latitudes, thereby decreasing the amount of highly reflective surface there, and (2) relative humidity and cloud cover increase in the lower stratosphere at all latitudes which reduces the loss of outgoing longwave radiation to space without significantly increasing the reflective surface due to the relatively low reflectivity assigned to high, thin clouds. Both of these changes act to create a positive cloud feedback process to the model atmospheric system in response to increases of CO₂. These two positive cloud feedback processes are offset somewhat by increases of low clouds in high latitudes but since these regions receive a relatively small amount of insola-

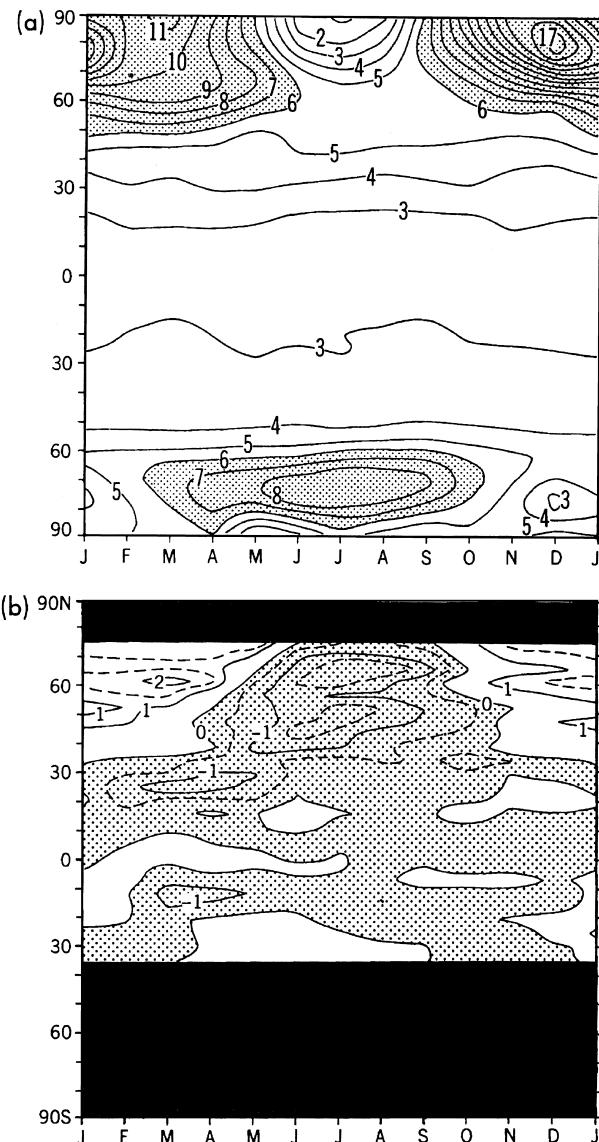


FIGURE 18 Latitude–time distribution of the zonal mean difference in (a) surface air temperature (K) over both continents and oceans and (b) soil moisture (cm) over the continents only in response to the doubling of CO₂. This model utilized cloud prediction as well as realistic geography. [From Manabe, S., and Wetherald, R. T. (1987). *J. Atmos. Sci.* **44**, 1211–1235.]

tion over a limited area as compared with the rest of the globe, the overall effect of this cloud increase is minimal. Therefore, the corresponding increase of global mean surface air temperature in response to a doubling of CO₂ is 2.0°C for the prescribed cloud model and 4.0°C for the predicted cloud model, respectively.

To illustrate the geographical distributions of climate obtained from this model, Figs. 20 and 21 are presented. The geographic distribution of CO₂-induced surface air temperature change for the December–February period

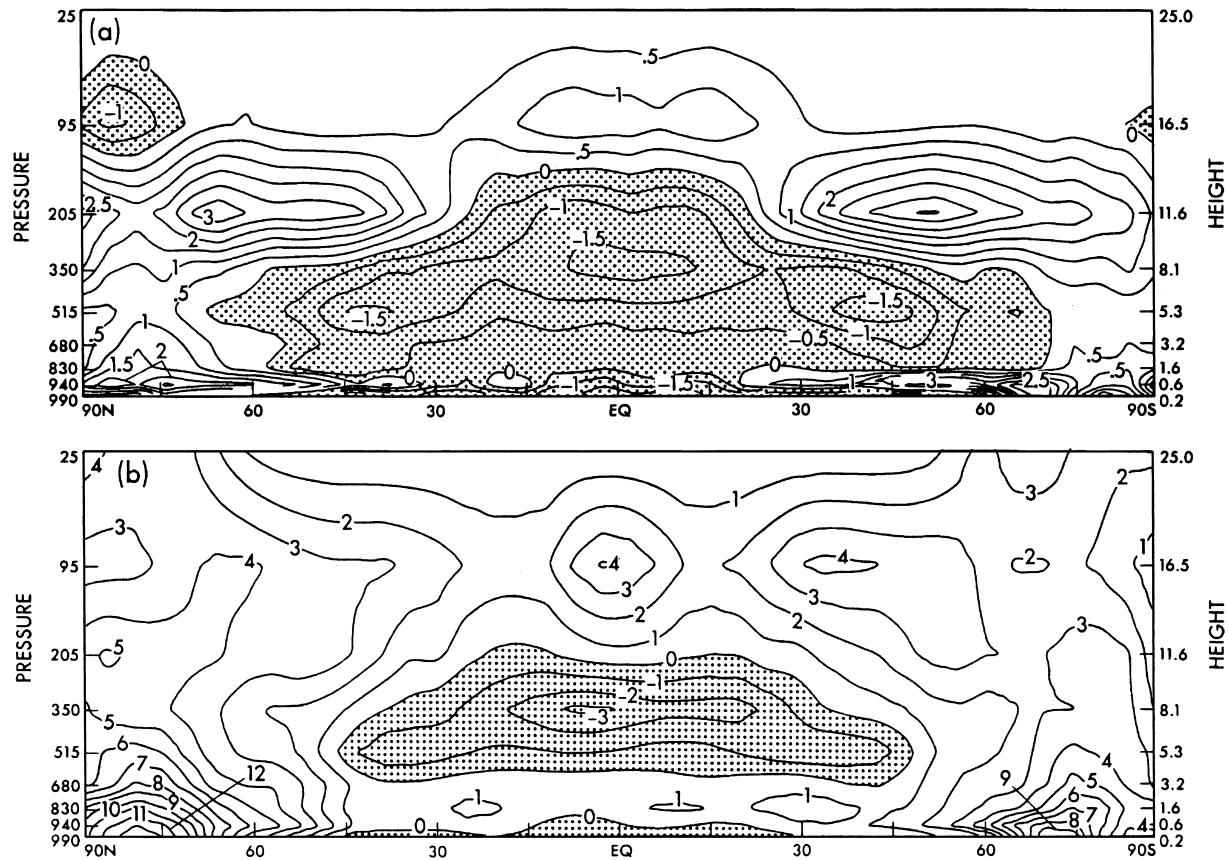


FIGURE 19 Latitude–height distribution of the zonal mean difference of (a) total cloud amount and (b) relative humidity (percent) over both continents and oceans in response to a doubling of CO₂. [From Wetherald, R. T., and Manabe, S. (1988). *J. Atmos. Sci.* **45**, 1397–1415.]

(Fig. 20, a) shows a relatively large response in middle to high latitudes which is not present in the June–August period (Fig. 20, b). This is due to the recession of continental snow cover and sea ice during the winter and spring seasons, mentioned previously, whereas these processes are relatively inactive during the summer season. As was also previously shown, temperature changes in the tropical and subtropical latitudes are smaller and practically invariant with season.

The corresponding changes of soil moisture are shown in Fig. 21. According to this figure, there is a general increase of soil moisture during December–February for most middle- and high-latitude regions and a decrease of soil moisture for the southern portion of the North American continent and Asia (Fig. 21, a). On the other hand, there was a general decrease of soil moisture for the June–August period for the entire continents of North America and Asia (Fig. 21, b). The magnitude of the summer dryness is particularly pronounced over the Great Plains of the United States. The only exception to this overall summer dryness pattern is an increase of soil moisture over India, which implies an increase of monsoonal rainfall there.

As was previously noted, the decreases of soil moisture are caused by an earlier disappearance of continental snow cover, increased evaporation, and changes of precipitation patterns associated with the middle-latitude rainbelt.

It should be noted that, for the summertime distribution of surface air temperature change, the maximum temperature increase is centered over the upper United States. This maximum is directly associated with the region of maximum soil moisture decrease described above and is, in fact, caused by this region of increased dryness. As the soil moisture is depleted during summer, it eventually becomes too dry to support further increases of evaporation in a warmer CO₂ simulation. An analysis of the heat-budget components over this region reveals that evaporation (and therefore, latent heat) actually decreases during the latter portion of the summer season. Therefore, a greater amount of the available energy at the model surface is realized as sensible rather than latent heat. Since sensible heat is a less efficient means of ventilating the surface as compared with latent heat, the surface temperature goes up accordingly as the bottom portion of Fig. 20 indicates. A reduction of low and upper tropospheric cloud cover

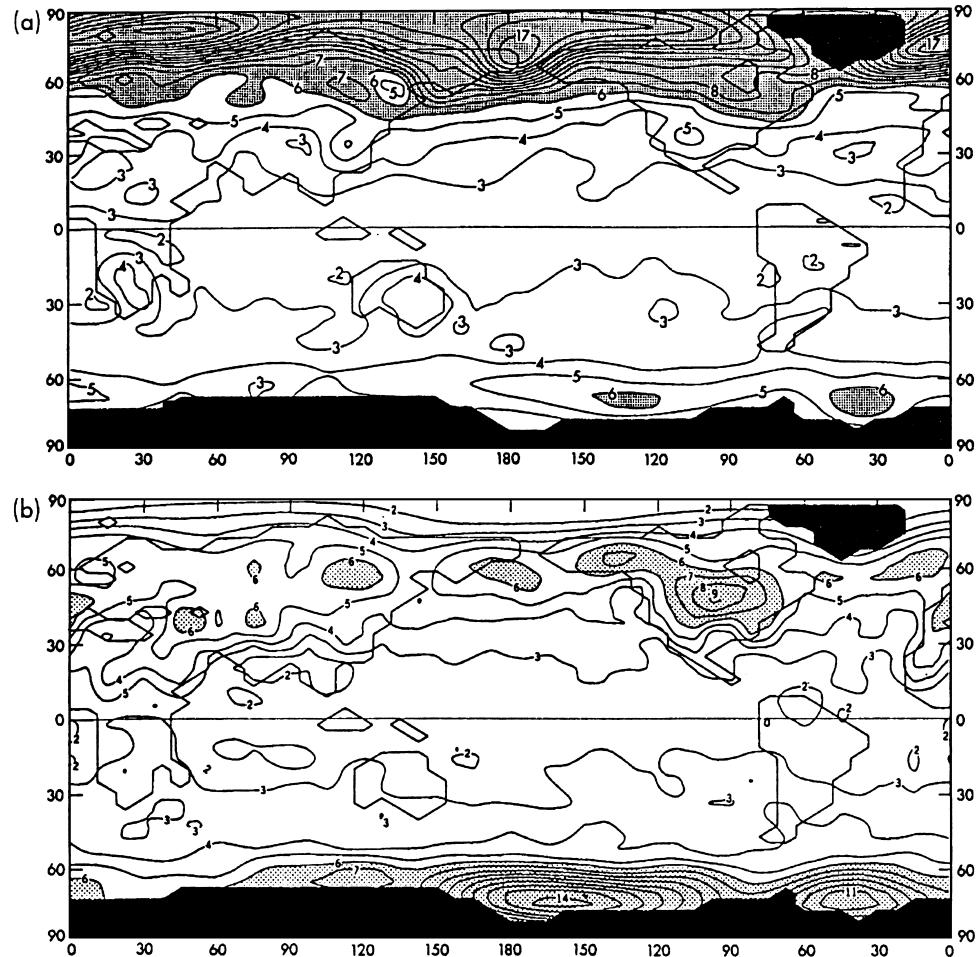


FIGURE 20 Geographical distribution of surface air temperature difference (K) during (a) December–February and (b) June–August in response to the doubling of CO₂. [From Manabe, S., and Wetherald, R. T. (1987). *J. Atmos. Sci.* 44, 1211–1235.]

(positive cloud feedback) was also found to exacerbate the summer dryness and resulting heating in this region during the summer season.

V. COMPARISON WITH OTHER MODELING GROUPS

In the 1980s to 1990s, GCMs by institutions other than the Geophysical Fluid Dynamics Laboratory (GFDL) were being used and analyzed for climate sensitivity investigations. These institutions included the Goddard Institute for Space Studies (GISS), the National Center for Atmospheric Research (NCAR), the United Kingdom Meteorological Office (UKMO), Oregon State University (OSU), Bureau Meteorology Research Centre (BMRC), the Canadian Climate Centre (CCC), and the European Centre for Medium-Range Weather Forecasts (ECMWF).

In 1987, a detailed investigation was made by M. E. Schlesinger and J. F. B. Mitchell of the results obtained from three of these institutions: GFDL, GISS, and NCAR. In addition to substantial regional differences of temperature and soil moisture, this comparison revealed that there was not universal agreement among the three models on the issue of middle-latitude continental summer dryness. A sample comparison is given in Fig. 22 which shows the latitude–time differences of soil moisture induced by a doubling of CO₂ for the three GCMs. According to this comparison, neither the GISS or the NCAR models produced a significant summer dryness pattern similar to that obtained by the GFDL model, whereas all three models yielded a tendency to produce wetter soil conditions during winter and early spring. All three models produce more consistent CO₂-induced hydrologic changes during the winter season than they do during summer.

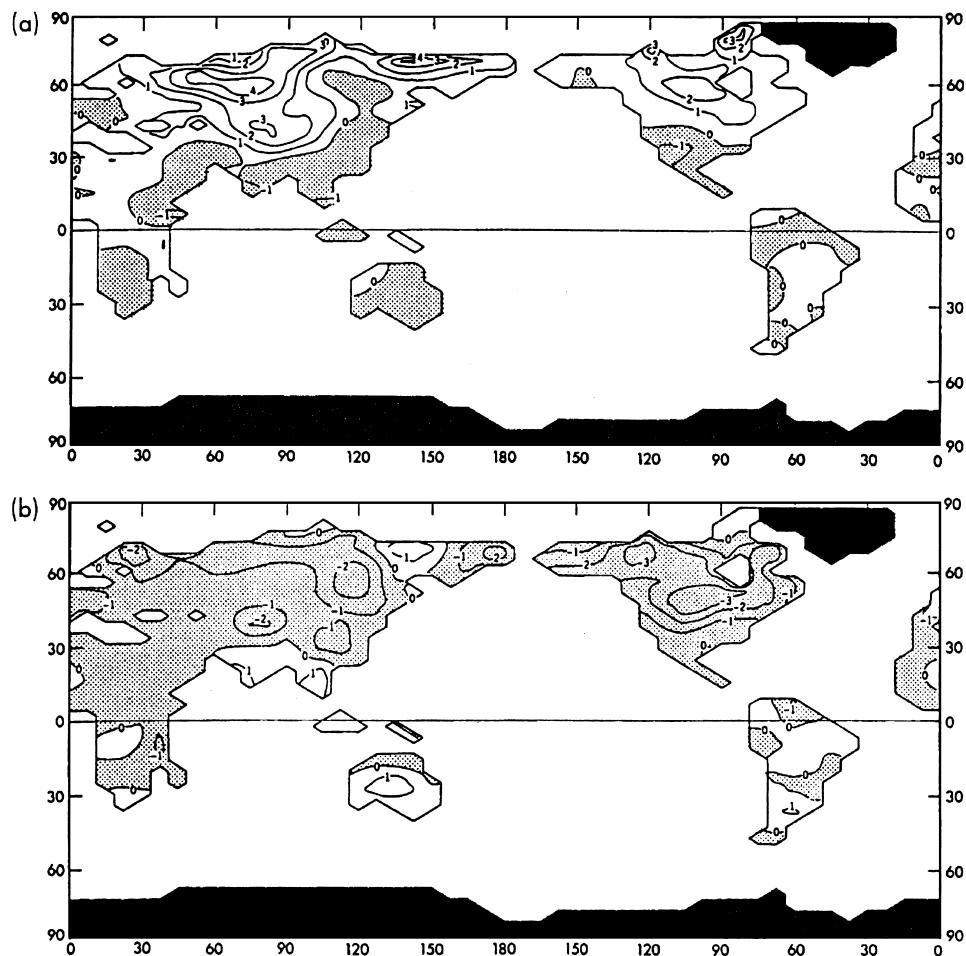


FIGURE 21 Geographical distribution of soil moisture difference (cm) during (a) December–February and (b) June–August in response to the doubling of CO₂. [From Manabe, S., and Wetherald, R. T. (1987). *J. Atmos. Sci.* **44**, 1211–1235.]

However, more recent GCM investigations appear to produce results that are consistent with the GFDL summertime scenario. These include models run at the UKMO, CCC, and BMRC, where substantial summer dryness patterns were obtained over North America, southern Europe, and Asia. These results are described in detail in the 1990 report by the Intergovernmental Panel for Climate Change (IPCC). These same models also produced, in varying degrees, the increased monsoonal conditions over India that were evident in the GFDL model.

With the aid of the researchers involved (G. A. Meehl, W. M. Washington, and D. Rind), an attempt was made to discover the reasons for the apparent discrepancy regarding the GISS and NCAR models. After an extensive analysis of the seasonal changes of hydrology, they determined that the amount of moisture in the soil moisture “buckets” in the control integration at the beginning of the summer season was too low to allow much further depletion to occur in either the GISS or NCAR models.

If this is the case, there exists the distinct possibility that all three general circulation models would have produced dryer soil conditions in the middle-latitude continental regions during the summer season provided enough moisture had been present in the soil to allow this to occur. A definitive conclusion on this issue must await future modeling studies that incorporate more realistic representations of surface hydrologic processes.

VI. COUPLED AIR-SEA MODELS

In the 1990s, the advent of main-frame supercomputers with greater speed and memory size allowed the development and integration of so-called “coupled air-sea models” (general circulation models which coupled together dynamic models of both the atmosphere and ocean). In our case, this involved the coupling of the previously described R15 GCM with a dynamical model of the world

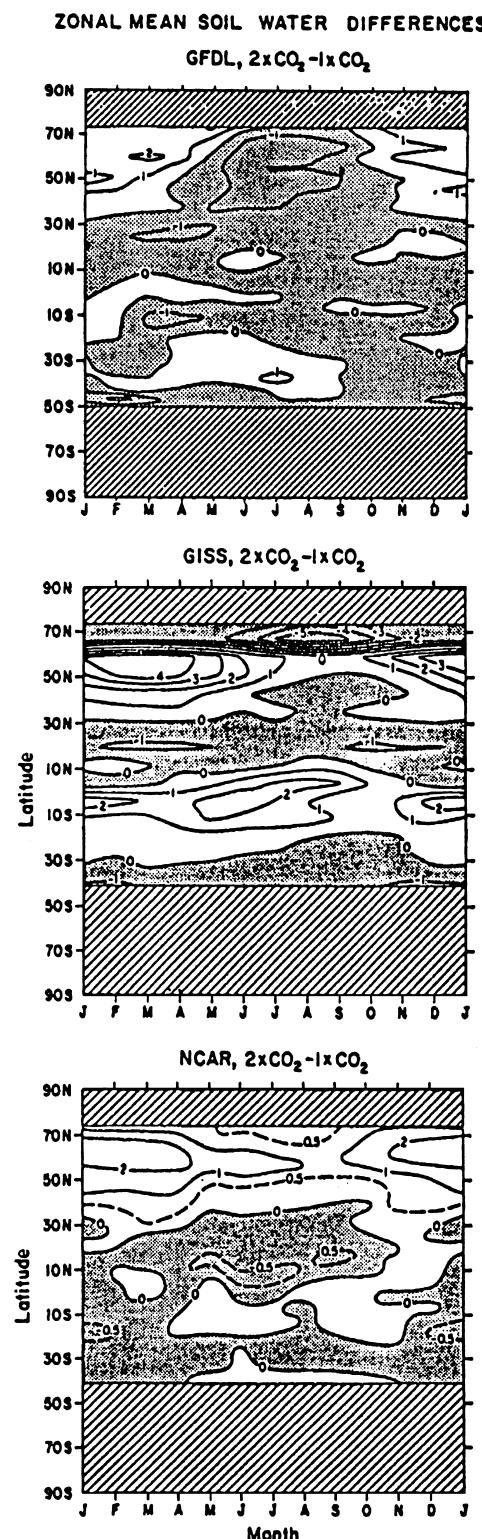


FIGURE 22 Latitude–time differences of zonal mean soil moisture (cm) as obtained from the GFDL model (upper), the GISS model (middle), and the NCAR model (lower) in response to a doubling of CO₂. [From Schlesinger, M. E., and Mitchell, J. F. B. (1987). *Rev. Geophys.* **25**, 760–798. Reproduced by permission of American Geophysical Union.]

oceans which computes explicitly the three-dimensional structure of the ocean currents, large scale eddies (gyres), and the corresponding heat transports resulting from these motions. Changes in salinity (saltiness) and sea ice are also included in the ocean model. To investigate the evolution of climate with respect to greenhouse warming, we incorporated a scheme devised by J. F. B. Mitchell and his colleagues of the United Kingdom Meteorological Office which included a past, present, and future estimation of both greenhouse gases and sulfate aerosols. Here, the equivalent carbon dioxide concentration (a combination of both CO₂ and other trace greenhouse gases, except water vapor) is prescribed for years 1765 to 1990 followed by a 1% per year increase compounded for years 1990 to 2065. The direct radiative forcing of sulfate aerosols is simulated by simply adjusting the surface reflectance of solar radiation without performing the radiative computations on the three-dimensional distributions of aerosols directly. The actual aerosol distribution for each model year is determined from separate loading patterns for years 1986 and 2050 following the scenario IS92a of the report IPCC-1992. It is interpolated from a zero loading pattern at 1765 to the loading pattern at 1986, linearly interpolated between the two loading patterns for years 1986 to 2050, then extrapolated from the loading pattern of 2050 to 2065. The same procedure for integrating the coupled air-sea model that was outlined at the beginning of this discussion was employed here, namely,

- A 1000-year control integration where the equivalent CO₂ concentrations of greenhouse gases (except water vapor) and sulfate aerosols are held fixed at year 1765 levels
- An integration in which both the equivalent carbon dioxide and sulfate aerosol forcings stated above are specified for years 1765 to 2065.

As implied above, there are large uncertainties in the nature and magnitude of the thermal forcing of sulfate aerosols as well as ignorance of other anthropogenic and natural factors which may have forced climate over the past 200 years. Nevertheless, the present model, which includes in a crude parameterization both sulfate aerosols and greenhouse gases, simulates the warming trend during this century quite well as shown in Fig. 23. Although this is not conclusive proof that the model will accurately forecast detailed changes of climate due to greenhouse warming, it can reasonably serve as a possible indicator on what type of large-scale changes might be expected to occur in the future.

To illustrate to the reader the nature of some of these changes, the geographical distributions of annual mean surface air temperature difference at years 2000 and 2050 are presented in Fig. 24. An inspection of Fig. 24 reveals

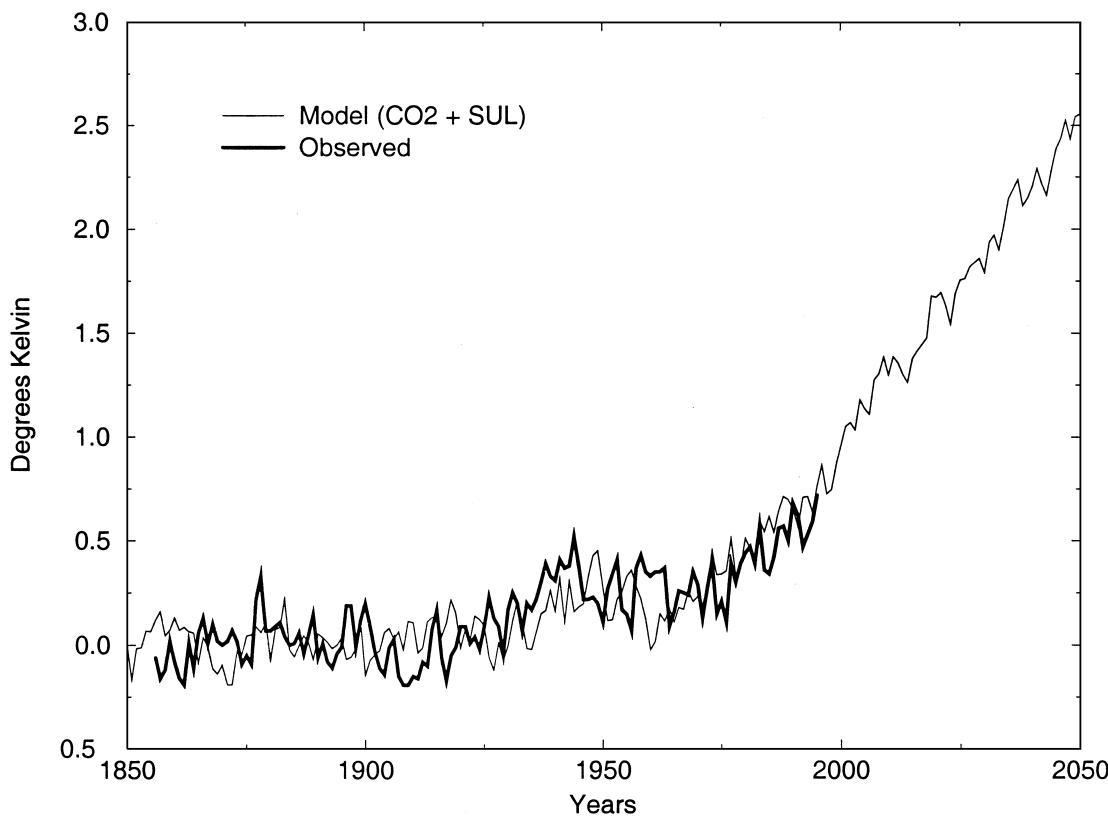


FIGURE 23 Time series of globally averaged annual mean surface air temperature anomaly (K) for the GFDL R15 coupled ocean-atmosphere model (TAOGCM) utilizing both greenhouse gas and sulfate aerosol radiative forcings (thin line) and corresponding observed anomaly defined as the deviation from the 1880–1920 average (thick line) (Jones and Briffa, 1992). [From Haywood, J. M., Stouffer, R. J., Wetherald, R. T., and Ramaswamy, V. (1997). *Geophys. Res. Lett.* **24**, 1335–1338. Reproduced by permission of American Geophysical Union.]

that, for a fully coupled air-sea model, the northern polar regions respond approximately as expected from the earlier studies but there is significantly less change in the southern polar regions in response to the increase of greenhouse gases. More specifically, this relatively slow response over the Southern Ocean and northeast portion of the Atlantic Ocean is due to the thermal inertia of the ocean model which has a much deeper mixed-layer (and hence a much larger heat capacity) than the 50-m oceanic mixed-layer that was assumed in the earlier studies. This feature has been noted not only in all of the recent GFDL investigations but also in many other institutions around the world where investigations have been carried out with the use of coupled air-sea GCMs. This result implies that the Antarctic region (and the accompanying ice sheet) will not experience a significant increase in surface air temperature for a very long time. It is evident that the response of surface air temperature is considerably greater at year 2050 than it is at year 2000 although both distributions indicate responses that are considerably less than the responses would have

been if the atmospheric-ocean system had been allowed to reach equilibrium with their respective radiative forcings due to the large thermal inertia of the oceans. This issue will be briefly dealt with in the succeeding discussion.

With regard to the geographical distributions of hydrologic changes, these are similar in nature to those already illustrated. In particular, both the United States and southern Europe have a tendency to experience longer, hotter, and dryer summers as time goes on. Winters, on the other hand, are estimated to be shorter and wetter with less percentage of total precipitation realized as snowfall. The mechanisms responsible for these hydrologic changes are identical to those already given for the mixed-layer GCMs.

VII. COMMITTED WARMING

As a final topic, it is worthwhile to briefly describe our latest research efforts on the issue of committed warming. Basically, this term is defined as the difference between

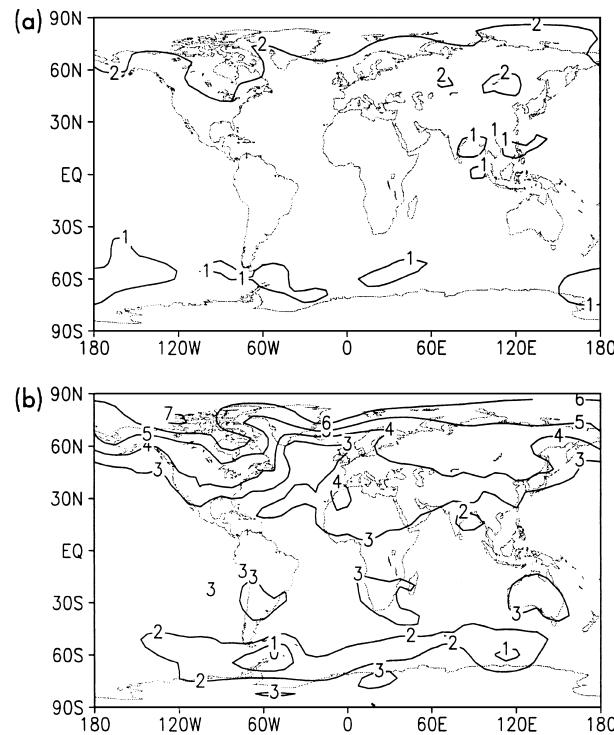


FIGURE 24 Geographical distribution of the annual mean surface air temperature response (K) as simulated in the TAOGCM experiment at (a) year 2000 and (b) year 2050. [From Wetherald, R. T., Stouffer, R. J., and Dixon, K. W. (2001). *Geophys. Res. Lett.* **8**, 1535–1538. Reproduced by permission of American Geophysical Union.]

the realized warming at a given time and the warming of climate that would occur if the climate system had an infinitely long time to adjust to that particular radiative forcing (i.e., the gap between the equilibrium and realized temperature change for a given forcing). Two main factors that determine the magnitude of the committed warming are the amount of oceanic heat uptake (heat given off by the ocean to the atmosphere) and the climate sensitivity. The efficiency with which the deep ocean mixes with the upper ocean affects oceanic heat uptake. In a coupled model having a dynamical ocean, ocean heat uptake is influenced by ocean circulation patterns, the static stability of the ocean, the ocean model's subgrid scale mixing parameterizations, and other factors. A model's sensitivity depends upon the nature of the various feedback processes present in the model and upon the model's state before carbon dioxide or other greenhouse gases are increased.

In this particular investigation, both a coupled air-sea model (AOGCM) and a mixed-layer (ML model) are used. In doing this, we are building upon work done previously by J. Hansen, T. Wigley, and their co-workers. The main improvement here is that three-dimensional GCMs were

utilized rather than the simpler energy balance or one-dimensional models used in the earlier research efforts.

Three separate experiments were integrated for this study: a transient AOGCM (TAOGCM), a transient ML model (TML), and an equilibrium ML model (EML). The AOGCM and ML model are identical to the respective models described earlier. Both transient experiments were started at year 1765 and integrated until year 2065. The equilibrium model was integrated until climate equilibrium was reached with the radiative forcings prescribed at 1980, 2000, 2020, 2040, 2050, and 2060, respectively. The results of these experiments are summarized by Fig. 25 which illustrates the time evolution of global mean surface air temperature difference in response to the greenhouse gas and sulfate aerosol forcing for the three models. It is immediately seen that the temperature response for the TAOGCM is considerably less than that for either the TML or EML models. This is due to the relatively large thermal inertia of the ocean system particularly in the Antarctic Ocean and the northern Atlantic which was noted previously. On the other hand, the temperature response for both the TML and EML models are quite similar which implies that the results of the both the TML and EML models can be used as a proxy for the TAOGCM's equilibrium response on a real-time basis. This is because the TML responds almost as quickly as the EML model in its approach to climatic equilibrium. These results have implications for future climate change, both realized and unrealized. For example, according to Fig. 25, the “committed warming” (the difference between the EML and TAOGCM results) at year 2000 is approximately 1.0 K. This is larger than the observed warming of 0.6 K that has taken place since 1900 (Fig. 23). At the same time, the global mean surface air temperature responses of all the numerical experiments are considerably greater at year 2060 than they were at year 2000. As the climate warms in response to the increasing greenhouse gases, the committed warming increases from about 1.0 K at year 2000 to nearly 2.0 K at year 2060. The realized warming increases from 0.6 K at year 2000 to 3.0 K at year 2060. In other words, the greater the greenhouse gas forcing, the greater the warming commitment.

There are large uncertainties in the magnitude of the radiative forcing of sulfate aerosols as well as other forcings which were not included in this study. However, it should be noted that when the TAOGCM is forced with both historical estimates of the sulfate aerosols and greenhouse gas concentrations, the observed warming of the twentieth century is simulated quite well. Therefore, it is reasonable to speculate that the current study could provide a viable estimate of the magnitude of the committed global warming for both the present day and the future.

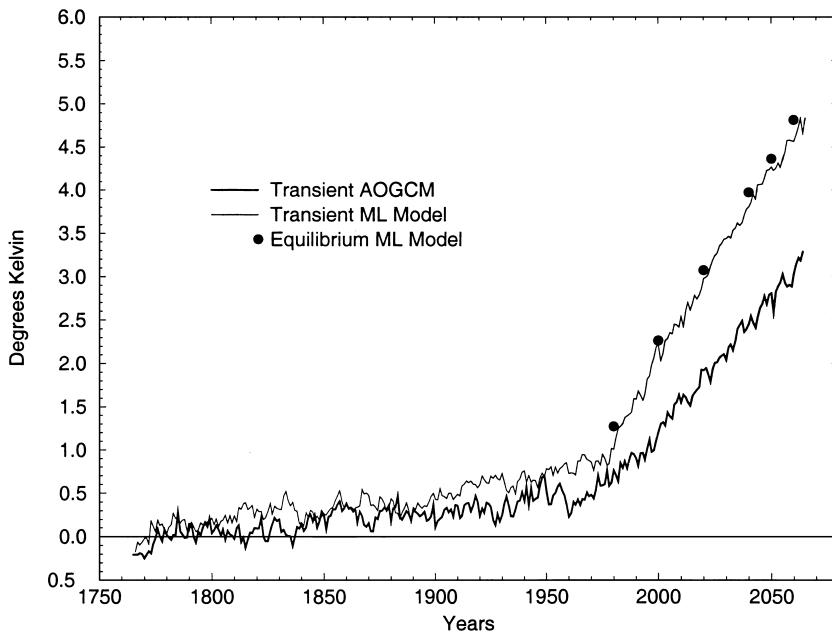


FIGURE 25 Model-simulated global mean surface air temperature anomalies (K). Thick and thin lines represent time series for the coupled TAOGCM and TML experiments, respectively. Large dots indicate EML model results. [From Wetherald, R. T., Stouffer, R. J., and Dixon, K. W. (2001). *Geophys. Res. Lett.* **28**, 1535–1538. Reproduced by permission of American Geophysical Union.]

VIII. SUMMARY

Although there are many areas of disagreement between the various models, it is worthwhile to highlight the areas of agreement. With regard to temperature, evaluation of the state-of-the-art GCMs (IPCC-2001) states that: (1) a projected increase of global surface air temperature due to a doubling of CO₂ lies in the range 1.4 to 5.8°C and (2) for all models, the increase of surface air temperature is much greater in higher latitudes than it is in the tropics. This polar amplification is greatest during the early winter and early spring seasons due to the conductive heat flux–sea ice and the snow cover–albedo feedback processes which operate mainly at these times, respectively. It should be noted that the projected temperature increase is greater than that estimated by earlier IPCC reports (IPCC-1990, IPCC-1992, IPCC-1995) due to the lower projected sulphur emissions in the IPCC-2001 scenarios relative to the IS92 scenarios rather than a change in more recent model results. Therefore, the rate of warming is estimated to be considerably larger than the observed historical changes during the twentieth century and is likely to exceed any warming which has occurred during the past 10,000 years, based upon paleoclimate data.

With regard to hydrology, a survey of the latest GCMs reveal that continental snow cover in midlatitudes is less extensive and shallower in depth for the higher CO₂ experiments. This implies that snow cover, in general, will appear later in fall and disappear earlier in spring and re-

sult in less spring runoff from snowmelt (although there will be greater runoff in the form of rainfall). Also, the soil surface will be exposed earlier in the winter season and, therefore, higher rates of evaporation will occur from it which will cause greater soil moisture loss from spring to summer. The same survey of model results also indicates a general consensus concerning the scenario of increased midlatitude continental dryness during the summer season although the geographical details of these regional patterns vary considerably from one model to another.

One of the largest uncertainties in climate sensitivity studies is the CO₂-induced response of precipitation over the continents during the period from early spring to late fall. Whether or not a given GCM will produce a summer dryness scenario appears to be dependent upon a poleward shift of the middle-latitude storm track (and accompanying rainbelt) and the state of the soil moisture of the control experiment for both early spring and summer. In the final analysis, a given GCM will produce a tendency for dryer summertime conditions if the projected rainfall is forecast to either decrease or remain approximately the same. Only if the rainfall is forecast to increase as much as the projected increase of evaporation will the desiccation of soil moisture be prevented.

Other significant uncertainties include modeling of cloud processes, treatment of aerosols, inclusion of active ocean currents, and the use of low horizontal resolution. For example, R. Cess and his colleagues have shown that

there are major differences among the various GCMs concerning the methods of cloud formation and their corresponding effects of cloud feedback. The method by which the effects of aerosols are incorporated into the model is another factor. Currently, relatively simple parameterizations are being used rather than explicitly computing the three-dimensional distributions of sulfate aerosols and their radiative effects. In addition, the manner in which ocean currents are explicitly included can significantly alter the transient or time-dependent phase of a climate sensitivity experiment. Until recently, the use of relatively large "grid boxes" has greatly hampered the successful simulation of climate particularly on a regional scale. However, the advent of larger and faster supercomputers is making it possible for modelers to repeat their experiments with a considerably higher computational resolution and, thus, achieve a correspondingly higher degree of credibility.

In any event, it appears certain that, if the earth's climate becomes warmer, the earth's hydrology will change. If the above theoretical hydrologic responses to greenhouse warming prove to be correct, they will have serious implications for water resource management and agricultural planning. Slow melting of snow cover is a much more efficient means for recharging the water table rather than increased rainfall which is more likely to simply run off and be lost to the watershed system. At the same time, increased dryness during the growing season could place severe demands on the available water supply which would require more irrigation. These factors could easily combine to cause serious water shortages at precisely the time when water is needed the most for growing crops.

Finally, it should be emphasized that the warming commitment discussion presented above describes the climate response for a constant radiative forcing of the earth. This requires greenhouse gas concentrations to remain constant over time at a particular level. The protocol reached by the Third Conference of Parties which met in Kyoto, Japan, in 1997, limits emissions to approximately present day (year 2000) levels. As shown by T. M. Wigley and his co-workers, this level of emissions does not provide a stable greenhouse gas concentration by 2100 or anything close to it. In fact, the CO₂ concentration continues to rise considerably beyond the doubling of the preindustrial values. Much more stringent controls on greenhouse gas emissions will be required to stabilize the greenhouse gas concentrations before 2100. Only after this stabilization occurs will the climate begin to come into equilibrium with that forcing, and only then will the total warming and associated climate changes be realized.

Based upon these observations and the information contained in the latest IPCC report, two main conclusions can be drawn.

1. Greenhouse warming and its attendant potential climate change is an issue that will not go away by simply trying to ignore it.
2. Greenhouse warming, once initiated, will set into motion climate changes which will persist for many centuries in the future due to the thermal inertia of the oceans and the atmospheric time scale of greenhouse gases.

We would, therefore, do well to seriously consider the possible options available to us while there is still time to act. These include energy and water conservation, recycling, development of alternate energy sources which do not involve fossil fuel burning (such as fusion, solar, geothermal, etc.), agricultural research into developing hardier crops that can withstand hotter and dryer conditions, and development of electric vehicles. The technology to accomplish many of these tasks is available today but it must be further encouraged by research and development programs. Such strategies would enable us to either take maximum advantage of the projected climate change or at least mitigate the adverse effects of that change. It should be kept in mind that we may well be embarking on the world's greatest "uncontrolled experiment" and if we do not act now, the results of this "experiment" could well become irreversible or, at the very least, persist for a long time to come.

IX. APPENDIX: ILLUSTRATION OF GCM PROCEDURE

The method of determining atmospheric motions can be demonstrated by a simple illustration. Suppose you have a transparent, thin box filled with air placed in the vertical position ([Fig. A1](#), upper left). Further suppose that this box is heated on the right side by a Bunsen burner and cooled on the left side by ice. This pattern of heating and cooling would then set up a "convection cell" rotating counterclockwise where the air on the right side rises due to the heating and sinks on the left side due to the cooling. Finally, assume that we are able to measure both the temperature and fluid motion (horizontal and vertical motion components) with appropriate sensors without disturbing the flow. We, then, are able to determine the temperature and flow conditions within the box by varying the number of measurements in a systematic fashion.

If we decide to take only one measurement in the exact middle of the box, we will obtain an extremely crude average of the conditions within the box. If, however, we divide the box up into two equal vertical portions and take our measurements in the middle of each separate portion (i.e., two sets of measurements), we obtain a slightly better approximation to the temperature and flow conditions

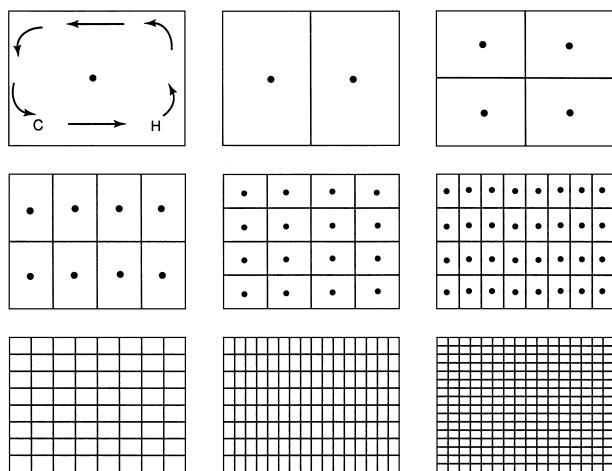


FIGURE A1 Diagram illustrating how a region may be divided into separate “boxes” to represent a fluid flow with increasingly higher resolution.

within the box (Fig. A1, upper middle). On the other hand, if we now divide the box into four equal parts (Fig. A1, upper right) we obtain an even better approximation to the flow inside the box. One can continue this procedure even further by dividing our box again into successive equal portions, eight (Fig. A1, middle left) and sixteen portions (Fig. A1, middle), respectively. By continuing this procedure, one may obtain an increasingly better approximation to the temperature and flow pattern within the box—the more subdivisions, the better the approximation (i.e., 32, 64, 128, 256 boxes, respectively; see rest of panels). This is, basically, how numerical weather forecasting is accomplished except instead of directly measuring the motions within the box by external instruments, they are calculated explicitly by certain mathematical equations which describe fluid flow. This method has been extended to actual weather forecasting except our sample box is replaced by a “grid network” of small “boxes” covering the entire earth and “stacked” vertically to provide three-dimensional coverage. For example, the horizontal coverage of the R15 (low resolution) and R30 (high resolution) spectral models currently used at GFDL consists of 1920 (48×40) and 7680 (96×80) grid boxes, respectively. The necessary calculations are, then, performed for each small box to determine the future atmospheric conditions as a function of both space and time for the entire globe.

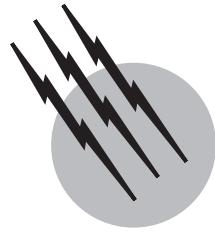
SEE ALSO THE FOLLOWING ARTICLES

CARBON CYCLE • CLIMATOLOGY • GAS HYDRATE IN THE OCEAN ENVIRONMENT • GREENHOUSE EFFECT AND

CLIMATE DATA • HEAT TRANSFER • HYDROLOGIC FORECASTING • METEOROLOGY, DYNAMIC • OCEAN-ATMOSPHERIC EXCHANGE • OZONE MEASUREMENTS AND TRENDS • RADIATION, ATMOSPHERIC

BIBLIOGRAPHY

- Cubasch, U., Meehl, G. A., Boer, G. J., Stouffer, R. J., Dix, M., Noda, A., Senior, C. A., Raper, S., and Yap, K. S. (2001). Projections of future climate change. “Climate Change 2001: The Scientific Basis,” IPCC Cambridge Univ. Press, Cambridge, UK.
- Hansen, J., Lacis, A., Rind, D., Russel, G., Stone, P., Fung, I., Ruedy, R., and Lerner, J. (1984). Climate sensitivity: Analysis of feedback mechanisms, Climate Processes and Climate Sensitivity. Amer. Geophys. Union. *Geophys. Mono.* **29**, 130–163.
- Harvey, D., Gregory, J., Hoffert, M., Jain, A., Lai, M., Leemans, R., Raper, S., Wigley, T., and de Wolf, J. (1997). An introduction to simple climate models used in the IPCC second assessment report. IPCC Technical Paper II, 47 pp. IPCC Cambridge Univ. Press, Cambridge, UK.
- Houghton, J. T. (1997). “Global Warming: The Complete Briefing,” 2nd ed. Cambridge Univ. Press, Cambridge, UK.
- Jones, P. D., Osborn, T. J., and Briffa, K. R. (1997). Estimating sampling errors in large-scale temperature averages. *J. Clim.* **10**, 2548–2568.
- Kattenburg, A., Giorgi, F., Grassl, H., Meehl, G. A., Mitchell, J. F. B., Stouffer, R. J., Tokioka, T., Weaver, A. J., and Wigley, T. M. (1996). Climate models—projections of future climate. “Climate Change 1995,” pp. 285–357. IPCC Cambridge Univ. Press, Cambridge, UK.
- Keeling, C. D., Adams, J. A., Ekdahl, C. A., and Guenther, P. R. (1976a). Atmospheric carbon dioxide variations at the South Pole. *Tellus* **28**, 552.
- Keeling, C. D., Bacastow, R. B., Bainbridge, A. E., Ekdahl, C. A., Guenther, P. R., Waterman, L. S., and Chin, J. S. (1976b). Carbon dioxide variations at Mauna Loa Observatory, Hawaii. *Tellus* **28**, 538.
- Legget, J., Pepper, W. J., and Swart, R. J. (1992). Emission scenarios for IPCC: An update. “Climate Change: Supplementary Report to the IPCC Scientific Assessment,” IPCC, Cambridge Univ. Press, Cambridge, UK.
- Manabe, S., Spelman, M. S., and Stouffer, R. J. (1992). Transient responses of a coupled ocean-atmosphere model to gradual changes of atmospheric CO₂, Part II: Seasonal response. *J. Clim.* **5**, 105–126.
- Mitchell, J. F. B., and Warrilow, D. A. (1987). Summer dryness in northern mid-latitudes due to increased CO₂. *Nature* **330**, 238–240.
- Mitchell, J. F. B., Manabe, S., Meleshko, V., and Tokioka, T. (1990). Equilibrium climate change and its implications for the future. “Climate Change: IPCC Scientific Assessment,” (J. T. Houghton, G. I. Jenkins, and E. J. Ephraums, eds.), pp. 131–164. Cambridge Univ. Press, Cambridge, UK.
- Mitchell, J. F. B., Johns, T. C., Gregory, J. M., and Tett, S. B. F. (1995). Climate response to increasing levels of greenhouse gases and sulfate aerosols. *Nature* **376**, 501–504.
- Washington, W. M., and Parkinson, C. L. (1986). “In Introduction to Three-Dimensional Climate Modeling,” Oxford University Press, Oxford, New York.
- Wigley, T. M. L. (1998). The Kyoto Protocol: CO₂, CH₄, and climate implications. *Geophys. Res. Lett.* **25**, 2285–2288.
- Wigley, T. M. L., Jain, L. A., Joos, F., Shukla, P. R., and Nyenzi, B. S. (1997). Implications of proposed CO₂ emission limitations. IPCC Technical Paper 4, 41 pp. IPCC, Geneva.



Imaging through the Atmosphere

N. S. Kopeika

Ben-Gurion University of the Negev

- I. Properties of the Atmosphere
- II. Modulation Transfer Function and Modulation Contrast Function
- III. Turbulence
- IV. Background Contrast
- V. Overall Resolution through the Atmosphere
- VI. Techniques to Correct for Atmospheric Blur
- VII. Active Imaging

GLOSSARY

Aerosol Airborne particulates such as dust, smoke particles, water droplets, and so on. The smallest aerosols are the atoms of the various atmospheric constituent gases.

Focal length The distance from the rear vertex of a thin lens to the point at which a parallel light beam normally incident on the front surface is focused to as small a point as possible.

Fourier transform Any of the various methods of decomposing a signal into a set of coefficients of orthogonal waveforms. For a spatial signal $f(r)$, its Fourier transform is $F(\omega_r) = \int_{-x}^x f(r)e^{-j\omega_r r} dr$, where r is spatial coordinate and ω_r is spatial Fourier frequency.

Frequency The number of crests of waves that pass a fixed point in a given time.

Refractive index Ratio of velocity of light at a given wavelength in air to that in a refractive medium.

Scattering Change of direction of a beam of radiation when it interacts with the surface of a particulate: this process involves no change in the energy of the incident radiation.

Wavelength Physical distance covered by one cycle of sinusoidal propagation of electromagnetic (including light) waves.

I. PROPERTIES OF THE ATMOSPHERE

Many properties of the atmosphere affect the quality of images propagating through it. There are atmospheric phenomena that give rise to attenuation of the irradiance of the propagating image, thus reducing the contrast of the final

image. There are also atmospheric phenomena that cause blurring of the detail. Both types of phenomena prevent small details from being resolved in the final image, thus degrading image quality.

Phenomena that give rise to attenuation are electromagnetic-wave (1) absorption by the constituent gases of the atmosphere and (2) scattering by airborne particulates. The particulates can also absorb light; however, the absorption by the particulates is often relatively negligible compared to that by atmospheric gas molecules. Scattering of photons by airborne particulates is manifested as deflections of the photons to directions other than that of original propagation. If such scattering causes the deflected photons to miss the imaging receiver, then the scattering is manifested as attenuation. The received irradiance of the image propagated through the atmosphere is correspondingly diminished from that in the object plane. However, if the light scattering is at very small angles (milliradian order or less) with respect to the original directions of propagation, and if many such small-angle scattering events take place, then forward-scattered radiation can take roundabout paths and still be received by the imaging system together with the unscattered radiation. The net effect is image blurring caused by a multitude of angles-of-arrival at the imaging receiver of radiation, emanating from the same point in the object plane, as illustrated in Figure 1 for one multiple forward-scattered ray and one unscattered ray. Many such multiple-scattering paths give rise to a relatively large blurred point image rather than a fine sharp-point image. Several adjacent object-plane points can then appear as a single blurred image-plane point, thus degrading image resolution.

Another effect of light scattering, particularly at large angles, is increased path radiance. This is illustrated in Figure 2. The atmospheric background irradiance H_A , at wavelengths less than $2\text{--}4 \mu\text{m}$, consists of scattered background light such as sunlight during the day. This atmospheric background radiation is imaged over the same image space as the received irradiance H_T from the target plane. The result is decreased contrast of the target plane scene. The effect is similar to turning on the lights in a movie theater. In the dark, almost all light reaching

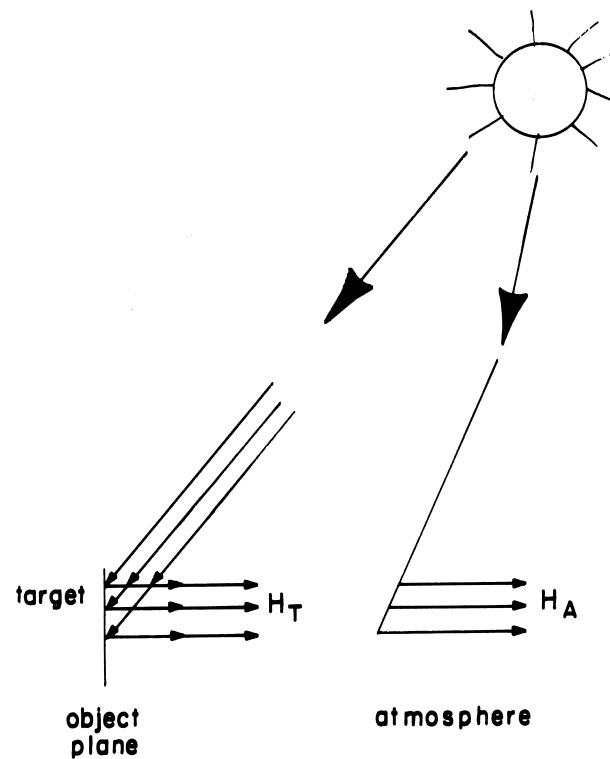


FIGURE 2 Imaging of atmospheric path luminance irradiation H_A together with target irradiance H_T over identical image-planes space.

the eye emanates from the movie screen, thus providing good contrast. When lights are turned on, the percentage of overall light reaching the eye from the movie screen is diminished because of the background light. The result is decreased contrast from the movie screen. Similarly, atmospheric background light or path radiance gives rise to decreased contrast of the target plane scene because the percentage of radiation emanating from the target plane relative to overall radiation reaching the receiver is diminished by the intervening atmosphere background radiation or path radiance. This makes it more difficult to resolve small detail. At wavelengths larger than $4 \mu\text{m}$, most of the path radiance is not scattered sunlight, as in Fig. 2, but rather thermal emission of the atmospheric

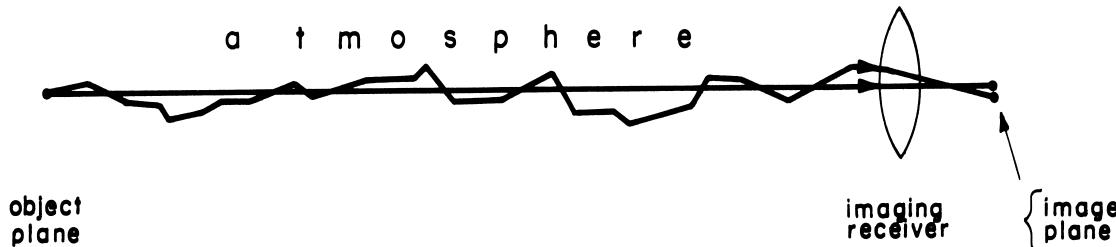


FIGURE 1 Unscattered radiation (straight line) and multiply forward-scattered radiation from the same object-plane point imaged at two different locations in the image plane.

constituents, gases in particular. This thermal path radiance decreases contrast in thermal imaging systems, especially in the 8–13- μm wavelength range.

Although multiple forward scatter gives rise to image blur, another significant blur mechanism caused by the atmosphere is turbulence. Turbulence results from variations in the atmospheric refractive index, which are random both in time and space. These refractive index fluctuations are caused by local fluctuations in atmospheric temperatures, pressure, humidity, and so on, all of which are affected by wind speed, which in itself is also random in time and space. Generally, the larger the temperature and humidity gradients, the more serious the image resolution degradation caused by turbulence. Physically, the blur resulting from turbulence can also be described by Fig. 1, if it is assumed that deflections of beam direction are now caused by random refractions instead of random scatterings. The net effect is still a variety of angles-of-arrival at the imaging receiver, thus causing image blur. However, scattering and turbulence are basically two very different mechanisms, each exhibiting quite different properties. Turbulence is very significant usually at low elevations close to the ground. Aerosol blur is usually significant at higher elevations, especially for optical depths on the order of unity or more. For large dynamic range imaging systems, aerosol blur becomes important at even smaller optical depths. Each of the various optical properties of the atmosphere described above will now be considered in turn.

A. Absorption

Absorption of radiation is the meachnism by which the atmosphere is heated. The solar irradiance intercepted as the mean earth–sun distance is about 1365–1372 $\text{W} \cdot \text{m}^{-2}$, most of which is in the visible wavelength range with a peak at about 0.49 μm wavelength. This corresponds to blue-green color. Part of this radiation is absorbed in the atmosphere, thus heating it, and part is transmitted to the ground. Part of the energy reaching the ground is absorbed, thus heating it, and part is reflected, thus further heating the atmosphere. For example, at a temperature of 288 K, the surface emits 390 $\text{W} \cdot \text{m}^{-2}$. Only 237 $\text{W} \cdot \text{m}^{-2}$, penetrates the atmosphere to space. The energy absorbed in the atmosphere is the 153 $\text{W} \cdot \text{m}^{-2}$ difference between surface emission and total energy loss. Because the atmosphere is generally colder than the ground, atmospheric gases absorb more energy than they emit. The infrared absorption by the atmosphere is due primarily to water vapor and CO_2 , with a smaller contribution from gases such as O_3 , N_2O , and CH_4 .

The atmospheric absorption is very wavelength-selective. Below about 0.2 μm , absorption by O_2 and O_3 is so great that there is essentially no propagation. This

spectral region is referred to as the vacuum UV because at such wavelengths propagation is virtually possible only under vacuum conditions. Imaging through the earth's atmosphere here is practically impossible. (It is also for this reason that destruction of ozone in the upper atmosphere is dangerous, for it permits ionizing radiation to reach Earth's surface.) On the other hand, there is hardly any absorption at visible wavelengths. This permits imaging using the human visual system. This absorption window continues out to about 1.3 μm .

Absorption windows also exist at 1.5–1.7 μm and 2.0–2.5 μm . The major thermal infrared absorption windows are 3.4–4.2 μm , 4.5–5.0 μm , and 8.0–13 μm . In between are absorption bands where transmission is essentially negligible and, as a result, imaging through the atmosphere is essentially impossible. Absorption attenuation decreases over the 16–22- μm range, permitting limited transmission there. However, from 22 μm down to nearly 1 mm wavelength (300 GHz), water vapor absorption of radiation is so strong that imaging and transmission through the atmosphere are virtually nonexistent.

B. Scattering

Light scattering is very much wavelength-dependent, according to the ratio of particulate radius to wavelength. Those particles that are small compared to the wavelength of the energy will produce Rayleigh scattering. The scattering coefficient in Rayleigh scattering is inversely proportional to the fourth power of the wavelength. For air molecules, scattering is negligible at wavelengths greater than about 3 μm . At a wavelength of 1 μm , such scattering is no longer negligible, and in the visible band it is quite pronounced. The classical demonstration of its effect is the blue color of the sky on a clear day, caused by selective atmospheric scattering of the broad solar spectrum so that nearly four times more blur light is scattered to the observer than red light.

The relationship between the scattering coefficient S_a , the absorption coefficient A_a , and transmission τ for homogeneous medium is

$$\tau = \exp[-(A_a + S_a)z] \quad (1a)$$

for a propagation path of length z . The atmosphere is essentially nonhomogeneous, so that

$$\tau = \exp\left[-\int_0^z [S_a(z') + A_a(z')] dz'\right]. \quad (1b)$$

The exponent is called optical depth. The visual range or “visibility” corresponds to the range at which light is attenuated to 0.02 times its transmitted level. Here Rayleigh scattering by molecules implies a visual range of about

340 km (213 mi). Thus, Rayleigh scattering theory applies only to extremely clear air. Such a condition seldom, if ever, exists near the surface of Earth. The Rayleigh scattering coefficient for molecular scattering is therefore an extreme limit that can never be exceeded in practice.

The primary reason why Rayleigh, or molecular, scattering theory has limited applicability to visible and infrared energy transmission is that the air normally carries in suspension many particles that are comparable to or larger than the wavelength of the transmitted energy. The effect of these particles on the energy is explained by the Mie theory of scattering. Mie theory predicts that, because of diffraction effects, the scattering cross-section of a particle will be equal to about two times its physical cross-section for all wavelengths, much smaller than the effective radius. Scattering losses decrease rapidly with increasing wavelength, approaching the Rayleigh scattering situation. The effects of Mie scattering are shown in Fig. 3, which illustrates the manner in which Mie scattering from water droplets affects scattering losses.

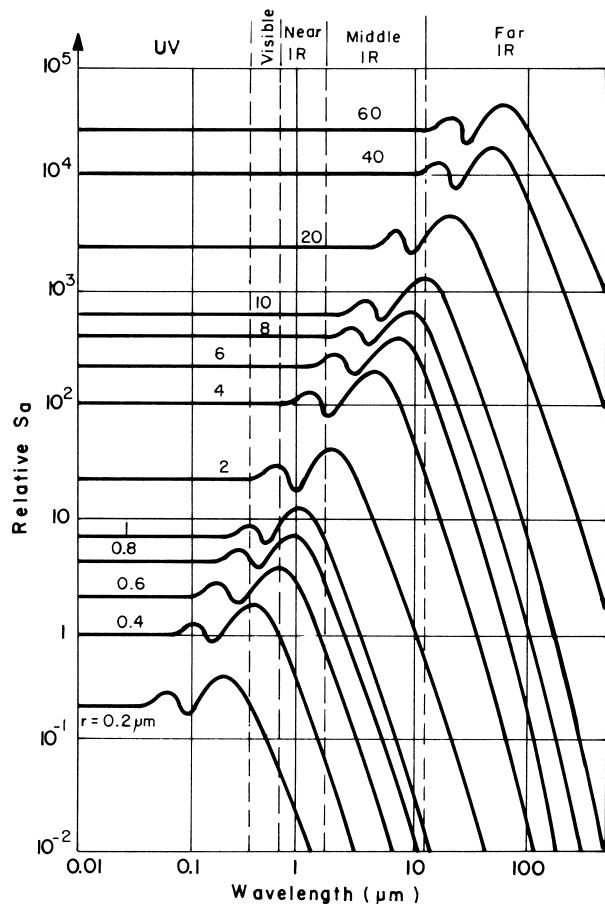


FIGURE 3 Relative values of scattering coefficient as a function of wavelength for different particulate radii (r). IR, Infrared; UV, Ultraviolet.

Some additional qualifications regarding Fig. 3 are in order. First it must be remembered that those data apply only in the “window” bands for which absorption due to water vapor and CO in the propagation path is negligible. Second, water droplets in haze, fog, and clouds are seldom uniform in size, so that the true characteristics will have less clearly defined “ripples” and “knees” than the curves shown in the figure, and will tend to be leveled out in most real situations. Third, no consideration has been given here to particles such as pollen, smoke, and combustion products that may be present in the air. Both scattering and absorption are involved in such conditions, and the absorption will tend to be selective according to wavelength.

It can be noticed in Fig. 3 that maximum scatter occurs at wavelengths essentially equal to particulate radius. As wavelength increases beyond that point, the scattering coefficient decreases with wavelength until for a given particulate size Rayleigh scattering conditions are obtained. The directionality of the scatter is very important in imaging applications. For Rayleigh scattering, the scattering phase function that describes the direction θ in which light is deflected is essentially proportional to $(1 + \cos^2 \theta)$. Such scattering is significant in all directions and is manifested in imaging as attenuation and resultant loss of contrast. However, as wavelength decreases toward maximum scatter conditions, and even beyond that point, the scatter becomes more and more directed into small angles in the forward direction. It is under these conditions that multiple scattering events significantly affect image quality by blurring, as shown in Fig. 1. Thus, image quality also depends largely on the ratio of particulate size to wavelength.

Detailed data on values of particulate size and resulting scattering and absorption coefficient for various weather and climatic conditions can be obtained from the literature, particularly that of the Air force Geophysics Laboratory, which is also summarized in the *RCA Electro-Optics Handbook* and in Kopeika and Bordogna (1970).

Although scattered sunlight determines the color of the sky, it is the transmitted sunlight that determines the apparent color of the sun. Thus, even though the solar spectrum peaks at blue-green wavelengths, in clear weather the sun appears to be orange or reddish in color because of improved atmospheric transmission or greatly reduced scatter at these longer wavelengths.

In haze and fog conditions, particulates increase in size because of water vapor adsorption and absorption. At visible wavelengths, the scattering becomes much less wavelength dependent and more neutral. As a result, both the sky (scattered light) and the sun (unscattered light) appear more and more “whitish” in color than in clear weather.

The scattered versus transmitted irradiances play a significant role in overall image contrast in imaging through the atmosphere, because the path radiance (Fig. 2) is

determined by the background radiation in passive imaging systems.

C. Background Radiation

As illustrated in Fig. 2, in passive imaging systems background light illuminates the object plane and is reflected by it toward the receiver. The transmitted light forms the image. On the other hand, scattered background light, or path radiance, is also received by the imager, while *not* emanating from the object plane. The greater the irradiance of this scattered background light, as compared to that of the light transmitted from the object to image plane, the poorer the contrast of the image. At visible and infrared wavelengths out to $2\text{--}4 \mu\text{m}$, the atmospheric background radiation that composes the transmitted and scattered light consists of sunlight, moonlight, star light, and so on. At longer wavelengths, the thermal emissions of the object and of the atmosphere give rise to the image and atmospheric background (path radiance) irradiances, respectively, since at such longer wavelengths they are of greater irradiance than sunlight.

Atmospheric background radiance peaks strongly at around $0.5\text{-}\mu\text{m}$ wavelength for scattered sunlight and $10\text{-}\mu\text{m}$ wavelength for thermal emission of atmospheric gases. A minimum between these two spectra exists at $2\text{--}4\text{-}\mu\text{m}$ wavelength, and it is here that atmospheric background degradation of contrast is least. At visible and near-infrared wavelengths this atmospheric background radiance is a strong function of turbidity, with wavelength variation depending on the size distribution of airborne particle scatterers. Generally, particulate size distributions peak at radii on the order of submicrometer size, which is often due to pollution and reactions between various atmospheric gases. These both decrease the solar irradiance reaching the earth's surface and increase the atmospheric background radiance, primarily at near-UV and visible wavelengths.

Changing aerosol mass in the atmosphere can cause considerable turbidity fluctuations. This is true particularly on warm days when, as the ground heats up, thermal-induced convective mixing causes aerosol particles to be carried to higher altitudes and to higher concentrations. Such heating effects can cause large variation in turbidity between different seasons of the year and between different locations. For example, farming communities are usually much more turbid in the spring, summer, and fall when bare soil is exposed and such thermal activity is relatively high. Since the size of such particulates is relatively small compared to wavelength (except in the short end of the visible), scattering by them is generally at large angles and thus tends to decrease transmission and increase background radiance of the atmosphere. This is also char-

acteristic of the early stages of fog formation, when there is a predominance of smaller droplets. However, as ambient humidity increases, water vapor may be adsorbed by aerosol particles suspended in the atmosphere. After ambient relative humidity has reached a threshold value such as 70%, which depends on the deliquescent property of the aerosol compound, the condensed water increases the size of the aerosol particles and tends to make them more spherical in shape. The size increase causes a greater proportion of the scattering to be small-angle, or in the forward direction, thus increasing transmission when wavelengths are small compared to particle size.

This increase in particle size and forward directionality of scattering with increased relative humidity is characteristic not only of common rural and urban aerosols but also of maritime salt-particle aerosols, desert dust, and fog particulates. Visibility is very closely linked to variations in relative humidity through these physical changes in aerosol size as the relative humidity changes. There is a general association of low transmittances with high relative humidity, and low relative humidity with high transmittances. Effects of ambient humidity and temperature changes are negligibly small when relative humidity is low. However, for a humid environment, a few percent increase in relative humidity or a few degrees decrease in temperature (which brings the atmosphere closer to 100% relative humidity) may cause significant increase in scattering coefficient from aerosol particles of even small sizes at even middle infrared ($3.4\text{--}5 \mu\text{m}$ and $10 \mu\text{m}$) wavelengths.

The main differences between infrared (IR) and visible light scattering can be explained by aerosol size distributions. The IR scattering is influenced mostly by aerosols with large sizes. When particulates are very large compared to optical wavelengths, the geometrical optics approximation suggests that scattering attenuation should be independent of wavelength, as seen on the left side of Fig. 3. However, numerous measurements have indicated that under rain, snow, and dust conditions there is less extinction in the visible than in the infrared. This difference of 10–20% has been attributed to phase functions being more sharply peaked in the forward direction for shorter wavelengths, thus increasing the forward directionality of their scatter. The increase in forward directionality of scatter is so pronounced that an increase in the field of view of the optical communication receiver during snowstorms has been known to generate order-of-magnitude increases in the signal-to-noise ratio at visible wavelengths.

D. Turbulence

The refractive index n of the atmosphere is described by

$$n = 1 + \frac{77.6p}{T} \left[1 + \frac{7.52 \times 10^{-3}}{\lambda} - 7733 \frac{q}{T} \right] 10^{-6}$$

$$= 1 + N, \quad (2)$$

where p is air pressure (millibars), T is temperature (K), q is specific humidity ($\text{g}\cdot\text{m}^{-3}$), and λ is wavelength. By plugging in typical values for p , T , and q , one can see that $N \ll 1$. Nevertheless, the fluctuations in N , which depend largely on temperature and humidity gradients, strongly affect image quality because of the large number of random refractions that a light beam undergoes while propagating a relatively large distance through the atmosphere. The randomness in both time and space cause the incident light to be received at a large variety of angles of incidence, similar to the scattered light in Fig. 1. An example of such effects can be obtained by looking along a line-of-sight over an operating toaster at home. Because of the thermal gradient in the air above the toaster, the fluctuations in N are so great that even nearby objects appear to be dancing randomly and exhibit blurring effects.

Since the ground is generally warmer than the surrounding air during the day and colder at night, a thermal gradient exists in the atmosphere, which is usually greatest near the ground. Air temperature generally decreases with height, thus continuing the thermal gradients, albeit with smaller values, to increasing altitude.

Water vapor is generally confined for the most part to regions near the ground. Thus, a humidity gradient also exists and basically decreases with increasing altitude.

Temperature and water-vapor fluctuations are known to be the primary mechanisms mediating the effects of atmospheric turbulence on the index of refraction. Pressure fluctuations can be neglected in the real atmosphere. Correlation coefficient studies at microwave frequencies, where the refractive index is much more sensitive to humidity than at optical frequencies, indicated, however, that wind speed was the dominant meteorological parameter affecting turbulence. In fact, increasing wind speed, because of the homogenizing effect on the index of refraction, was found to be the best meteorological predictor of reduced monthly accumulated microwave fading time. Higher wind speed can thus be expected to give improved imaging resolution as affected by turbulence, within certain limits. Nighttime clear skies have been associated with increased atmospheric degradation because of reduced vertical air mixing. The presence of clouds, on the other hand, tends to increase the mixing of air (decrease vertical stabilization) and thus to decrease the degradation of atmospheric imaging.

Around sunrise and before sunset, ground and surrounding air temperatures are closest. At these times, when temperature gradients are least, turbulence has the least degradation effect on image quality.

II. MODULATION TRANSFER FUNCTION AND MODULATION CONTRAST FUNCTION

A convenient engineering tool with which to characterize the quality of an imaging system is transfer function. The concept and implications in imaging systems are similar to those in electronics. The wider the Fourier frequency bandwidth, the better the fidelity between output and input. In imaging systems, the output is the image and the input is the object. As in electronics, the overall system transfer function is equal to the product of the transfer functions of each individual component, provided they are independent of one another. Hence, a useful approach will be described which characterizes imaging through the atmosphere and quantitatively the different atmospheric effects on quality of image propagation.

A. Modulation Transfer Function

Light propagating from a point source should ideally be imaged into a point image. However, forward scatterings and refractions of the beam, which are random in both time and space, cause the beam finally incident on the receiver to arrive from many different angles, thus smearing or spreading the image of the point source. This spreading is called the system “spread function”; it characterizes the resolving capability of the imaging system, including that of the intervening atmosphere.

Thus, if $i(r_i)$ is the image (or system output) as a function of spatial coordinate r_i , and if $o(r_0)$ is the object (or system input) as a function of spatial coordinate (r_0) , then according to convolution theory

$$i(r_i) = \int_{-x}^x s(r_i - r_0)o(r_0)dr_0, \quad (3)$$

where $s(r_i)$ is the spread function or intensity impulse response. With regard to imaging systems, a delta function or impulse input signal is represented by a point object. If the imaging system, including the atmosphere, were ideal, then a point image would be obtained. In other words, if $o(r_0) = \delta(r_0)$, then according to Eq. (3) the impulse response or spread function is $i(r_i) = s(r_i)$.

In the Fourier domain, if the Fourier transform of the object is $O(\omega_r)$, of the image is $I(\omega_r)$, and of the spread function is $S(\omega_r)$, then

$$I(\omega_r) = S(\omega_r)\theta(\omega_r), \quad (4)$$

where ω_r is radian spatial (Fourier) frequency. In electronics, where information is amplitude as a function of time, Fourier frequency is in units of cycles per second or hertz (Hz). In imaging, information is intensity as intensity is a function of position. Therefore, in a manner analogous to

electronics, Fourier frequency is in cycles per unit picture width.

In electronics, the Fourier transform of the impulse response is defined as the transfer function. Analogously in imaging, $S(\omega_r)$, should be imaging or optical transform function (OTF). However, it is customary to normalize the OTF by its maximum value, so that the normalized OTF varies between unity and zero. Except for the human visual system, $S(\omega_r)$ is maximum when $\omega_r = 0$ and decreases as spatial frequency increases. The magnitude of the optical transform function is called the modulation transfer function (MTF). In other words,

$$\text{OTF}(\omega_r) = \frac{S(\omega_r)}{S_{\max}} = \frac{S(\omega_r)}{S(0)} = \text{MTF} e^{j\text{PTF}}, \quad (5)$$

where PTF is phase transfer function.

B. Modulation Contrast Function

For a sinewave object (rather than a bar or squarewave resolution chart), the system MTF can be shown to be equal to the system modulation contrast function (MCF), which is the ratio of image-plane modulation contrast to object-plane modulation contrast. Modulation contrast in each plane is defined as $(I_{\max} - I_{\min})/(I_{\max} + I_{\min})$, I signifying irradiance. For a square-wave or bar chart. MCF is not equal to MTF, but is an approximation. Square-wave response (MCF) is slightly higher than sine-wave response ($\text{MTF} \leq \text{MCF}$). The decrease of MTF or MCF with increasing spatial frequency signifies contrast degradation at higher spatial frequencies. At some relatively high spatial frequency, system MTF or MCF has decreased to such a low value of contrast that it is below the threshold contrast function of the observer. In the case of a black-and-white bar chart, for example, wide bars represent low spatial frequencies, since the number of them that fit into a unit picture width is small. Narrow stripes or lines represent high spatial frequencies. As black-and-white line pairs decrease in width (or, equivalently, as spatial frequencies increase), the white lines appear to blend into the black and vice versa, and contrast decreases. Narrow black-and-white lines appear to be different shades of gray rather than black and white. At a sufficiently high spatial frequency, the “black” and “white” shapes of gray appear to be identical, in which case the black-and-white lines can no longer be resolved because of the poor contrast. The spatial frequency at which system MTF or MCF is just equal to the threshold contrast of the observer defines the maximum useful spatial frequency content of the system. We will call it here $f_{r\max}$. It is related to maximum useful radian frequency ω_r via $f_{r\max} = \omega_{\max}/2\pi$. Often the threshold contrast of the observer is taken to be 2%, although there is some evidence that it really varies with spatial frequency, as shown in Fig. 4. If Δx and $\Delta x'$ are

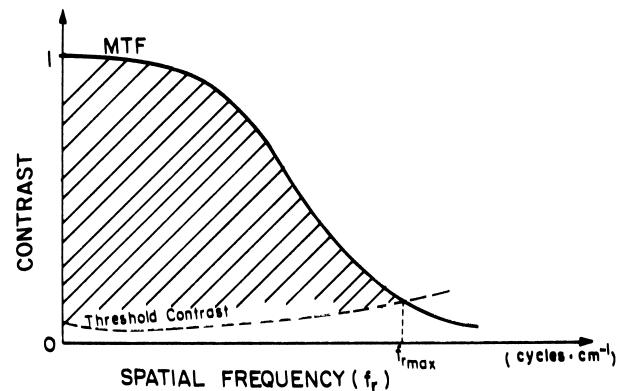


FIGURE 4 Typical modulation transfer function (MTF) and threshold contrast plots.

uncertainties in the object and image planes, respectively, and if s and s' are object and image distances, respectively, then on the basis of similar triangles as in Fig. 5

$$\frac{\Delta x}{s} = -\frac{\Delta x'}{s'} \approx \frac{1}{f_{r\max}s'}, \quad (6)$$

where the minus sign refers to image inversion and a large target range s is assumed. Since optical magnification M is equal to the ratio of s' to s , the limit of detectable resolvable detail in the object plane is

$$\Delta x \propto (f_{r\max} M)^{-1}. \quad (7)$$

The greater the usable spatial frequency content $f_{r\max}$ provided by it the imaging system, the smaller the object-plane detail that can be resolved and the better the imaging system To image an ideal point object ($\Delta x \rightarrow 0$) clearly would require an infinite Fourier bandwidth ($f_{r\max} \rightarrow \infty$). This is equivalent to obtaining a point image for a point object.

For relatively long propagation paths through the atmosphere, system spread function and MTF are limited primarily by atmosphere—rather than instrumentation—phenomena. Here, we are concerned with the MTF or MCF of the atmosphere. Each of the various types of image degradation produced by the atmosphere will be considered now quantitatively from the standpoint of MTF or MCF. Use of MTF or MCF theory in imaging system design is discussed in the literature (see Kopeika, 1998, for example).

III. TURBULENCE

A. Long-Exposure and Short-Exposure MTFs

For a diagonal path looking down through the atmosphere, the MTF for atmospheric turbulence can be described by

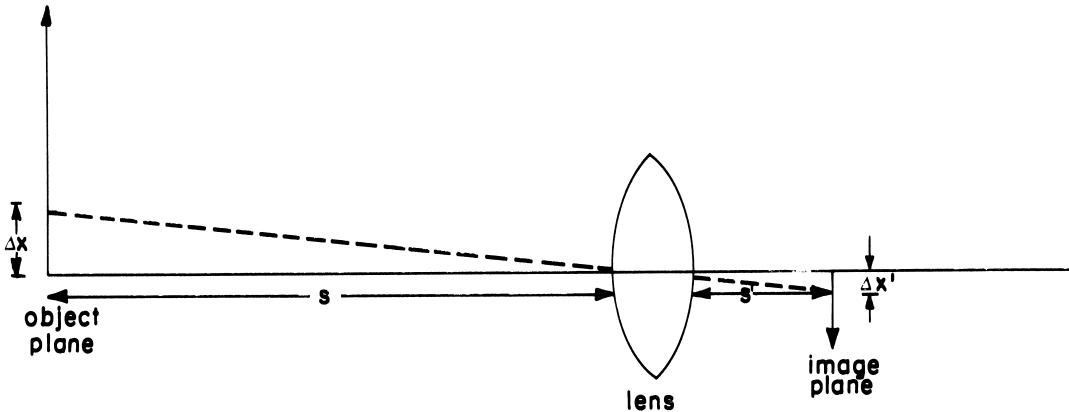


FIGURE 5 Relation between resolvable detail in object plane (Δx) and image plane ($\Delta x'$).

$$M_T = \exp \left[-57.44(f_l f_r)^{5/3} \lambda \sec \alpha \times \int_{h_t}^H C_n^2(h) \left(\frac{h}{H-h_t} \right)^{5/3} dh \right], \quad (8)$$

for plane wave imaging with an imaging system of focal length f_l . In Eq. (8) f_r is spatial frequency, α is viewing angle relative to a vertical path, H is elevation of the imaging system. C_n is the refractive index structure coefficient that characterizes the turbulence, and the lower limit of the integral, h_t corresponds to the elevation of that part of the target being imaged. For short exposures ($f_l \ll .04$ s), the degradation is less and the MTF is

$$M_{TS} = \exp \left\{ -57.44(f_l f_r)^{5/3} \lambda^{1/3} \sec \alpha \times \int_{h_t}^H C_n^2 \left(\frac{h}{H-h_t} \right)^{5/3} \times \left[1 - \frac{1}{b} \left(\frac{f_l f_r}{D} \right)^{1/3} \right] dh \right\}, \quad (9)$$

where $b = 1$ at the image center and $b = 1$ at the edges. It can be seen from Eqs. (1) and (2) that turbulence favors imaging at longer wavelengths, although the improvement is limited.

The difference between long- and short-exposure degradation is that short exposures involve interacts with large turbulence eddies, the light tends to be deflected, whereas for small turbulence eddies objects tend to appear broadened. The longer the exposure, the greater the deflection or beam wander, thus causing further blurring of the short-exposure image. A limitation involved in short-exposure imaging is that although there are fewer and perhaps even only one angle-of-arrival at a given location, that angle itself is apt to vary randomly over the receiving aperture, thus causing blurring as a result of multiple images. Conse-

quently, the probability of obtaining a good quality “lucky shot” image decreases as receiver area increases.

For a horizontal path of range R , the product of the integral and $\sec \alpha$ in Eqs. (8) and (9) is replaced by $\int_0^R C_n^2(z') (\frac{z}{R})^{5/3} dz'$.

B. Refractive Index Structure Coefficient

The strength of the turbulence is described by the integral in Eqs. (8) and (9), which, in turn is determined by the value of C_n and the path length whose differential value is $dh \sec \alpha$. The square of the refractive index structure coefficient is equal to

$$C_n^2 = D_n(\rho) \rho^{-2/3} = \langle [\Delta n(r + \rho) - \Delta n(r)]^2 \rangle \rho^{-2/3}, \quad (10)$$

where D_n is the refractive index structure function, which is defined as the mean square difference between refractive index fluctuations at two points spaced ρ apart. Equation (10) holds when $D_n(\rho)$ is stationary over a small locally homogeneous region. The larger the variance of refractive index over the distance ρ , the larger the value of C_n^2 . Good imaging conditions require C_n to be as small as possible.

C. Environmental Factors on Turbulence

The following environmental factors affect turbulence.

1. *Time of day:* C_n^2 is minimum at sunrise and sunset when air temperature is closest to ground temperature. Temperature gradient is generally greatest at midday and midnight, when ground is warmer than overlying air, thus increasing C_n^2 .
2. *Cloud cover:* During daytime, cloud cover limits surface heating by the sun, thus decreasing C_n^2 . At night, cloud cover limits ground cooling, thereby increasing temperature gradient and thus C_n^2 .

3. *Wind*: Wind produces more air mixing, thus decreasing temperature and humidity gradients, except for very strong winds on the order of 15 knots and higher.
4. *Elevation*: C_n^2 is maximum near ground.
5. *Latitude*: Length of day and amount of solar heating are determined by latitude. There is less turbulence at higher latitudes.
6. *Surface conditions*: Surface moisture limits the humidity gradient, and thus C_n^2 . On the other hand, surface roughness increases temperature gradient and thus C_n^2 .

It is possible to predict C_n^2 in advance according to weather forecast ([Kopeika, 1998](#)).

D. Effect of Wavelength

The wavelength dependence of image resolution through turbulence is very weak. As shown in Eqs. (8) and (9), M_T does favor imaging at longer wavelengths, but only slightly. To observe the same turbulence MTF value at two different wavelengths would require the exponents in Eq. (8) to be equal at each wavelength. This requires

$$\frac{f_{r1}}{f_{r2}} = \left(\frac{\lambda_1}{\lambda_2} \right)^{1/2}, \quad (11)$$

where f_{r1} and f_{r2} are respective spatial frequencies at which $M_T(\lambda_1) = M_T(\lambda_2)$. If one compares imaging at $\lambda_1 = 10 \mu\text{m}$ to $\lambda_2 = 0.55 \mu\text{m}$, then $f_{r1}/f_{r2} \cong 1.8$. The resolution improvement is small. (It is even smaller when the wavelength dependence of C_n^2 is also considered, although that is beyond the scope of this text.)

E. Variation with Altitude

The dependence of C_n on height (h) has been investigated many times in many different locations. Vertical profiles have been observed to change considerably over even a few hours. Experimentally, there is no unique vertical profile for C_n . Nevertheless, to aid in system design, several mathematical models have been suggested. Experiments up to about 100 m in elevation indicate that the most relevant are the models of Tatarski and Brookner, which are, respectively,

$$C_n^2(h) = C_{n0}^2 h^{-4/3} \quad (12)$$

$$C_n^2(h) = C_{n0}^2 h^{-5/6} \exp\left(-\frac{h}{h_0}\right), \quad (13)$$

where C_{n0} is the refractive index structure coefficient at elevation h_t and h_0 is 320 m. Turbulence is strongest near the ground, as for imaging from low elevations. Eqs. (12)

and (13) can be quite useful in system design. Equation (2) is based on theoretical considerations, whereas Eq. (13) is based on empirical data.

For the region above the boundary layer, several different height profiles for C_n^2 have also been suggested.

The so-called submission laser communication (SLC) models represent median values, averaged over a year above the AMOS observatory on Mt. Haleakala, Maui, Hawaii. There are no parametric dependences. The values of C_n^2 derive from stellar scintillation measurements, and are

$$\begin{aligned} C_n^2 &= 8.4 \times 10^{-15}, & h \leq 18.5 \text{ m} \\ &= 2.87 \times 10^{-12} h^{-2}, & 18.5 < h < 110 \text{ m} \\ &= 2.55 \times 10^{-16}, & 110 < h < 1500 \text{ m} \\ &= 8.87 \times 10^{-7} h^{-3}, & 1.5 < h < 7.2 \text{ km} \\ &= 2 \times 10^{-16} h^{-0.5}, & 7.2 < h < 20 \text{ m}, \end{aligned} \quad (14)$$

where h is in meters above mean sea level (MSL) and units of C_n^2 are $\text{m}^{-2/3}$. This model is for a subtropical atmosphere with a high (17 km MSL) tropopause. The constant value of C_n^2 for $110 < h < 1500$ m derives from lack of measurements over such heights. Usually there is significant variability over this height range at Maui.

A daytime SLC model extension was also developed, and is

$$\begin{aligned} C_n^2 &= 1.70 \times 10^{-14}, & h \leq 18.5 \text{ m} \\ &= 3.13 \times 10^{-13} h^{-1}, & 18.5 < h \leq 110 \text{ m} \\ &= 1.305 \times 10^{-15}, & 240 < h < 880 \text{ m} \\ &= 8.87 \times 10^{-7} h^{-3}, & 0.8809 < h < 7.20 \text{ km} \\ &= 2.00 \times 10^{-16} h^{-0.5}, & 7.2 < h < 20 \text{ km}. \end{aligned} \quad (15)$$

By using balloons, thermal measurements of C_n^2 from the top of Mt. Haleakala were performed and a high resolution refinement of the SLC nighttime model was developed. This model, known as the AFGL AMOS model, included a wide range of weather conditions, and is

$$\begin{aligned} \log_{10} C_n^2 &= -12.2 - 0.4713z - 0.0906z^2, & 3.052 < z \leq .5.2 \\ &= -17.1273 - 0.0301z - 0.0010z^2 \\ &\quad + 0.5061 \exp[(z - 15.0866)/6.5954]^2, & 5.2 < z \leq 30, \end{aligned} \quad (16)$$

where z is height in kilometers (MSL).

These models may be site-dependent and are not suitable for continental, midlatitude locations. Also, they do not model vertical layers.

Similar models but for the New Mexico desert in the summer were developed by the U.S. Air Force Geophysics Lab, using procedures similar to those for the AMOS model. This latter model, known as CLEAR 1 is, for nighttime,

$$\begin{aligned} \log_{10} C_n^2 &= -10.7025 - 4.3507z + 0.8141z^2, \\ &\quad 1.23 < z < 2.13 \\ &= -16.2897 + 0.0335z - 0.0134z^2, \\ &\quad 21.3 < z < 10.34 \\ &= -17.0577 - 0.0449z - 0.0005z^2, \\ &\quad + 0.6181 \exp\{(z - 15.5617)/6.9332\}^2, \\ &\quad 10.34 < z \leq 30, \quad (17) \end{aligned}$$

where z is again in km. This model was developed for fairly homogeneous meteorological conditions (no jet streams or fronts). Both Eqs. (16) and (17) show similar results above 10 km, although for the boundary layer and the lower troposphere the CLEAR 1 model indicates more average turbulence. However, the similarity shown between the previous models in the stratosphere has also been substantiated by other analysis and data. CLEAR 1 is a good model of average or nominal conditions in the troposphere and stratosphere and is useful for comparing sites and seasons. It is a fairly simple model with which to evaluate optical effects and to use in propagation codes. The CLEAR 1 night model and the AFGL model are compared in Fig. 6.

The Huffnagel-Valley model was developed on the basis of stellar scintillations and balloon measurements. Developed for altitudes from the ground to 24 km, the model is

$$\begin{aligned} C_n^2 &= 8.2 \times 10^{-16} W^2 z^{10} e^{-z} + 2.7 \\ &\quad \times 10^{-16} e^{-z/1.5} + A e^{-z/0.1}. \quad (18) \end{aligned}$$

where again height z is specified in km,

$$W^2 = (1/15) \int_5^{20} V^2(z) dz \quad (19)$$

and $V(z)$ is the rms wind speed in range of 5–20 km above ground. The most popular version of this model is termed the Huffnagel-Valley 5/7 model, because parameters are selected so that the C_n^2 profile yields a coherence length of 5 cm and isoplanatic angle of 7 μrad at 0.5 μm wavelength. For such conditions, $A = 1.5 \times 10^{-4}$ and $W = 21 \text{ m} \cdot \text{s}^{-1}$. This model is in widespread use. The exponential falloff [especially the last term in Eq. (18)]

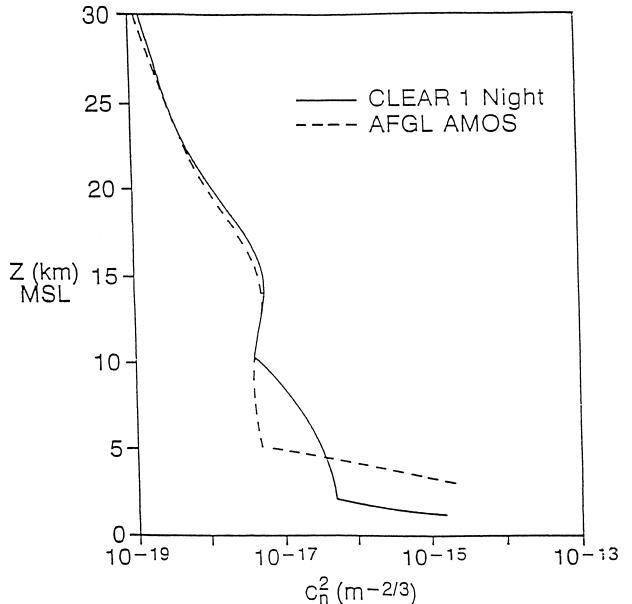


FIGURE 6 Comparison between AFGL AMOS and CLEAR 1 night profiles of C_n^2 .

in the first 3 km is unrealistic in light of boundary-layer modeling. This model too might be site-dependent. In general, performance in the stratosphere is limited because of the dependence of C_n^2 on tropospheric winds.

F. Effective Atmospheric Aperture

Fried has shown that the long-exposure resolution permitted by the turbulent atmosphere is essentially equivalent to that permitted by a diffraction-limited lens of diameter r_0 without turbulence. For short exposure, the turbulent atmosphere behaves as a diffraction-limited lens of size approximately $3.4 r_0$. For exceptionally good seeing vertically, $r_0 \cong 20 \text{ cm}$, for moderately poor seeing, $r_0 \cong 5 \text{ cm}$. For horizontal paths near ground, r_0 is considerably less, depending on path length. The analysis relates only to resolution, not to light-gathering power.

IV. BACKGROUND CONTRAST

A. MCF

It is not only photons emanating from the target that are received by the imaging system but also those emanating from the background of the target plane and from the intervening atmospheric background between target and imaging system planes. The atmospheric background causes glare that reduces the contrast emanating from the target plane. The reason for the glare is that the atmospheric background is imaged throughout the image plane, thus

contributing to the irradiance of the images of both the target and the target–plane background. The greater the atmospheric background, the smaller the percentage of photons received that derive from the target. If ρ_t and ρ_b are target and target-plane background reflection coefficients, the modulation contrast in the target plane is, for short wavelengths ($\lambda < 2\text{--}4 \mu\text{m}$) (see Section II.B),

$$C_0 = \frac{H_o - H_b}{H_o + H_b} = \frac{H_s(\rho_t - \rho_b)}{H_s(\rho_t + \rho_b)}, \quad (20)$$

where H_s is solar irradiance incident on the object plane, and H_o and H_b are target and object–plane background irradiances, respectively. Equation (20) describes the modulation contrast in the object plane prior to propagation through the atmosphere. This analysis is based on Fig. 2, but with the added complication that the object plane irradiance in Fig. 2 now consists not only of target, but also of background in the target plane, such as, for example, a vehicle (H_o) with vegetation or sky background (H_b). The apparent irradiances reaching the imager are τH_o and τH_b , where τ is atmospheric transmission defined in Eqs. (1a) and (1b). In addition, the atmospheric background irradiance or path radiance (H_A in Fig. 2) is assumed to fall uniformly over both target and object–plane background spaces in the image plane. As a result, modulation contrast is

$$\begin{aligned} C_i &= \frac{(\tau H_o + H_A) - (\tau H_b + H_A)}{(\tau H_o + H_b) + (\tau H_b + H_A)} \\ &= \frac{\tau H_s(\rho_t - \rho_b)}{\tau H_s(\rho_t + \rho_b) + 2H_A} \\ &= \frac{\rho_t - \rho_b}{\rho_t + \rho_b + 2H_A/(\tau H_s)}. \end{aligned} \quad (21)$$

By definition, MCF is equal to C_i/C_0 . Atmospheric background irradiance contrast reduction is

$$M_B = \frac{C_i}{C_0} = \frac{\rho_t + \rho_b}{\rho_t + \rho_b + 2H_A/(\tau H_s)}. \quad (22a)$$

Were there no path radiance ($H_A = 0$), then Eq. (22) would be unity and there would be no degradation of contrast resulting from background atmospheric radiation.

At longer wavelengths, atmospheric irradiation is primarily thermal in nature, rather than transmitted and scattered sunlight. In this case,

$$M_B = \frac{H_o + H_b}{H_o + H_b + 2H_A/\tau}, \quad (22b)$$

where H_0 , H_b , and H_A are essentially thermal emissions.

It should be noted that the atmospheric background MCF, Eq. (22a) or (22b), is not a function of spatial frequency. However, for a sinewave target, such an MCF also is an MTF. It represents a constant damping of the overall

system MTF, thus causing higher spatial frequency components of the overall imaging system MTF to be at contrasts below the threshold contrast required at the output, as shown in Fig. 4.

B. Seeing Limit

The resolution impairment caused by the atmospheric background radiation can be understood in the following way. The decrease of the MTF with increasing spatial frequency signifies contrast degradation at higher spatial frequencies. At some relatively high spatial frequency, the system MTF has decreased to such a low contrast that it is below the threshold contrast function required by the observer at the output. This means that such a high spatial frequency content of an image cannot be resolved by the observer because of its poor contrast.

The contrast degradation caused by atmospheric background causes the overall system MTF to be damped uniformly across the spatial frequency spectrum. As shown in Fig. 7, this leads to a reduction in $f_{r\max}$ or an increase in Δx in Eq. (6), thus impairing resolution. For orientation, recognition, and identification requirements, as opposed to simple detection, $f_{r\max}$ should be divided by the required number of TV lines or spatial frequencies (discussed in Biberman, 1973, or Kopeika, 1998, concerning the Johnson chart).

C. Wavelength dependence

Good contrast through the atmosphere requires decreasing the third term in the denominator in Eq. (22), for without

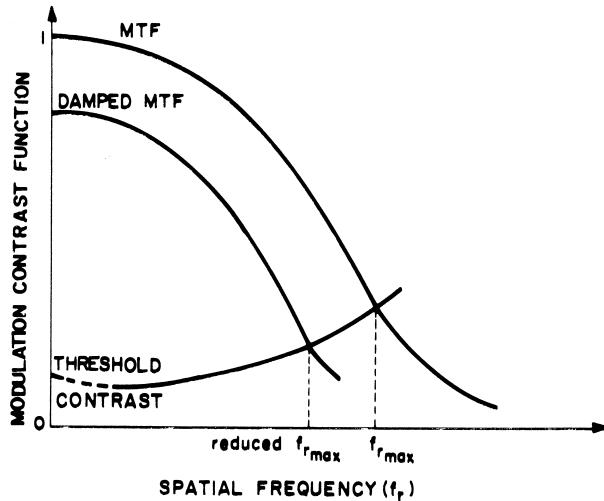


FIGURE 7 Reduction in usable resolution imposed by uniform damping of modulation contract function (MCF) by atmospheric background. MTF, modulation transfer function.

this term (22) would be unity. The wavelength dependences of atmospheric radiance, sea-level solar irradiance, and atmospheric scattering coefficients are well known. In general, depending on weather conditions, this term is at a minimum over the wavelength intervals of approximately 2–2.4 and 3.1–4.1 μm ; an increase of wavelength from 0.5 μm to only 0.9 μm , for example, can result in quite a significant improvement in contrast and range. However, this wavelength dependence can be reversed in desert and marine atmospheres. The reason is that aerosol distributions in desert and marine atmospheres are generally trimodal rather than bimodal because of the large quantities of windblown dust airborne from the dry soil, or salt particles and sea spray windblown from the sea surface. In continental nondesert atmospheres, aerosol radii peak at sizes on the order of 0.1–0.4 μm , or even less in clear weather. In desert atmospheres aerosol radii peak also at larger sizes in clear weather and on the order of 30 μm in dust storms. In marine atmospheres, peaks are usually also at larger sizes, according to wind speed and humidity. Since scattering is at a peak when the aerosol radius is on the order of a wavelength, the third term in the denominator in Eq. (22) often peaks in the near-IR in clear-weather desert atmospheres, thus favoring visible and/or far-IR (8–12 μm) wave imaging over near-IR wave imaging.

The effect of the soil-derived aerosols on the wavelength dependence of imaging is significant for airborne imaging because the relative aerosol size distributions in clear weather can be fairly constant up to altitudes on the order of 5 km. Thus, with a CCD TV system, for example, in clear weather Eq. (22a) favors near-IR imaging in nondesert atmospheres and visible wave imaging in desert atmospheres. Since turbulence exhibits very weak wavelength dependence, the wavelength dependence of Eq. (22) is significant. In other words, if near-IR sensors, which are advantageous in nondesert atmospheres, are utilized in desert atmospheres, they are most likely to be disadvantageous in the dry season. In haze and high-uniformity conditions, the aerosol size increases, and in fog it tends toward a more uniform distribution, thus decreasing the benefits of wavelength filtering.

D. Environmental Dependences

The wavelength dependence of the scattering coefficient in Eq. (9) is very much a function of meteorological conditions with regard to both soil and the atmosphere. If the soil contains moisture, for example, the resulting adhesiveness of the soil particles to one another prevents many particles, particularly the larger and heavier ones, from being airborne in the wind. Good correlations between aerosol MTF and the time integral of the wind strength have been

observed under dry soil conditions in experiments involving imaging through desert atmospheres in both horizontal and near-vertical directions. The time integral of the wind strength plays a large role in determining both the size distribution and the concentration of soil-derived particles that are airborne. Aerosol MTF is particularly relevant to image quality propagated through the atmosphere in areas where the soil is dry and bare, so that particulates can be uplifted by the wind. Once uplifted, the soil-derived particulates are carried by the wind to very distant locations, even thousands of miles.

Although the aerosol MTF was described above primarily from the standpoint of a desert atmosphere, recent evidence shows it to be quite relevant in nondesert atmospheres in which particulate radii also reach sizes on the order of micrometers or more under high humidity conditions.

Normal civilian vehicular use of gravel roads with relatively high silt content is also known to contribute significantly to airborne dust concentrations. Dust under combat conditions has been studied extensively with regard to its effect on optical transmission. Equation (23) suggests that under such conditions not only would image brightness be attenuated but image quality and resolution would be strongly degraded according to image wavelength and particulate sizes.

[Figure 3](#) suggests that in desert atmospheres under conditions where aerosols are present, distribution can peak at radii on the order of 2–4 μm ; image resolution in the 3–5 μm window in particular should be adversely affected. Furthermore, recent evidence indicates additional peaks for Sahara aerosol scatter at around 10- μm wavelength, thus suggesting that resolution also is impaired in the 8–14- μm window, although not to the same extent as in the 3–5 μm window. Another adverse factor to contend with in thermal imaging in the desert is the higher background radiance level in the daytime.

V. OVERALL RESOLUTION THROUGH THE ATMOSPHERE

To a first approximation, the overall atmospheric MTF can be considered as a cascading of all three forms of MTF [Eqs. (8) or (9), (22), and (23)]

$$\text{MTF}_{\text{atm}} = M_{\text{T}} M_{\text{B}} M_a \quad (24)$$

and is essentially depicted in [Fig. 8](#). It is assumed in [Fig. 10](#) that for spatial frequencies higher than f_{ac} , the roll-off derives from turbulence.

The multiplication of MTFs in Eq. (24) is only an approximation because those component MTFs are not necessarily independent. Winds, for example, cause changes

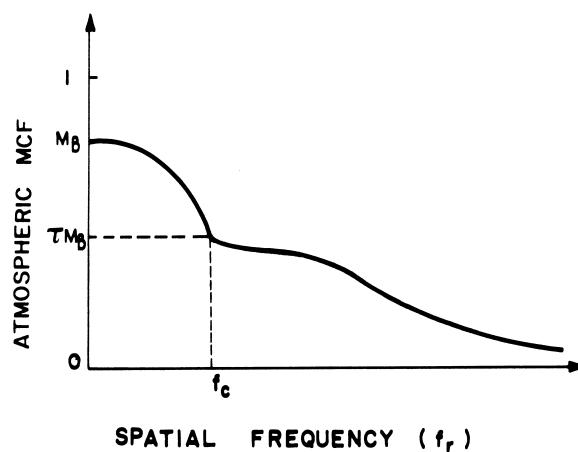


FIGURE 8 Overall atmospheric modulation transfer function to a first approximation.

in the atmospheric refractive index as well as in the spatial gradients of airborne particulate concentrations and sizes. Air temperature affects not only the atmospheric refractive index but also airborne particulate size through adsorption and absorption of water vapor, the latter process depending on relative humidity, which in turn depends on air temperature. Adsorption by aerosols increases atmospheric warming, thus increasing C_n^2 . Nevertheless, the approximation is quite accurate in most cases.

For horizontal imaging, turbulence will in general usually dominate overall atmospheric MTF at elevations near the ground, especially in the middle of hot summer days. At elevations of several meters and more above the ground, aerosol blur becomes more and more significant, and often becomes dominant at elevations on the order of 10–20 m and higher.

For vertical imaging, imaging downward from satellites is dominated by aerosol MTF, which is often referred to as the *adjacency effect*, since such small angle forward scatter by aerosols causes photons to be imaged in pixels adjacent to where they should have been imaged. Imaging upward from elevations near the ground is often dominated by turbulence blur. However, in general both aerosol and turbulence blur are usually evident, and should be considered in system design and in image restoration from atmospheric blur.

A. MTF

As described previously, turbulence gives rise to image degradation as a result of wavefront tilt, and to random image detail displacement deriving from random changes in refractive index of the propagation channel. Because of the randomness, a statistical rather than deterministic approach characterizes turbulence. However, forward light

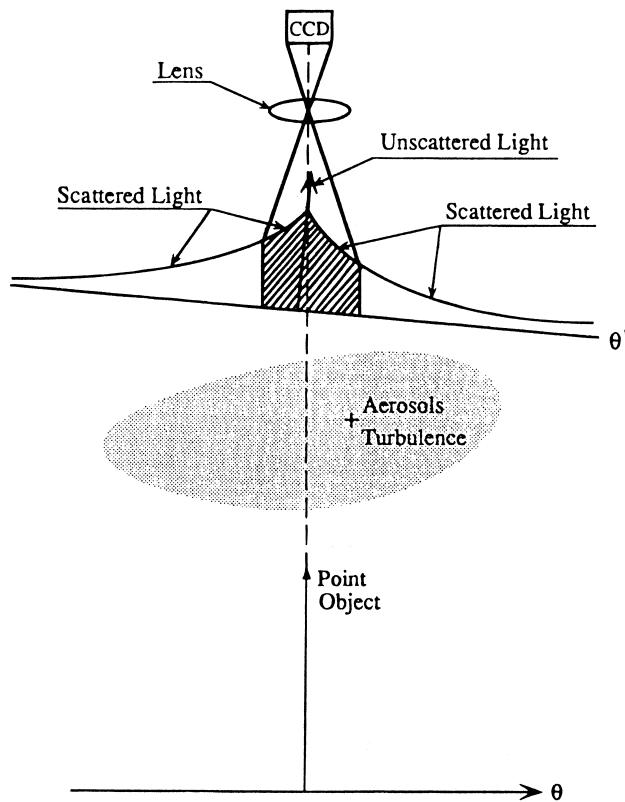


FIGURE 9 Propagation of angular point object through turbulent and scattering atmosphere to image sensor, such as charge-coupled array (CCD).

scatter by aerosols and atmospheric particulates in general cause relatively very little wavefront distortion and tilt. Rather, they cause broad diffusion of details in the propagating image, thus generating image blur in a different fashion. These latter processes are rather steady and lack the time dynamic properties of turbulence.

A general configuration describing the problem of imaging through the atmosphere is illustrated in Fig. 9.

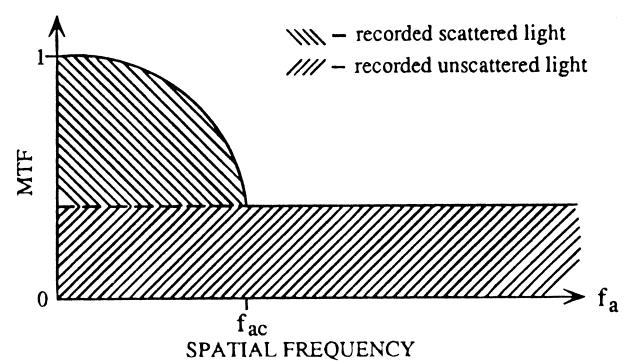


FIGURE 10 Shape of aerosol modulation transfer function (MTF).

In the object plane is the source of radiation which, in this example, is a point source producing a spatial delta function. The propagation channel is the atmosphere that is a random medium containing both turbulence and scattering and absorbing particles. At the other end of the atmospheric path is the imaging system, including both the optics and electronics. The main effect of the turbulent medium is to produce a wavefront tilt, as illustrated in Fig. 9, which causes image shifts in the image plane. Those tilts are so random, and their temporal power spectra are usually limited to several tens up to a few hundred hertz under ordinary atmospheric conditions. The image distortions caused by the wavefront tilt (typically on the order of tens or maybe hundreds of microradians) can be compensated either by adaptive optics techniques or by means of software (described below) if the exposure time is short enough, less than the characteristic fluctuation time (usually a few milliseconds). There is also a blur effect related to turbulence (inner scale wavefront distortions that displace image plane detail) not illustrated in this figure, which usually affects very high spatial frequencies. This blur effect can be characterized by the turbulence short exposure MTF, described previously. In addition, scintillations cause noticeable degradation of image quality as well, particularly for horizontal imaging near the ground.

In addition to turbulence, there are scattering and absorption effects produced by molecules and aerosols. Scattering gives rise to very wide diffusion of the point object radiation, as shown in Fig. 9. This causes a wide variance in angle of arrival according to the particulate scattering phase function, and thus also produces blur at the image plane. Unlike the short-exposure turbulence case, the blur due to aerosols is very wide, usually on the order of radians, which is typically orders of magnitude wider than the imaging system field of view.

An example would be the large blurred diffuse image of the moon through clouds. Even in clear weather, if one searches for the moon through a telescope he will find it by moving the instrument in the direction of increasing skylight or moonlight even if the moon is not within the field-of-view. Particulate scattering in the atmosphere is over a very broad angular region. The whole “sky” is simply scattered light—from the sun during the day and principally from the moon during the night. The image of a point is broadened greatly by such scatter. This implies a wide point-spread function (PSF). However, there are some practical instrumentation-based limitations that should be applied to the scattered light actually recorded in an image. The angular spread of the incident scattered light is on the order of several *radians*, and will usually be truncated by the imaging system field of view, which may be typically on the order of *milliradians*. Therefore, as shown in Fig. 9, not all the scattered light will reach

the sensor in the image plane. Furthermore, even if it does reach the sensor, the whole quantity of received scattered radiation will not necessarily be detected by it. The reason is that every sensor has a limited dynamic range, and scattered radiation of very low irradiance compared to the unscattered light will not be detected at all (Kopeika, 1998).

In addition, practical limitations must be applied also to the *unscattered* component of the radiation too. The finite angular spatial frequency bandwidth of the imager truncates the high angular spatial frequency radiation. This is because unscattered light from a point object (delta function) is equivalent to a constant of infinite spectrum in the angular spatial frequency plane. Limited spatial frequency bandwidth of the sensor limits the amount of unblurred or unscattered light actually recorded in the image from that incident to the imager, because the higher spatial frequency components are not within the spatial frequency bandwidth of the hardware. This causes broadening of the unscattered light PSF. Because of conservation of energy, such widening of the unscattered light PSF gives rise to a decrease in amplitude of unscattered light in the image, often very significantly. The scattered light PSF is so broad that it is hardly—if at all—affected by the spatial-frequency bandwidth limitations of the instrumentation. This causes the amplitudes of the scattered and unscattered light portions of the PSF to approach each other by orders of magnitude.

Consider Fig. 10. In Fig. 9 the unscattered light is a delta function with respect to the angular spatial domain, since it has no angular spread arising from scatter. Therefore, in the spatial frequency domain (Fig. 10) in accordance with its Fourier transform it is constant for all spatial frequencies up to, in principle, infinite spatial frequency. The scattered light of Fig. 9 (upper part) has wide angular spread, and the spatial frequency domain (Fig. 10) is a narrow function with cut-off angular frequency f_{ac} .

For limited absorption, a version of the practical instrumentation-based aerosol MTF approximate to the numerical calculation is

$$M_a(f_a) = \begin{cases} \exp[A_a L - S_a^* L (f_a/f_{ac})^2] & f_a \leq f_{ac} \\ \exp[-(A_a + S_a^*)L] & f_a > f_{ac} \end{cases}, \quad (23)$$

where $\exp[-(A_a + S_a^*)L]$ is the high spatial frequency asymptote of the practical aerosol MTF obtained either experimentally or by calculation (Kopeika, 1998). Coefficient S_a^* is an effective scattering coefficient that can be evaluated from the high spatial frequency practical aerosol MTF asymptote, the absorption coefficient A_a , and pathlength L . The practical aerosol MTF cut-off angular spatial frequency f_{ac} is approximately reciprocal to the limiting angle of light scatter actually recorded

in the image. It depends on aerosol size distribution, receiver field-of-view, dynamic range, and spatial frequency bandwidth (Kopeika, 1998) and may easily reach many tens of cycles per milliradian. However, for extreme aerosol loading conditions such as fog or dust storms, the instrumentation has little effect. For clear weather conditions, instrumentation has a dominant effect because it truncates the scattering pattern recorded in the image.

B. Absorption

Absorption by the atmosphere also affects aerosol MTF. As seen in Fig. 10, aerosol MTF is composed of two images superimposed. One is a sharp broad spatial frequency bandwidth image derived from unscattered light. The other is a blurred narrow spatial frequency bandwidth image generated by the received scattered light. When a photon interacts with a particulate (even a gas molecule or atom), the photon may be either absorbed or scattered. It cannot undergo both processes. Therefore, it is the sharp unscattered light image that is primarily subject to atmospheric absorption, rather than the blurred scattered light image. Thus, atmospheric absorption enhances the blur caused by small angle forward scatter by aerosols. Such phenomena are particularly important in thermal imaging, which is characterized by noticeable atmospheric absorption, and causes atmospheric blur to usually be caused primarily by aerosols rather than by turbulence. The high spatial frequency asymptote of aerosol MTF is equal approximately to atmospheric transmission, including both scattering and absorption effects, as seen in Eq. (23).

Since aerosol scatter of light does not exhibit time-dynamic properties, such as image dancing and scintillations, which are characteristic of turbulence, there is very little difference between long- and short-exposure aerosol MTFs.

VI. TECHNIQUES TO CORRECT FOR ATMOSPHERIC BLUR

There are basically two techniques used to correct for atmospheric degradation. One is adaptive optics, whose purpose is to prevent turbulence-derived distortions from being recorded in the image. The other technique is image restoration typically using digital computers. In the latter case, turbulence-derived blur has already been recorded in the image but is removed through image processing.

Adaptive optics is an expensive process not yet readily available, applicable primarily to astronomical imaging. Briefly, the concept in adaptive optics is to sense the wavefront tilt and distortion using an incoming laser beam or the image of a distant point source, such as a star. These make

it most convenient to apply when looking upward. Telescopes used in adaptive optics use mirrors as objectives. The thickness of each area of the mirror is varied electronically using transducers such as piezoelectric elements. The information obtained from the wavefront sensors is used to control piezoelectric elements so as to alter the surface of the mirror to conform to the incoming wavefront shape. In this way, if the wavefront from a distant point source such as a star would be planar because of the long distance but is distorted by turbulence, the mirror surface distortions by the piezoelectric elements are such that the wavefront recorded is now planar despite the fact that turbulence causes the incident wavefront to be nonplanar. Adaptive optics have been used successfully with large telescopes to produce very fine quality images through the atmosphere. To work successfully, adaptive optics requires each element in the mirror to be no larger than the isoplanatic patch size incident to the mirror. Adaptive optics are intended to prevent turbulence blur from being recorded in the image. This improves the measurement SNR at high spatial frequencies, and can then allow post-processing techniques such as digital image restoration in order to correct too for aerosol blur.

However, digital image restoration using an atmospheric Wiener filter can also be amazingly successful in removing turbulence blur. Restoration also has the advantage of being much cheaper and can be done with readily available personal computers. Essentially, all atmospheric blur, including that derived from both turbulence and aerosols, as well as path radiance is removed. A high-resolution image is obtained as if there were no atmospheric channel. Such image restoration is applicable to any contrast-limited imaging situation, including terrestrial imaging, in *any* imaging direction.

Use of the ordinary Wiener filter for correction of atmospheric blur is often not effective because, although aerosol MTF is rather deterministic, turbulence MTF is random. The atmospheric Wiener filter is one method for overcoming turbulence jitter, which is random changes of turbulence MTF from its average value. In this approach, the atmospheric MTF variance, which derives essentially from turbulence, is treated as an additional noise source since noise by definition is random. Aerosol MTF, on the other hand, since it is fairly constant as long as atmospheric conditions do not vary too much, contributes primarily to an average atmospheric MTF. Turbulence MTF changes with time due to its tilt jitter characteristic. These tilts are random and their temporal power spectra are usually limited to several 10, up to a few 100 Hertz under ordinary atmospheric conditions. The image distortions caused by overall atmospheric MTF can thus be regarded as the sum of a deterministic and a random filter. The deterministic filter includes aerosol and average turbulence MTFs,

while the random filter includes the noise component induced on the imaging system, both by the turbulence MTF variance and hardware. Stated mathematically, the atmospheric Wiener filter is determined by

$$s' = s + n_1, \quad (25)$$

where s' is the instantaneous atmospheric PSF, s is the average atmospheric PSF, and n_1 is an additive random component with zero expectation. Using this model, Eq. (25) is altered to yield the image received at the imaging system after propagating through the atmosphere as

$$i(x', y') = [s(x, y) + n_1(x, y)]^* o(x, y) + n_2(x, y), \quad (26)$$

where $*$ signifies convolution and n_2 is an additive noise imposed by the instrumentation, including camera, digitization, electronics, and so on, but not by the atmosphere. In other words, n_1 describes turbulence MTF random jitter, whereas n_2 describes the usual instrumentation noise. Fourier transforming Eq. (26) yields:

$$\mathbf{I}(f_x, f_y) = [\tau(f_x, f_y) + \mathbf{N}_1(f_x, f_y)] \cdot \mathbf{O}(f_x, f_y) + \mathbf{N}_2(f_x, f_y), \quad (27)$$

where \mathbf{I} , τ , \mathbf{O} , \mathbf{N}_1 , and \mathbf{N}_2 are Fourier transforms of i , s , o , n_1 and n_2 respectively. The received image thus is a sum of a deterministic part \mathbf{I}_1 and a random part \mathbf{N}

$$\mathbf{I} = \mathbf{I}_1 + \mathbf{N}, \quad (28)$$

where

$$\mathbf{I}_1(f_x, f_y) = \tau(f_x, f_y) \cdot \mathbf{O}(f_x, f_y) \quad (29)$$

and

$$\mathbf{N}(f_x, f_y) = \mathbf{O}(f_x, f_y) \cdot \mathbf{N}_1(f_x, f_y) + \mathbf{N}_2(f_x, f_y). \quad (30)$$

The atmospheric Wiener filter is defined similarly to the standard Wiener filter, differing by the noise component, which includes an additional term imposed by the turbulent atmosphere

$$\mathbf{M}(f_x, f_y) = \frac{|\tau(f_x, f_y)|^2}{\tau(f_x, f_y) \cdot \{|\tau(f_x, f_y)|^2\} + [\mathbf{S}_{n_1 n_1}(f_x, f_y) + \mathbf{S}_{n_2 n_2}(f_x, f_y)/\mathbf{S}_{n_o n_o}(f_x, f_y)]}, \quad (31)$$

where \mathbf{M} is the restoring filter, $\tau(f_x, f_y)$ is the average atmospheric MTF, $\mathbf{S}_{n_1 n_1}(f_x, f_y)$, $\mathbf{S}_{n_2 n_2}(f_x, f_y)$, and $\mathbf{S}_{o o}(f_x, f_y)$ are the power spectral densities of \mathbf{N}_1 , \mathbf{N}_2 and \mathbf{O} , and n_1 and n_2 are the inverse Fourier transforms of \mathbf{N}_1 and \mathbf{N}_2 in Eq. (30). (If there is no turbulence jitter and no instrumentation noise, then Eq. (31) assumes the form of a simple inverse filter.)

Assuming independence between aerosol and turbulence effects, the term $\tau(f_x, f_y)$ can be measured or calculated by a multiplication of the turbulence MTF (for either

the short- or long-exposure case) and aerosol MTF. Turbulence MTF can be evaluated with a knowledge of standard meteorological parameters using a C_n^2 prediction model (Kopeika, 1998) or IMTURB or PROTURB (U.S. Army Atmospheric Sciences Laboratory). The aerosol MTF can be evaluated (Kopeika, 1998) according to knowledge of particle-size distribution. The term $\mathbf{S}_{o o}(f_x, f_y)$ can be estimated by using the Fourier transform of the received image $\mathbf{I}(f_x, f_y)$. As with the standard Wiener filter, the term $\mathbf{S}_{n_2 n_2}(f_x, f_y)$ can be assumed to be constant for all spatial frequencies since the additive noise \mathbf{N}_2 is assumed to be white noise. This assumption is commonly used and very practical, and it has a relatively weak effect on the Wiener filter. The term $\mathbf{S}_{n_1 n_1}(f_x, f_y)$ is very important, since it includes the random part of the atmospheric distortions. One way of estimating $\mathbf{S}_{n_1 n_1}(f_x, f_y)$ is by a direct measurement. By using the relation

$$\mathbf{S}_{n_1 n_1}(f_x, f_y) = E\{\mathbf{N}_1^2(f_x, f_y)\} \quad (32)$$

and the Fourier transform of Eq. (25)

$$\mathbf{N}_1(f_x, f_y) = \tau'(f_x, f_y) - \tau(f_x, f_y), \quad (33)$$

it follows that $S_{n_1 n_1}$ equals the variance of the instantaneous atmospheric MTF

$$S_{n_1 n_1}(f_x, f_y) = E\{\tau'(f_x, f_y) - \tau(f_x, f_y)\}. \quad (34)$$

Here, τ' and τ represent instantaneous and average overall atmospheric MTFs.

Because the contribution to the random part of the atmospheric MTF is due mainly to turbulence rather than aerosols, in Eqs. (32)–(34) atmospheric MTF $|\tau(f_x, f_y)|$ refers to turbulence only. However, in Eq. (31), $|\tau(f_x, f_y)|$ includes aerosol MTF in addition to turbulence MTF, because it refers to the average atmospheric MTF, which is the product of the two. The variance of overall atmospheric MTF can be evaluated by calculating both terms

on the right-hand side of Eq. (33). This can be carried out by measuring a series of instantaneous atmospheric MTFs, and evaluating the average of both the MTF and its square. For example, in high-resolution astromonical imaging it has been found that short exposures that lack the temporal smear of long exposures can be combined in such a way that a useful signal-to-noise ratio is available at higher spatial frequencies. This technique is called *speckle interferometry*. It is, however, not a very practical way for *real-time* image restoration.

An alternative way of obtaining $\mathbf{S}_{n_1 n_1}(f_x, f_y)$ is to evaluate both terms inside the brackets on the right-hand side of Eq. (34). The second term τ is the square of the turbulence MTF, which can be predicted or measured. The first term $E\{\tau'^2\}$ was evaluated analytically to yield

$$\begin{aligned} E\{\tau'^2(f_x, f_y)\} &\propto \tau^2(f_x, f_y) \\ &\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tau^2(f'_x, f'_y) \tau(f'_x + f_x, f'_y + f_y) \\ &\times \tau(f'_x - f_x, f'_y - f_y) df'_x df'_y, \end{aligned} \quad (35)$$

which determines the expected value of the squared MTF, or in other words the OSF's power spectral density. The integral in Eq. (35) can be evaluated numerically with the use of the average turbulence MTF only. Hence, the input required is simply average turbulence and aerosol MTF at the time the image is recorded. These can be evaluated according to weather (Kopeika, 1998).

In the standard Wiener filter of Eq. (35), $\mathbf{S}_{nn}(f_x, f_y)$ refers to white noise only. Restorations with atmospheric Wiener filters can deblur essentially all atmospheric blur, and resolution is limited by hardware only. Whereas the conventional Wiener filter optimizes restoration at those spatial frequencies where signal-to-instrumentation noise is highest, the atmospheric Wiener filter optimizes restoration at those spatial frequencies where the *turbulence jitter is also minimum*. The improved atmospheric Wiener filter is thus *most advantageous when $\mathbf{S}_{n_1 n_1}$ is not negligible when compared to $\mathbf{S}_{n_2} \mathbf{S}_{oo}$ in Eq. (31)*.

The MTF after restoration is both broader and higher than that before restoration. The increase in spatial frequency bandwidth permits resolution of smaller details, as can be seen from Eqs. (6) and (7). The increase in MTF at higher spatial frequencies permits improvement in contrast of small detail. The system MTF broadening using the improved Wiener filter is considerable, particularly at low contrasts, and is limited by hardware MTF and spatial frequency bandwidth. Essentially, all atmospheric blur can be removed with atmospheric Wiener filter correction. The broadening indicates considerable increase in $f_{r\max}$ and therefore decrease in the size of resolvable detail.

VII. ACTIVE IMAGING

Active imaging systems usually involve scene illumination with a laser beam that scans the scene. Often, the laser beam is pulsed. At the receiving end, a computer then assigns each return pulse to its relevant image pixel. There are several advantages to active imaging.

1. Increased transmitter power can be used to overcome poor atmospheric transmission.

2. By using pulsed radiation, range information is obtained as well. Thus, the resulting computer image is three-dimensional.

3. By using range-gating techniques in which the detector is "closed" during times corresponding to returns from ranges closer than the object scene, path luminance background radiation is greatly reduced. This improves contrast, often considerably.

4. By using optical heterodyne detection if possible, essentially all scattered light is removed from the image because the efficiency of I-F conversion is drastically reduced as the angle between the received signal and local oscillator radiant powers increases. This means that aerosol MTF, deriving from scattered light recorded in the image, is of greatly reduced significance if heterodyne detection is feasible. Thus, active imaging can greatly improve image quality if the number of image pixels is large. At present, much effort is being invested in this direction, particularly as regards increasing the number of pixels.

A potential disadvantage of active imaging with lasers is multiple images of edges stemming from diffraction effects in the object plane, as well as speckle arising from diffraction effects by atmospheric particulates. However, such coherence effects are negligible if the atmospheric coherence diameter of the return beam is much smaller than the receiving aperture diameter. In this way, since the atmosphere affects coherence adversely, lower bounds for the receiving aperture diameter can be established so as to produce a better quality image deriving from incoherent light. If multimode lasers are used, which comprise the great bulk of commercial lasers, coherence length is typically only tens of centimeters, so that this too should pose no problem in obtaining incoherent active images.

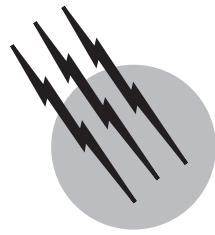
SEE ALSO THE FOLLOWING ARTICLES

AEROSOLS • ATMOSPHERIC DIFFUSION MODELING • ATMOSPHERIC TURBULENCE • FLOW VISUALIZATION • IMAGE RESTORATION, MAXIMUM ENTROPY METHOD • IMAGING OPTICS • MESOSCALE ATMOSPHERIC MODELING • RADIATION, ATMOSPHERIC • WAVE PHENOMENA

BIBLIOGRAPHY

Beland, R. R. (1993). Propagation through atmospheric optical turbulence. In "Atmospheric Propagation of Radiation," (F. G. Smith, ed.), pp. 157–232; also *In The Infrared and Electrooptical Systems*

- Handbook*, **2** (J. S. Accetta and D. L. Shumaker, exec. eds.). Environmental Research Institute of Michigan, Ann Arbor, and SPIE, Bellingham, Washington.
- Ben-Yosef, N., Tirosh, E., Weitz, A., and Pinsky, E. (1979). Refractive-index structure constant dependence on height. *J. Opt. Soc. Am.* **64**, 1616–1618.
- Biberman, L. M. (1973). “Perception of Displayed Information,” Plenum Press, New York.
- Dror, I., and Kopeika, N. S. (1995). Experimental comparison of turbulence MTF and aerosol MTF through the open atmosphere. *J. Opt. Soc. Am.* **12**, 970–980.
- Elterman, E. (1965). Atmospheric Optics. In “Handbook of Geophysics and Space Environment” (S. L. Valley, ed.). Sec. 7.1. U.S. Air Force Ambridge Research Labs, Cambridge, Massachusetts.
- Fried, D. L. (1966). Optical resolution through a randomly inhomogeneous medium for very large and very short exposures. *J. Opt. Soc. Am.* **56**, 1372–1379.
- Fried, D. L. (1966). Limiting resolution looking down through the atmosphere. *J. Opt. Soc. Am.* **56**, 1380–1384.
- Good, R. E., Beland, R. R., Murphy, E. A., Brown, J. H., and Dewan, E. M. (1988). Atmospheric models of optical turbulence. In “Modeling of the Atmosphere.” *Proc. SPIE 928*, SPIE, Bellingham, Washington, 165–186.
- Goodman, J. W. (1985). “Statistical Optics,” Wiley, New York.
- Ishimaru, A. (1978). “Wave Propagation and Scattering in Random Media, Vol. 1 and 2,” Academic Press, San Diego.
- Kneizys, F. X., Shettle, E. P., Abreu, L. W., Shetwynd, J. H., Anderson, G. P., Gallery, W. O., Selby, J. E. A., and Clough, S. A. (1988). “Users Guide to LOWTRAN 7.” Air Force Geophysical Laboratory, Environmental Research papers, No. 1010, AFGL-TR-88-0177, 16 August, 1988.
- Kopeika, N. S., and Brodagna, J. (1970). Background noise in optical communication systems. *Proc. IEEE* **58**, 1571–1577.
- Kopeika, N. S. (1984). Effects of aerosols on imaging through the atmosphere: a review of spatial frequency and wavelength dependent effects. *Opt. Eng.* **24**, 707–712.
- Kopeika, N. S. (1998). “A System Engineering Approach to Imaging,” SPIE Press, Bellingham, WA.
- Kopeika, N. S., Dror, I., and Sadot, D. (1998). The causes of 2 atmospheric blur: Comment on atmospheric scattering effect on spatial resolution of imaging systems. *J. Opt. Soc. Am.* **15**, 3047–3106.
- Miller, W. B., and Ricklin, J. C. (1990). IMTURB: A module for imaging through optical turbulence. Report ASL-TR-0221-27, U.S. Army Atmospheric Sciences Laboratory, White Sands Missile Range, New Mexico.
- Richter, J. H., and Hughes, N. G. (1991). Marine atmospheric effects on electro-optical systems performance. *Opt. Eng.* **30**, 1804–1820.
- Roggemann, M. C., and Welsh, B. (1996). “Imaging through Turbulence,” CRC Press, Boca Raton, FL.
- Sadot, D., Rosenfeld, R., Shuker, G., and Kopeika, N. S. (1995). High resolution restoration of images distorted by the atmosphere, based upon average atmospheric MTF. *Opt. Eng.* **34**, 1799–1807.
- Sadot, D., Shamriz, S., Sasson, I., Dror, I., and Kopeika, N. S. (1995). Prediction of overall atmospheric MTF with standard weather parameters: Comparison with measurements with two imaging systems, *Opt. Eng.* **34**, 3239–3248.
- Shettle, E. P., and Fenn, R. W. (1979). “Models for the Aerosols of the Lower Atmosphere and the Effects of Humidity Variations on their Optical Properties,” Air Force Geophysics Lab., Environmental Research Papers, No. 676, AFGL-TR-79-0214, 20 September, 1979.
- Zardecki, A., Gerstel, S. A. W., Tam, W. G., and Embury, J. F. (1986). Image-quality degradation in a turbid medium under partially coherent illumination. *J. Opt. Soc. Am. A.* **3**, 393–400.



Ionosphere

A. F. Nagy

University of Michigan

R. W. Schunk

Utah State University

- I. Background
- II. Basic Theory
- III. The Low- and Midlatitude Ionosphere
- IV. The High-Latitude Ionosphere
- V. Ionospheric Weather
- VI. Planetary Ionospheres

GLOSSARY

- Auroral oval** Oval-shaped region in the high-latitude atmosphere into which energetic particles from deep space penetrate.
- Conjugate ionospheres** Ionospheric regions on both sides of the hemispheres connected by a magnetic field line.
- Diffusion** Net plasma transport due to forces such as pressure gradient and gravity.
- D-region** Region of the ionosphere extending from about 70 to 90 km above the earth, where there are significant positive and negative ion populations.
- Dynamo electric field** Electric field created in the E and F regions by plasma that is dragged along with the neutral wind across magnetic field lines.
- E region** Region of the ionosphere extending from about 90 to 145 km above the earth, where molecular positive ions dominate.
- F region** Region of the ionosphere extending from about 145 to 1000 km above the earth, where atomic ions play a major role.

Ionopause Boundary separating the ionospheric plasma from the shocked and decelerated solar wind plasma; this transition region has also been called the contact discontinuity in cometary environments.

Midlatitude trough Region of low-electron densities located just equatorward of the nocturnal auroral oval.

Photochemistry Chemical processes influenced by sunlight.

Photoelectrons Electrons created in the ionosphere by photoionization.

Plasma Gas containing a significant fraction of free electrons and ions.

Plasma convection Large-scale motion of charged particles in the ionosphere, driven by electric fields created at high latitudes by the interaction of the solar wind with the geomagnetic field.

Polar cap High-latitude region of the atmosphere, poleward of the auroral oval, from which magnetic field lines extend deep into space.

Polar wind High-speed plasma outflow from the high-latitude ionospheres.

Solar wind Supersonic plasma flow from the sun.

THE IONOSPHERE is understood to be that region of the upper atmosphere where significant numbers of free electrons and ions are present. In general, the ionosphere of the earth is the region from about 70 to over 1000 km above the earth, where the electron and ion densities vary from below 10^3 to over 10^6 cm^{-3} .

I. BACKGROUND

The first suggestions for the presence of charged particles in the upper atmosphere were made more than 150 years ago. Gauss, Lord Kelvin, and Balfour Stewart hypothesized the existence of electric currents in the atmosphere to explain the observed variations of the magnetic field at the surface of the earth. In 1901, Marconi succeeded in sending radio signals across the Atlantic, which implied that the radio waves were deflected around the earth in a manner not immediately understood. The following year, working independently, Heaviside in England and Kennelly in the United States proposed that a layer of free electrons and ions in the upper atmosphere is responsible for the reflection of these radio waves. The first experimental proof of this reflecting layer did not come until 1925 when Appleton and Barnett demonstrated the existence of down-coming waves. Nearly simultaneously, Breit and Tuve devised a technique, known today as the ionosonde method, in which pulses of radio waves are vertically transmitted and the reflected signals analyzed; the electron densities present in the reflection region can be deduced from the characteristics of the received signal. The name *ionosphere* for the region of the upper atmosphere containing a significant population of free electrons and ions was coined by R. A. Watson-Watt in 1926. Experimental observations of the ionosphere were limited to remote sensing by radio waves until the end of World War II, when sounding rockets first became available to allow in situ measurements. The International Geophysical Year in 1959 provided the next large impetus to ionospheric research. The introduction of satellites and powerful ground-based radar systems capable of measuring a variety of the important parameters resulted in tremendous advances in our understanding of the physical and chemical processes that control the behavior of our terrestrial ionosphere. Beginning with the flyby of Venus by *Mariner 5* in 1965, the ionospheres of other bodies in the solar system also began to receive a great deal of attention.

II. BASIC THEORY

Ionosphere research during the last four decades has shown that the earth's ionosphere exhibits significant

variations with altitude, latitude, longitude, universal time, solar cycle, season, and geomagnetic activity. These variations are a consequence of the competition among the forces acting within and on the ionosphere. Of particular importance are the forces that result from the coupling to the denser neutral atmosphere. At high latitudes, the ionosphere also strongly couples to the overlying hot, tenuous plasma that extends deep into space. Consequently, before presenting the basic theory of ionospheric behavior, it is useful to describe the ionospheric environment and the average plasma conditions in the ionosphere.

A. Ionospheric Environment

The earth's magnetic field and the different flow regimes in the ionosphere are shown schematically in Fig. 1. The earth possesses a relatively strong intrinsic magnetic field that has a dipole character in the ionosphere. However, far from the earth the magnetic field configuration is distorted by the interaction of the earth's intrinsic field with the hot plasma that is continually emitted from the sun (solar wind). At high latitudes, the magnetic field lines extend deep into space in the antisunward direction. Along the so-called open field lines, ions and electrons are capable of escaping from the ionosphere in a process termed the polar wind, in analogy to the solar wind. This loss of plasma can have an appreciable effect on the temperature and density structure in the high-latitude ionosphere. Also, the hot plasma that exists in deep space is capable of penetrating to ionospheric altitudes at high latitudes, and this affects ambient densities and temperatures. In addition, the interaction of the solar wind with the earth's magnetic field sets up an electrical potential difference across

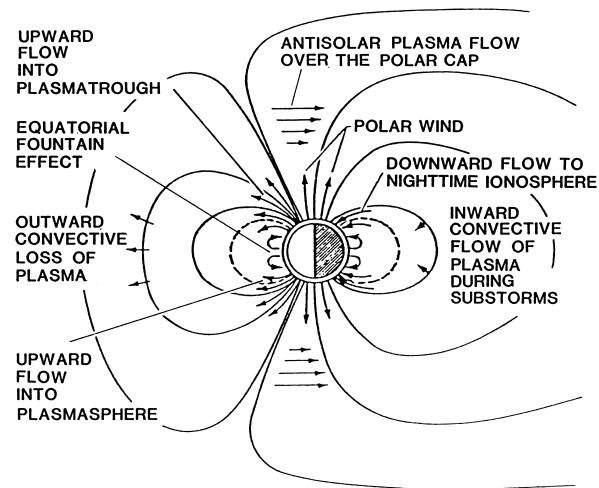


FIGURE 1 Schematic diagram showing the earth's magnetic field and the flow regimes in the ionosphere. (Adapted from a figure by C. R. Chappell.)

the polar region, and the resulting electric field causes the high-latitude ionospheric plasma to flow in an antisunward direction over the polar region.

At mid- and low-latitudes, the ionospheric plasma is not appreciably affected by external electric fields, and consequently the plasma tends to corotate with the earth. However, the plasma can readily flow along magnetic field lines like beads on a string. As a consequence, the plasma can escape from the topside ionosphere in one hemisphere, flow along geomagnetic field lines, and then enter the conjugate ionosphere. This interhemispheric flow of plasma is a source of ionization for one hemisphere and a sink for the other hemisphere.

At mid- and low-latitudes, the ionospheric plasma is strongly affected by the motion of the earth's upper atmosphere (neutral wind). In general, this neutral wind blows away from the subsolar point on the dayside and around to the nightside. The ionospheric plasma at midlatitudes is confined to move along magnetic field lines; therefore the meridional wind exerts a major influence on the ionosphere. On the dayside, this wind blows toward the poles, and the ionization is driven downward. On the nightside, the meridional wind blows toward the equator, and the ionization is driven up field lines.

At low latitudes, the geomagnetic field lines are nearly horizontal, which introduces some unique transport effects. First, the meridional neutral wind can very effectively induce an interhemispheric flow of plasma along geomagnetic field lines. At solstice, the dayside wind blows across the equator from the summer to the winter hemisphere. As the ionospheric plasma rises on the summer side of the equator, it expands and cools, while on the winter side it is compressed and heated as it descends.

Another interesting transport effect at low latitudes is the so-called equatorial fountain. In the daytime equatorial ionosphere, eastward electric fields associated with neutral wind-induced ionospheric currents drive a plasma motion that is upward. The plasma lifted in this way then diffuses down the magnetic field lines and away from the equator due to the action of gravity. This combination of electromagnetic drift and diffusion produces a fountain-like pattern of plasma motion.

B. Plasma Conditions

One of the early (1925–1930) discoveries in ionospheric research was that the ionosphere is stratified into regions (D, E, F₁, and F₂). This layered structure is shown schematically in Fig. 2 for typical daytime, midlatitude conditions. The corresponding neutral gas-density profiles are shown in Fig. 3. The E region (≈ 120 km alt.) is dominated by molecular species, with NO⁺, O₂⁺, and N₂⁺, being the major ions and N₂ and O₂ being the dominant neutrals.

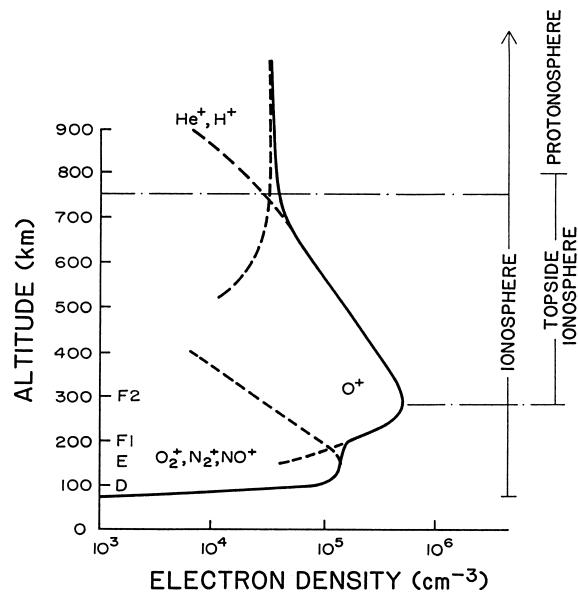


FIGURE 2 Schematic diagram of the midlatitude ionosphere showing the various regions: D, E, F₁, and F₂. (From Banks, P. M., Schunk, R. W., and Raitt, W. J. (1976). *Ann. Rev. Earth and Planet Sci.* **4**, 381–440.)

The total ion density is of the order of 10^5 cm^{-3} , while the neutral density is greater than 10^{11} cm^{-3} . Therefore the E region plasma is weakly ionized, and collisions between charged particles are not important. In the F region (≈ 200 –1000 km alt.), the atomic species dominate, with O⁺ and O being the major ion and neutral species, respectively. The peak ion density in the F region is roughly a factor of 10 greater than that in the E region, while the neutral density is about two to three orders of magnitude higher. The plasma in this region is partially ionized, and collisions between the different charged particles and between the charged particles and neutrals must be considered. The topside ionosphere is generally defined to be the region above the F region peak, while the protonosphere is the region where the lighter atomic ions (H⁺, He⁺) dominate. Although the neutrals still outnumber the ions in the protonosphere, the plasma is effectively fully ionized owing to the long-range nature of charged particle collisions. The D region (~ 70 –90 km alt.) differs from the E region in that both negative and positive molecular ions exist and three-body collisions are important.

The bulk of the following discussion will be devoted to the F region, where the main ionization peak occurs.

C. Photochemistry

Solar extreme ultraviolet (EUV) radiation photoionizes the neutral constituents of the upper atmosphere, producing free electrons and ions. The photoionization process

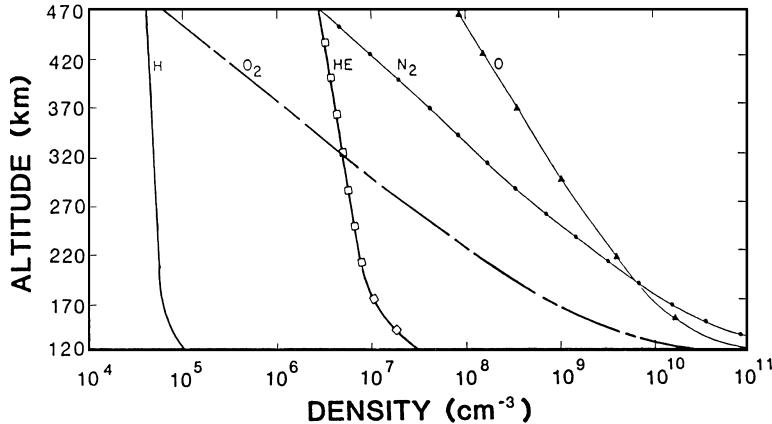
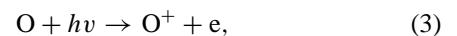
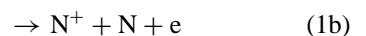
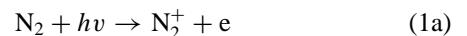


FIGURE 3 Altitude profiles of the earth's daytime neutral gas densities at midlatitudes. (From Schunk and Nagy, 1980; based on the empirical model of Hedin *et al.*, 1977.)

occurs predominantly at the lower levels of the ionosphere, where the neutrals are abundant. Typically, the peak in the ionization rate occurs at about 150 km, owing mainly to the absorption of radiation with wavelengths $\lambda < 796 \text{ \AA}$ (the ionization threshold of N_2). The ionization thresholds of the more common atmospheric species are given in [Table I](#). Photons (denoted as $h\nu$ in all equations appearing in this chapter), with wavelengths in the range

of 796 to 1027 \AA , penetrate down into the E region. For the E and F regions of the ionosphere, the most important photoionization processes are



where Eq. (1b) is produced with an efficiency of about 21%.

The calculation of ion production rates due to photoionization requires knowledge of the solar radiation flux incident upon the top of the atmosphere, the neutral gas densities as a function of altitude, and the absorption and ionization cross-sections of the neutral species as a function of wavelength. Generally, these quantities can be obtained from satellite, ground-based, and laboratory measurements.

Although the calculation of photoionization rates usually requires a computer, a simplified expression can be obtained for the F region, where the neutral atmosphere is optically transparent and atomic oxygen is the dominant neutral. For overhead sun, the O^+ production rate $P(\text{O}^+)$ is simply proportional to the atomic oxygen number density $N(\text{O})$,

$$P(\text{O}^+) = \beta N(\text{O}), \quad (4)$$

where β is a constant. However, in the F region, the atomic oxygen density decreases exponentially with altitude (see [Fig. 3](#)):

$$N(\text{O}) = N_0(\text{O}) \exp\left[-\frac{z - z_0}{H}\right], \quad (5)$$

TABLE I Ionization Threshold Potentials^a

Neutral constituent	Ionization potential	
	(eV)	(nm)
C	11.260	110.100
CH_4	12.550	98.790
CO	14.010	88.490
CO_2	13.770	90.040
H	13.600	91.160
H_2	15.430	80.350
H_2O	12.620	98.240
He	24.590	50.420
Mg	7.460	162.200
N	14.530	85.330
N_2	15.580	79.580
NH_3	10.170	121.900
NO	9.264	133.800
Na	5.139	241.300
O	13.620	91.030
O_2	12.060	102.800
OH	13.180	94.070
S	10.360	119.700

^a Information obtained from Rosenstock, N. H., et al., "Ion Energetic Measurements," Washington DC: U.S. Department of Commerce, National Bureau of Standards, 1980.

where z is altitude and $N_0(O)$ the density at some reference altitude z_0 . H , the neutral-gas scale height, is given by the following expression:

$$H = \frac{kT}{mg}, \quad (6)$$

where k is the Boltzmann constant, T the neutral-gas temperature, m the mass of the neutral gas, and g the acceleration due to gravity. Substituting Eq. (5) into Eq. (4) yields

$$P(O^+) = P_0(O^+) \exp\left[-\frac{z - z_0}{H}\right], \quad (7)$$

where $P_0(O^+) = \beta N_0(O)$, a constant. If the sun is not overhead but at an angle with respect to the zenith, then Eq. (7) becomes a function of the solar zenith angle χ .

After the ions are produced by photoionization, they can undergo chemical reactions with the neutral constituents in the upper atmosphere. These chemical reactions act to create and destroy ions. Some of the important chemical reactions in the E and F regions are



$$k_8 = 1.2 \times 10^{-12}$$



$$k_9 = 2.1 \times 10^{-11}$$



$$k_{10} = 5 \times 10^{-11}$$



$$k_{11} = 9.8 \times 10^{-12}$$



$$k_{12} = 1.3 \times 10^{-11},$$

where k_s is the rate constant for reactions in $\text{cm}^3 \text{ sec}^{-1}$. The chemical reaction rates are usually measured in the laboratory and are typically functions of temperature (cf. Schunk and Nagy, 2000). The values given above are for a representative temperature of 300 K.

The molecular ions can also recombine with electrons,



$$\alpha_{13} = 2.2 \times 10^{-7} (300/T_e)^{0.39}$$



$$\alpha_{14} = 1.95 \times 10^{-7} (300/T_e)^{0.7} \text{ (for } T_e < 1200 \text{ K)}$$

$$\alpha_{14} = 7.38 \times 10^{-8} (1200/T_e)^{0.56} \text{ (for } T_e > 1200 \text{ K)}$$



$$\alpha_{15} = 4.0 \times 10^{-7} (300/T_e)^{0.5},$$

where the recombination rates α_s are in units of $\text{cm}^3 \text{ sec}^{-1}$ and the electron temperature, T_e , dependence is specifically indicated. These recombination reactions are fairly

rapid, and they account for the main loss of ionization in the E region.

In the E and the lower F regions of the ionosphere, photochemical processes dominate, and the electron density can be calculated simply by equating local production and loss rates,

$$P_e = L_e N_e, \quad (16)$$

where N_e is the electron density and L_e is the loss frequency. For example, if NO^+ is the dominant ion in the E region, then

$$L_e = \alpha_{15} N_e \quad (17)$$

and therefore:

$$P_e = \alpha_{15} N_e^2. \quad (18)$$

D. Plasma Diffusion

In the F region the ionospheric plasma behavior is controlled by chemical processes below about 300 km, but at higher altitudes diffusive and other transport processes dominate. The electron density peaks in the region where the transition from chemical to transport control takes place; this density peak is called the F_2 peak. Above the F_2 peak where both wind-induced plasma drifts and magnetic field-aligned plasma diffusion are important in addition to photochemical reactions, it is not possible to calculate the electron density simply by equating local production and loss processes. In this case, a more general conservation equation governs the spatial and temporal variation of the electron density:

$$\frac{\partial N_e}{\partial t} + \frac{\partial (N_e U_e)}{\partial s} = P_e - L_e N_e, \quad (19)$$

where $\partial/\partial t$ is the time derivative, $\partial/\partial s$ the spatial derivative in the magnetic field direction, and U_e the bulk flow velocity of the electron gas along the magnetic field direction. Equation (19) indicates that in a given region of space, a temporal variation of the electron density occurs in response to electron production P_e , electron loss $L_e N_e$, and a nonuniform flow of electrons into or out of that region of space.

The field-aligned flow of electrons is influenced by gravity, the neutral wind, and density and temperature gradients. Owing to the small electron mass, the effect of gravity is to cause a charge separation, with light electrons trying to settle on top of the heavy ions. However, a charge-separation electric field develops that acts to prevent a large charge separation. Once this electric field develops, the ions and electrons move together as a single gas under the influence of gravity, the neutral wind, and the density and temperature gradients. Such a motion is called ambipolar diffusion. For motion along the

magnetic field, the diffusion equation takes the following form:

$$U_e = U_n - D_a \left(\frac{1}{N_e} \frac{\partial N_e}{\partial s} + \frac{1}{T_p} \frac{\partial T_p}{\partial s} + \frac{1}{H_p} \right), \quad (20)$$

where the ambipolar diffusion coefficient D_a , the plasma scale height, H_p , and the plasma temperature T_p are given by

$$D_a = \frac{k(T_e + T_i)}{M_i v_{in}} \quad (21)$$

$$H_p = \frac{2kT_p}{M_i g_s} \quad (22)$$

$$T_p = \frac{T_e + T_i}{2}, \quad (23)$$

where U_n is the component of the neutral wind along the magnetic field, g_s , the component of gravity along the magnetic field, k Boltzmann's constant, T_e the electron temperature, T_i the ion temperature, M_i the ion mass, and v_{in} the momentum transfer collision frequency between ions and neutrals. Note that in the F region the most abundant ion is O^+ , and the neutral is O. The collision frequency is given in Schunk and Nagy (2000). Here we merely note that $v_{in} \propto N(O)$; the more O atoms there are, the greater is the frequency of collisions between O^+ and O.

To obtain the variation of the electron density and bulk-flow velocity along the magnetic field, Eqs. (19) and (20) must be solved simultaneously, which generally requires a computer. However, at altitudes above the F region peak density, a simplified expression for N_e can be obtained. First, it is useful to express Eq. (20) in the form

$$\frac{1}{N_e} \frac{\partial N_e}{\partial s} = -\frac{1}{H_p} - \frac{1}{T_p} \frac{\partial T_p}{\partial s} + \frac{U_n - U_e}{D_a}. \quad (24)$$

Note that $D_a \propto (1/v) \propto [1/N(O)]$. Since $N(O)$ decreases rapidly with altitude (see Fig. 3), D_a increases rapidly with altitude. Consequently, above the F region peak, the last term in Eq. (24) is negligible. Also, above the F region peak, the plasma temperature is nearly constant. Therefore Eq. (24) reduces to

$$\frac{1}{N_e} \frac{\partial N_e}{\partial s} = -\frac{1}{H_p}. \quad (25)$$

If the small variation of gravity with altitude is ignored, Eq. (25) can be easily integrated to yield

$$N_e = [N_e]_r \exp \left[-\frac{s - s_r}{H_p} \right], \quad (26)$$

where the subscript r corresponds to some reference altitude. Equation (26) indicates that above the F region peak the electron density decreases exponentially with altitude, as shown in Fig. 2. The electron density variation given by Eq. (26) is called a diffusive equilibrium distribution.

E. Thermal Structure

The flow of energy in the earth's upper atmosphere is shown schematically in Fig. 4. The main source of energy for the ionosphere is EUV radiation from the sun. The absorption of EUV radiation by the neutral atmosphere results in both photoionization and excitation of the neutral gas. The resulting excited atoms and molecules lose their energy either by radiation or in quenching collisions with electrons and neutral particles. Photoionization produces energetic photoelectrons, since the energy carried by the ionizing photons, in general, exceeds the energy required for ionization. Typically, photoionization produces photoelectrons with initial energies of some 10s of electron volts.

Only a relatively modest amount of the initial photoelectron energy is deposited directly in the ambient electron gas. Most of the excess kinetic energy is lost in elastic and inelastic collisions with neutral particles and in Coulomb collisions with the ambient ions. If the photoelectrons lose their energy near the altitude at which they were produced, the heating is said to be local, whereas if the photoelectrons lose their energy over a distance greater than about a neutral scale height, the heating is termed nonlocal. Nonlocal heating effects occur mainly at high altitudes where ambient densities are low and at high photoelectron energies. Photoelectrons with sufficient energy can even escape from the ionosphere, travel along geomagnetic field lines, and deposit their energy in the conjugate ionosphere. Typically, photoelectron energy deposition above about 300 km constitutes nonlocal heating.

Shortly after creation a photoelectron undergoes a number of inelastic and elastic collisions. Briefly, for photoelectron energies greater than about 50 eV, ionization and optically allowed excitation of the neutral constituents are the dominant energy loss processes. At energies of about

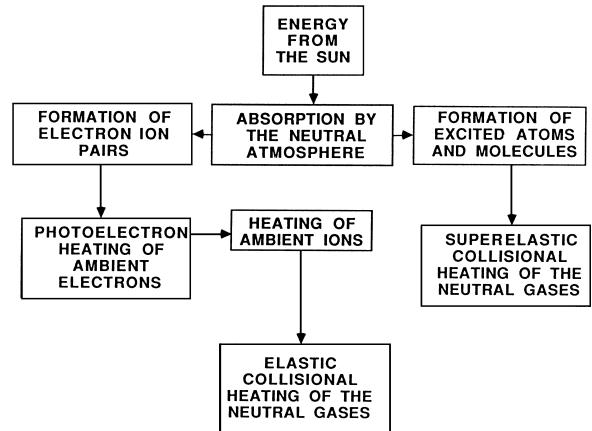


FIGURE 4 Block diagram showing the flow of energy in the earth's upper atmosphere.

20 eV, the energy loss via excitation of metastable levels of the major constituents is comparable to the energy loss through allowed transitions, becoming increasingly important as the energy decreases. At photoelectron energies below about 5 eV, energy loss through excitation of the vibrational levels of molecular neutral species, such as N₂, becomes important. Finally, below about 2 eV, energy loss to the ambient thermal electrons through elastic collisions is the dominant photoelectron energy loss process, although loss due to excitation of the rotational levels of molecules (e.g., N₂) is not entirely negligible.

Although photoelectrons are the primary source of heat for the ambient electron gas, other heat sources exist. At high latitudes, the hot electrons that exist in deep space can penetrate to ionospheric altitudes, and these hot electrons can heat the ambient electrons in a manner similar to that of the photoelectrons. However, these hot electrons reach ionospheric altitudes only in a narrow latitudinal band known as the auroral oval. (This region will be discussed in more detail later.) Another important heat source for the ambient electrons arises as a result of thermal conduction. As noted earlier, some of the photoelectrons can escape the ionosphere and hence can lose energy at high altitudes. This energy can then be conducted down along geomagnetic field lines to the ionosphere, thereby providing an additional heating mechanism.

A number of processes are effective in cooling the ambient electron gas. In the lower F region where the molecular species are abundant, rotational excitation of N₂ and O₂ and excitation of the fine structure levels of atomic oxygen are the most important cooling processes. However, at electron temperatures greater than about 1500 K, vibrational excitation of N₂ and O₂ and electronic excitation of O and O₂ have to be considered. At high altitudes, Coulomb collisions with the ambient ions are an important energy loss mechanism for thermal electrons.

At low altitudes, below about 250 km, the electron temperature can be obtained simply by equating local heating and cooling rates. However, above this altitude, electron thermal conduction is important, and the equation governing the electron temperature becomes more complicated (see Schunk and Nagy, 1978, 2000).

The primary heat source for the ion gas in the ionosphere is the ambient electron gas. Although additional ion heat sources exist, such as electric field heating, heating by exothermic chemical reactions, and frictional heating by means of neutral winds, these sources are usually characteristic of certain regions of the ionosphere and are seldom the primary heat source for the ions. The heat gained by the ions is sufficient to raise the ion temperature above the neutral gas temperature. The heated ions then lose energy through collisions with the colder neutrals. To a good approximation, the ion temperature in the E and

F regions can be calculated simply by equating the heat gained from the ambient electrons to the heat lost to the neutrals.

III. THE LOW- AND MIDLATITUDE IONOSPHERE

The midlatitude ionosphere has been extensively studied during the last four decades using rockets, satellites, and ground-based radar and optical facilities. These studies have shown that the ionosphere at midlatitudes displays a marked variation with altitude, local time, season, and solar cycle. Figure 5 shows altitude profiles of the ion densities in the daytime ionosphere, as measured by the Atmosphere Explorer C satellite. The density profiles were obtained by averaging 3 years of data gathered by a variety of instruments on the satellite. These data indicated that NO⁺ and O₂⁺ are the dominant ions below 175 km and that O⁺ dominates above this altitude. Note, however, that the O⁺ profile in the F region does not have a sharp peak as shown in the schematic diagram presented earlier (Fig. 2). The height of the density peak moves up and down depending on the strength of the neutral wind, and the effect of averaging data over several years is to broaden the peak. Altitude profiles obtained at a given instant of time by the incoherent scatter radar technique and by rockets are very similar to that shown in Fig. 2.

Altitude profiles of the electron and ion temperatures in the daytime midlatitude ionosphere are shown in Fig. 6. These profiles were obtained from a rocket flight in 1962 and were the earliest measurements that clearly showed that the electron temperature is greater than the ion temperature in the daytime ionosphere. As noted in Section II,

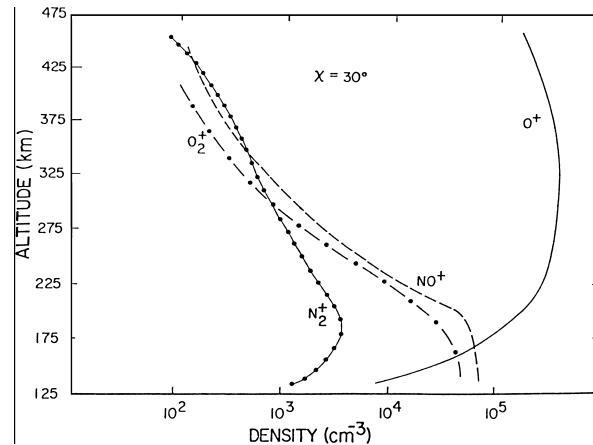


FIGURE 5 Ion density for the mid-latitude ionosphere at a solar zenith angle of 30 deg. (Courtesy D. G. Torr, private communication.)

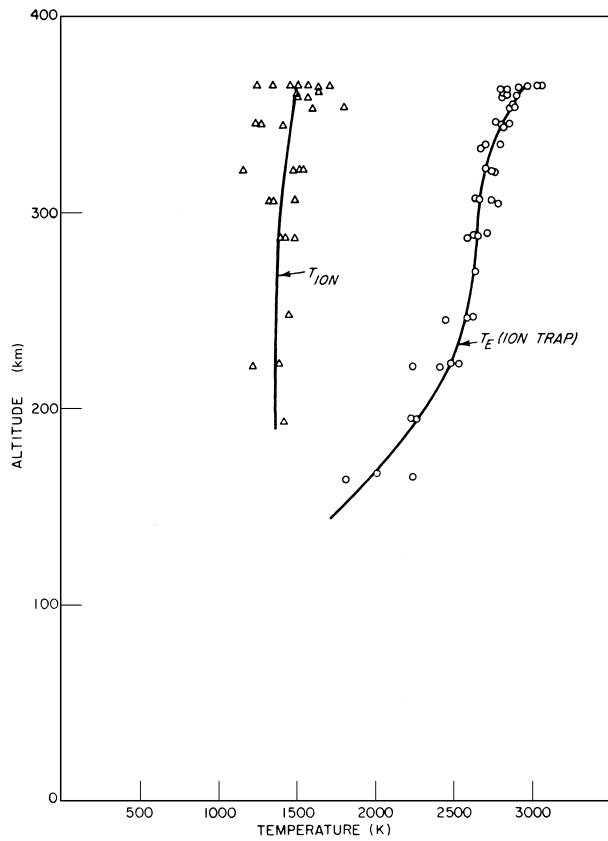


FIGURE 6 Electron (O) and ion (Δ) temperature profiles from the downleg portion of a daytime rocket flight at mid-latitudes. (From Nagy, A. F., Brace, L. H., Carignan, G. R., and Kanal, M. (1963). *J. Geophys. Res.* **68**, 6401.)

the energy lost by photoelectrons to the thermal electrons is the main heat source responsible for raising T_e above T_i .

Figures 7, 8, and 9 show the diurnal variations of the electron density, electron temperature, and ion temperature measured by the Millstone Hill incoherent scatter radar on March 23–24, 1970. Also shown in Fig. 7 is the height of the F region peak. The physical processes that control the diurnal variation of the electron density change with altitude and local time. After sunrise, the ionization increases rapidly from its nighttime minimum, due to photoionization. The ionization in the F region below about 300 km is under strong solar control, peaking at noon when the solar zenith angle is smallest and then decreasing symmetrically as the sun decreases. Ionization above 300 km, however, is influenced by other effects, such as neutral winds, electron and ion temperatures, neutral composition, and plasma flow into and out of the ionosphere. Therefore the electron density contours in this region do not show a strong solar zenith angle dependence, and the maximum ionization occurs late in the afternoon near the time when the neutral temperature peaks. At night, the

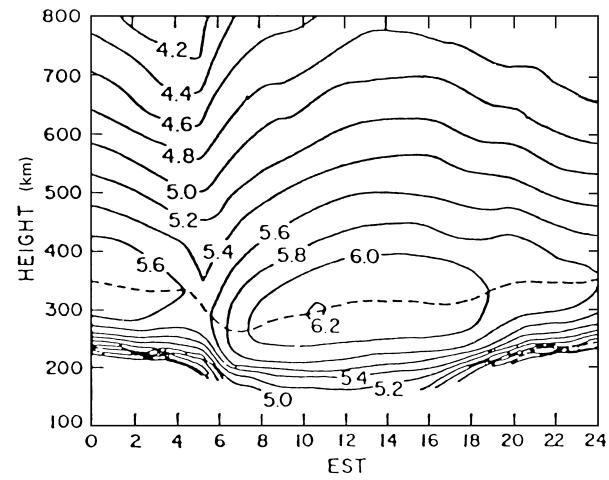


FIGURE 7 Contours of the electron density (N_e , cm^{-3}) measured by the Millstone Hill incoherent scatter radar on March 23–24, 1970. Dashed line is the height of the F region peak. (From Roble, R. G. (1975). *Planet. Space Sci.* **23**, 1017.)

electron density at the F region peak is controlled by several processes, including a downward-directed ionization flux, neutral winds, and ambipolar diffusion in the lower F region (below 300 km). However, the height of the F region peak is controlled primarily by the neutral winds forcing ionization up and down the inclined geomagnetic field lines. During the day, the wind is toward the poles, and the F region is driven downwards; at night the wind is toward the equator, and the F region is driven upwards. Consequently the F region peak is higher at night than during the day.

The ambient electrons are heated by photoelectrons that are created in the photoionization process and by a

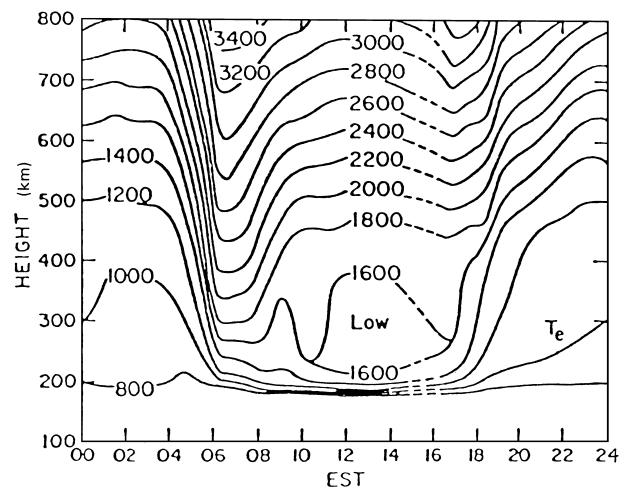


FIGURE 8 Contours of the electron temperature (K) measured by the Millstone Hill incoherent scatter radar on March 23–24, 1970. [From Roble, R. G. (1975). *Planet. Space Sci.* **23**, 1017.]

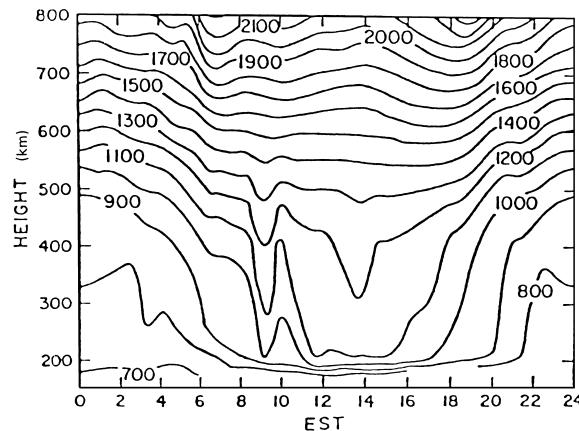


FIGURE 9 Contours of the ion temperature (K) measured by the Millstone Hill incoherent scatter radar on March 23–24, 1970. [From Roble, R. G. (1975). *Planet. Space Sci.* **23**, 1017.]

downward flow of heat from high altitudes. During the day, photoelectron heating dominates, and the electron temperature is greatest at noon when the solar zenith angle is the smallest. When the sun sets (18:00 to 19:00 local time), the electron temperature decreases rapidly due to the decrease in photoelectron production. At night, photoelectron heating is absent, and the electron temperature is maintained by a downward heat flow from high altitudes, which produces the positive gradient in the nocturnal electron temperature above 200 km.

The ions gain energy from the warmer electrons and lose energy to the colder neutrals. Below about 400 km, the ions strongly couple to the neutrals, and the ion temperature variation merely reflects the variation in the neutral temperature, which is not as dramatic as the electron temperature variation. Above 400 km, the ion temperature increases with altitude due primarily to the increased thermal coupling to the warm electrons; there is also a small downward ion heat flow from high altitudes. Note that above 400 km, the ion temperature displays a very small variation from day to night.

The dominant ionospheric features seen at low latitudes are the equatorial fountain and the Appleton density peaks. Near the equator, atmospheric winds in the E region generate electric fields by a dynamo action. These electric fields are transmitted along the curved geomagnetic field lines to the equatorial F region. During the daytime, these dynamo electric fields cause the equatorial F region to drift upward across magnetic field lines. The plasma lifted in this way then diffuses down the magnetic field lines and away from the equator because of the action of gravity. This combination of electromagnetic drift and diffusion produces a fountainlike pattern of plasma motion, as shown in Fig. 10.

A result of the equatorial fountain is that ionization peaks are formed in the subtropics on each side of the mag-

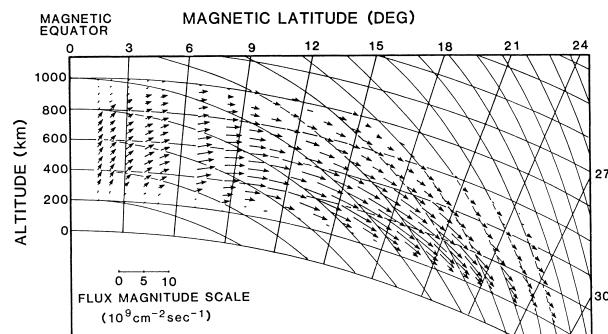


FIGURE 10 The pattern of plasma drift at low latitudes due to the combined action of an electromagnetic drift across magnetic field lines and plasma diffusion along field lines. The magnetic field lines are shown every 200 km above the equator. [From Hanson, W. B., and Moffett, R. J. (1966). *J. Geophys. Res.* **71**, 5559.]

netic equator. These so-called Appleton peaks are shown in Fig. 11 for 14:00 local time (the usual time for maximum peak development). At this time, the Appleton peaks are the most developed over the east Asia region. Typically, the Appleton peaks lie between 10° and 20° from the magnetic equator. With regard to diurnal variations, the peaks form around noon and disappear during the night.

IV. THE HIGH-LATITUDE IONOSPHERE

The ionosphere at high latitudes exhibits a more complex behavior than at low and midlatitudes owing to a variety of physical processes that are unique to this region.

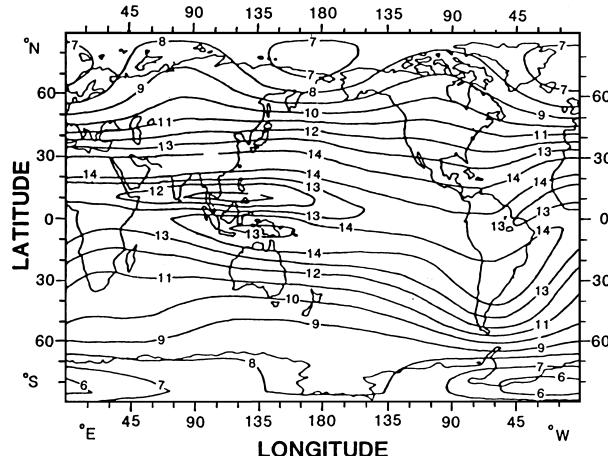


FIGURE 11 Contours of the ionospheric critical frequency foF_2 (in megahertz) as measured by the topside sounding satellite ISS-B. The F region peak electron density is proportional to $(foF_2)^2$. [From Matuura, N. (1979). "Atlas of ionospheric critical frequency (foF_2) obtained from ionosphere sounding satellite-b observation." Part 1, August to December 1978. Radio Research Lab., Ministry of Posts and Telecommunications, Japan.]

Foremost among these are the electric fields generated by the solar wind–geomagnetic field interaction, the presence of intense particle precipitation within the auroral oval, the polar wind escape of thermal plasma, and other features resulting from the auroral disturbance of the neutral atmosphere.

As noted earlier, the interaction of the solar wind with the earth's magnetic field generates electric fields that are mapped down along geomagnetic field lines to the high-latitude ionosphere. These electric fields cause the high-latitude ionosphere to drift (or convect) horizontally across the polar region. In the F region, the electric field-induced convection pattern is a two-cell pattern with antisunward flow over the polar cap and sunward flow at slightly lower latitudes. However, the ionospheric plasma at high latitudes also has a tendency to co-rotate with the earth. When this co-rotation velocity is added to the two-cell convection pattern, the resulting drift pattern for the F region plasma is as shown schematically in Fig. 12; this figure is a polar diagram with the magnetic pole at the center and magnetic local time displayed on the outer circle. The plasma flow is basically antisunward over the polar cap (white area poleward of the auroral oval) and sunward outside of this region. At dusk (18 magnetic local time [MLT]) and near 65° invariant latitude, the sunward convection of plasma driven by the solar wind–geomagnetic field interaction opposes the co-rotation velocity, and near stagnation of plasma occurs. This stagnation region plays

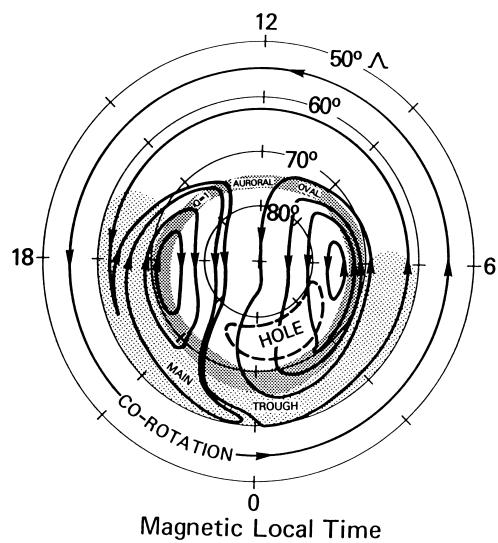


FIGURE 12 Schematic illustration of the polar ionosphere showing the plasma convection trajectories in the F region (solid lines) in a magnetic local time (MLT), magnetic latitude reference frame. Also shown are the locations of the high-latitude ionization hole, the main plasma trough, and the quiet time auroral oval. [From Brinton, H. C., Grebowsky, M. M., and Brace, L. H. (1978). *J. Geophys. Res.* **83**, 4767.]

an important role in the formation of the main electron density trough, which will be discussed below.

An important region shown in Fig. 12 is the auroral oval. In this region, energetic electrons from deep space can penetrate to ionospheric altitudes, and these energetic electrons can ionize the neutral atmosphere and heat the ambient electrons and ions. The exact effect that these energetic electrons have on the ionosphere depends not only on the intensity of these electrons but also on the length of time the ambient plasma spends in the auroral oval, since the plasma is continually circulating through this region.

The effect of the plasma motion (Fig. 12) on the electron density depends on both the speed of the flow and the seasonal conditions. For summer, most of the polar region is sunlit; consequently, the electron densities are elevated due to photoionization. In winter, on the other hand, most of the polar region is in darkness, and the electron densities tend to be low.

In Fig. 12, two electron density features are shown: a polar hole and the main trough. Both of these features occur in winter when the plasma flow speeds are low. The polar hole is a low-density, low-temperature region that is situated in the polar cap just poleward of the nocturnal auroral oval. This polar hole results from the slow antisunward drift of the plasma in the dark polar cap, during which time the ionosphere decays. The lowest densities are obtained just before the plasma enters the nocturnal auroral oval; the F region peak density in the polar hole can be as low as $5 \times 10^2 \text{ cm}^{-3}$. Upon entering the nocturnal auroral oval, the densities begin to build up owing to electron-ion production from precipitating energetic electrons. For a slow traversal through the oval, the F region peak density can increase to 10^5 cm^{-3} . The main electron density trough is situated just equatorward of the nocturnal auroral oval. This region has a limited latitudinal extent but is extended in local time (16–6 MLT). The trough is composed of plasma that has stagnated near dusk, decayed in the darkness, and then either co-rotated around the nightside or moved slowly back toward the dayside. The F region peak density in the trough can get as low as 10^3 cm^{-3} .

The electron density features shown in Fig. 12 occur only if the plasma flow speed is low. When the solar wind–geomagnetic field interaction is strong, large plasma flows can occur ($\sim 1 \text{ km/sec}$). In this case, the polar hole does not form, and instead a tongue of ionization extends over the polar cap, as shown in Fig. 13. In this figure, contours of the F region peak electron density $N_m F_2$ are plotted for the winter Southern hemisphere. The contours numbered 31, 45, and 61 on the dayside show the increase in ionization for a decreasing solar zenith angle. Near noon, the rapid antisunward flow of plasma carries the high dayside densities into the dark polar cap. Since the plasma flow

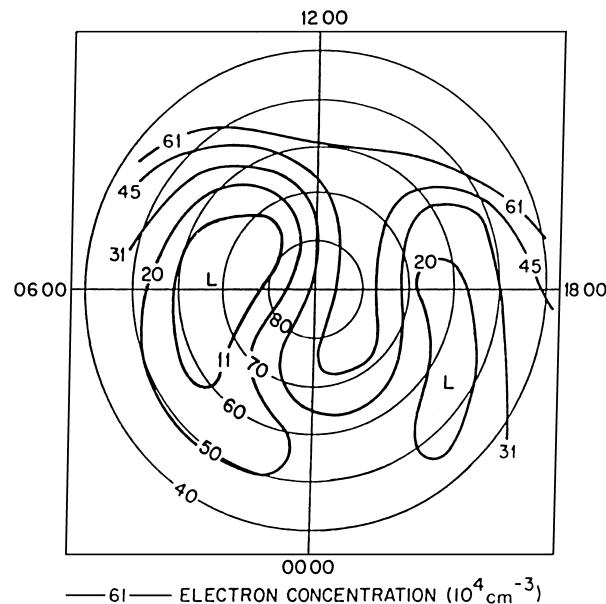


FIGURE 13 Synoptic NmF_2 contours for the Antarctic polar region in geographic coordinates. [From Knudsen, W. C., Banks, P. M., Winningham, J. D., and Klumpar, D. M. (1977). *J. Geophys. Res.* **82**, 4784.]

speed is high, the ionosphere does not have sufficient time to decay, and a tongue of ionization extends across the polar region.

V. IONOSPHERIC WEATHER

The basic ionospheric features discussed in the previous sections correspond to the climatology of the ionosphere. However, the terrestrial ionosphere can vary significantly from hour to hour and from day to day. It also displays a considerable amount of spatial structure. These weather features occur because the external forces acting on the ionosphere can be localized, spatially structured, and unsteady. In addition, there are time delays with regard to when external forces begin to act on the ionosphere and the subsequent ionospheric response. The density structure in the ionosphere typically varies from less than a meter to more than 1000 km, and it can appear in the form of propagating plasma patches, auroral blobs, plasma density holes, as well as density irregularities due to plasma instabilities. Such weather disturbances are particularly severe during geomagnetic storms and substorms, which occur when the solar wind and interplanetary magnetic field exhibit large temporal variations. The ionospheric weather disturbances can adversely affect numerous civilian and military systems, including over-the-horizon (OTH) radars, high-frequency (HF) communications, surveying and navigation systems that

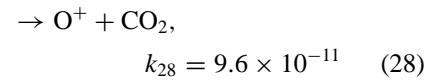
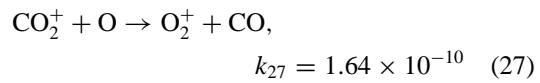
use global positioning system (GPS) satellites, surveillance, satellite tracking, power grids, and pipelines (e.g., Schunk and Sojka, 1996; and references therein).

VI. PLANETARY IONOSPHERES

Any body in our solar system that has a surrounding neutral gas envelope, due either to gravitational attraction (e.g., planets) or some other processes such as sputtering (e.g., Europa) or sublimation (comets), also has an ionosphere. The very basic processes of ionization, chemical transformation, and diffusive as well as convective transport are analogous in all ionospheres; the major differences are the result of the background neutral gas compositions, the nature or lack of a magnetic field, and the differences in some of the important processes (e.g., photo versus impact ionization). The remainder of this chapter describes the characteristics of the Venus ionosphere as a representative example of the so-called inner or terrestrial planets, the ionosphere of Jupiter as representative of the outer or major planets, and finally the ionosphere of Titan to represent one of the moons in our solar system.

A. Venus

The ionosphere of Venus is the most explored and best understood one in our solar system besides that of the earth. The atmosphere at the surface of Venus consists of approximately 96.5% CO_2 and 3.5% N_2 . Photodissociation of CO_2 results in atomic oxygen becoming the major atmospheric constituent above about 150 km. The behavior of the ionosphere of Venus is controlled by chemical processes below an altitude of about 180 km. This region of the ionosphere is analogous to the terrestrial E-region from the point of view that the main ion is molecular and is under chemical control. However, unlike the earth the maximum plasma density peaks near an altitude of 140 km (see Fig. 14), and is the result of a peak in the photoionization rate. Venus is an excellent example of the importance of chemical processes in establishing the nature of some of the important aspects of an ionosphere. The ion with the largest density is O_2^+ , and yet there is practically no neutral O_2 in the upper atmosphere. As mentioned earlier, the major neutral gas constituents in the upper atmosphere are CO_2 and O . The photoionization of these neutral gas species is followed by the reactions indicated below, which very effectively turn these initial ions into O_2^+ :



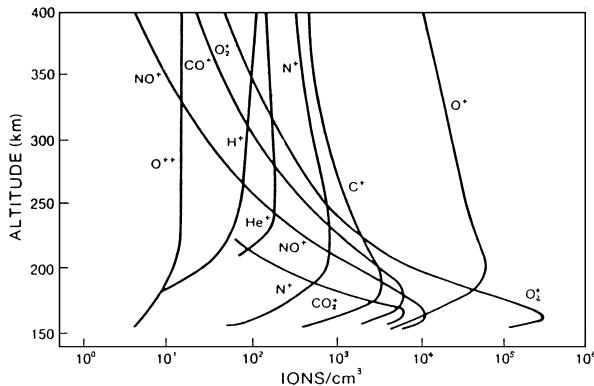
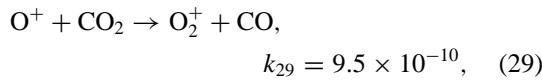


FIGURE 14 Ion densities measured by the ion mass spectrometer during one orbit of the Pioneer Venus orbiter spacecraft. [Bauer, S. J., Donahue, T. M., Hartle, R. E., and Taylor, H. A. (1979). *Science* **205**, 109.]



where the rate constants k_s are for a temperature of 300 K, and the units are $\text{cm}^3 \text{ sec}^{-1}$.

At altitudes above about 200 km, transport processes control the distribution of the electron and ion densities. Venus does not have an intrinsic magnetic field; therefore, horizontal transport at mid- and high latitudes is not impeded as is the case on the earth. The lack of an intrinsic field also means that processes associated with transport between magnetically conjugate ionospheres do not take place. The density of O^+ ions has a peak near 200 km, which is an F_2 type of peak, and is the major ion above this altitude, as shown in Fig. 14.

The lack of an intrinsic magnetic field also means that the solar wind must interact directly with the ionosphere/atmosphere system. A bow shock is formed and the shocked solar wind is deflected around the planet. There is a transition between the flowing shocked solar wind and the ionospheric plasma. This transition region, commonly called the ionopause, is generally narrow, of the order of a few 10s of km. The altitude of the mean ionopause height rises from about 350 km at the subsolar location to about 900 km at the terminator. Given that the ionopause is at an altitude where the ionospheric thermal pressure balances the solar wind dynamic pressure, its location must change as solar wind and ionospheric conditions change. For example, as the solar wind pressure increases the subsolar ionopause height decreases, but actually levels off at around 250–300 km, when the pressure exceeds about 4×10^{-8} dyne cm^{-2} .

The effective night on Venus lasts about 58 earth days (solar rotation period is 117 earth days), during which time the ionosphere could be expected to disappear, because no new photoions and electrons are created to replace the ones lost by recombination. Therefore it was very surprising, at first, when *Mariner 5* found a significant nightside ionosphere at Venus. Subsequent, extensive measurements have confirmed the presence of a significant, but highly variable nightside ionosphere with a peak electron density of around $2 \times 10^4 \text{ cm}^{-3}$. Plasma flows from the dayside, along with impact ionization, caused by precipitating electrons, are responsible for the observed nighttime densities; their relative importance for a given ion species depends on the solar wind pressure and solar conditions.

Observed solar cycle maximum ion and electron temperatures are shown in Fig. 15. These plasma temperatures are significantly higher than the neutral gas temperatures and cannot be explained in terms of EUV heating and classical thermal conduction, as is the case for the mid-latitude terrestrial ionosphere. The two suggestions that lead to model temperature values, consistent with the observations, are (1) an ad hoc energy input at the top of the ionosphere and/or (2) reduced thermal conductivities. The latter causes reduced downward heat flow and the eventual energy loss to the neutrals at the lower altitudes. There are reasons to believe that both mechanisms are present, but there is insufficient information available to establish, which is dominant, when and why. The model fit to the data, shown in Fig. 15, is achieved by assuming reasonable, but ad hoc, topside heat inflows into the ionosphere. Good summaries of our present knowledge of the Venus ionosphere can be found in Schunk and Nagy (2000), Fox and Kliore (1997), and Nagy and Cravens (1997).

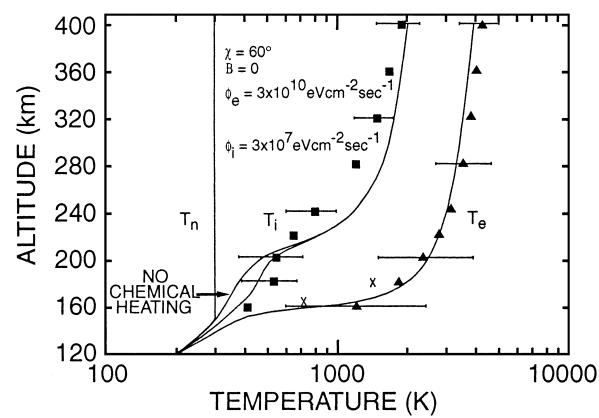


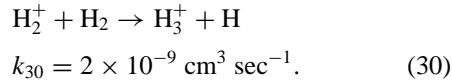
FIGURE 15 Measured (solid squares and triangles) and calculated (solid lines) electron and temperatures for zero magnetic field and 60° solar zenith angle. The assumed heat inputs at the upper boundary are indicated in the Figure. [Cravens, T. E., et al., (1980). *J. Geophys. Res.* **85**, 7778.]

B. Jupiter

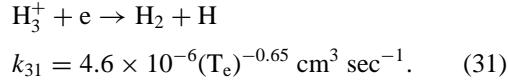
The upper atmosphere of Jupiter consists mainly of molecular hydrogen and some lesser amounts of helium and atomic hydrogen; the latter becomes dominant at the very high altitudes. There is also methane and other hydrocarbons present at the lower altitudes. Jupiter has no solid surface; therefore, the altitude scales are referred to a given pressure, usually that of 1 mb (millibar; bar = $10^5 N m^{-2}$). Presently the only available direct information regarding the ionosphere of Jupiter is based on the *Pioneer 10* and *11*, *Voyager 1* and *2* and *Galileo* radio occultation measurements. Some indirect information, mainly auroral remote sensing observations, also provide some insight into certain ionospheric processes.

Given that Jupiter's upper atmosphere consists mainly of molecular hydrogen, the major primary ion, which is formed by either photoionization or particle impact, is H_2^+ . H^+ ions are also created by either dissociative ionization of H_2 or by direct ionization of atomic hydrogen. At high altitudes, H^+ can only recombine directly via radiative recombination, which is a very slow process. It was suggested some time ago that H^+ could charge exchange, with H_2 excited to a vibrational state $v > 4$. The vibrational distribution of H_2 is not known, but recent calculations indicate that the vibrational temperature is elevated at Jupiter, but it is not clear how important this effect is.

H_2^+ is very rapidly transformed to H_3^+ , especially at the lower altitudes where H_2 is dominant:



H_3^+ is likely to undergo dissociative recombination:



Significant uncertainties have been associated with the dissociative recombination rate of H_3^+ . However, recent measurements have shown that the rate is rapid, even if the ion is in its lowest vibrational state. The main primary ions in the topside ionosphere can be rapidly lost by reactions with upflowing methane. However, the importance of this process depends on the rate at which methane is transported up from lower altitudes, which in turn depends on the eddy diffusion coefficient, which is not well known. Direct photoionization of hydrocarbon molecules at the lower altitudes can lead to a relatively thin, about 50 km broad, hydrocarbon ion layer around 300 km.

The early, hydrogen-based models predicted an ionosphere, which is predominantly H^+ , because of its long lifetime ($\sim 10^6$ sec). In these models H^+ is removed by downward diffusion to the vicinity of the homopause

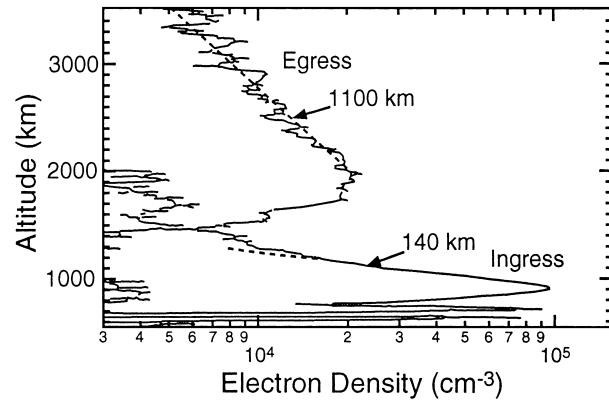


FIGURE 16 Measured electron density profiles of Jupiter's ionosphere near the terminator. [Hinson, D. P., et al. (1997). *Geophys Res. Lett.* **24**, 2107.]

{ ~ 1100 km}, where it undergoes charge exchange with heavier gases, mostly hydrocarbons such as methane, which in turn are lost rapidly via dissociative recombination. The *Voyager* and more recently the *Galileo* electron density profiles indicated the presence of an ionosphere, with peak densities between 10^4 and 10^5 cm^{-3} , as indicated in Fig. 16. These electron density profiles seem to fall in two general classes: one where the peak electron density is located at an altitude around 2000 km, and another group with the electron density peak near 1000 km. The two groups also exhibit different topside scale heights, with the high-altitude peaks associated with the larger scale heights. There appears to be no clear latitudinal nor temporal association with these separate group of profiles. The different peaks may be the result of a combination of different major ionizing sources (EUV versus particle impact) and different ion chemistries. A number of different models of the ionosphere have been developed since the *Voyager* encounters. The limited observational database, combined with the large uncertainties associated with such important parameters as the relevant reaction rates, drift velocities, degree of vibrational excitation, and the magnitude and nature of the precipitating particles means that there are too many free parameters to allow a unique and definitive model of the ionosphere to be developed.

C. Titan

Titan, the largest satellite of Saturn, is surrounded by a substantial atmosphere and therefore one expects a correspondingly significant ionosphere. To date the only opportunity for a radio occultation measurement of an ionosphere occurred when *Voyager 1* was occulted by Titan. The initial analysis of that data could only provide upper limits of $3 \times 10^3 \text{ cm}^{-3}$ and $5 \times 10^3 \text{ cm}^{-3}$ on the peak electron densities at the evening and morning terminators,

respectively. However, a careful reanalysis of the data indicates the presence of an electron density peak of about $2.7 \times 10^3 \text{ cm}^{-3}$ at around 1190 km for a solar zenith angle of near 90° .

The UV emissions from Titan's upper atmosphere appears to indicate that the electron energy deposition rate is significantly greater than the solar EUV rate. However, the situation is not totally clear, because the UV emissions are not observed on the darkside of the atmosphere. Thus the main ionization sources that may be responsible for the formation of Titan's ionosphere are solar extreme ultraviolet radiation, photoelectrons produced by this radiation, and magnetospheric electrons, and therefore the calculations to date have concentrated on EUV and magnetospheric electron impact ionization. Calculations indicate that photoionization is the main source for the dayside ionosphere, followed by photoelectron impact and finally magnetospheric electron sources. Of course magnetospheric electrons must dominate in the nightside ionosphere.

A variety of one-dimensional, photochemical calculations have been carried out, and they all lead to electron density values consistent with the *Voyager* results. Until recently it was believed that the major ion is HCN^+ . The most important initial ion is N_2^+ up to about 1800 km and CH_3^+ at the higher altitudes; N^+ and CH_4^+ is also an important initial ion. These initial ions quickly undergo a number of ion-neutral reactions leading to HCN^+ , which then will undergo disssociative recombination or proton transfer, leading to more complex hydrocarbon ions. Figure 17 shows the results of a representative set of calculations.

A variety of different studies looked at the issue of the transition from chemical to transport control in the ionosphere. Simple time constant considerations, as well as more detailed model solutions, have indicated that the transition from chemical to diffusive control takes place in the altitude region around 1500 km. The magneto-

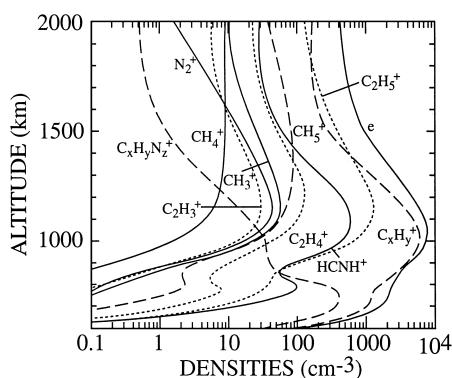


FIGURE 17 Calculated ion densities for the dayside of Titan. [Fox, J. L., and R. V. Yelle (1997). *Geophys. Res. Lett.* **24**, 2179.]

spheric plasma velocity (\sim 120 km/s) is subsonic (sound speed is \sim 210 km/s) and superalfvenic (Alfven speed is \sim 64 km/s); therefore no bow shock is formed, and the plasma is gradually slowed, as it enters Titan's exosphere, by massloading. The magnetic field strength increases, piles up, and eventually drapes around Titan. This piled-up magnetic field, similar to the so-called magnetic barrier at Venus, is expected to be the dominant source of pressure against the ionosphere.

The question of plasma temperatures in the ionosphere of Titan has also been studied. No observational constraints concerning these parameters exist; therefore, at best one can set some range of reasonable values through model calculations. It is expected that the temperatures will be very different on the ramside from those on the wakeside of Titan. This comes about because the draped magnetic field on the ramside is expected to be nearly horizontal, thus reducing vertical heat flow, whereas on the wakeside the field is expected to be nearly radial. The ion and electron temperatures have been estimated to be around 400 and 1000 K in the wakeside, respectively, and in the range of 300 to thousands of degrees on the ramside, respectively.

SEE ALSO THE FOLLOWING ARTICLES

AURORA • GEOMAGNETISM • PHOTOCHEMISTRY BY VUV PHOTONS • PLANETARY ATMOSPHERES • RADIO PROPAGATION • SOLAR PHYSICS • SOLAR SYSTEM, MAGNETIC AND ELECTRIC FIELDS • SOLAR TERRESTRIAL PHYSICS • SPACE PLASMA PHYSICS

BIBLIOGRAPHY

- Atreya, S. K., Waite, J. H., Donahue, R. M., Nagy, A. F., and McConnell, J. C. (1984). In "Saturn" (T. Gehrels and M. S. Mathews, eds.). University of Arizona Press, Tucson.

Atreya, S. K. (1986). "Atmospheres and Ionospheres of the Outer Planets and Their Satellites," Springer-Verlag, New York.

Bauer, S. J. (1973). "Physics of Planetary Ionospheres," Springer-Verlag, New York.

Brace, L. H., and Kliore, A. J. (1991). *Space Sci. Rev.* **55**, 81.

Brace, L. H., Gombosi, T. I., Kliore, A. J., Knudsen, W. C., Nagy, A. F., and Taylor, H. A. (1983). In "Venus" (D. M. Hunten, L. Colin, and T. M. Donahue, eds.), pp. 779–840. University of Arizona Press, Tucson.

Chamberlain, J. W., and Hunten, D. M. (1987). "Theory of Planetary Atmospheres," Academic Press Inc, New York.

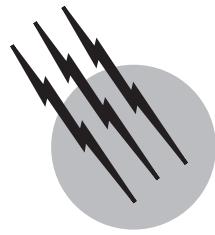
Cravens, T. E. (1987). *Adv. Space Res.* **7**(12), 147.

Cravens, T. E., and Nagy, A. F. (1983). *Rev. Geophys. and Space Phys.* **21**, 263.

Cravens, T. E., et al. (1980). *J. Geophys. Res.* **85**, 7778.

Fox, J. L., and Kliore, A. J. (1997). In "Venus II" (S. W. Bougher, D. M. Hunten, and R. J. Phillips, eds.), The University of Arizona Press, Tucson.

- Fox, J. L., and Yelle, R. V. (1997). *Geophys. Res. Lett.* **24**, 2179.
- Kar, J. (1996). *Space Sci. Rev.* **77**, 193.
- Kelly, M. C. (1989). "The Earth's Ionosphere," Academic Press, San Diego.
- Mahajan, K. K., and Kar, J. (1988). *Space Sci. Rev.* **47**, 193.
- Majeed, T., and McConnell, J. C. (1991). *Planet. Space Sci.* **39**, 1715.
- Mendis, D. A., Houpis, H. L. F., and Marconi, M. L. (1985). *Cosmic Phys.* **10**, 1–380.
- Miller, K. L., and Whitten, R. C. (1991). *Space Sci. Rev.* **55**, 165.
- Nagy, A. F. (1987). *Adv. Space Res.* **7**(12), 89–99.
- Nagy, A. F., and Cravens, T. E. (1997). In "Venus II" (S. W. Bougher, D. M. Hunten, and R. J. Phillips, eds.). University of Arizona Press, Tucson.
- Nagy, A. F., Cravens, T. E., and Gombosi, T. I. (1983). In "Venus" (D. M. Hunten, L. Colin, T. M. Donahue, and V. I. Moroz, eds.), pp. 841–872, University of Arizona Press, Tucson.
- Nagy, A. F., Cravens, T. E., and Waite, J. H. (1995). *Rev. Geophys., Suppl.*, 525.
- Schunk, R. W., and Nagy, A. F. (1978). *Rev. Geophys. and Space Phys.* **16**, 355–399.
- Schunk, R. W., and Nagy, A. F. (1980). *Rev. Geophys. Space Phys.* **18**, 813–852.
- Schunk, R. W., and Nagy, A. F. (2000). "Ionospheres," Cambridge University Press, Cambridge.
- Schunk, R. W., and Sojka, J. J. (1996). *J. Atmos. Terr. Phys.*, 1527.
- Strobel, D. F., and Atreya, S. K. (1983). "Physics of the Jovian Magnetosphere" (A. J. Dessler, ed.). Cambridge University Press, Cambridge.
- Waite, J. H., and Cravens, T. E. (1987). *Adv. Space Sci.* **7**(12), 119.



Mesoscale Atmospheric Modeling

Roger Pielke, Sr.

Colorado State University

- I. Model Equations
- II. Model Assumptions
- III. Grid-Volume Averaging
- IV. Coordinate Transformations
- V. Grid Meshes
- VI. Parametrizations
- VII. Solution Techniques
- VIII. Initial Top and Lateral Boundary Conditions
- IX. Surface Boundary Conditions
- X. Model Performance Evaluation
- XI. Examples of Models

GLOSSARY

Boussinesq approximation Approximation in which fluctuations of density are ignored except when multiplied by gravity.

Conservation equation Equation that describes the local change of a quantity in terms of resolvable and subgrid-scale advection and sources and sinks.

Grid-volume averages Integral over space and time of the model grid increments.

Hydrostatic Pressure uniquely determined by the weight of the air overhead.

Parametrization Representation of a physical process which is not a fundamental basic physics concept.

Resolution In a model, at least four grid increments in each spatial direction.

ATMOSPHERIC MESOSCALE SYSTEMS are identified as those in which the instantaneous pressure field can be determined accurately by the temperature field, but the winds, even in the absence of surface frictional effects, are out of balance with the horizontal pressure gradient force. The pressure field under this situation is said to be hydrostatic. Larger scale atmospheric features, in contrast, have a wind field that is close to a balance with the horizontal pressure gradient force. These large-scale winds are said to be near “gradient wind balance.”

Atmospheric features which are smaller than the mesoscale have pressure fields in which wind acceleration is a significant component (which is referred to as the dynamic wind). The pressure gradient which causes this dynamic wind is called the nonhydrostatic pressure.

Atmospheric mesoscale models are based on a set of conservation equations for velocity, heat, density, water, and other trace atmospheric gases and aerosols. The equation of state used in these equations is the ideal gas law. The conservation-of-velocity equation is derived from Newton's second law of motion ($\vec{F} = m\vec{a}$) as applied to the rotating earth. The conservation-of-heat equation is derived from the first law of thermodynamics. The remaining conservation equations are written as a change in an atmospheric variable (e.g., water) in a Lagrangian framework where sources and sinks are identified.

I. MODEL EQUATIONS

Each of these conservation equations can be written as

$$d\phi/dt = S_\phi,$$

where $\phi = \vec{V}, \theta, q_i, X_n$, and M for the velocity, potential temperature (entropy), water in its three phases, other atmospheric gases and aerosols, and mass, respectively. In atmospheric science, the conservation equation of mass is written in the form of the conservation of density ρ . The source/sink terms S_ϕ for each conservation equation include the following:

- $S_{\vec{V}}$: pressure gradient force, gravity, molecular diffusion of air motion
- S_θ : change of heat due to phase change of water, radiative flux divergence, change of heat due to chemical reactions, molecular diffusion of heat
- S_{q_i} : change of phase of water, molecular diffusion of water
- S_{X_n} : change of phase, chemical changes, molecular-scale diffusion
- S_ρ : zero

Models seldom express the conservation relations in a Lagrangian framework. The chain rule of calculus is used to convert to an Eulerian framework,

$$\frac{\partial \phi}{\partial t} = -\vec{V} \cdot \vec{\nabla} \phi + S_\phi,$$

where $\partial \phi / \partial t$ is the local tendency.

II. MODEL ASSUMPTIONS

Several assumptions are typically made in these conservation equations beyond those that are implicitly present

already (such as the validity of the ideal gas law). These include:

1. The neglect of small-scale fluctuations of density except when multiplied by gravity (the so-called Boussinesq approximation).
2. The neglect of vertical acceleration and the Coriolis effect relative to the differences between the vertical pressure gradient force and gravity (the hydrostatic assumption).
3. The neglect of all molecular transfers.

Assumptions 1 and 2 have not been made in recent years in the models, however, since the numerical equations are actually easier to solve without these assumptions. Nonetheless, the spatial and temporal scales of mesoscale systems result in the two assumptions being excellent approximations with respect to mesoscale-size systems. Assumption 3 is justified since the advection of ϕ is much more significant in transferring ϕ than is molecular motion on the mesoscale.

III. GRID-VOLUME AVERAGING

These conservation relations, which are written as a set of simultaneous, nonlinear differential equations, unfortunately cannot be used without integrating them over defined volumes of the atmosphere. These volumes are referred to as the model grid volumes. The region of the atmosphere for which these grid volumes are defined is called the model domain. The integration of the conservation relations produces grid-volume averages, with point-specific values of the variables called subgrid-scale values. The resolution of data is limited to two grid intervals in each spatial direction, as illustrated in Fig. 1.

The result of the grid-volume averaging produces equations in the form

$$\frac{\partial \bar{\phi}}{\partial t} = \overline{-\vec{V}'' \cdot \vec{\nabla} \phi''} + \bar{S}_\phi,$$

where $\overline{-\vec{V}'' \cdot \vec{\nabla} \phi''}$ is called the subgrid flux, with ϕ'' and V'' the subgrid-scale fluctuations. The variable $\bar{\phi}$ is called the grid-volume resolved variable. An assumption that is routinely made in all mesoscale models (usually without additional comment) is that $\overline{\vec{V}''} \equiv 0$ and $\overline{\phi''} \equiv 0$. This assumption, often referred to as "Reynolds averaging," is actually only accurate when there is a clear spatial scale separation between subgrid-scale and grid-volume resolved quantities so that the grid-volume average does not change rapidly (in time or space) compared with the subgrid-scale variables (Fig. 2).

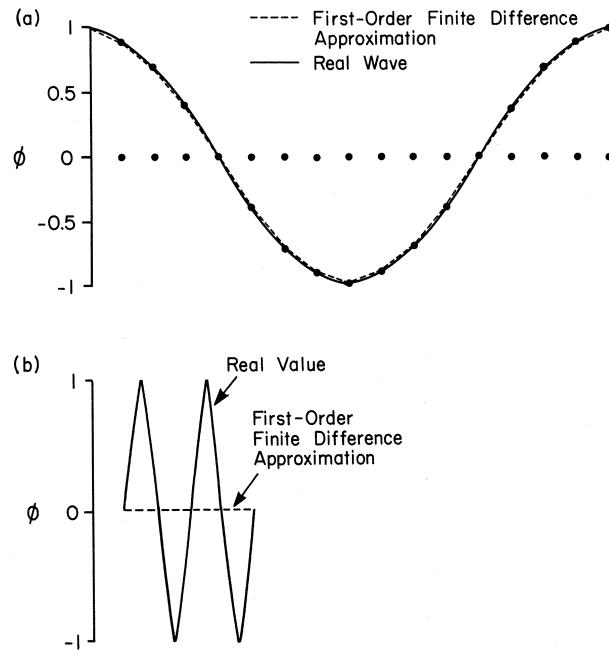


FIGURE 1 Resolution of a 16-grid-increment wavelength wave (top) and a 2-grid-increment wavelength wave (bottom). [From Pielke (1984), Fig. 10.2.]

IV. COORDINATE TRANSFORMATIONS

Mesoscale model equations have been solved in a Cartesian coordinate framework. Each coordinate in this system is perpendicular to the other two coordinates at every location. Most mesoscale models, however, transform to a generalized vertical coordinate. The most common coordinates are some form of terrain-following transformation, where the bottom coordinate surface is terrain height or terrain surface pressure. The result of these transformations is that the new coordinate system is not

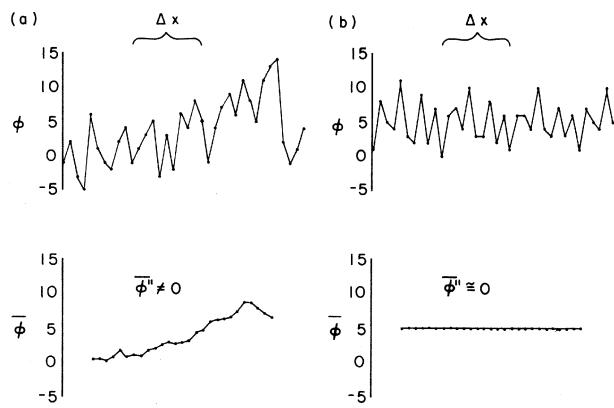


FIGURE 2 One-dimensional schematic illustration of the situation when (a) $\bar{\phi}'' = 0$ and (b) $\bar{\phi}'' \neq 0$. The averaging length is Δ_x .

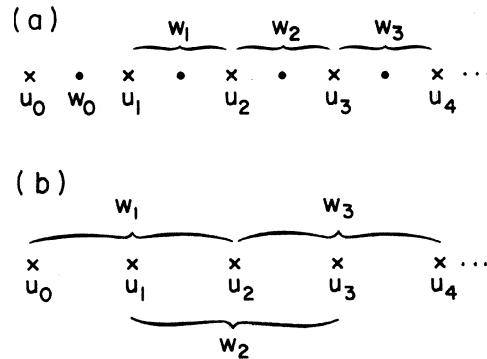


FIGURE 3 Schematic one-dimensional (a) unstaggered and (b) staggered grids for the computation of the variables u and w . The braces indicate which values of u are used in the computation of w . [From Pielke (1984), Fig. 11.8.]

orthogonal, in general. Unless this nonorthogonality is small, the correct treatment of nonhydrostatic pressure effects in mesoscale models requires the use of tensor transformation techniques as opposed to the separate use of the chain rule on each component of velocity separately. The use of generalized coordinate systems introduces additional sources for errors in the models since the interpolation of variables to grid levels becomes more complicated.

V. GRID MESHES

The model variables also need to be defined on a specified grid mesh. When all dependent variables are defined at the same grid points, the grid is said to be nonstaggered. When dependent variables are defined at different grid points, the grid is called a staggered grid. Examples of a nonstaggered grid and a staggered grid in one dimension are shown in Fig. 3. The grid meshes can also be defined with smaller grid increments in one region surrounded by coarser grid increments (Fig. 4). Such a grid is referred to as a nested grid. If the grid increments vary in size at all locations, the grid is called a stretched grid.

VI. PARAMETRIZATIONS

A. Subgrid-Scale Fluxes

The subgrid-scale fluxes in mesoscale models are parametrized in terms of resolvable variables. Turbulence theory, as observed from atmospheric field campaigns over horizontally homogeneous landscapes and for undisturbed atmospheric conditions, is the basis for all mesoscale model representations of the vertical subgrid-scale flux

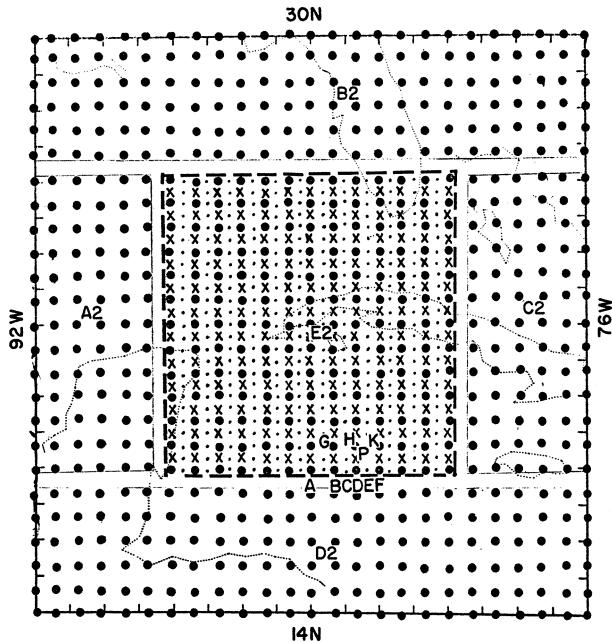


FIGURE 4 Example of a staggered grid with a fine mesh embedded within a coarse mesh. [From Mathur (1974).]

terms. The vertical fluxes are parametrized differently when the lowest 50 m or so of the atmosphere is unstably stratified and when it is stably stratified. The planetary boundary is typically represented by three layers: a thin layer of a few centimeters near the surface where laminar fluxes are important (called the laminar layer), a layer above which extends upward tens of meters where wind direction with altitude is ignored (referred to as the surface layer), and the remainder of the boundary layer where the winds approach the free atmospheric value (referred to as the transition layer). Disturbed (unsteady) boundary layers are not parametrized accurately, however, by existing parametrizations. The effect of land surface heterogeneity has been included on the subgrid scale only as a weighting of the surface layer fluxes by the fractional coverage of each land surface type. This technique is called the mosaic or tile subgrid-scale surface flux parametrization.

In contrast to the vertical fluxes, horizontal subgrid-scale fluxes in mesoscale models are not realistically represented. They are included only to smooth the model calculations horizontally.

B. Source/Sink Parametrizations

The representations of the source/sink terms \bar{S}_ϕ can be separated into two types: those that are derived from basic concepts and those that are parameterized. The only \bar{S}_ϕ terms in mesoscale models that are derived from fundamental physical concepts are those in $\bar{S}_{\bar{v}}$, which are

the pressure gradient force and gravity. Neither of these two forces involves adjustable coefficients, which is one method to separate fundamental terms in the conservation equation from a parametrization.

The remainder of the \bar{S}_ϕ terms need to be parametrized.

1. Radiative Flux Parametrization

The radiative flux terms (in \bar{S}_ϕ) are typically separated into short-wave and long-wave fluxes. The short-wave fluxes, also called solar fluxes, are separated into direct and diffuse irradiance. The direct irradiance is the non-scattered flux, while the diffuse irradiance is the scattered radiative flux from the sun. The direct irradiance is sometimes further separated into visible and near-infrared components. In cloudy model atmospheres, parametrizations based on cloud liquid water content, or more crudely on arbitrary attenuation based on relative humidity, are used. Typically only diffuse irradiance is permitted for overcast model conditions. Some models weight the fluxes for partly cloudy skies, using separate parametrizations for clear and overcast sky conditions. Polluted atmospheres also require parametrization of their effect on solar irradiance, although only a few mesoscale models have explored this issue.

Long-wave irradiance is from the earth's surface and from within the atmosphere. Scattering of long-wave radiative fluxes is typically ignored, such that only upwelling and downwelling irradiances are parametrized. This type of parametrization is called a two-stream approximation. The major absorbers and emitters represented in mesoscale model parametrizations are liquid and ice clouds, water vapor, and carbon dioxide. Clouds are usually parametrized as black bodies to long-wave irradiance. The water vapor and carbon dioxide are represented by the path length through the atmosphere and their concentrations along their path. As with solar radiative fluxes, mesoscale models seldom include parametrizations of long-wave irradiance due to pollution. This neglect is partially a result of the dependence of the absorption, transmissivity, and scattering of both solar and long-wave irradiance on the specific chemical composition and size spectra of the pollution.

2. Parametrizations of the Phase Change of Water

The phase changes of water and their effect on the conservation of the heat source/sink term are separated into stable cloud and cumulus convective cloud and precipitation parametrizations. Stratiform cloud parametrizations range in complexity from algorithms which instantaneously precipitate rain (or snow) when the model relative humidity

exceeds a user-specified relative humidity (referred to as a “dump bucket” scheme) to individual conservation equations for several categories of hydrometeors (e.g., cloud water, rain water, ice crystals, snow, graupel, hail). For the larger hydrometeors, a nonzero, finite terminal fall velocity is usually specified. More detailed microphysical representations, where cloud hydrometeor spectra are classified into more size class intervals (called microphysical bin parametrizations), are also used.

The parametrization of cumulus cloud rainfall utilizes some form of one-dimensional cloud model. These are called cumulus cloud parametrization schemes. Their complexity ranges from instantaneous readjustments of the temperature and moisture profile to the moist adiabatic lapse rates when the relative humidity exceeds saturation, to representations of a set of one-dimensional cumulus clouds with a spectra of radii. These parametrizations typically focus on deep cumulus clouds, which produce the majority of rainfall and diabatic heating associated with the phase changes of water. Cumulus cloud parametrizations remain one of the major uncertainties in mesoscale models since they usually have a number of tunable coefficients, which are used to obtain the best agreement with observations. Also, since mesoscale-model resolution is close to the scale of thunderstorms, care must be taken so that the cumulus parametrization and the resolved moist thermodynamics in the model do not “double count” this component of the \bar{S}_ϕ and \bar{S}_{q_i} .

VII. SOLUTION TECHNIQUES

The grid-volume averaged conservation equations are nonlinear (i.e., terms for advection, the subgrid-scale averaged fluxes) and therefore must be solved using numerical approximation schemes. The solution techniques include finite-difference, finite-element, interpolation (also called semi-Lagrangian), and spectral methods. These approximation schemes must represent temporal ($\partial\phi/\partial t$), spatial ($\vec{V} \cdot \vec{\nabla}\phi$; pressure gradient force), and vertical (subgrid-scale fluxes) terms, and the source/sink terms \bar{S}_ϕ . An important aspect of mesoscale models is that only advection and the pressure gradient force involve horizontal gradients explicitly (i.e., \vec{V}). All other model terms, including each of the source/sink terms, are one-dimensional column models or point values.

Finite-difference schemes involve some form of truncated Taylor series expansion. The finite-element technique uses a local basis function to minimize numerical error, while the spectral method utilizes global basis functions. A spectral method has the advantage that differential relations are converted to algebraic expressions. The semi-Lagrangian scheme is based on fitting interpolation

equations to data at a specific time and advecting the data with model winds.

Mesoscale models have predominately utilized the finite-difference and (for advection), the semi-Lagrangian approaches. A few groups have applied the finite-element method, but its additional computational cost has limited its use. The spectral method, which is most valuable for models without boundaries and for simulation without sharp spatial gradients, has not generally been used since mesoscale models have lateral boundaries.

The use of numerical approximations introduces errors. Linear stability analyses show that it is impossible to create a numerical solution scheme which accurately represents both amplitude change over time and speed of motion (advection, gravity wave propagation) for features that are shorter than about four grid increments in each spatial direction. Figure 1 illustrated the difficulty of resolving a wave feature with just a few grid points, as contrasted with a better resolved feature. Thus a necessary condition for accurate model resolutions is at least four spatial increments in each coordinate direction. Weidman and Pielke (1983) proposed separating model equations into linear and nonlinear components, so that only the nonlinear terms need to be computed using numerical approximation techniques.

Products of the variable (which is a nonlinear term) produce transfers of spatial scales to larger and smaller scales. The inability of the numerical model to sample the smallest scales (less than two grid intervals) results in the spatial scale appearing at a large scale. This error is called aliasing and unless corrected can result in an erroneous accumulation of atmospheric structure at the wrong spatial scale. Figure 5 illustrates how a short-wave feature is aliased to a longer wavelength. For this reason the term “model resolution” should be reserved for features that are at least four grid intervals in each direction (Pielke, 1991; Laprise, 1992; Grasso, 2000).

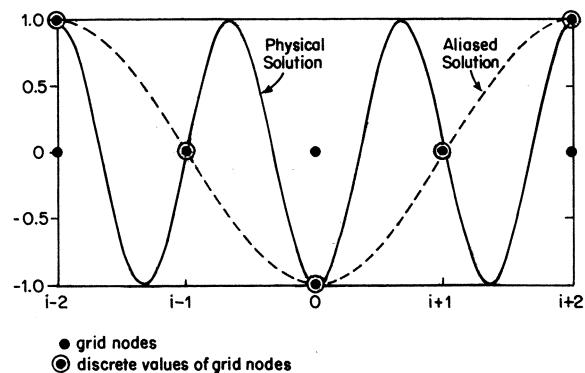


FIGURE 5 Illustration of the aliasing of an inadequately resolved short-wave into a longer wavelength feature.

VIII. INITIAL TOP AND LATERAL BOUNDARY CONDITIONS

To integrate the models forward in time, the dependent variables must be initialized. These values are called initial conditions. Observed data, or a combination of observed data and previous model calculations, are typically used to initialize mesoscale models. The insertion of data during a model calculation is called four-dimensional data assimilation (4DDA). Lateral, top, and bottom boundary conditions are also needed for the duration of the model calculations. Lateral boundary conditions in mesoscale models can be idealized for theoretical studies (e.g., cyclic boundary conditions) or derived from large-scale observations such as the NCEP reanalysis or from larger scale model simulations. Mesoscale models are often strongly influenced by the lateral boundary conditions ([Anthes and Warner, 1978](#)), such that their accurate representation is a necessary condition for an accurate mesoscale simulation.

The top boundary conditions are similar to the lateral boundary condition and must be accurately represented. Most mesoscale models extend into the stratosphere in order to minimize the effect of the model top on the mesoscale simulation. Damping zones at the model top (referred to as an absorbing layer) are usually inserted so that upward-propagating model-simulated gravity waves do not erroneously reflect from the artificial model top.

IX. SURFACE BOUNDARY CONDITIONS

The surface boundary is the only surface of a mesoscale model which is physically based. This surface is typically separated into ocean (and freshwater lakes) and land surfaces. Ocean and lake surfaces can be represented simply as prescribed sea surface temperatures or surface subgrid-scale fluxes, or using mesoscale atmospheric models coupled to mesoscale ocean, lake, and/or sea ice models. Over land, the ground is separated into bare soil and vegetated land. Soil–vegetation/atmosphere transfer schemes (SVATS) have been introduced to represent the fluxes of velocity, heat, moisture, and other trace gases between the atmosphere and the surface. Most SVATS include the effect on water flux of transpiration. Recently, vegetation dynamical processes such as plant growth have been included in longer term (months to seasons) mesoscale model calculations.

X. MODEL PERFORMANCE EVALUATION

Model performance is assessed in several ways. The comparison of observations with model results using statistical

skill tests is a major assessment tool. A complication of these evaluations is that observations have a different sampling volume (e.g., a point) than the model grid volume. Comparisons of simplified (usually linearized) versions of numerical models with analytic theory have been completed to test the accuracy of linear components of the model. Several models can be intercompared to assess which features they have in common and which they do not. The mass and energy budgets of the mesoscale models, if they are each calculated in two separate manners, provide an opportunity to check the internal consistency of the model. Peer-reviewed scientific publications and the availability for scrutiny of the model source code provide two additional valuable procedures to assess the value of the mesoscale model and the degree to which the programmed model logic agrees with the mathematical formulations presented in the literature. Proposals have been made to standardize model computer codes in order to assist in their more general use ([Pielke and Arritt, 1984](#); [Pielke et al., 1995](#)).

XI. EXAMPLES OF MODELS

[Table I](#) presents a list of several mesoscale models along with recent papers where details on the model can be found.

These models have been applied to two basic types of mesoscale systems: those found primarily using initial and lateral boundary conditions (referred to as synoptically forced mesoscale systems) and those found using surface boundary conditions (referred to as surface-forced mesoscale systems). Of the latter type, there are mesoscale systems which are caused when terrain is an obstacle to the flow (referred to as terrain-forced or orographic mesoscale systems) and those generated by horizontal gradients in

TABLE I Atmospheric Mesoscale Models

Model	Reference
RAMS	Pielke et al. (1992)
BLFMESO	Daggupaty et al. (1994)
ARPS	Xue et al. (1995, 2000, 2001)
MM5	Dudhia (1993)
TVM/URBMET	Bornstein et al. (1994)
Hot Mac	Yamada (2000)
FITNAH	Gross (1992)
MRI	Saito (1997), Kato et al. (1998)
Eta	Mesinger (1988), Mesinger et al. (1998), Black (1994)
COAMPS	Hodur (1997)
MC2	Benoit et al. (1997), Laprise et al. (1997)

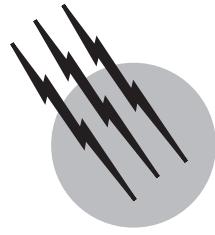
sensible heating of the surface (called thermally forced mesoscale systems).

SEE ALSO THE FOLLOWING ARTICLES

METEOROLOGY, DYNAMIC (STRATOSPHERE) • METEOROLOGY, DYNAMIC (TROPOSPHERE) • THERMODYNAMICS

BIBLIOGRAPHY

- Anthes, R. A., and Warner, T. T. (1978). "Development of hydrodynamic models suitable for air pollution and other mesometeorological studies," *Monthly Weather Rev.* **106**, 1045–1078.
- Benoit, R., Desgagne, M., Pellerin, P., Pellerin, S., Desjardins, S., and Chartier, Y. (1997). "The Canadian MC2: A semi-Lagrangian, semi-implicit wide band atmospheric model suited for five-scale process studies and simulation," *Monthly Weather Rev.* **125**, 2382–2415.
- Black, T. L. (1994). "The new NMC mesoscale Eta model: Description and forecast examples," *Weather Forecasting* **9**, 265–277.
- Bornstein, R., Thunis, P., and Schayes, G. (1994). Observation and simulation of urban-topography barrier effects on boundary layer structure using the three dimensional TVM/URBMET model. "Air Pollution and Its Application X," Plenum Press, NY, 101–108.
- Daggupaty, S. M., Tangirala, R. S., and Sahota, H. (1994). "BLFMESO—A 3-dimensional mesoscale meteorological model for microcomputers," *Boundary Layer Meteorol.* **71**, 81–107.
- Dudhia, J. (1993). "A nonhydrostatic version of the Penn State/NCAR mesoscale model: Validation tests and simulations of an Atlantic cyclone and cold front," *Monthly Weather Rev.* **121**, 1493–1513.
- Grasso, L. D. (2000). "The differentiation between grid spacing and resolution and their application to numerical modeling," *Bull. Am. Met. Soc.* **81**, 579–580.
- Gross, G. (1992). "Results of supercomputer simulations of meteorological mesoscale phenomena," *Fluid Dyn. Res.* **10**, 483–490.
- Hodur, R. M. (1997). "The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS)," *Monthly Weather Rev.* **125**, 1414–1430.
- Kato, T., Kurihara, K., Seko, H., Saito, K., Kazuo, and Goda, H. (1998). "Verification of the MRI-nonhydrostatic-model predicted rainfall during the 1996 Baiu season," *J. Meteor. Soc. Japan* **76**, 719–735.
- Laprise, R. (1992). "The resolution of global spectral models," *Bull. Am. Meterol. Soc.* **73**, 1453–1454.
- Laprise, R., Caya, D., Bergeron, G., and Giguere, M. (1997). "The formulation of the Andre Robert MC² (mesoscale compressible community) model," *Atmos. Ocean* **XXXV**, 195–220.
- Mathur, M. B. (1974). "A multiple grid primitive equation model to simulate the development of an asymmetric hurricane (Isabel, 1964)," *J. Atmos. Sci.* **31**, 371–393.
- Mesinger, F. (1998). "Comparison of quantitative precipitation forecasts by the 48- and by the 29-km Eta model: An update and possible implications." In "12th Conference on Numerical Weather Prediction," pp. J22–J23, American Meteorological Society.
- Mesinger, F., Janjic, Z. I., Nickovic, S., Gavrilov, D., and Deaven, D. G. (1988). "The step mountain coordinate: Model description and performance for cases of alpine cyclogenesis and for a case of an Appalachian redevelopment," *Monthly Weather Rev.* **116**, 1493–1518.
- Pielke, R. A. (1984). "Mesoscale Meteorological Modeling," Academic Press, New York.
- Pielke, R. A. (1991). A recommended specific definition of "resolution," *Bull. Amer. Meteor. Soc.* **72**, 1914.
- Pielke, R. A. (2000). "Mesoscale Meteorological Modeling," 2nd ed., Academic Press, New York.
- Pielke, R. A., and Arritt, R. W. (1984). "A proposal to standardize models," *Bull. Amer. Meteor. Soc.* **65**, 1082.
- Pielke, R. A., Cotton, W. R., Walko, R. L., Tremback, C. J., Lyons, W. A., Grasso, L. D., Nicholls, M. E., Moran, M. D., Wesley, D. A., Lee, T. J., and Copeland, J. H. (1992). "A comprehensive meteorological modeling system—RAMS," *Meteorol. Atmos. Phys.* **49**, 69–91.
- Pielke, R. A., Bernardet, L. R., Fitzpatrick, P. J., Gillies, S. C., Hertenstein, R. F., Jones, A. S., Lin, X., Nachamkin, J. E., Nair, U. S., Papineau, J. M., Poulos, G. S., Savoie, M. H., and Vidale, P. L. (1995). "Standardized test to evaluate numerical weather prediction algorithms," *Bull. Amer. Meteor. Soc.* **76**, 46–48.
- Saito, K. (1997). "Semi-implicit fully compressible version of the MRI mesoscale nonhydrostatic model: Forecast experiment of the 6 August 1993 Kagoshima torrential rain."
- Weidman, S., and Pielke, R. A. (1983). "A more accurate method for the numerical solution of nonlinear partial differential equations," *J. Computational Phys.* **49**, 342–348.
- Xue, M., Droege, K. K., Wong, V., Shapiro, A., and Brewster, K. (1995). "ARPS Version 4.0 User's Guide," ...
- Xue, M., Droege, K. K., and Wong, V. (2000). "The Advanced Regional Prediction System—A multiscale nonhydrostatic atmospheric simulation and prediction model. I: Model dynamics and verification," *Meteorol. Atmos. Phys.* **75**, 161–193.
- Xue, M., Droege, K. K., Wong, V., Shapiro, A., Brewster, K., Can, F., Weber, D., Liu, Y., and Wang, D.-H. (2001). "The Advanced Regional Prediction System—A multiscale nonhydrostatic atmospheric
- Yamada, T. (2000). "Numerical simulations of airflows and tracer transport in the southwestern United States," *J. Appl. Meteor.* **39**, 399–411.



Meteorology, Dynamic (Stratosphere)

Rolando R. Garcia

National Center for Atmospheric Research

- I. Introduction
- II. Radiative Processes
- III. Circulation Systems
- IV. Stratospheric Dynamics
- V. Transport of Minor Species
- VI. Current Research

GLOSSARY

Coriolis force An apparent force, due to the rotation of the Earth, that acts at right angles to the velocity vector of moving objects; it is directed to the right of the direction of motion in the Northern Hemisphere, and to the left in the Southern Hemisphere.

Critical layer The locus of points in the latitude–height plane where the phase speed of a wave approaches the speed of the background (zonal-mean) flow.

Downward-control principle An analytical result stating that the steady-state, zonal-mean vertical velocity at a given level is a function of the Eliassen–Palm flux divergence integrated vertically between the level in question and the top of the atmosphere.

Eliassen–Palm flux A measure of the pseudomomentum associated with vertically propagating atmospheric waves. Its divergence gives the acceleration of the zonal-mean zonal wind by atmospheric waves.

Gravity wave A wave that occurs in a stable, density-

stratified fluid under the influence of gravity; also known as a buoyancy wave.

Kelvin wave An eastward-propagating gravity wave modified by the Earth’s rotation. Its structure is symmetric about the equator and depends on the change of sign of the Coriolis force across the equator.

Mean meridional circulation The zonal-mean, or zonally averaged, circulation in the latitude–height plane.

Meridional wind The component of the wind along the meridional, or north–south, direction.

Mesosphere The layer of the atmosphere immediately above the stratosphere; the mesosphere extends from about 50 to 80 km.

Non acceleration theorem A statement of the conditions under which the divergence of the Eliassen–Palm flux vanishes. According to the theorem, the EP flux divergence is nonzero only for transient or dissipating waves.

Normal mode A free, or resonant, oscillation of the atmosphere; the solution to the homogeneous wave equation with homogeneous (unforced) boundary conditions.

Planetary wave A very large-scale Rossby wave (horizontal wavelength of 10,000–40,000 km). Planetary waves are the dominant type of wave in the extratropical stratosphere during winter.

Polar night jet Strong eastward wind jet that dominates the zonal-mean circulation of the wintertime stratosphere. It averages some 80 m s^{-1} in the Northern Hemisphere and as much as 100 m s^{-1} in the Southern Hemisphere, at about 50 km of altitude.

Quasibiennial oscillation An approximately periodic oscillation of the zonal-mean zonal wind in the tropical stratosphere, between the tropopause and about 45 km. The period of the oscillation is somewhat irregular, but averages about 27 months.

Rossby wave A westward-propagating wave that originates from the variation of the Coriolis parameter with latitude.

Semiannual oscillation A 6-month oscillation in the tropical zonal-mean zonal wind that occurs above 35 km; the semiannual period is a result of regulation by the seasonal cycle. The stratospheric oscillation is coupled to a similar oscillation in the mesosphere.

Troposphere The layer of the atmosphere between the ground and the stratosphere; the troposphere is the layer where all “weather” phenomena (fronts, thunderstorms, hurricanes, etc.) take place.

Stratosphere The layer of the atmosphere between about 10 and 50 km, characterized by a monotonic increase of temperature with height; the stratosphere lies between the troposphere below and the mesosphere above.

Vorticity The curl of the vector velocity field; a measure of the angular momentum of the flow.

Zonal mean The mean of a meteorological quantity obtained by averaging with respect to longitude along an entire latitude circle.

Zonal wind The component of the wind along the zonal, or east–west, direction.

THE STRATOSPHERE is the layer of the Earth’s atmosphere immediately above the troposphere; it extends from a lower boundary (the tropopause) whose altitude varies between about 8 and 16 km to an upper boundary (the stratopause) near 50 km. The stratosphere is characterized by increasing temperature with altitude, which is due primarily to the absorption of solar ultraviolet radiation by ozone. The meteorology of the stratosphere is governed mainly by the seasonal variation of heating due to absorption of solar radiation, and by the upward propagation of atmospheric waves originating in the troposphere.

I. INTRODUCTION

High-altitude balloon measurements made in France at the turn of the century by Léon Teisserenc de Bort showed that, at altitudes of 11–13 km, the temperature of the atmosphere ceases to decline with altitude and instead begins to increase. This marks the tropopause, the boundary between the troposphere (the layer of the atmosphere closest to the ground) and the stratosphere. In the tropics, the tropopause is found at higher altitude (16–17 km), while in the polar regions it can be as low as 7–8 km. After World War II, rocket temperature soundings began to be made up to altitudes of about 100 km and global mean vertical temperature profiles or “standard atmospheres” were produced. [Figure 1](#) shows one such standard atmosphere, which indicates the locations of the troposphere, the stratosphere, and the upper layers of the atmosphere (the mesosphere and the thermosphere).

In the troposphere, the principal energy source is visible solar radiation absorbed at the Earth’s surface. To some extent, this energy is redistributed throughout the troposphere by meteorological processes such as convection and large-scale weather systems. Nonetheless, the result still is a temperature profile that decreases with altitude up to the tropopause. On the other hand, the stratosphere is heated directly by solar ultraviolet (UV) radiation absorbed by ozone, the triatomic form of oxygen. Although present in minute amounts in the stratosphere—less than 11 parts per million by volume (ppmv)—ozone is a very efficient absorber of UV radiation. It provides a large internal heat source that peaks in the upper stratosphere and gives rise to a temperature profile that increases monotonically with altitude up to the stratopause (50 km). The absorption of UV radiation by ozone not only heats the stratosphere but also reduces the amount of biologically damaging ultraviolet radiation reaching the Earth’s surface.

The meteorology of the stratosphere is affected directly by solar heating due to ozone absorption, since the distribution of heat varies greatly with the solar zenith angle and is not moderated by the thermal capacity of oceans and continents as in the troposphere. In addition to direct solar input, energy can be transmitted from the troposphere to the stratosphere by a variety of processes. Large-scale weather systems in the troposphere extend their influence into the lower stratosphere, especially during winter. More importantly, atmospheric waves generated in the troposphere can carry energy and momentum into the stratosphere. This makes the stratosphere a unique geophysical laboratory not only for the meteorologist but for the fluid dynamicist as well.

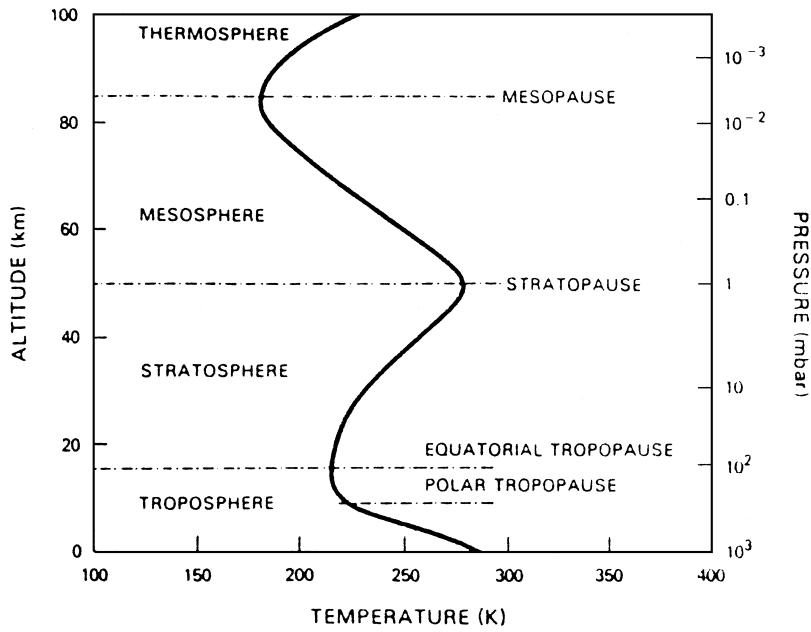


FIGURE 1 The global mean temperature structure of the Earth's atmosphere, with conventional nomenclature.

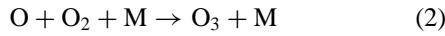
II. RADIATIVE PROCESSES

A. Solar Heating

The absorption of solar UV radiation in the stratosphere is almost entirely controlled by ozone (O_3), with minor contributions from molecular oxygen (O_2) and nitrogen dioxide (NO_2). Ozone is formed in a two-step process, beginning with the dissociation of molecular oxygen



at wavelengths less than about 2400 Å, followed by reaction of oxygen atoms with molecular oxygen



where M is a third molecule, usually one of the major atmospheric constituents (O_2 or N_2).

The distribution of atmospheric ozone based on data obtained by the Nimbus 4 BUV satellite is shown in Fig. 2. Ozone is most abundant in the tropical middle stratosphere (~30 km), where the solar zenith angle is small and the atmospheric density is sufficiently large to produce O_3 efficiently via reactions (1) and (2). The decrease in ozone toward the poles is due to a decrease in production due to the steeper solar zenith angles at high latitudes. The important solar absorption bands for ozone and oxygen are listed in Table I. Figure 3 shows the altitudes where the stratosphere and mesosphere are heated by these absorption bands. The peak in the heating is nearly coincident with the stratopause temperature maximum shown in Fig. 1.

B. Infrared Cooling

Infrared radiative cooling of the Earth's stratosphere takes place through collisional excitation followed by radiative relaxation of molecular vibrational and rotational states, resulting in the emission of infrared and microwave radiation. In addition, these radiatively active molecules may intercept and reemit infrared radiation from the surface before it escapes to space. The result is a complex exchange of radiation between atmospheric layers and the surface. The effect of solar infrared radiation on this process is nearly negligible, since there is almost no overlap between solar and terrestrial blackbody surface emission.

In the troposphere, atmospheric radiative cooling is dominated by water vapor. The stratosphere, however, is very dry (less than 7 ppmv H_2O) because tropospheric air enters the stratosphere mainly through the very cold tropical tropopause (~190 K), where most of its water content is lost through freezing. Stratospheric infrared radiation to space is thus controlled mostly by carbon dioxide (which is present at a mixing ratio of about 340 ppmv) and ozone. Carbon dioxide has vibrational absorption bands at 15 and 4.3 μm. The latter plays almost no role in infrared radiative transfer, since both terrestrial and solar emissions are weak in this range of wavelength. The 15-μm band, on the other hand, is located near the peak of the terrestrial emission spectrum, and thus plays a critical role. Ozone has three important bands, 9.016, 9.597, and 14.27 μm; the last overlaps the CO_2 15-μm band.

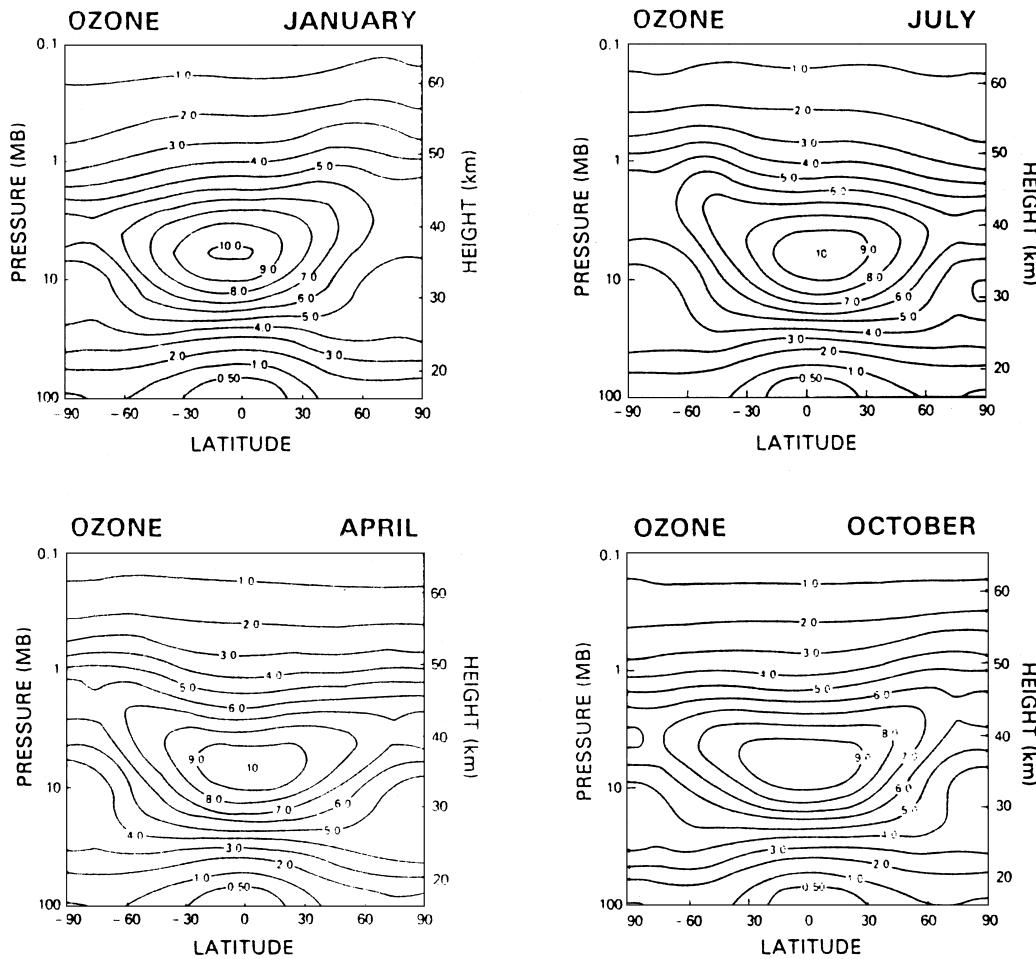


FIGURE 2 The distribution of atmospheric ozone as a function of altitude and latitude for different times of the year, from Nimbus-4 BUV satellite observations. Units are parts per million by volume. [From J. Rosenfield *et al.* (1987). *J. Atmos. Sci.* **44**, 859–876.]

Figure 4 shows the global-mean vertical profiles of cooling and heating by CO_2 , H_2O , and O_3 . Although ozone plays a secondary role to CO_2 , it cannot be neglected. Of special importance is the heating (shown as negative cooling in the figure) provided by ozone in the tropical lower stratosphere near the cold tropopause. In this region, blackbody radiation from the surface can be directly ab-

sorbed by ozone molecules and so heat the lower stratosphere. Higher up in the stratosphere, where temperature increases, ozone's infrared bands cool the atmosphere, as does CO_2 .

C. Net Thermal Drive and Temperature Structure

On a temporal and global average, the net incoming flux of solar radiation must be balanced by outgoing infrared (IR) flux, that is, the atmosphere must be in a state of global radiative equilibrium. Locally, however, solar heating and infrared cooling rarely balance because local temperature changes can occur through advection or by expansional cooling or compressional warming during adiabatic motion (see Section IV). (An adiabatic process is one wherein no internal heat is exchanged between the air parcel and its surroundings; the temperature of the parcel may change during an adiabatic process as long as the parcel's internal energy remains constant.)

TABLE I Stratospheric Absorbers of UV and Visible Solar Radiation

Wavelength range (Å)	Absorber
1250–1750	Schumann–Runge continuum of O_2
1750–2000	Schumann–Runge bands of O_2
2060–2425	Herzberg continuum of O_2 and O_3
2425–2775	Hartley band of O_3
2775–3800	Huggins band of O_3
3800–8500	Chappuis band of O_3

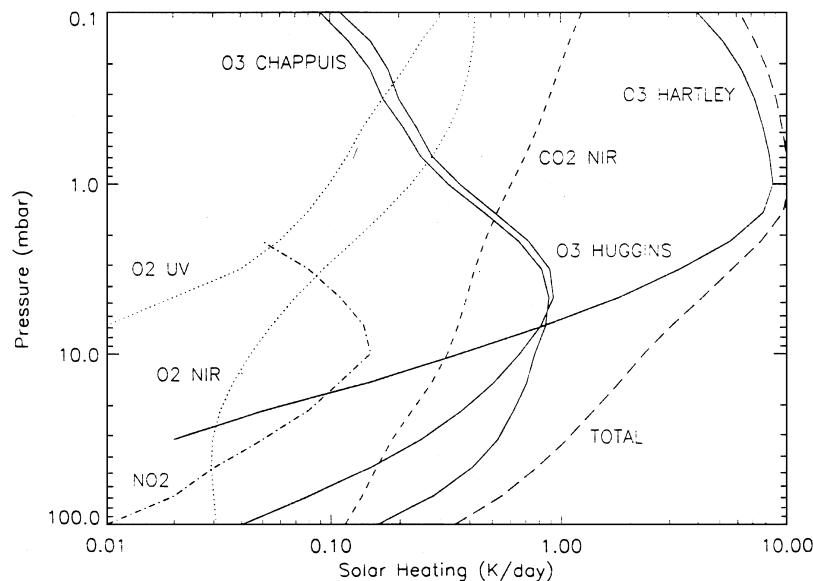


FIGURE 3 Solar heating rates at equinox for absorbers important in the stratosphere. [From C. J. Mertens *et al.* (1999). *J. Geophys. Res.* **104**, 6021–6038.]

Figure 5 shows observations of stratospheric temperatures averaged along latitude circles for Northern Hemisphere winter and summer, together with calculations of net heating rates (the sum of UV heating and IR heating and cooling). It is evident that most of the stratosphere is not in local radiative equilibrium. It turns out that departures from radiative equilibrium are due mainly

to adiabatic warming in regions of downwelling, and cooling in regions where upwelling occurs. Therefore, the sense of the vertical component of the stratospheric circulation can be inferred from the sign of the net radiative heating in Fig. 5. For example, downward motion (hence compressional warming) must prevail at polar latitudes during winter since these are regions of net radiative

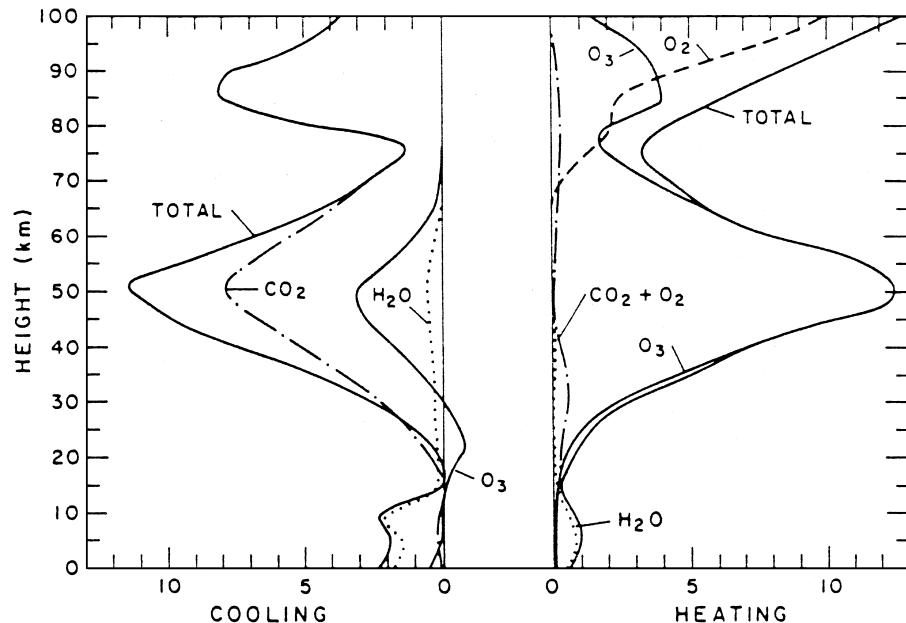


FIGURE 4 Vertical distribution of solar short-wave heating rates and atmospheric long-wave cooling rates due to radiatively active gases in the stratosphere. [From J. London (1979). *Proc. NATO Advanced Institute on Atmospheric Ozone*, 703–721.]

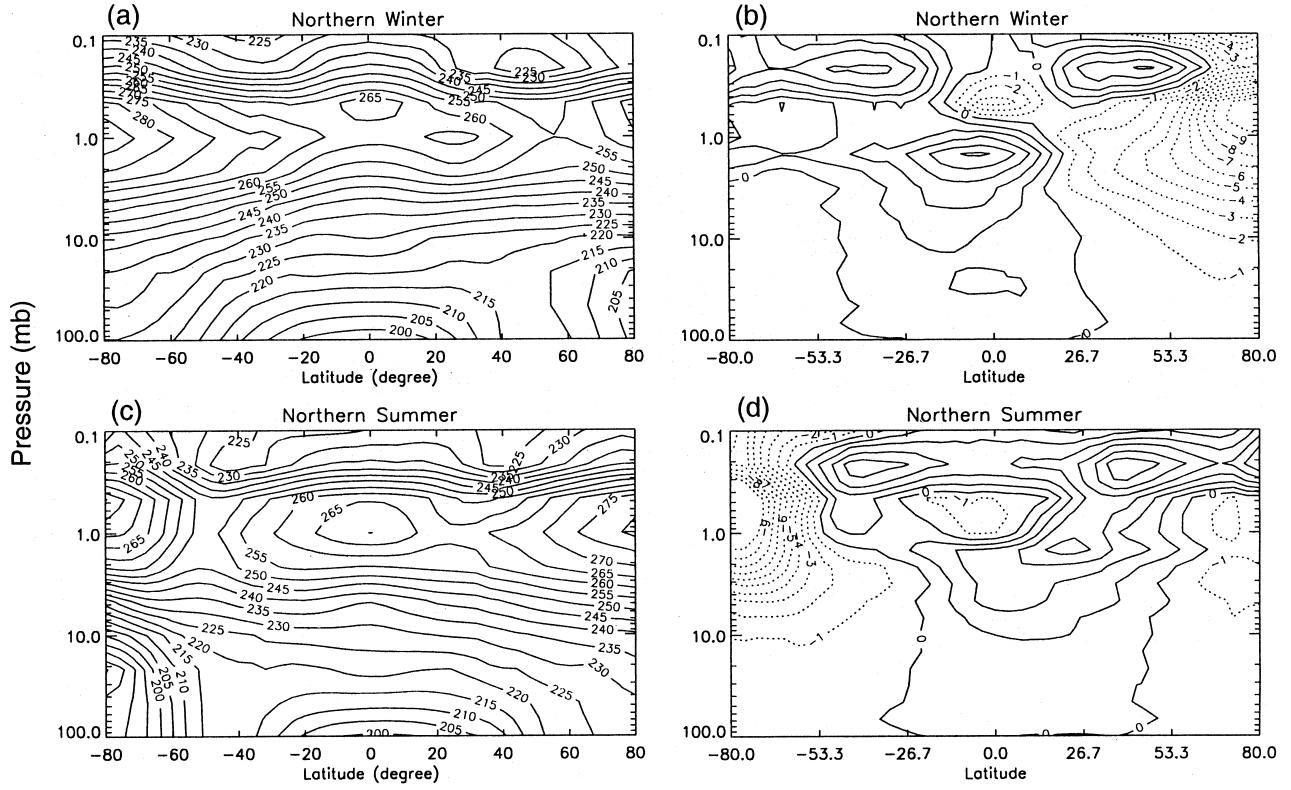


FIGURE 5 Monthly mean, zonal-mean temperature distributions observed by the Upper Atmosphere Research Satellite (UARS) (panels a, c) and calculated net zonal-mean heating rates (panels b, d). [Courtesy of M. G. Mlynczak, NASA/Langley Research Center.]

cooling. Similarly, in the lower stratosphere there must be upwelling (and expansional cooling) in the tropics, while downwelling (and compressional warming) must dominate the middle and high latitudes.

The global circulation of the stratosphere thus plays a major role in determining its temperature distribution; calculations indicate that, in the absence of adiabatic warming, the winter polar stratosphere would cool to temperatures of 150–170 K, much colder than what is actually shown in Fig. 5. The stratospheric circulation also plays a crucial role in determining the distribution of chemical species. In fact, well before the stratospheric thermal budget was understood, the sense of the global circulation of the lower stratosphere was deduced by A. W. Brewer and G. B. Dobson from observations of the distribution of stratospheric water vapor and ozone, respectively.

III. CIRCULATION SYSTEMS

As shown earlier, the variation of solar heating with latitude gives rise to an equator-to-pole temperature difference, which in turn produces a pressure difference. For example, the 100-mb pressure surface is located at about 16 km

in the tropics but at about 14 km at the poles. However, because the atmosphere of the Earth is in rapid rotation, its motion is constrained by the requirement of conservation of angular momentum. Conservation of angular momentum manifests itself in the Earth's frame of reference as the Coriolis force, an apparent force that deflects moving air parcels to the right of their velocity vector in the Northern Hemisphere, and to the left in the Southern Hemisphere.

Since the motion of the Earth's atmosphere is largely parallel to the Earth's surface, conservation of angular momentum has a small effect on air motion in the equatorial region, where the vertical component of the planetary angular rotation vector tends to vanish. At higher latitudes, the Coriolis force is much stronger because the vertical component of the rotation vector is large. What this means in practice is that the general circulation of the atmosphere is not simply a matter of flow from high to low pressure; instead, the large-scale circulation systems outside the tropics are in geostrophic equilibrium, wherein pressure gradients are balanced by the Coriolis force. These concepts will be given precise meaning in Section IV; for the moment it is sufficient to bear in mind that the circulation of the tropical and extratropical stratosphere are

profoundly different, and that these differences arise from the varying strength of the Coriolis force.

A. The Extratropical Circulation

The extratropical circulation tends to follow latitude circles, that is, the zonal (east–west) circulation is usually much stronger than the meridional (north–south) circulation, and also more persistent. In fact, if temporal and zonal averages are taken, the resulting zonal winds are on the order of 100 times stronger than the meridional winds, even though the large-scale pressure differentials produced by radiative heating are aligned in the meridional direction. This is a manifestation of the angular momentum constraint mentioned previously.

The concept of zonal mean, or zonal average, is especially useful for describing the meteorology of the stratosphere, so a brief definition is in order: The zonal mean of an atmospheric variable, Ψ , is the average taken in the east–west (longitudinal) direction. Denoting longitude by x , the zonal mean $\bar{\Psi}$ is defined as

$$\bar{\Psi}(y, z; t) = \frac{1}{L} \int_0^L \Psi(x, y, z; t) dx \quad (3)$$

where y , z , and t are latitude, height, and time, respectively, and L is the distance around a latitude circle.

[Figure 6](#) shows the seasonal evolution of the zonal-mean zonal wind in the stratosphere. In the lowermost stratosphere, the winds are eastward in all months, but above about 20 km the winds in summer are westward,

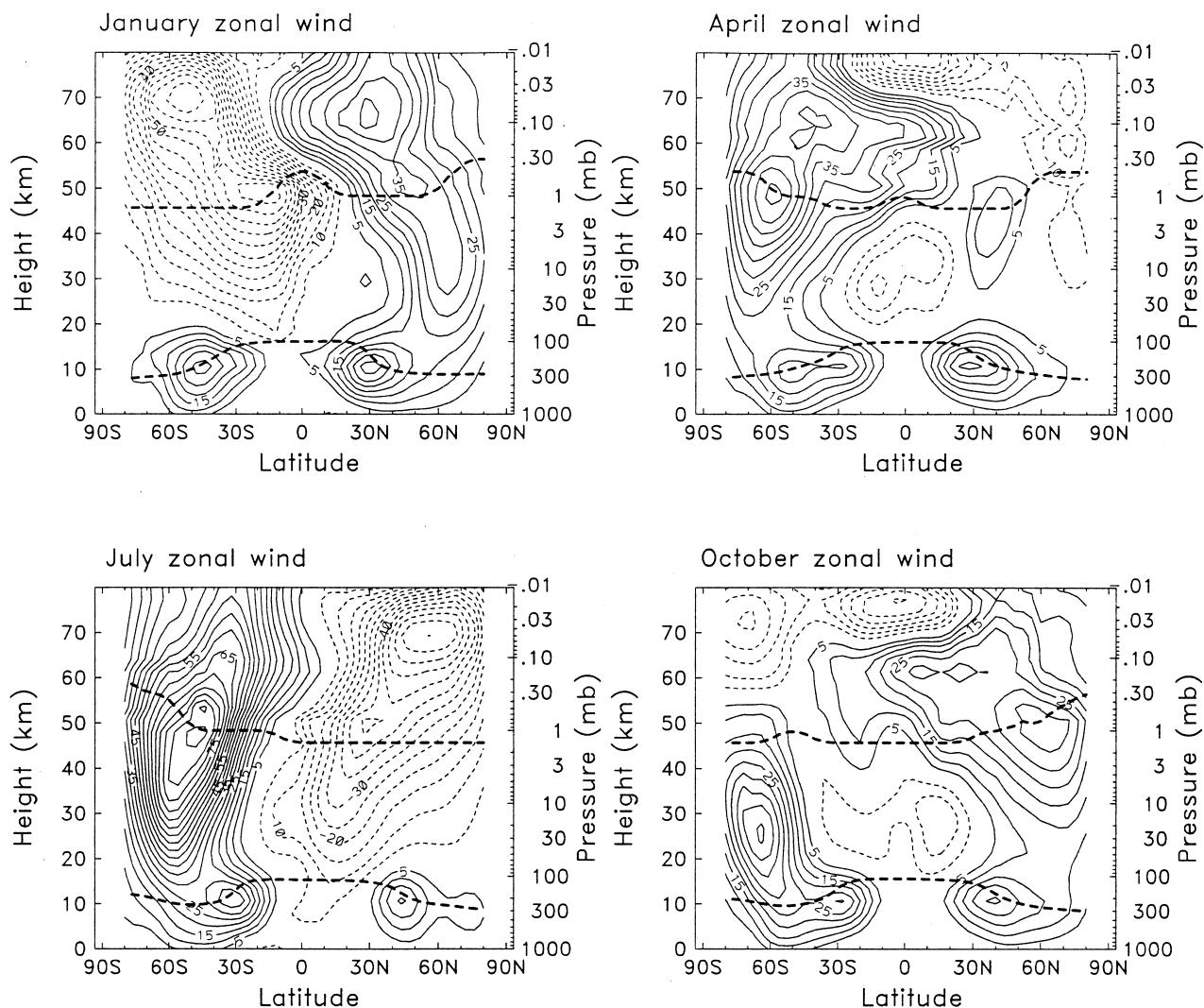


FIGURE 6 The zonal-mean distribution of zonal (east–west) winds in the stratosphere for January, April, July, and October. The dashed lines indicate the approximate position of the tropopause and stratopause. [Courtesy of W. Randel, National Center for Atmospheric Research.]

while those in winter remain eastward. (By convention, longitude is taken to increase in the eastward direction, so eastward winds are positive, and westward winds negative.) The summertime wind regimes are of very nearly the same intensity in the northern and southern hemispheres, but the wintertime winds above 20 km are not. The wintertime stratospheric wind system is known as the “polar night jet,” and it is considerably stronger in southern winter (July) than in northern winter (January).

The zonality of the stratospheric circulation is extremely marked during summer, as illustrated in Fig. 7a, which shows a typical summertime circulation pattern on the 30-mb pressure surface. During winter, on the other hand, the flow is often perturbed by large-scale disturbances, so that there is significant motion in the north-south direction, as shown in Fig. 7b. The meridional flow is more comparable in this case to the zonal wind. However, because the meridional flow meanders back and forth across latitude circles, its zonal average is much smaller than that of the zonal wind, which nearly always blows from west to east in winter.

The disturbances in the stratospheric circulation seen in Fig. 7b are caused by atmospheric waves that propagate upward from the troposphere. The dominant type of wave in the extratropical stratosphere is the Rossby wave (see Section IV.B.2). Rossby waves found in the stratosphere are of very large scale, having horizontal wavelengths of tens of thousands of kilometers; for these reason, these Rossby waves are often referred to as “planetary waves.”

On occasion, planetary waves can grow to very large amplitudes over a few days, displacing or splitting the polar night jet and contributing to a spectacular polar temperature rise of 80 K or more. Figure 8 shows the development of a so-called “sudden warming”; the contours show the height in meters of the 10-mb pressure surface above sea level (the height of a pressure surface is proportional to the vertical integral of the temperature in the underlying levels). The high geopotential (warm) feature over the Aleutian Islands amplifies and pushes toward the pole, splitting the cold circumpolar region of low geopotential. The warmer air inside the geopotential high displaces the cold polar air, and the polar temperature increases over a wide range of altitude. This can be seen in the lower panel of Fig. 8, which shows the zonally averaged temperature as a function of latitude and altitude. Comparison with Fig. 5 reveals an increase of some 30–40 K throughout the middle and upper stratosphere at high latitudes.

Sudden warmings are usually observed in mid- to late winter in the Northern Hemisphere. During early winter, the polar night jet builds in strength until late December or early January. A series of small warmings often begin, with larger warmings usually occurring later in the winter. In late February or in March, the last or final warming

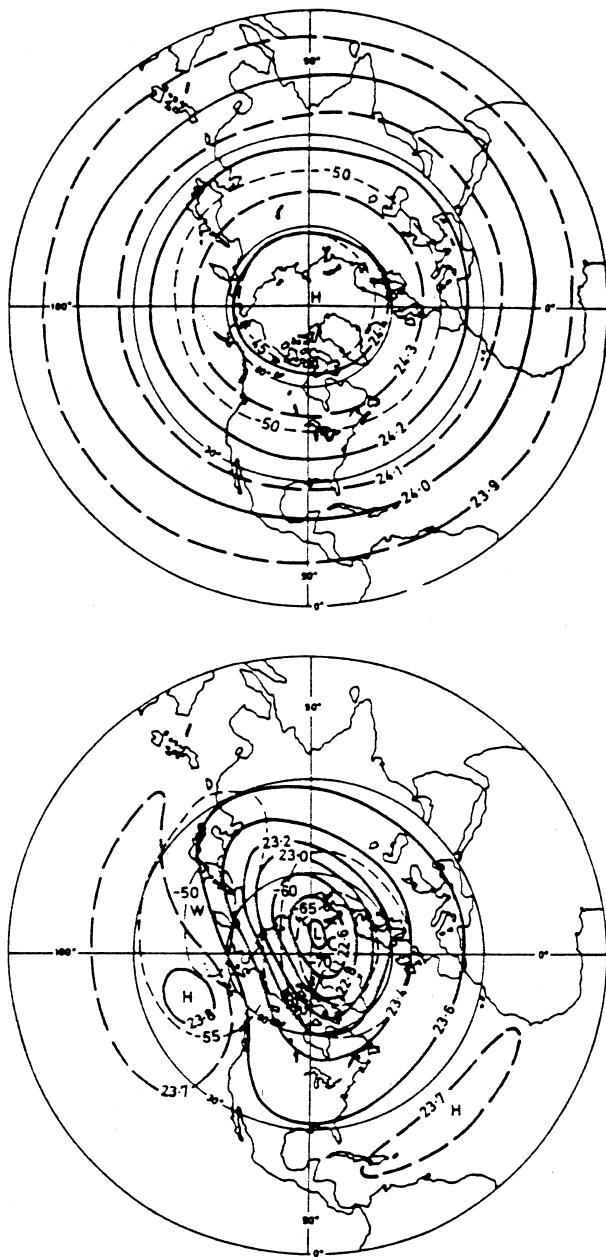


FIGURE 7 Stratospheric circulation on the 30-mb pressure surface during (top) summer and (bottom) winter. The stratospheric winds are approximately parallel to the geopotential contours (solid lines, in geopotential km); the temperature distribution (K) is denoted by the dashed contours. [After F. K. Hare (1968). Q. J. R. Meteorol. Soc. **94**, 439–459.]

occurs, which marks the transition from winter to summer circulation. The eastward winds of winter are replaced by weak westward winds, which eventually gain strength and produce the westward summer wind systems seen in Fig. 6.

Although major midwinter warmings occur in the Northern Hemisphere every other year on average, they are

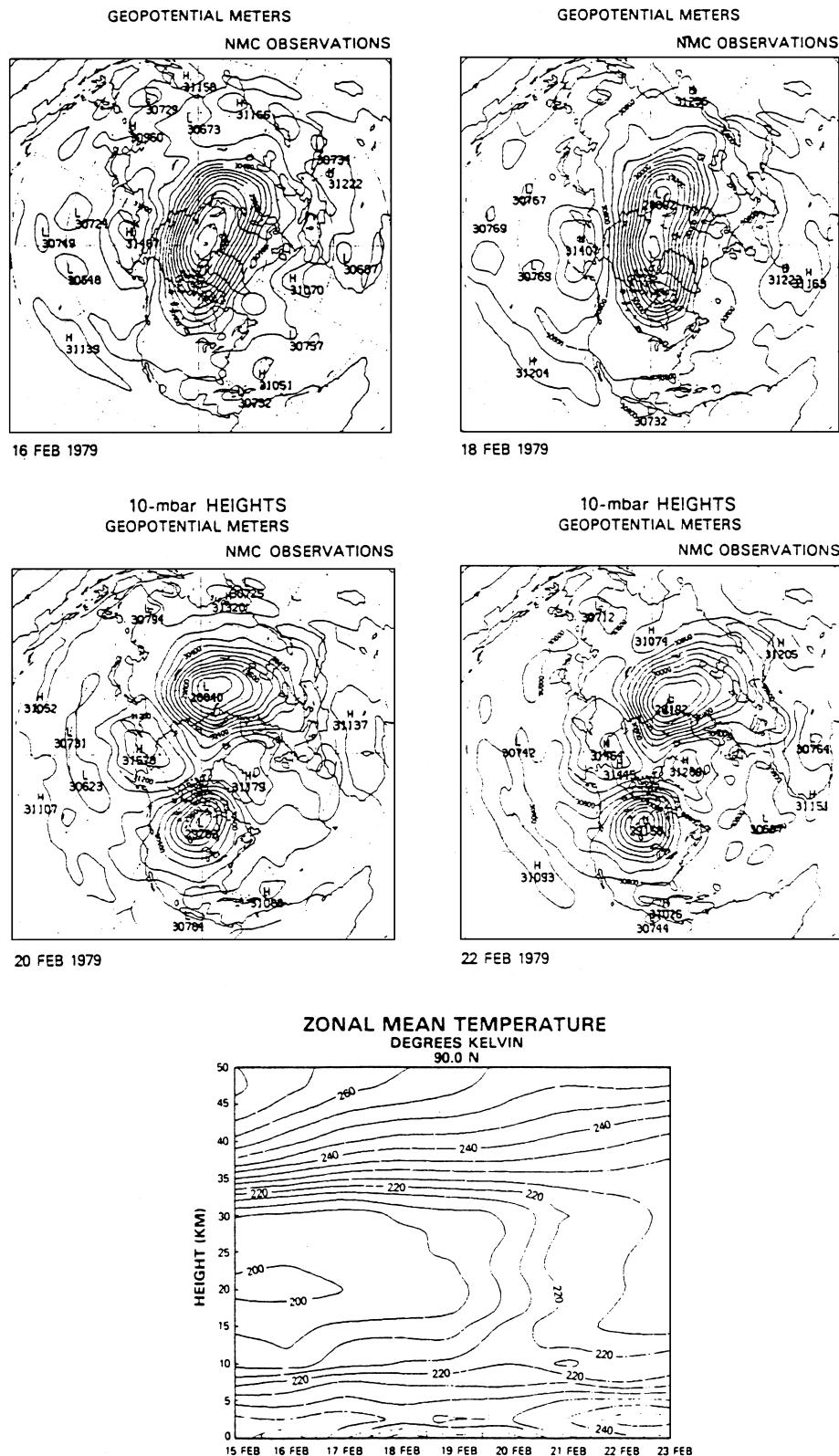


FIGURE 8 The development of a stratospheric sudden warming. The four upper panels show the distribution of geopotential height on the 10-mb pressure surface as the warming progresses; the lower panel shows the zonal mean temperature toward the end of the warming event. [Data courtesy of National Meteorological Center.]

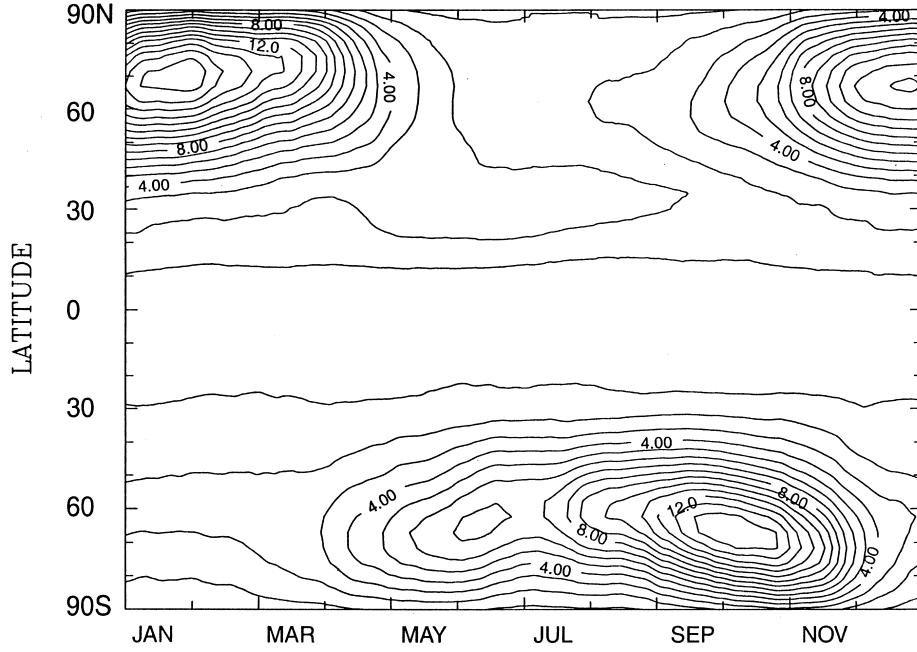


FIGURE 9 Seasonal climatology of planetary wave temperature amplitude (K) at 50 mb (~21 km). Note the large amplitudes throughout Northern Hemisphere winter. In the Southern Hemisphere, wave amplitudes remain relatively small until late winter. [Data courtesy of W. Randel, National Center for Atmospheric Research.]

absent in Southern Hemisphere winter. The final warming, or seasonal spring transition, in the Southern Hemisphere occurs in November, almost a month and a half after solstice, and much later than in the Northern Hemisphere, where it occurs in March. Stratospheric warmings are occasionally seen in midwinter in the Southern Hemisphere, but their intensity is weaker than in the Northern Hemisphere and they tend to be confined to altitudes above 30 km; these events are known as “minor warmings.”

The planetary Rossby waves responsible for the distortion of the zonal jets in winter, and for the sudden warming phenomenon, are quasistationary; that is, these waves move only slightly with respect to the Earth’s surface. The marked differences between the wintertime circulation of the Northern and Southern hemispheres is due to the weaker amplitude of quasistationary planetary waves in the latter, as illustrated in Fig. 9. This, in turn, is probably related to the concentration of continental land masses in the Northern Hemisphere. Major orographic features are known to excite planetary waves when the tropospheric jet streams blow strongly across them, as is the case during winter; the thermal contrast between continents and oceans is also known to excite planetary waves.

A variety of rapidly moving disturbances, known as traveling waves, also exist in the extratropical stratosphere. A traveling wave is one whose phase at a fixed location changes in time. The most commonly observed waves are listed in Table II. Outside the tropics, they appar-

ently arise from weakly unstable flows, or as Rossby normal modes of the Earth’s atmosphere (see Section IV.B.3). Most traveling waves have small amplitudes in the stratosphere compared to the quasistationary planetary waves. The traveling waves listed in Table II attain amplitudes of at most a few hundred geopotential meters, whereas quasistationary wave amplitudes can be as large as 1000 geopotential meters in the upper stratosphere during Northern Hemisphere winter. On the other hand, traveling waves are often observed in the summertime stratosphere, when vertical propagation of quasi-stationary waves is prohibited (see Section IV.B.2).

B. The Tropical Circulation

The circulation of the tropical stratosphere is characterized by regimes of westward and eastward winds that vary regularly on seasonal to interannual time scales, succeeding each other throughout the entire altitude range of the stratosphere. This behavior is fundamentally different from that seen outside the tropics, where a westward jet in summer, and an eastward one in winter, dominate the circulation. The regularly varying wind systems of the tropical stratosphere owe their existence to the effect of wave dissipation in a region where the Coriolis force is relatively weak.

A time series of the equatorial zonal-mean zonal wind, from 1953 through 1996, is shown in Fig. 10. The

TABLE II Traveling Waves Observed in the Extratropical Stratosphere

Period (days)	Wave number	Maximum amplitude	Remarks
2	3	0.4 K (1 mb)	Antisymmetric ^a Rossby normal mode ^b
4	2	75 m (1 mb)	Symmetric Rossby normal mode
5	1	75 m (1 mb)	Symmetric Rossby normal mode
10	1	200 m (1 mb)	Antisymmetric Rossby normal mode
16	1	160 m (3 mb)	Symmetric Rossby normal mode
4	1	Variable	Instability of the winter polar vortex? Found at high latitudes

^a“Symmetric” and “antisymmetric” refer to symmetry about the equator.

^b Rossby modes propagate westward with respect to the background flow.

observations reveal a remarkable long-term oscillation, such that successive regimes of westward and eastward winds propagate from higher to lower altitudes. This quasiregular switching of the zonal-mean zonal wind is known as the quasibiennial oscillation (QBO) and has a mean period of about 27 months. Note, however, that there is considerable variability among QBO cycles, some be-

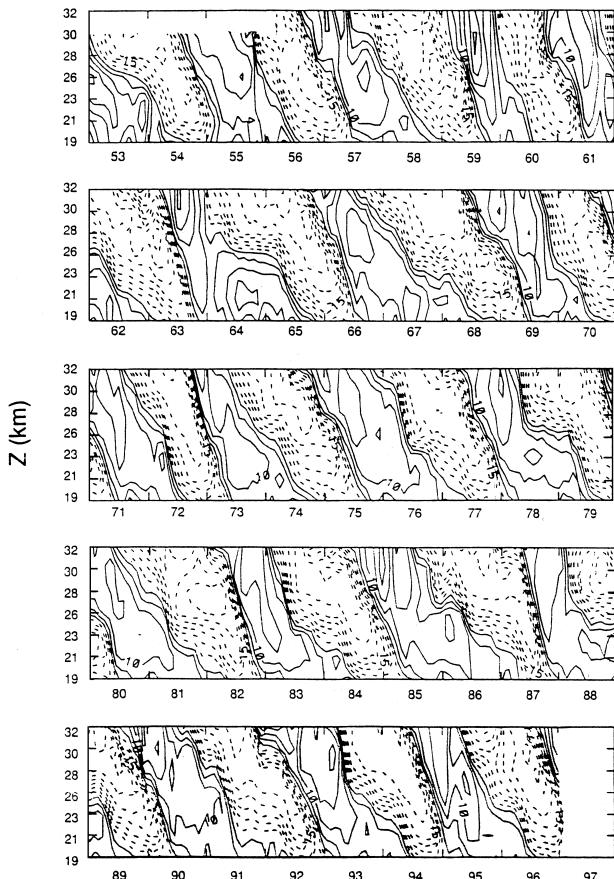


FIGURE 10 The QBO in mean zonal wind at the equator, 1953–1996. The mean period is 27 months, but individual cycles vary considerably in length. [Data courtesy of Free University of Berlin.]

ing considerably longer and others shorter than the mean period.

Above the range of altitudes where the QBO dominates the tropical circulation (~ 35 km), a semiannual oscillation (SAO) is present in the zonal-mean zonal wind field. The average behavior of the SAO is shown in [Fig. 11](#), which is a composite of rocketsonde observation taken over several years in the 1970s. One important difference between the SAO and the QBO is that the former has a regular period of 6 months, whereas the QBO has an irregular period, as we have seen. The regular behavior of the SAO is known to be governed by the influence of the seasonal cycle in the tropics.

Wave motions on a variety of scales are common in the tropics and, as noted earlier, some of these are thought to play a central role in bringing about the oscillations of the zonal-mean wind system. Best documented among these are the large-scale Kelvin and Rossby-gravity waves. [Figure 12](#) shows the horizontal structure of an equatorial Kelvin wave; the dynamics of the wave are discussed in [Section IV.B.4](#).

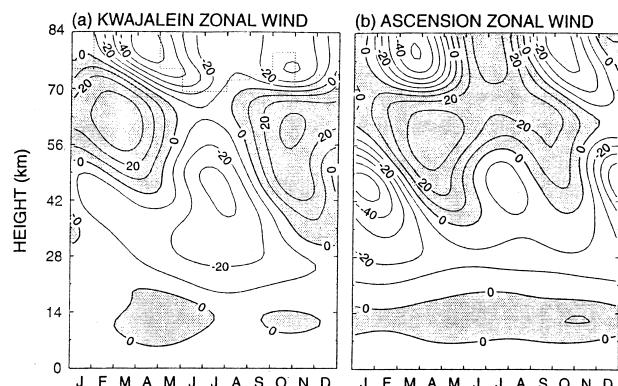


FIGURE 11 The SAO in zonal mean wind determined by rocketsonde observations. In contrast to the QBO, the SAO repeats regularly every 6 months. [From Garcia et al. (1997). *J. Geophys. Res.* **102**, 26019–26032.]

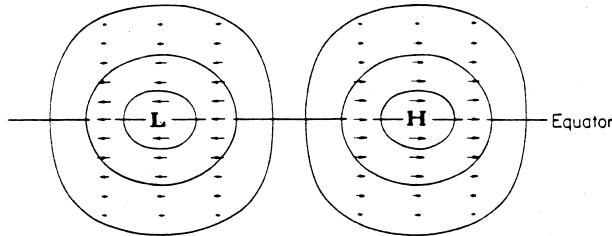


FIGURE 12 Schematic representation of the horizontal structure of an equatorial Kelvin wave. The contours denote geopotential anomalies, while the arrows denote zonal wind perturbations. Note that the Kelvin wave has no meridional wind perturbation. [From Andrews *et al.* (1987). “Middle Atmosphere Dynamics,” Academic Press, New York.]

C. The Mean Meridional Circulation

It was noted in Section II.C that departures from zonal-mean radiative equilibrium in the stratosphere are balanced mainly by adiabatic heating and cooling due to vertical motions. This fact can be used to deduce the vertical zonal-mean vertical motion of the stratosphere from its net heating rate; the zonal-mean meridional velocity can then be obtained by continuity.

Figure 13 shows the zonal-mean meridional circulation during Northern Hemisphere winter and summer calculated from the net heating rate. The latter was in turn obtained from observations of temperature and ozone made by the Upper Atmosphere Research Satellite (UARS). In the lower stratosphere, the mean meridional circulation is upward in the tropics and downward in extratropical latitudes throughout the year (this is the Brewer–Dobson circulation mentioned earlier). Above about 30 mb (25 km), the circulation exhibits a single-cell structure, directed mainly from the summer to the winter hemisphere.

In the tropical regions of the upper stratosphere a local circulation cell can be seen, with upward motion below about 1 mb (45 km), and downward motion above; poleward meridional motion is maximum near the vertical convergence level at 1 mb. This circulation cell, which is present during both solstices, is associated with the semi-annual oscillation in zonal wind described previously.

The magnitude of the mean meridional circulation is small, especially in comparison to the zonal-mean zonal winds discussed earlier. Meridional velocities are less than 1 m s^{-1} in the lower stratosphere and at most a few meters per second in the upper stratosphere. Vertical velocities are smaller by a factor of at least 100; they never exceed 1 cm s^{-1} , and can be as small as 0.1 mm s^{-1} in the lower stratosphere.

The mean meridional circulation is strongest in the winter hemisphere; in fact, the summertime circulation below about 5 mb in midlatitudes is so weak that this region of the stratosphere is close to radiative equilibrium, as can be appreciated from Fig. 5. Note also that the wintertime

circulation is stronger in the northern winter than in southern winter. The reasons for these seasonal and interhemispheric differences in the mean meridional circulation are addressed in Section IV.

IV. STRATOSPHERIC DYNAMICS

Much of the observed structure of the stratosphere can be understood in terms of elementary wave propagation, momentum and heat transport by waves, and *in situ* forcing by radiatively active trace gases. The dynamical aspects of these interactions can often be described satisfactorily in analytical terms, although detailed calculations of the stratospheric circulation require the use of numerical (computer) methods. In what follows, a brief introduction to stratospheric dynamic meteorology is presented. The conceptual development will be oriented toward the interpretation of the stratospheric observations discussed in the previous sections.

Because wave propagation and dissipation is a crucial feature of stratospheric dynamics, it is also convenient to distinguish between zonal-mean fields and perturbations thereto:

$$\Psi = \bar{\Psi} + \Psi' \quad (4)$$

where $\bar{\Psi}$ is the zonal mean defined earlier [Eq. (3)] and Ψ' is the perturbation, or eddy, component of the field. Perturbation fields describe deviations from zonal symmetry; in the stratosphere such deviations are usually associated with wave motions.

We will consider separately the equations that govern the zonal-mean flow and the eddies. To obtain zonal-mean and eddy equations, all variables in the equations of motion and thermodynamics are decomposed as in (4); zonal averaging of the results then produces the zonal-mean governing equations. Subtraction of the latter from the equations for the total fields yields the equations that govern the behavior of the eddies.

A. The Zonal Mean Circulation

In this article, the zonal-mean equations are written in the Transformed Eulerian Mean (TEM) formalism, whose derivation can be found in several of the works listed in the bibliography. The vertical coordinate adopted is log-pressure altitude, an isobaric (constant-pressure) coordinate defined by

$$z = H \ln(p_0/p) \quad (5)$$

where p_0 is a reference pressure and H is a scale height, usually taken to be 7 km for the stratosphere. The use of (5) is common in stratospheric work because it simplifies the form of the governing equations. Such a simplification

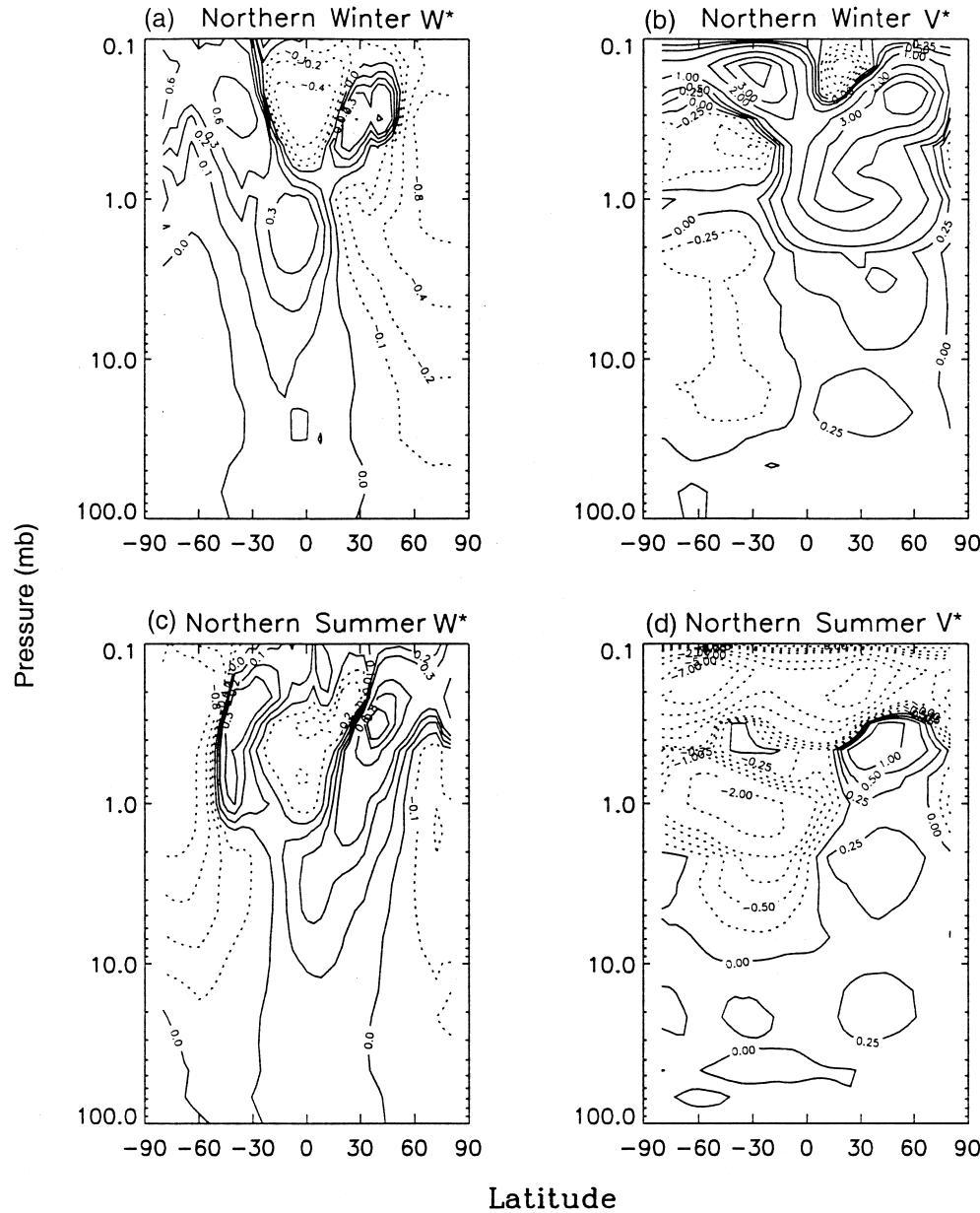


FIGURE 13 The mean meridional circulation of the stratosphere derived from temperature and chemical constituent observations made by instruments on-board UARS. [Courtesy of M. G. Mlynczak, NASA Langley Research Center.]

is possible because the Earth's atmosphere is largely in hydrostatic equilibrium, that is, the force of gravity is very nearly balanced by the vertical pressure gradient:

$$\frac{\partial p}{\partial h} = -\rho g \quad (6)$$

where g is the acceleration of gravity and ρ is the density. Thus, there exists a single-valued relationship between height h and pressure p that allows the two variables to be used interchangeably.

When (5) and (6) are combined with the ideal gas law $p = \rho RT$, where R is the gas constant for dry air, the

hydrostatic equation can be written in log-pressure coordinates as follows:

$$\frac{\partial \Phi}{\partial z} = \left(\frac{R}{H} \right) T \quad (7)$$

Where $\Phi = gh$ is the geopotential. The version of the TEM equations given hereafter is simplified further by omitting factors related to the sphericity of the Earth and neglecting certain terms that are known to be small.

In the TEM system the zonal-mean momentum equation is

$$\frac{\partial \bar{u}}{\partial t} + \bar{v}^* \left(\frac{\partial \bar{u}}{\partial y} - f \right) = \frac{1}{\rho_0} \nabla \cdot \mathbf{F} \quad (8)$$

where y, z, t are latitude, log-pressure altitude, and time; u, \bar{v}^*, \bar{w}^* denote the zonal-mean zonal, meridional, and vertical velocities; ρ is the density; f is the Coriolis parameter,

$$f = 2\Omega \sin \theta \quad (9)$$

Ω being the angular velocity of the Earth and θ the latitude; and \mathbf{F} is the Eliassen–Palm (EP) flux, a quantity that describes the acceleration of the zonal mean wind due to wave transience or dissipation. The divergence of the EP flux plays a central role in determining both the circulation and the temperature structure of the stratosphere, as shown in detail in Sections IV.A.2 and IV.A.3.

The meridional momentum equation can be shown to reduce to a balance between the Coriolis force on the zonal-mean zonal wind and the meridional pressure gradient:

$$f \bar{u} = - \frac{\partial \bar{\Phi}}{\partial y} \quad (10)$$

which is known as geostrophic equilibrium. Note that, in log-pressure coordinates, the pressure gradient is expressed in terms of the geopotential, Φ . Geostrophic equilibrium is a very good approximation for the zonal-mean circulation of the stratosphere on scales larger than about 1000 km.

The evolution of the temperature field is described by the thermodynamic energy equation:

$$\frac{\partial \bar{T}}{\partial t} + \bar{w}^* S = \bar{Q}_{\text{UV}} + \bar{Q}_{\text{IR}} \quad (11)$$

where $S = HN^2/R$; R is the gas constant for dry air; and N^2 is the Brunt–Väisälä, or buoyancy, frequency, a measure of the vertical stability of the atmosphere. Finally, the meridional and vertical velocities are related via the continuity equation

$$\frac{\partial \bar{v}^*}{\partial y} + \frac{1}{\rho} \frac{\partial (\rho \bar{w}^*)}{\partial z} = 0 \quad (12)$$

This relatively simple set of equations allows one to interpret the stratospheric observations described in Section III.

1. Thermal Wind Balance

For most meteorologically significant motions in the extratropical stratosphere, the Coriolis force is nearly balanced by the pressure gradient. This type of flow is called geostrophic. Under such conditions, the longitudinally averaged east–west wind is given by (10). With the aid of the hydrostatic equation (7), this can also be written as

$$\frac{\partial \bar{u}}{\partial z} = - \left(\frac{R}{fH} \right) \frac{\partial \bar{T}}{\partial y} \quad (13)$$

This alternative formulation of geostrophic equilibrium is known as the thermal wind equation; it relates the vertical shear of the zonal-mean zonal wind to the horizontal gradient of the zonal-mean temperature. If the temperature decreases toward the poles (as observed in the wintertime stratosphere), then a positive vertical shear will develop so that the zonal-mean zonal wind becomes increasingly eastward with altitude. On the other hand, if the poleward temperature gradient is positive (as is the case in the summer stratosphere), then the zonal wind will become more westward with altitude. This is precisely what is observed, as may be ascertained by comparing the mean temperature distributions of Fig. 5 with the zonal-mean zonal wind distributions of Fig. 6.

The thermal wind equation relates the appearance of the summer westward and winter eastward jets in the stratosphere to the temperature distribution, which is in turn a result of the latitudinal variation of heating due to absorption of ultraviolet radiation by ozone and large-scale motion of air parcels.

2. The Extratropical Mean Meridional Circulation

The constraint imposed by geostrophic equilibrium implies that the zonal-mean circulation of the stratosphere tends to follow longitude circles; zonal-mean motion in the meridional and vertical directions is then due to small departures from geostrophy. As noted previously, zonal-mean zonal winds are typically 10–100 times larger than zonal-mean meridional winds, and the latter in turn exceed vertical motions by another a factor of at least 100 (compare Fig. 6 and 13).

The zonal-mean circulation in the latitude–height plane is known as the mean meridional circulation. Although small compared to the zonal-mean zonal flow, the mean meridional circulation is of crucial importance for the thermal and chemical budgets of the stratosphere since the principal source terms in both cases vary mainly in the latitude–height plane. For example, the solar UV heating distribution has strong equator-to-pole and summer-to-winter gradients; similarly, production rates of, say, ozone peak in the tropics and diminish towards high latitudes. The effect of the mean meridional circulation is then to redistribute energy and minor chemical species in the latitude–height plane.

Under quasi-steady-state conditions, the existence of a mean meridional circulation outside of the tropics depends on the transport of momentum by atmospheric waves. Away from the equator f is much larger than \bar{u}_y , so the latter can be neglected in Eq. (8); if, in addition, one assumes

that the zonal-mean zonal wind evolves slowly, the equation reduces to

$$\bar{v}^* = -\frac{1}{\rho f} \nabla \cdot \mathbf{F} \quad (14)$$

which states that the steady-state meridional velocity, \bar{v}^* , vanishes if $\nabla \cdot \mathbf{F} = 0$. Since \bar{v}^* and \bar{w}^* are linked by the equation of continuity, it follows that \bar{w}^* also vanishes under these circumstances. In fact, it is easy to show from (12) and (14) that

$$\bar{w}^*(y, z) = -\frac{1}{\rho} \frac{\partial}{\partial y} \left[\frac{1}{f} \int_z^\infty \nabla \cdot \mathbf{F}(y, z') dz' \right] \quad (15)$$

This expression is often referred to as the “downward control principle,” because it states that the zonal-mean vertical velocity at altitude z depends on the integrated divergence of the $\nabla \cdot \mathbf{F}$ from the top of the atmosphere to the altitude in question.

Equations (14) and (15) constitute a remarkable result; they imply that, under quasistationary conditions, the extratropical mean meridional circulation is wave-driven. That is, (\bar{v}^*, \bar{w}^*) depend directly on $\nabla \cdot \mathbf{F}$, which, as noted earlier, is a measure of the acceleration of the zonal-mean flow due to wave dissipation. This explains the principal seasonal characteristics of the extratropical mean meridional circulation discussed in Section III.C. For example, the weakness of the meridional circulation during summer is a direct consequence of the inability of planetary waves to propagate into the westward zonal wind regime that prevails at that time of the year. Similarly, the greater strength of the wintertime meridional circulation in the Northern Hemisphere can be ascribed to the fact that planetary wave excitation is considerably stronger in the Northern than in the Southern Hemisphere (see Fig. 9).

Furthermore, because the net radiative drive is balanced primarily by adiabatic heating and cooling (see Section II.C), it follows that any departures of the zonal-mean temperature field from radiative equilibrium are also wave driven. This can be expressed analytically by writing the net radiative drive as a relaxation toward radiative equilibrium,

$$\bar{Q}_{\text{UV}} + \bar{Q}_{\text{IR}} = -\frac{1}{\tau} (\bar{T} - \bar{T}_e) \quad (16)$$

where τ is the time scale for radiative relaxation and \bar{T}_e is the radiative equilibrium temperature. Then, from (11), (15), and (16),

$$(\bar{T} - \bar{T}_e) = \frac{\tau S}{\rho} \frac{\partial}{\partial y} \left[\frac{1}{f} \int_z^\infty \nabla \cdot \mathbf{F}(y, z') dz' \right] \quad (17)$$

where, as before, $\partial \bar{T} / \partial t$ is neglected under the quasi-steady-state assumption. Equation (17) shows explicitly

the connection between departures from \bar{T}_e and the divergence of the EP flux.

The foregoing results depend on the assumption of slowly varying conditions, but in practice this detracts little from their general applicability. On seasonal time scales, and especially around the solstices, the stratospheric circulation is for the most part slowly varying, so the conclusion that the circulation is wave-driven is a robust one. Other processes can also drive a mean meridional circulation; the most important among these are the seasonal variation of the radiative drive, and the sudden warming phenomenon. The former is most important around the equinoxes, when solar heating rates are changing rapidly; the latter is a recurrent feature of the winter stratosphere, especially in the Northern Hemisphere.

3. Sudden Stratospheric Warmings

The stratospheric sudden warming phenomenon is the result of wave–mean flow interaction following the rapid amplification of extratropical planetary waves. This phenomenon has been intensively studied and has also been simulated with a variety of numerical models, ranging from simple mechanistic ones to fully interactive general circulation models.

The evolution of the sudden warming can be summarized as follows: Under normal conditions, planetary waves propagate from the troposphere toward the tropical stratosphere, as shown in Fig. 14a, where the heavy lines indicate the direction of propagation of the EP flux, \mathbf{F} . As planetary waves amplify, the direction of propagation switches toward the pole in the sequence shown in Fig. 14b–e. A rapid increase in the magnitude of the EP flux divergence ensues, which decelerates the zonal-mean zonal wind while producing rapid adiabatic warming at high latitudes via the wave-driven circulation shown in Fig. 15. Incidentally, this circulation also lowers tropical temperatures (due to upwelling, and hence expansional cooling). However, the warming at high latitudes is much more intense and gives the phenomenon its name.

The sense of the departures from radiative equilibrium that accompany a sudden warming is consistent with Eq. (17). Note that, because $\nabla \cdot \mathbf{F} < 0$, the RHS of the equation is negative poleward of the latitude where $|\nabla \cdot \mathbf{F}|$ is a maximum, and positive equatorward. Thus, $\bar{T} - \bar{T}_e < 0$ in the tropics and $\bar{T} - \bar{T}_e > 0$ at high latitudes, as observed. However, Eq. (17) cannot be used to estimate the magnitude of the departures from radiative equilibrium because it was derived under the slowly varying assumption, which is not applicable to the sudden warming phenomenon.

It turns out that the response of the stratosphere to $\nabla \cdot \mathbf{F}$ during a typical sudden warming is about equally divided

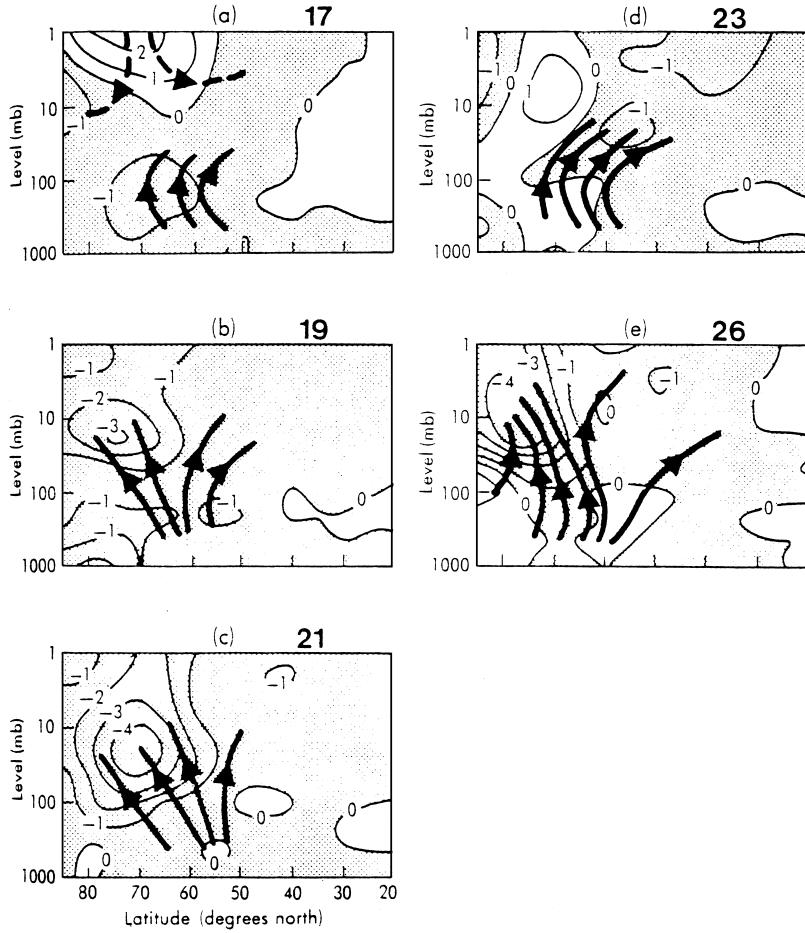


FIGURE 14 The propagation of the EP flux due to planetary waves (heavy arrows) and its divergence (contours, in units of 10^4 m s^{-2}) during the sudden warming of February 1979. [Adapted from T. N. Palmer (1981). *J. Atmos. Sci.* **38**, 844–855.]

between the deceleration of the zonal wind and the generation of a wave-driven circulation. Thus, during a sudden warming,

$$\bar{u}_t \simeq -f \bar{v}^* \simeq \frac{1}{2} \left[\frac{1}{\rho} \nabla \cdot \mathbf{F} \right] \quad (18)$$

This result can be obtained from simple scaling arguments of the zonal-mean equations (8)–(12) and is consistent with detailed computer simulations.

4. The Tropical Oscillations

The analytical approach used in Section IV.A.2 is not applicable to the tropical circulations, even though these are low-frequency phenomena, with time scales comparable to or longer than that of the extratropical circulation. The weakness of the Coriolis force in the tropics tends to decouple the zonal and meridional wind fields, so even at very low frequencies the EP-flux divergence acts mainly

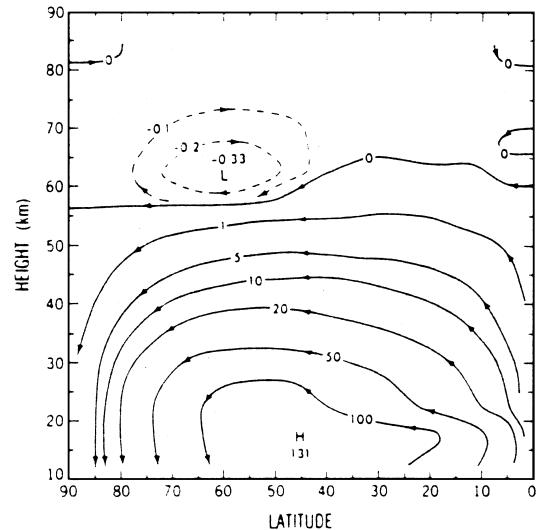


FIGURE 15 The mean meridional circulation induced by wave driving during a stratospheric sudden warming. [From T. J. Dunkerton *et al.* (1981). *J. Atmos. Sci.* **38**, 819–843.]

to accelerate the zonal flow rather than to drive a mean meridional circulation, that is,

$$\bar{u}_t \simeq \frac{1}{\rho} \nabla \cdot \mathbf{F} \quad (19)$$

As a consequence, the zonal-mean zonal winds of the tropical stratosphere exhibit downward-propagating oscillations (see Figs. 10 and 11), which can only be understood in terms of alternating absorption of waves that deposit westward and eastward momentum in the stratosphere.

Figure 16 illustrates schematically the mechanism responsible for the oscillatory wind regimes of the tropical stratosphere. The mechanism was originally proposed to explain the QBO, but it applies equally well to the SAO with some modifications. In panel (a) of the figure, the winds are westward throughout most of the domain (which can be thought to correspond to the stratosphere below

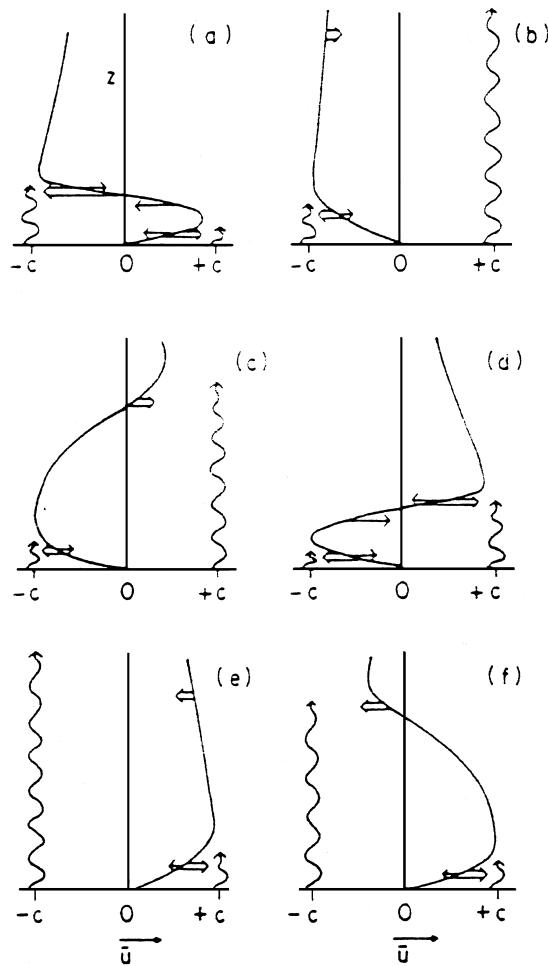


FIGURE 16 A schematic diagram of the sequence of wave forcing and zonal-mean zonal flow evolution that gives rise to the tropical wind oscillations. [After R. A. Plumb (1984). In "Dynamics of the Middle Atmosphere," pp. 217–251, Terrapub, Tokyo.]

about 45 km); a small region of eastward winds is present at the bottom of the domain. Waves propagating from below and having both positive (eastward) and negative (westward) phase velocities are denoted by the wave arrows labeled $+c$ and $-c$, respectively.

Eastward-propagating waves ($c > 0$) are absorbed at their critical levels (where $\bar{u} = c$), while eastward waves propagate to higher altitudes where they force westward winds. The result, shown in panel (b), is that the westward phase descends and occupies nearly the entire domain; this in turn allows eastward waves ($c < 0$) to propagate to higher altitudes, where they begin to force a new eastward phase. As the westward phase descends further [panel (c) of the figure], a new eastward phase develops near the top of the domain. In this way, alternating eastward and westward regimes are generated and propagate downward.

The period of oscillation of these alternating regimes is about 27 months for the QBO, but precisely 6 months for the SAO. Computer simulations suggest that the QBO period is sensitive to the strength of the wave forcing, such that stronger forcing decreases the period and vice versa. Interannual variations in the strength of wave forcing presumably account in part for the variability of the QBO cycle.

The period of the SAO, on the other hand, is regulated by the seasonal cycle because of one crucial difference between the QBO and the SAO: The westward wind phase of the SAO in the stratosphere (see Fig. 11) is produced, not by wave forcing, but by the advection of westward zonal-mean momentum by the mean meridional circulation. In the tropics, the term $\bar{v}^* \bar{u}_y$ in Eq. (8) cannot be neglected with respect to the Coriolis force, which vanishes at the equator. As shown in Fig. 13, the meridional velocity \bar{v}^* is relatively large near the stratopause and its direction is such as to produce negative (westward) zonal-mean zonal winds.

The presence of westward winds at the solstices leads to selective absorption of westward waves in the stratosphere and propagation of eastward waves to the mesosphere, where they force eastward winds. Near the equinoxes, on the other hand, eastward winds can be generated in the stratosphere by dissipation of eastward waves, while westward waves propagate to the mesosphere and force a mesospheric westward phase. In this way the seasonal cycle determines the timing of the SAO wind regimes.

The waves responsible for driving the SAO in the stratosphere are mainly large-scale ($k = 1-3$) Kelvin waves, although small-scale gravity waves apparently also play a role. As for the QBO, it remains unclear whether the main driver are large-scale Kelvin and inertia-gravity waves, or small-scale gravity waves. Recent computer models and observational evidence suggest that the latter are likely to be most important.

The oscillating wind regimes depicted schematically in Fig. 16 (and from observations in Figs. 10 and 11) are accompanied by secondary circulations in the meridional plane, and by temperature anomalies in balance with the adiabatic heating and cooling associated with the secondary circulations. The magnitude of these anomalies can be estimated because, even in the tropics, the zonal-mean zonal wind and temperature are in thermal wind balance. Near the equator, Eq. (13) can be written as follows:

$$\beta y \bar{u}_z = -\frac{R}{H} \bar{T}_y \quad (20)$$

where the Coriolis parameter has been approximated as $f = \beta y$, and β is the rate of change of f at the equator. Ignoring variations of \bar{u} with respect to y , differentiation of Eq. (20) yields

$$\bar{T}_{yy} = -\frac{H\beta}{R} \bar{u}_z \quad (21)$$

Assuming a meridional scale L for \bar{T} , the equatorial temperature anomaly associated with the tropical wind oscillation \bar{u} is approximately

$$\bar{T} = \frac{L^2 H \beta}{R} \bar{u}_z \quad (22)$$

From the thermodynamic equation (11) it can be deduced further that

$$\bar{w}^* = -\frac{L^2 \beta}{N^2 \tau} \bar{u}_z \quad (23)$$

where the net heating rate has been written as relaxation toward radiative equilibrium, with time scale τ .

Figure 17 shows the temperature and vertical velocity anomalies associated with a tropical zonal-mean zonal wind oscillation. Maximum temperature anomalies occur at the altitude where the zonal-mean zonal wind changes sign, that is, where \bar{u}_z is largest. The temperature anomalies are accompanied by vertical velocity anomalies consistent with Eq. (23). Where $\bar{u}_z > 0$ (top panel of the figure) there is downwelling ($\bar{w}^* < 0$) and a positive temperature anomaly; the reverse occurs where $\bar{u}_z < 0$. Typical values of the temperature and vertical velocity anomalies for the QBO are about 3 K and 0.1 mm s⁻¹; these follow from (22) and (23) if one takes $L \sim 1000$ km, $\bar{u}_z \sim 5$ m s⁻¹ km⁻¹, and $\tau \sim 30$ days.

B. Waves in the Stratosphere

In the preceding sections we have seen that wave dissipation, expressed by the EP flux divergence $\nabla \cdot \mathbf{F}$, plays a central role in determining the stratospheric circulation. We now present a brief description of the principal types of wave found in the stratosphere. Atmospheric waves can

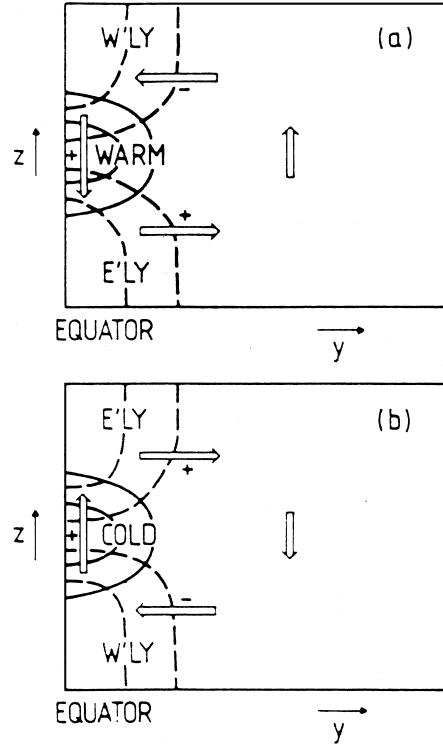


FIGURE 17 Temperature and vertical velocity anomalies associated with the zonal-mean zonal wind in the QBO or SAO. [From R. A. Plumb and R. C. Bell (1982). *Q. J. R. Meteorol. Soc.* **108**, 335–352.]

be thought of as originating from constraints imposed on air motions by density stratification and angular momentum conservation. As noted earlier, the atmospheric density stratification is generally stable. As a consequence, air parcels displaced vertically from their equilibrium positions tend to oscillate about those positions. Similarly, angular momentum conservation acts through the Coriolis force to restore moving air parcels to their equilibrium positions in latitude.

Most waves observed in the stratosphere are excited in the troposphere and propagate into the stratosphere. As they propagate vertically, conservation of energy requires that their amplitude grow as the atmospheric density decreases. For this reason, waves that appear to be insignificant in the troposphere, such as many of the traveling waves, will have a much larger amplitude in the upper atmosphere.

A full description of atmospheric waves is beyond the scope of this work; however, it is possible to gain some insights into the mechanisms responsible for wave motion by considering a simplified set of equations that govern the motions of the eddies, or perturbations, to the zonal-mean flow. This set consists of the momentum equations in the zonal and meridional directions,

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right) u' - f v' = - \frac{\partial \Phi'}{\partial x} \quad (24)$$

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right) v' + f u' = - \frac{\partial \Phi'}{\partial y} \quad (25)$$

the thermodynamic equation,

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right) T' + w' S = 0 \quad (26)$$

the hydrostatic equation,

$$\frac{\partial \Phi'}{\partial z} = \left(\frac{R}{H} \right) T' \quad (27)$$

and the continuity equation, which links the horizontal and vertical components of the motion field,

$$u'_x + v'_y + \rho^{-1}(\rho w')_z = 0 \quad (28)$$

1. Gravity Waves

According to (24) and (25), eddy accelerations are due to buoyancy forces (represented by the gradient of the geopotential) and to the Coriolis force. It turns out that the latter can be neglected when the horizontal scale of the motion is sufficiently small (a few hundred kilometers). Then, differentiation of (24) with respect to x and of (25) with respect to y yields an equation for the divergence, $(u'_x + v'_y)$:

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right) (u'_x + v'_y) = -\nabla^2 \Phi' \quad (29)$$

With the aid of (26)–(28) this reduces to an equation for Φ' alone,

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right)^2 \left(\Phi'_{zz} - \frac{\Phi'_z}{H} \right) + N^2 \nabla^2 \Phi' = 0 \quad (30)$$

which describes the behavior of small-scale gravity waves.

Equation (30) has solutions of the form

$$\Phi' = \Phi_0 \exp[i(kx + ly + mz - \omega t)] \exp(z/2H) \quad (31)$$

where k , l , and m are the zonal, meridional, and vertical wavenumbers, respectively, $\omega = kc$ is the frequency, and c is the phase speed. (The zonal wavenumber, k , is related to the zonal wavelength λ_x by $k = 2\pi/\lambda_x$, and similarly for the meridional wavenumber, l , and vertical wavenumber, m .) The factor $\exp(z/2H)$ describes the vertical amplitude growth that results from the density stratification of the atmosphere. Wave energy is proportional to $\rho|\Phi'|^2$, while $\rho = \exp(-z/H)$, so the solution (31) represents a wave that propagates conservatively in the vertical.

Substituting (31) into (30) yields the dispersion relationship for small-scale gravity waves:

$$(k\bar{u} - \omega)^2 = \frac{(k^2 + l^2) N^2}{\left(m^2 + \frac{1}{4H^2} \right)} \quad (32)$$

Gravity waves affect the zonal-mean state of the stratosphere when they dissipate, usually through wave breaking. Wave breaking occurs when vertically propagating waves reach sufficiently large amplitude that their temperature perturbations produce a locally unstable temperature stratification. This is bound to happen as the small-scale gravity waves propagate vertically because they are not significantly attenuated by other processes and therefore grow exponentially with height, as mentioned earlier.

Wave temperature perturbations also grow rapidly in the vicinity of critical layers, that is, regions where $(k\bar{u} - \omega) \rightarrow 0$. It can be seen from (32) that, when this happens,

$$m^2 = \frac{(k^2 + l^2) N^2}{(k\bar{u} - \omega)^2} - \frac{1}{4H^2} \rightarrow \infty \quad (33)$$

so that the vertical temperature gradient $\partial T'/\partial z = imT'$ becomes very large, leading to unstable stratification.

Small-scale gravity waves generated in the troposphere are known to be crucial for the momentum budget and the transport of minor species in the mesosphere. In the stratosphere, their role is still being debated, although it is certainly more modest than in the mesosphere. Nevertheless, breaking gravity waves may be responsible for much of the small-scale vertical mixing in the stratosphere, and of the clear-air turbulence found in the upper troposphere.

2. Rossby Waves

The gravity waves just described occur at high frequencies and small scales. For motions with intrinsic periods greater than a day, and horizontal scales of the order of 1000 km or larger, a distinct type of wave motion appears, the Rossby wave (named after Carl Gustav Rossby, who first pointed out their meteorological importance in the 1930s). The Rossby wave owes its existence to the variation of the Coriolis parameter, f , with latitude.

A governing equation for Rossby waves can be derived by noting, first of all, that at low frequencies and outside the tropics, the eddy momentum equations (24) and (25) reduce to the condition of geostrophic equilibrium:

$$-f v' = -\Phi'_x \quad (34)$$

$$f u' = -\Phi'_y \quad (35)$$

so that the motion field (u', v') is directly related to the gradient of the geopotential Φ' . When this condition holds,

a vorticity equation can be obtained by taking the curl of the momentum equations, that is, $\partial(25)/\partial x - \partial(24)/\partial y$:

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right) (v'_x - u'_y) + v' \beta + f(u'_x + v'_y) = 0 \quad (36)$$

where $(v'_x - u'_y)$ is the vorticity of the flow, $(u'_x + v'_y)$ is its divergence, and $\beta = \partial f / \partial y$.

An equation in terms of Φ' can then be obtained by writing the vorticity in terms of the geopotential by means of the geostrophic equations (34) and (35). Note that the divergence term would be zero if expressed in terms of the geostrophic velocities. Instead, slight departures from geostrophy are represented by eliminating the divergence term $(u'_x + v'_y)$ by means of the thermodynamic (26)–(28). The result is the quasigeostrophic potential vorticity equation:

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right) q' + v' \beta = 0 \quad (37)$$

where

$$q' = \left(\nabla^2 \Phi' + \frac{f^2}{N^2} \frac{1}{\rho} \frac{\partial \rho \Phi'_z}{\partial z} \right) \quad (38)$$

is the potential vorticity.

Substitution into (37) of a solution of the form (31) leads immediately to the dispersion relation for Rossby waves,

$$\omega = k\bar{u} - \frac{\beta k}{k^2 + l^2 + \frac{f^2}{N^2} \left(m^2 + \frac{1}{4H^2} \right)} \quad (39)$$

The dependence of the vertical wavenumber m on other parameters follows from (39);

$$m^2 = \frac{N^2}{f^2} \left(\frac{\beta k}{k\bar{u} - \omega} - k^2 - l^2 \right) - \frac{1}{4H^2} \quad (40)$$

whence the conditions necessary for vertical propagation of Rossby waves can be deduced. A vertically propagating wave must have real vertical wavenumber, and thus $m^2 > 0$; otherwise m is imaginary and the wave decays with altitude. It can be seen from (39) that vertical propagation requires

$$0 \leq (\bar{u} - c) \leq \frac{\beta}{\left(k^2 + l^2 + \frac{f^2}{N^2} \frac{1}{4H^2} \right)} \quad (41)$$

This simple condition can be used to explain both the seasonal behavior and the horizontal scale of Rossby waves found in the stratosphere. The first inequality on (41) implies that quasi-stationary ($c \rightarrow 0$) waves will not be found in the summer stratosphere. This prediction is borne out by observations, as shown in Fig. 9. In the stratosphere, where $\bar{u} < 0$ in summer, quasistationary waves are absent; in fact, the summertime stratospheric circulation essentially follows latitude circles, as seen earlier in connection with Fig. 7.

The second inequality determines the scale of the waves that can propagate vertically in winter, when the zonal-mean zonal wind is eastward ($\bar{u} > 0$). For example, the midlatitude troposphere is dominated by slow westward-moving Rossby waves with horizontal scales on the order of a few thousand kilometers that arise from instability of the tropospheric flow. However, these waves are not found in the stratosphere because their horizontal wavenumbers ($k = 2\pi/\lambda_x$) are relatively large so the second inequality cannot be satisfied for typical values of \bar{u} . On the other hand, very large scale waves (horizontal wavelengths of tens of thousands of kilometers) have smaller horizontal wavenumbers and can propagate into the stratosphere. This accounts for the fact that the wave field in the winter stratosphere is dominated by planetary-scale Rossby waves, with wavenumber k in the range 1–3.

3. Rossby Normal Modes

It was noted in Section III.A that, in addition to quasistationary Rossby waves, traveling waves are also present in the stratosphere. For a wave of the form (31), a traveling wave is simply one for which the phase speed c is nonzero; the perturbation associated with such a wave moves with respect to the space coordinates.

Normal modes are solutions to the homogeneous wave equation, (37) in the case of Rossby waves, with homogeneous (unforced) boundary conditions. For the upper boundary condition it is required that upward propagating waves radiate outward from the upper boundary (radiation condition) or, in the case of trapped waves, that their energy remain finite. A suitable lower boundary condition for Rossby waves is obtained by applying the thermodynamic equation (26) at some material lower boundary and requiring that the vertical velocity in height coordinates vanish. Ignoring the zonal-mean zonal wind \bar{u} , it can be shown that this leads to the equation

$$\frac{\partial}{\partial t} \left(\Phi'_z - \frac{\kappa}{H} \Phi' \right) = 0 \quad (42)$$

where the approximation $N^2 = \kappa g / H$ has been made; $\kappa = R/c_p$; and c_p is the specific heat of air at constant pressure. Using (31), this gives

$$i\omega \left(im + \frac{1}{2H} - \frac{\kappa}{H} \right) = 0 \quad (43)$$

whence

$$m = i \left(\frac{1}{2H} - \frac{\kappa}{H} \right) \quad (44)$$

Substituting Eq. (44) into (31) yields the characteristic vertical structure of normal modes in the Earth's atmosphere,

$$\Phi' = \Phi_0 \exp[i(kx + ly - \omega t)] \exp(\kappa z / H) \quad (45)$$

Note that normal modes do not propagate vertically; the vertical wavenumber m is imaginary, per Eq. (44), and the vertical structure is given simply by the exponential growth factor $\exp(z/2H)$. This means that the amplitude of normal modes grows with altitude, so it can become substantial in the upper atmosphere (see Table II). On the other hand, the energy behaves as $\rho|\Phi'|^2 \propto \exp(2\kappa - 1)/H$ and remains bounded since $\kappa \simeq 2/7$.

4. The Equatorial Kelvin Wave

In the tropics, where f is small, the distinction between inertia-gravity waves and Rossby waves becomes blurred. The gravity and Rossby waves present at midlatitudes also occur in the tropics where, depending on their frequency, wave energy may be trapped. These so-called equatorial waves include Rossby and inertia-gravity waves, as well as a type of wave unique to the equatorial region, the Kelvin wave. The theoretical description of large-scale equatorial Rossby and inertia-gravity waves is beyond the present scope; instead we present a brief description of the Kelvin wave, which plays a major role in driving the SAO (Figure 11), and which illustrates the peculiar interplay of forces that can occur in the tropics.

The salient characteristic of the Kelvin wave is that it has no meridional motion field; air motions induced by this type of wave take place exclusively in the longitude–altitude plane. The equations governing the Kelvin wave are thus a subset of (24)–(28):

$$\left(\frac{\partial}{\partial t} + \bar{u}\frac{\partial}{\partial x}\right)u' = -\Phi'_x \quad (46)$$

$$\beta y u' = -\Phi'_y \quad (47)$$

$$\left(\frac{\partial}{\partial t} + \bar{u}\frac{\partial}{\partial x}\right)\Phi'_z + w'N^2 = 0 \quad (48)$$

$$u'_x + \rho^{-1}(\rho w')_z = 0 \quad (49)$$

where the thermodynamic equation (48) has been written in terms of the geopotential, making use of the hydrostatic relationship (7). The Coriolis force in the (47) is given approximately as $f \simeq \beta y$. Note that the absence of meridional motion v' decouples (46) and (47), which otherwise would be linked by the Coriolis force [as is the case in the full eddy equations (24) and (25)]. As a consequence, it is possible to eliminate u' between (46) and (47) to obtain an equation for the meridional structure of the Kelvin wave,

$$\frac{1}{\Phi'}\frac{\partial\Phi'}{\partial y} = -\frac{k\beta y}{\omega - k\bar{u}} \quad (50)$$

so that the meridional structure of the wave is given by the Gaussian

$$\Phi' = \exp\left(-\frac{y^2}{2L^2}\right) \quad (51)$$

with characteristic width

$$L = \sqrt{\frac{k\beta}{2(\omega - k\bar{u})}} \quad (52)$$

The characteristic Gaussian structure of the Kelvin wave was illustrated in Fig. 12. In other respects, the Kelvin wave resembles the gravity waves discussed in Section IV.B.1; in particular, its dispersion relation is

$$(k\bar{u} - \omega)^2 = \frac{k^2 N^2}{\left(m^2 + \frac{1}{4H^2}\right)} \quad (53)$$

which is identical to that of small-scale gravity waves (32) except that there is no meridional wavenumber, l , because the motion is limited to the longitude–altitude plane, as noted previously.

Kelvin waves figure prominently in observations of the stratospheric wave field in the tropics. They are known to play a major role in driving the eastward phase of the SAO in the upper stratosphere (see Section IV.A.4).

5. Wave Dissipation

The Eliassen–Palm flux has figured prominently in the examples of wave–mean flow interaction discussed earlier. It was noted throughout that its divergence, $\nabla \cdot \mathbf{F}$ decelerates the wintertime zonal-mean zonal jets, drives the stratosphere away from radiative equilibrium, makes possible the existence of a mean meridional circulation outside the tropics, and provides the forcing necessary to account for the tropical oscillation systems. It was also emphasized that $\nabla \cdot \mathbf{F} = 0$ unless the wave field is unsteady or undergoing dissipation; steady, conservative waves propagate through the atmosphere without interacting with the zonal mean flow.

The relationship between wave dissipation and the zonal-mean circulation is known as the nonacceleration theorem. It can be illustrated by means of the following simple example for planetary waves. The EP flux divergence for planetary Rossby waves can be shown to be given by

$$\nabla \cdot \mathbf{F} = \rho \overline{v' q'} \quad (54)$$

where q' is the eddy potential vorticity. The dependence of $\overline{v' q'}$ upon dissipative processes can be investigated using the potential vorticity equation (37), which governs the behavior of Rossby waves. Assuming that linear damping, with dissipation coefficient δ , operates upon both the velocity and temperature fields, the latter can be written as

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right) q' + v' \beta = -\delta q' \quad (55)$$

From (55) it is easy to show that

$$\nabla \cdot \mathbf{F} = \rho \bar{v}' q' = -\frac{\delta v' v' \beta}{(k \bar{u} - \omega)^2 + \delta^2} \quad (56)$$

which states that $\nabla \cdot \mathbf{F}$ vanishes in the absence of dissipation. Similar simple demonstrations of the nonacceleration theorem can be given for other types of wave, the only difference being the particular form assumed by $\nabla \cdot \mathbf{F}$.

The study of the mechanisms responsible for wave dissipation in the stratosphere is a topic of much current interest. It has long been known that planetary waves can be dissipated through thermal damping of the temperature field (via IR heating or cooling); in fact, the earliest theoretical treatments of the sudden warming phenomenon assumed that thermal dissipation alone accounted for the EP flux divergence that gives rise to the sudden warming. More recently, however, it has become clear that wave breaking is a major dissipation mechanism for both planetary Rossby waves and gravity waves. An atmospheric wave can break when it induces unstable perturbations in the medium. The simplest case to visualize is that of gravity waves, whose temperature perturbations can occasionally become large enough to make the atmosphere convectively unstable.

Gravity wave breaking resembles the familiar breaking of ocean waves on a beach. Figure 18 shows a computer simulation of gravity wave breaking; in this case the wave breaks at very high altitude because the initial perturbation (near the ground) is quite small. However, gravity waves excited by strong winds blowing over mountain barriers, and by the passage of strong frontal zones in the troposphere, are known to break in the lower stratosphere.

Rossby wave breaking is apparently due to barotropic instability, an instability of the potential vorticity field that arises when the wave perturbation to the total potential vorticity gradient becomes sufficiently large. Figure 19 shows an example of Rossby wave breaking in the stratosphere. Rossby wave breaking appears to be the main dissipation mechanism in the subtropical and midlatitude stratosphere (that is, between about 20 and 50 degrees of latitude), while thermal damping dominates at higher latitudes. The range of latitudes where Rossby wave breaking usually takes place has come to be known as the “stratospheric surf zone”; as the name implies, dissipation through breaking also leads to efficient lateral mixing, a process that is important for redistributing minor species in the stratosphere.

V. TRANSPORT OF MINOR SPECIES

The circulation systems described above have a large impact on the distribution of minor chemical species in the

stratosphere. In this regard, transport in the meridional plane is most important because the sources and sinks of minor chemical species vary most strongly with latitude and height. The effect of transport on chemical composition is illustrated dramatically by the case of ozone, a minor species that has been intensively studied in recent years in connection with the polar ozone hole phenomenon.

Ozone is produced in the tropics by the process outlined in Section II; it is destroyed by recombination with atomic oxygen, and by several catalytic cycles, including those involving chlorine and bromine of anthropogenic origin. The principal region of ozone production is in the tropics, and yet observations of the ozone content of the atmosphere show that the largest number of ozone molecules is found in the polar regions. This implies that ozone must be transported away from its source; this transport is carried out by the combined effect of the mean meridional circulation and quasihorizontal mixing due to wave breaking.

Figure 20 shows schematically the effect of transport processes on the distribution of ozone in the stratosphere. The mean meridional circulation carries ozone poleward and downward to the high-latitude, lower stratosphere, where it can accumulate because its destruction rate there is usually slower than in the middle and upper stratosphere.

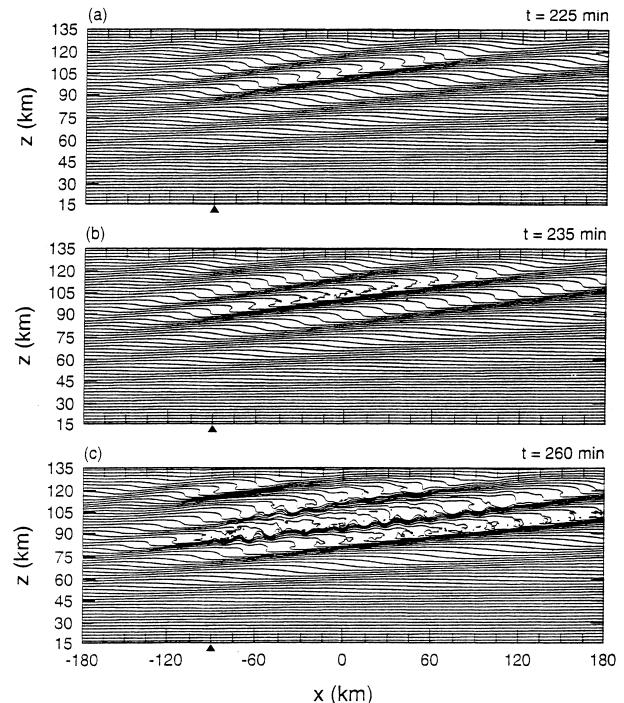


FIGURE 18 Gravity wave breaking simulated by a computer model. The three panels depict the evolution of the potential temperature field perturbed by a gravity wave excited at the location of the black triangle on the lower boundary. At 235 m after the initial excitation of the wave, potential temperature surfaces steepen and begin to overturn; full breaking is in progress by 260 m. [From Prusa *et al.* (1996). *J. Atmos. Sci.* **53**, 2186–2216.]

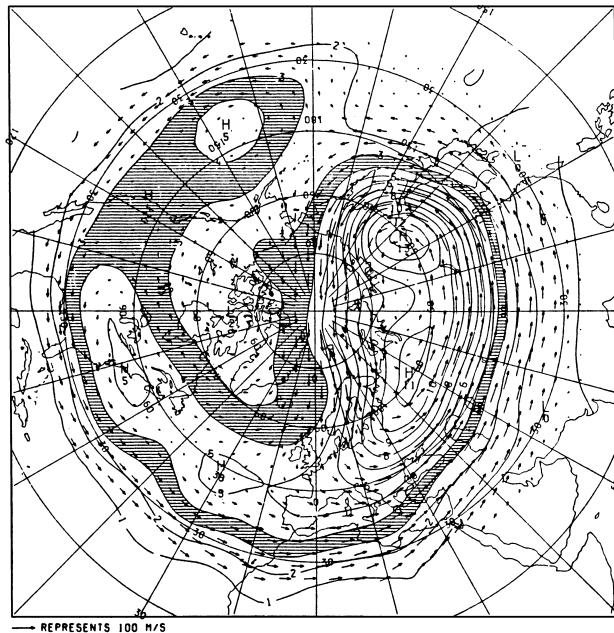


FIGURE 19 Polar stereographic map of potential vorticity near 30 km altitude, showing the breaking of a Rossby planetary wave. During Rossby wave breaking potential vorticity contours roll up and eventually break down. Breaking takes place on a quasihorizontal (constant potential temperature) surface. [From Andrews *et al.* (1987). “Middle Atmosphere Dynamics,” Academic Press, New York.]

(The exception is during the spring over the polar caps, where activation of anthropogenic chlorine and bromine can lead to rapid ozone loss.) The mean meridional circulation in the stratosphere is driven mainly by planetary wave dissipation (Section IV.B.5); insofar as dissipation occurs through wave breaking, the planetary waves also mix ozone in the north–south direction, increasing its abundance at high altitudes.

A. Polar Ozone Depletion

The mechanisms outlined in Fig. 20 operate in both the Northern and Southern Hemispheres; however, the evolution of ozone throughout the year shows large interhemispheric differences that are directly attributable to the difference in the behavior of planetary waves shown in Fig. 9.

In the late 1980s, aircraft were flown into the austral and boreal polar stratosphere to investigate the depletion of lower stratospheric ozone at high latitudes. The aircraft measurements revealed that stratospheric air poleward of the wintertime zonal-mean jet core is highly isolated from midlatitude air during winter. This isolation, combined with the cold temperatures and subsequent polar stratospheric cloud appearance, allows the chemical environment to become highly perturbed. Stratospheric temperatures below about 195 K at 20 km are required for

the formation of polar stratospheric clouds. Polar stratospheric cloud particles form the surfaces for reactions that activate chlorine. Upon the return of sunlight to the polar caps in spring, chlorine takes part in ozone-destroying catalytic cycles.

In the Southern Hemisphere, chlorine activation leads to a remarkable springtime decrease in ozone that has come to be known as the “Antarctic ozone hole”; on the other hand, polar ozone depletion is considerably smaller in the Arctic. The difference arises from the weaker wave driving of the stratospheric circulation of the Southern Hemisphere. As seen in Fig. 9, wave amplitudes are smaller in the Southern Hemisphere, and so is wave driving, to the point that the southern polar jet persists much longer into the spring than its northern counterpart. The delayed breakdown of the southern polar vortex allows ozone loss to continue throughout the months of September and October. In the Northern Hemisphere, on the other hand, the conditions necessary for ozone loss usually disappear in mid- to late March, when the northern polar vortex breaks down.

The validity of this explanation is borne out by the occurrence of several large Arctic ozone depletion events during the 1990s. Normally, the relatively warm boreal polar vortex covers about half the area of the austral vortex. Polar stratospheric clouds are also observed during the Arctic winter, and in several years during the 1990s, clouds were abundant and persisted well into Arctic spring, leading to the observed large depletion in Arctic ozone.

VI. CURRENT RESEARCH

Through the mid-1960s, knowledge of the stratosphere was derived mainly from relatively sparse ground-based observations, theoretical treatments of stratospheric dynamics were incomplete, and computer modeling was rudimentary. The advent of polar-orbiting satellite observations in the late 1960s, and especially in the 1970s, provided a global view of the stratosphere and stimulated the development of comprehensive theories. At the same time, rapid increases in computing power and advances in computational fluid dynamics have led to the development of ever more detailed numerical models.

Today, it may be said that stratospheric meteorology is a mature field, and that the fundamental dynamical and transport mechanisms that operate in the region are well understood. This is no doubt due in part to the fact that the basic processes of radiative heating and wave dissipation are relatively simple and amenable to theoretical and numerical treatment. Still, the stratosphere has provided some surprises in recent years, the best known being the development of severe ozone loss during Antarctic

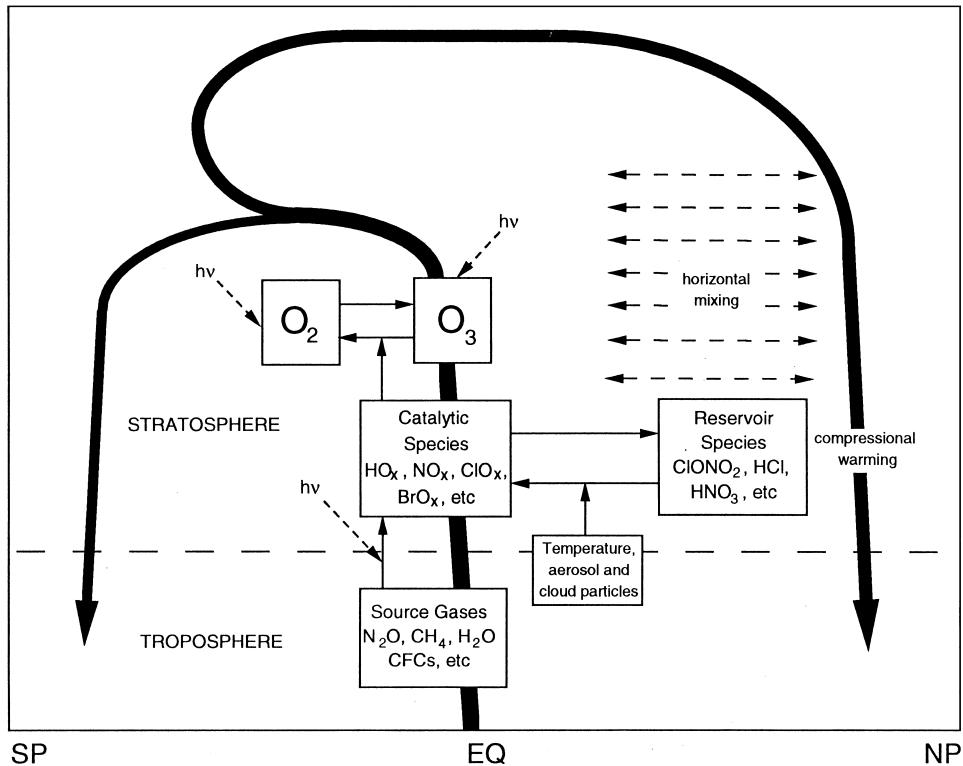


FIGURE 20 Schematic of the mechanisms that affect ozone distribution in the stratosphere. The mean meridional circulation is denoted by the heavy arrows; it transports ozone poleward and downward from the source region in the tropical middle stratosphere. The circulation also carries into the stratosphere compounds of anthropogenic origin that contribute to ozone loss. The effect of quasihorizontal mixing by breaking planetary waves is denoted by the dashed lines with arrows. [From R. R. Garcia (1994). *Phys. World* 7, 49–55.]

spring. The phenomenon involves subtle and hitherto unsuspected interactions among the circulation, thermal budget, cloud microphysics, and photochemistry of the lower stratosphere. On the other hand, the rapid development and validation of a theory for polar ozone loss attests to the basic soundness of current knowledge of the stratosphere.

Nonetheless, several areas of stratospheric dynamics pose problems that are incompletely understood at present. The nature of the waves that drive the quasi-biennial oscillation is one of them. Although the basic mechanism of the QBO (wave-mean flow interaction in a region of weak Coriolis force; see Section III.B) is undoubtedly correct, the oscillation has not been simulated satisfactorily with current computer models. The results of partially successful simulations suggest that the QBO is probably driven by gravity waves of scales small enough that they are not represented properly in existing models.

Another area of current interest is that of global change. Anthropogenic emissions of trace gases such as carbon dioxide, methane, nitrous oxide, and fluorocarbons are believed to be partly responsible for the observed warming of the troposphere in the past century. In the stratosphere,

where IR radiation mostly escapes to space, increases in these gases should lead to temperature decreases. Lower temperatures would be expected to produce stronger, more stable polar jets, greater isolation of polar regions during winter and spring, and thus larger spring ozone depletion. This is still a concern because the burden of chlorine released in the past 50 years will not be cleansed from the stratosphere until well into the 21st century, even if current protocols for the elimination of chlorofluorocarbon emissions are universally obeyed. The possibility of large ozone “holes” in the heavily populated Northern Hemisphere thus remains a threat. These and perhaps other presently unknown phenomena are likely to keep fluid dynamicists and observationalists busy for years to come.

SEE ALSO THE FOLLOWING ARTICLES

ATMOSPHERIC DIFFUSION MODELING • ATMOSPHERIC TURBULENCE • CLIMATOLOGY • CLOUD PHYSICS • COASTAL METEOROLOGY • FLUID DYNAMICS • METEOROLOGY, DYNAMIC (TROPOSPHERE) • OZONE MEASUREMENTS AND TRENDS (TROPOSPHERE) • PLANETARY

WAVES (ATMOSPHERIC DYNAMICS) • RADIATION, ATMOSPHERIC • TROPOSPHERIC CHEMISTRY • WEATHER PREDICTION, NUMERICAL

BIBLIOGRAPHY

Andrews, D. G., Holton, J. R., and Leovy, C. B. (1987). "Middle Atmosphere Dynamics," Academic Press, San Diego.

Austin, J., Butchart, N., and Shine, K. P. (1992). Possibility of an Arctic ozone hole from a double-CO₂ climate, *Nature* **360**, 221–225.

Brasseur, G. P., Orlando, J. J., and Tyndall, G. S. (eds.) (1999). "At-

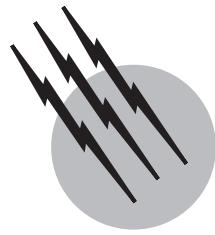
mospheric Chemistry and Global Change," Oxford University Press, Oxford.

Holton, J. R. (1975). "The Dynamical Meteorology of the Stratosphere and Mesosphere," *Meteor. Monogr.* **15**, American Meteorological Society.

Houghton, J. T., and Thomas, L. (eds.) (1980). "The Middle Atmosphere Observed from Balloons, Rockets and Satellites," The Royal Society of London.

McIntyre, M. E., and Palmer, T. N. (1983). Breaking planetary waves in the stratosphere, *Nature* **305**, 593–600.

Visconti, G., and Garcia, R. R. (eds.) (1987). "Transport Processes in the Middle Atmosphere," NATO ASI Series, D. Reidel, The Netherlands.



Meteorology, Dynamic (Troposphere)

Duane E. Stevens

Francis X. Crum

University of Hawaii at Manoa

- I. Physical and Mathematical Foundations
- II. Fundamental Simplifications and Approximations
- III. Theory of Atmospheric Motion

GLOSSARY

Available potential energy Difference between the actual potential energy of a system and the potential energy obtained by an adiabatic adjustment to a ground state with no horizontal variation of potential temperature. It is that part of the potential energy that is available for conversion to kinetic energy.

Baroclinic State of a fluid in which the surfaces of constant density and constant pressure do not coincide.

Barotropic State of a fluid in which surfaces of constant density and constant pressure coincide.

Beta-plane Geometric approximation in which the only effect of the earth's sphericity in an otherwise cartesian model is a linear variation of the Coriolis parameter with latitude.

Coriolis force Apparent force acting on a moving object due to the rotation of the coordinate system in which the object's velocity is measured. It is directed perpendicular to the velocity (toward the right in the northern hemisphere and toward the left in the southern) with

magnitude $2\Omega V \sin \phi$, where Ω is the rotation rate of the earth, V is the speed, and ϕ is the latitude.

Equivalent depth Separation constant arising from the separation of variables into their horizontal and vertical structure in the linearized governing equations. It is so named because it fulfills a role in the horizontal structure equations analogous to the mean fluid depth in the shallow water equations.

Geostrophic wind Wind resulting from a balance between the horizontal components of the pressure gradient and Coriolis forces.

Gradient wind Wind resulting from a balance between the horizontal components of the pressure gradient, Coriolis, and centrifugal forces.

Hydrostatic approximation Approximation to the vertical momentum equation that balances the vertical component of the pressure gradient force, which is directed upward, with the downward gravitational force.

Potential temperature Temperature a parcel of air would have if it were compressed or expanded adiabatically to a reference pressure (usually taken to be 1000 mbar).

Potential vorticity Scalar field that combines temperature and motion into an often-conserved quantity. As the scalar product of the absolute vorticity vector with the gradient of potential temperature, it is unique as a conservative quantity involving both dynamic and thermodynamic quantities.

Quasi-geostrophic theory Theory for large-scale atmospheric motion in which the quasi-geostrophic approximation is selectively used to simplify the governing equations with the result that the flow is completely determined from the time evolution of the geopotential or pressure fields.

Q-vector A vector which describes the Lagrangian rate at which the quasi-geostrophic flow tends to change horizontal temperature gradient. Its analysis aids in the inference of large-scale vertical motion.

Rossby radius of deformation Fundamental parameter for rotating fluids subject to gravitational restoring forces. It is defined as the gravity-wave phase speed divided by the Coriolis parameter. When a disturbance displaces the atmosphere away from an equilibrium state, the ratio of the Rossby radius to the horizontal length scale of the disturbance determines the character of the adjustment toward equilibrium.

Rossby wave Low-frequency, westward propagating wave whose restoring force arises from the variation of the Coriolis parameter with latitude.

Scale analysis Method for estimating the magnitudes of terms in the governing equations and simplifying the equations for a given phenomenon by neglecting terms that are small.

Stratosphere Portion of the atmosphere extending upward from the tropopause to approximately 50 km. It is characterized by a nearly constant temperature with height in its lower part and a temperature increase with height in its upper part. Vertical motion and water vapor content are small.

Synoptic meteorology Branch of meteorology that deals with phenomena having a length scale and a timescale of approximately 1000 km and 1–3 days, respectively. Its original usage referred to the analysis of weather observations taken over a wide area at approximately the same time.

Thermal wind Vector difference between the geostrophic wind at two pressure levels. Its magnitude is proportional to the horizontal gradient of mean temperature in the layer, and its direction is parallel to lines of constant mean temperature (isotherms) with the colder air on the left (in the northern hemisphere).

Troposphere Lowest 10–20 km of the atmosphere, characterized by decreasing temperature with height, large vertical motion, and large water-vapor content. Most of what is popularly thought of as weather occurs in this

region. The top boundary of the troposphere is called the tropopause.

Vorticity Local measure of the rotation of a fluid defined by the curl of the velocity.

DYNAMIC METEOROLOGY is the study of the motion of the atmosphere and the physical laws that govern that motion. Dynamic meteorology produces theories for atmospheric motion by applying basic principles from thermodynamics and classical mechanics. These principles are expressed mathematically as a set of partial differential equations. The goals of dynamic meteorology are twofold: to understand the various types of atmospheric motion and to provide a basis for quantitative prediction of atmospheric phenomena.

This article first reviews the physical and mathematical foundations of dynamic meteorology. Then the fundamental simplifications and approximations that contribute to our understanding of atmospheric motions are discussed. Finally, these approximations are applied to gain insight into several atmospheric phenomena. Throughout the article, the emphasis will be on large-scale, tropospheric dynamics.

I. PHYSICAL AND MATHEMATICAL FOUNDATIONS

A. Preliminary Mathematics

Understanding atmospheric motion requires the application of basic principles from thermodynamics and mechanics. These principles can be applied in an intuitive manner, using everyday language, perhaps combined with graphical aid or empirical insight, to make deductions. However, reasoning with everyday language is prone to errors in logic and ambiguous statements, especially for complicated problems. The language of mathematics helps to avoid these errors by providing a formal way to express concepts and a formal system of logic to manipulate these concepts. This section introduces some basic mathematical notation and concepts useful in dynamic meteorology.

1. Atmospheric Variables

The independent variables in dynamic meteorology represent time and position in three-dimensional space. The mathematical representation of the independent variables depends on the coordinate system chosen as a reference frame, but most commonly, x represents distance in the east–west direction, y distance in the north–south

direction, and z distance in the vertical direction. Time is denoted by t . The dependent variables that characterize the atmosphere are relatively few in number. For any variable f , the functional dependence on the independent variables may be written

$$f = f(x, y, z, t). \quad (1)$$

The motion of the air implies that the position (x, y, z) , of an individual air parcel changes with time. The velocity components in the eastward, northward, and upward directions are then defined as $u = dx/dt$, $v = dy/dt$, and $w = dz/dt$, called the zonal, meridional, and vertical wind components, respectively. In vector form, the velocity is given as $\mathbf{V} = u\hat{\mathbf{i}} + v\hat{\mathbf{j}} + w\hat{\mathbf{k}}$, where $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ are unit vectors pointing east, north, and up, respectively. The pressure is commonly denoted by p , the total density by ρ , and the temperature by T . In the course of studying dynamic meteorology many new quantities are introduced and are denoted by new symbols; however, they may all be derived from the basic dependent variables \mathbf{V} , p , ρ , and T . When phase changes of water are important, then ρ_e (partial density) is an additional variable.

2. Partial and Total Derivatives

Because the basic variables are functions of spatial position and time, their rate of change with respect to position or time is given by the partial derivatives $\partial f/\partial x$, $\partial f/\partial y$, $\partial f/\partial z$, and $\partial f/\partial t$. The partial derivative $\partial f/\partial t$ measures how the variable f changes with time at a fixed location (i.e., when x , y , and z are held constant). However, as a parcel of air moves about in the atmosphere, the velocity, pressure, temperature, etc., of that parcel can change with time as the parcel's position changes. This suggests the concept of a derivative with respect to time following the parcel. The relationship between this derivative and partial derivatives may be obtained as follows. For any variable f expressed as in Eq. (1), the total differential is

$$df = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz. \quad (2)$$

Dividing by dt and using the definitions of u , v , and w above yields

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + u\frac{\partial f}{\partial x} + v\frac{\partial f}{\partial y} + w\frac{\partial f}{\partial z} \quad (3)$$

or, in vector form,

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \mathbf{V} \cdot \nabla f, \quad (4)$$

where ∇ is the three-dimensional gradient operator defined by

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}. \quad (5)$$

The derivative df/dt is referred to as the total derivative (sometimes the material or substantial derivative). Because u , v , and w are associated with parcel velocities, this derivative measures the total variation of f following a parcel. The relation between partial and total derivatives is given by Eq. (4). The rate of change of f with time at a given point consists of the time rate of change in the parcel moving over that point and the advective rate of change $-\mathbf{V} \cdot \nabla f$ caused by the advection of different values of f over that point. For example, the local temperature change $(\partial T/\partial t)$ is caused by (1) advection of warmer or colder air over that point $(-\mathbf{V} \cdot \nabla T)$ and (2) heating or cooling of air parcels moving over that point (dT/dt) .

B. Fundamental Forces

The laws of mechanics were first formulated by Isaac Newton in the middle of the seventeenth century. His second law states that the total rate of change of momentum is equal to the net applied force and is in the same direction as that force. Forces are thus inherent to changes in motion. The net force appearing in Newton's second law consists of the vector sum of component forces due to all sources. This section considers each of those component forces important for atmospheric motion.

1. Pressure Gradient

The pressure is the normal force per unit area due to molecular collisions. If there is a difference in pressure across a volume of air in some direction, a macroscopic force, called the pressure gradient force, will act on the volume accelerating the air toward the lower pressure. The pressure gradient force per unit mass is written

$$\mathbf{f} = -\frac{1}{\rho}\nabla p. \quad (6)$$

(Throughout this section \mathbf{f} will be used to designate a force per unit mass.) Thus, the pressure gradient force is proportional to the spatial gradient of pressure and not to the pressure itself. On a horizontal surface, this force is directed normal to lines of constant pressure and points toward lower pressure.

2. Gravitational Force

Newton's law of gravitation states that the force exerted by one mass on another is directly proportional to the

product of the masses and inversely proportional to the square of the distance between them. If M is the mass of the earth, G the universal gravitational constant, r the distance between the earth's center of mass and a local air parcel (consisting of a specified set of molecules) in the atmosphere, and $\hat{\mathbf{r}}$ a unit vector pointing from the earth's center toward that air parcel, then the gravitational force per unit mass exerted by the earth is

$$\mathbf{f} = -\frac{GM}{r^2}\hat{\mathbf{r}}. \quad (7)$$

Because $r = a + z$, where a is the radius of the earth and z is the height above mean sea level, this can also be written

$$\mathbf{f} = \frac{\mathbf{g}_0}{(1 + z/a)^2}, \quad (8)$$

where $\mathbf{g}_0 = -GM\hat{\mathbf{r}}/a^2$. For 99.999% of the atmosphere, z is much less than a , so that the gravitational force can be treated as a constant given by \mathbf{g}_0 .

3. Frictional Force

Experience shows that objects in motion tend to slow down and stop. This is attributed to the force of friction. Ultimately, this force results from the collisions between molecules; however, friction is usually manifested in the atmosphere by the turbulence resulting from large wind shears on very small spatial scales. There are common mathematical expressions for the frictional force in terms of the gradient of the wind and coefficients of viscosity, but often it suffices to refer to friction by a dissipative time scale (Rayleigh friction) or to neglect it entirely.

4. Apparent Forces: Centrifugal and Coriolis

Because the earth rotates about its axis, it is natural and convenient to formulate the laws of dynamics in a coordinate system rotating with the earth. Writing the laws this way requires the inclusion of two "fictitious" forces not present in an inertial coordinate system fixed in space. To see this, consider two coordinate systems with the same origin, one fixed in space and the other rotating relative to it with the angular velocity of the earth. Now consider an object at rest in the rotating coordinate system. To an observer in this system the object is unaccelerated, but to an observer in the fixed system the object moves in a circle and is thus accelerated. This acceleration is called the centripetal acceleration. Now consider an object moving with constant velocity in the rotating system. Again, to an observer in this system the object is unaccelerated, but to an observer in the fixed system, the path of the object appears curved, thereby implying an acceleration. This acceleration caused by motion in the rotating system

is identified with the nineteenth century French scientist G. G. de Coriolis.

The preceding argument shows that observers in different coordinate systems perceive different accelerations. Mathematically, this is expressed in the following way. First of all, the relationship between the derivative of the position vector \mathbf{r} in an inertial reference frame denoted by subscript I and its derivative in a reference frame rotating with angular velocity $\boldsymbol{\Omega}$ relative to the inertial frame is

$$\left(\frac{d\mathbf{r}}{dt}\right)_I = \frac{d\mathbf{r}}{dt} + \boldsymbol{\Omega} \times \mathbf{r}. \quad (9)$$

This states that $\mathbf{V}_I = \mathbf{V} + \boldsymbol{\Omega} \times \mathbf{r}$, where \mathbf{V}_I is the velocity relative to the inertial frame and \mathbf{V} is the velocity relative to the rotating frame. Now, Eq. (9) holds for any vector, so that applying it to the vector \mathbf{V}_I yields

$$\left(\frac{d\mathbf{V}_I}{dt}\right)_I = \frac{d\mathbf{V}}{dt} + 2\boldsymbol{\Omega} \times \mathbf{V} + \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}), \quad (10)$$

assuming that $\boldsymbol{\Omega}$ is constant. Here is the relationship between accelerations in the two coordinate systems. They are not equal but differ by the sum of the Coriolis acceleration $2\boldsymbol{\Omega} \times \mathbf{V}$ and the centripetal acceleration $\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r})$.

The concept of apparent forces now follows. Newton's second law is valid in an inertial reference frame and therefore says

$$\left(\frac{d\mathbf{V}_I}{dt}\right)_I = \sum_i \mathbf{f}_i, \quad (11)$$

where \mathbf{f}_i is an individual force per unit mass. The truth of this law is not altered in a rotating frame of reference. However, if the law is expressed entirely in terms of quantities observable in the rotating frame, its mathematical form changes. Substituting Eq. (11) into Eq. (10) yields

$$\frac{d\mathbf{V}}{dt} = \mathbf{f} - 2\boldsymbol{\Omega} \times \mathbf{V} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}). \quad (12)$$

Thus the acceleration in the rotating frame equals the sum of the net force per unit mass that would be present in an inertial system and the two apparent forces due to the rotation of the coordinate system. When Newton's law is expressed in a rotating coordinate system, the Coriolis and centripetal accelerations are seen as additional forces per unit mass.

The second term on the right-hand side of Eq. (12) is called the Coriolis force. Its vector form shows that it is directed perpendicular to the velocity (toward the right in the northern hemisphere and toward the left in the southern) and hence can change only the direction and not the speed of an air parcel. The third term on the right-hand side of Eq. (12) is called the centrifugal force. It is perpendicular to the earth's axis and points outward.

C. Atmospheric Thermodynamics

Thermodynamics is the branch of science concerned with the interrelationships among the macroscopic properties of a system: specifically, how heat, work, and energy effect changes in temperature, pressure, and density (the so-called thermodynamic properties). While atmospheric dynamics studies the motions in the fluid that constitutes the atmosphere, atmospheric thermodynamics studies the relationships between the thermodynamic properties of the fluid itself.

Ultimately, it is the sun's radiative heating of the atmosphere, ocean, and earth that causes the energy changes necessary for the motions of the atmosphere. This heating produces the spatial and temporal distributions of pressure, temperature, and density, which in turn produce the forces which create motion. Thus, through relating heat, work, and energy to pressure, temperature, and density, atmospheric thermodynamics is the link between the driving mechanism of the sun's radiation and the motions of the atmosphere.

In the study of atmospheric thermodynamics, it is frequently necessary to take into account the water substance of the earth's atmosphere, particularly the changes of phase between the solid (ice), liquid (water), and gaseous (vapor) forms.

1. Ideal Gas Law

An equation of state is a relationship between the pressure, volume, and temperature of a gas. A gas obeying both Boyle's law (at constant temperature, the pressure and volume of a gas are inversely proportional) and Charles's law (at constant pressure, the volume of a gas is directly proportional to its absolute temperature) is called an ideal gas. The mixture of gases that constitute the earth's atmosphere obeys these laws only approximately, yet the agreement is sufficient that the atmosphere is considered to be an ideal gas.

The equation of state for an ideal gas takes a particularly simple form. If the specific volume (the inverse of density) is denoted by α , then

$$p\alpha = RT, \quad (13)$$

where R is called the gas constant, the value of which depends on the gas under consideration.

2. First Law of Thermodynamics

The first law of thermodynamics states that the change in internal energy of a system is equal to the difference between the heat added to the system and the work done by the system. This law is usually derived by considering the

principle of conservation of energy applied to a system at rest. Although the atmosphere is not at rest and therefore contains energy in other forms than internal energy, it may be shown that the first law of thermodynamics as stated earlier is still valid. Mathematically, an incremental change in internal energy per unit mass for an ideal gas is given by $c_v dT$, where c_v is the specific heat at constant volume. The incremental work done by the system may be written $p d\alpha$. The heat added may be denoted by $\dot{q} dt$ where \dot{q} is the heating rate per unit mass. Dividing by dt , the first law may be written

$$c_v \frac{dT}{dt} + p \frac{d\alpha}{dt} = \dot{q}. \quad (14)$$

An alternate and particularly convenient form of the first law,

$$c_p \frac{dT}{dt} - \alpha \frac{dp}{dt} = \dot{q} \quad (15)$$

is obtained by differentiating the ideal gas law of Eq. (13) and using the relationship for an ideal gas, $c_p = c_v + R$, where c_p is the specific heat at constant pressure. Note that $c_p dT$ is the incremental change in enthalpy.

3. Potential Temperature

An adiabatic process is a thermodynamic change whereby no heat is exchanged between a system and its surroundings ($\dot{q} = 0$). For an ideal gas undergoing an adiabatic process, the first law of thermodynamics may be written, from Eq. (15),

$$d \ln T - \frac{R}{c_p} d \ln p = 0. \quad (16)$$

Denoting R/c_p by κ and integrating this from a given pressure and temperature to a reference pressure and temperature p_{00} and θ , respectively, yields Poisson's equation

$$\theta = T \left(\frac{p_{00}}{p} \right)^\kappa. \quad (17)$$

The quantity θ is called the potential temperature. It is by definition the temperature a parcel would have if it were compressed or expanded adiabatically to a reference pressure p_{00} (usually taken to be 1000 mbar). Because it may be shown from Eqs. (16) and (17) that

$$\frac{d\theta}{dt} = 0 \quad (18)$$

the potential temperature is a conservative quantity for adiabatic motion. This characteristic makes it a useful quantity for many problems in dynamic meteorology.

4. Hydrostatic Pressure, Lapse Rates, and Static Stability

Thus far, pressure has been considered to be the normal force per unit area arising from molecular collisions. However, pressure can be thought of in other ways. For example, the product of the mass per unit area above a certain height with the acceleration caused by gravity defines a force per unit area that is called the hydrostatic pressure. Mathematically, this pressure is governed by the hydrostatic equation

$$\frac{\partial p_h}{\partial z} = -\rho g, \quad (19)$$

where p_h is the hydrostatic pressure and g is the acceleration caused by gravity. The hydrostatic pressure p_h equals the usual pressure p only in the absence of vertical accelerations. However, the difference between these two pressures is often small and Eq. (19) is often written with p instead of p_h . When this is done the result is known as the hydrostatic approximation—an accurate and powerful approximation useful in obtaining significant insight into the behavior of the atmospheric fluid. (Further discussion about this approximation and its validity is found in Section II.B)

The temperature throughout most of the troposphere is observed to decrease with height. This temperature gradient, called the lapse rate, is denoted by Γ and is defined by

$$\Gamma = -\frac{\partial T}{\partial z}, \quad (20)$$

where the negative sign is chosen so that the lapse rate is positive under most conditions. (In a stable temperature *inversion*, temperature *increases* with height and $\Gamma < 0$.) The lapse rate for an atmosphere in which the potential temperature is constant with height is called the dry adiabatic lapse rate and is denoted by Γ_d . An expression for Γ_d may be derived by taking the logarithm of Poisson's equation [(Eq. (17)], differentiating with respect to height, and applying the hydrostatic approximation. This yields

$$\frac{1}{\theta} \frac{\partial \theta}{\partial z} = \frac{1}{T} \left(\frac{g}{c_p} - \Gamma \right). \quad (21)$$

Thus, when θ is constant with height ($\partial \theta / \partial z = 0$), the lapse rate takes its dry adiabatic value of

$$\Gamma_d = g/c_p, \quad (22)$$

which is approximately 10 K/km. If θ is a function of height, then the actual lapse rate differs from its dry adiabatic value. This difference determines the stability to vertical displacements of the dry atmosphere. (For a moist atmosphere, the heat exchanges during condensation and evaporation of water vapor must be taken into account, thereby modifying the stability.) The vertical gradient of

θ also determines the time scale of vertical parcel motion. Specifically, it may be shown that for a stable atmosphere a parcel with an initial vertical velocity will oscillate with a frequency $N = [(g/\theta) \partial \theta / \partial z]^{1/2}$. This buoyancy frequency is called the Brunt–Väisälä frequency.

D. Mass Conservation

A fundamental conservation principle is the conservation of mass, which essentially states that matter is nowhere created or destroyed. The so-called continuity equation expresses this physical constraint. There are two alternate ways to derive this equation. The first, an Eulerian derivation, considers a volume element fixed in space for which the net rate of mass inflow through the sides must equal the rate of accumulation of mass within the volume. The net rate of mass inflow into a unit volume can be shown to be $-\nabla \cdot (\rho \mathbf{V})$, while the rate of accumulation of mass per unit volume is $\partial \rho / \partial t$. Therefore, the continuity equation is

$$\partial \rho / \partial t + \nabla \cdot (\rho \mathbf{V}) = 0. \quad (23)$$

Alternatively, a Lagrangian derivation considers a fixed mass moving with the fluid and constantly changing its volume. This method yields an alternate form of the continuity equation,

$$\nabla \cdot \mathbf{V} = \frac{1}{\alpha} \frac{d\alpha}{dt} = -\frac{1}{\rho} \frac{d\rho}{dt}, \quad (24)$$

which states that the velocity divergence in a parcel equals its fractional rate of change of volume per unit mass. The two forms of the continuity equation are equivalent, as may be shown by expanding the total derivative in Eq. (24) and employing the vector identity $\nabla \cdot \rho \mathbf{V} = \rho \nabla \cdot \mathbf{V} + \mathbf{V} \cdot \nabla \rho$.

E. Governing Equations of Atmospheric Motion

1. Effective Gravity

An object at rest in a coordinate system rotating with the earth is subject to a centrifugal force directed outward from the earth's axis and a gravitational force directed toward the earth's center. An observer or an instrument on the earth cannot distinguish between these two forces. Therefore they are combined into a resultant force

$$\mathbf{g} = \mathbf{g}_0 - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}) \quad (25)$$

known as the effective gravity or simply gravity. The local vertical is defined to be in the direction of this force. Note that it points toward the center of the earth only at the poles and equator.

2. Summary of Governing Equations

The governing equations of atmospheric motion consist of (1) Newton's second law in a form appropriate for a rotating coordinate system, (2) the first law of thermodynamics, (3) the continuity equation, and (4) the ideal gas law. These equations are summarized here:

$$\frac{d\mathbf{V}}{dt} = -\frac{1}{\rho} \nabla p - 2\Omega \times \mathbf{V} + \mathbf{g} + \mathbf{f}_r \quad (26a)$$

$$c_p \frac{dT}{dt} - \frac{1}{\rho} \frac{dp}{dt} = \dot{q} \quad (26b)$$

$$\frac{1}{\rho} \frac{d\rho}{dt} + \nabla \cdot \mathbf{V} = 0 \quad (26c)$$

$$p = \rho RT \quad (26d)$$

Equation (26a) is obtained by substituting formulas for the individual forces into Eq. (12), with the frictional force denoted by \mathbf{f}_r , and applying the definition of gravity. Because the vector velocity \mathbf{V} has scalar components in each of the three spatial directions, Eq. (26) constitutes a set of six equations in the six unknowns u , v , w , p , ρ , and T . These are the basic equations used in dynamic meteorology.

3. Scalar Equations in Spherical Coordinates

The vector form of the governing equations [Eqs. (26)] is compact and useful for physical interpretation of some terms, but application of the equations, especially in cases where solutions need to be obtained numerically, requires depiction in terms of a particular coordinate system. The obvious and most natural system is spherical polar coordinates (λ, ϕ, r) , where λ is longitude, ϕ is latitude, and r is radial distance from the center of the earth. (More rigorously, the coordinate system used is one in which the surfaces ($r = \text{constant}$) are everywhere perpendicular to the gravity force. To an excellent approximation the equations take the same form as those in spherical polar coordinates, the only difference being that " $r = \text{constant surfaces}$ " are interpreted differently.) Incremental elements of distance are defined as $dx = r \cos \phi d\lambda$, $dy = r d\phi$, $dz = dr$, so that velocity components can be defined in the obvious manner: $u = dx/dt = r \cos \phi d\lambda/dt$, $v = dy/dt = r d\phi/dt$, and $w = dz/dt = dr/dt$. The expansion of the total derivative is then given by

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \frac{u}{r \cos \phi} \frac{\partial}{\partial \lambda} + \frac{v}{r} \frac{\partial}{\partial \phi} + w \frac{\partial}{\partial r}. \quad (27)$$

In these coordinates, the fact that the unit vectors $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ in the λ , ϕ , and r directions, respectively, are a function of position must be taken into account when differentiating a vector quantity. Upon expanding the component forces into spherical coordinates, Eq. (26a) becomes

$$\begin{aligned} \frac{du}{dt} - \frac{uv \tan \phi}{r} + \frac{uw}{r} \\ = \frac{-1}{\rho r \cos \phi} \frac{\partial p}{\partial \lambda} + 2\Omega v \sin \phi - 2\Omega w \cos \phi \end{aligned} \quad (28a)$$

$$\frac{dv}{dt} + \frac{u^2 \tan \phi}{r} + \frac{uw}{r} = \frac{-1}{\rho r} \frac{\partial p}{\partial \phi} - 2\Omega u \sin \phi \quad (28b)$$

$$\frac{dw}{dt} - \frac{u^2 + v^2}{r} = \frac{-1}{\rho r} \frac{\partial p}{\partial r} - g + 2\Omega u \cos \phi, \quad (28c)$$

where the relationship $\mathbf{g} = -g\hat{\mathbf{k}}$ (g is the acceleration of gravity) has been used. The terms proportional to $1/r$ on the left-hand side are called curvature terms, because they arise from the variation of the unit vectors with position on the curved surface of the earth. Because $r = a + z$, where a is the radius of the earth and z is much smaller than a , r is often replaced by a in these equations as well as in the definitions of the velocity variables, and hence also in the total derivative expansion. When this is done it may be shown that several terms need to be neglected to ensure that the system of equations has the proper conservation properties and is consistent with vector invariant form. The result is called the traditional approximation. The complete set of equations with this approximation is

$$\frac{du}{dt} - \frac{uv \tan \phi}{a} = \frac{-1}{\rho a \cos \phi} \frac{\partial p}{\partial \lambda} + fv \quad (29a)$$

$$\frac{dv}{dt} + \frac{u^2 \tan \phi}{a} = \frac{-1}{\rho a} \frac{\partial p}{\partial \phi} - fu \quad (29b)$$

$$\frac{dw}{dt} = -\frac{1}{\rho} \frac{\partial p}{\partial z} - g \quad (29c)$$

$$c_p \frac{dT}{dt} - \alpha \frac{dp}{dt} = \dot{q} \quad (29d)$$

$$\frac{1}{\rho} \frac{d\rho}{dt} + \frac{1}{a \cos \phi} \left(\frac{\partial u}{\partial \lambda} + \frac{\partial v \cos \phi}{\partial \phi} \right) + \frac{\partial w}{\partial z} = 0 \quad (29e)$$

$$p = \rho RT, \quad (29f)$$

where $f = 2\Omega \sin \phi$ is called the Coriolis parameter. The continuity equation takes this form when the divergence is written consistently with the traditional approximation. Equations (29a)–(29c) resulting from Newton's second law are frequently called the momentum equations because they describe how forces act to change momentum.

F. Alternate Vertical Coordinates

1. Generalized Vertical Coordinates

The geometric height z is not always the most convenient vertical coordinate. Quite frequently, the governing equations take a simpler form when expressed in an

alternate vertical coordinate. Furthermore, observations are usually taken at specified pressure levels, so that for ease in evaluating terms in the equations using observational data, or for comparing theoretical results with observations, it is advantageous to use a coordinate other than z .

Transforming to an alternate vertical coordinate requires a knowledge of how partial derivatives in the two coordinate systems are related. Consider a general vertical coordinate s related to z by a single-valued monotonic function. In this coordinate system z is now a dependent variable so that $z = z(\lambda, \phi, s, t)$. Thus for any scalar function A of the independent variables (λ, ϕ, s, t) it is permissible to write

$$A(\lambda, \phi, s, t) = A(\lambda, \phi, z(\lambda, \phi, s, t), t), \quad (30)$$

which leads by application of the chain rule to

$$\left(\frac{\partial A}{\partial v} \right)_s = \left(\frac{\partial A}{\partial v} \right)_z + \frac{\partial A}{\partial z} \left(\frac{\partial z}{\partial v} \right)_s$$

and

$$\frac{\partial A}{\partial s} = \frac{\partial A}{\partial z} \frac{\partial z}{\partial s}, \quad (31)$$

where v is λ , ϕ , or t and subscripts indicate which vertical coordinate is held constant during the differentiation. The total derivative in the s coordinate is

$$\frac{d}{dt} = \left(\frac{\partial}{\partial t} \right)_s + \frac{u}{a \cos \theta} \left(\frac{\partial}{\partial \lambda} \right)_s + \frac{v}{a} \left(\frac{\partial}{\partial \phi} \right)_s + \dot{s} \left(\frac{\partial}{\partial s} \right), \quad (32)$$

where $\dot{s} = ds/dt$ is the vertical velocity in the s system. These expressions may be used to transform the governing equations from the z coordinate into an alternative s coordinate.

2. Isobaric Coordinates

A natural choice for s is the pressure p . The resulting coordinate system is referred to as isobaric coordinates. Horizontal derivatives contained in the pressure gradient force are transformed using Eq. (31). Anticipating the validity of hydrostatic balance for large scales of motion, the vertical momentum equation is replaced by the hydrostatic approximation, which in turn can be transformed using Eq. (31). The continuity equation is transformed using the above relations or rederived from first principles in isobaric coordinates (with the hydrostatic approximation, and g assumed constant). Total derivatives need not be transformed, only interpreted in light of Eq. (32). The vertical velocity in isobaric coordinates dp/dt is denoted by ω and now appears directly in the thermodynamic equation. The complete set of governing equations in isobaric coordinates, analogous to Eq. (29), is then

$$\frac{du}{dt} - \frac{uv \tan \phi}{a} = \frac{-1}{a \cos \phi} \frac{\partial \Phi}{\partial \lambda} + fv \quad (33a)$$

$$\frac{dv}{dt} + \frac{u^2 \tan \phi}{a} = -\frac{1}{a} \frac{\partial \Phi}{\partial \phi} - fu \quad (33b)$$

$$\frac{\partial \Phi}{\partial p} = -\frac{RT}{p} \quad (33c)$$

$$c_p \frac{dT}{dt} - \alpha \omega = \dot{q} \quad (33d)$$

$$\frac{1}{a \cos \phi} \left(\frac{\partial u}{\partial \lambda} + \frac{\partial v \cos \phi}{\partial \phi} \right) + \frac{\partial \omega}{\partial p} = 0, \quad (33e)$$

where all horizontal derivatives are now taken with p held constant. The quantity $\Phi = \int_0^z g dz$ ($z = 0$ is the mean sea level) is called the geopotential and is defined as the work required to raise a unit mass to height z from mean sea level. Note that the geopotential has replaced the pressure as one of the dependent variables. There are two primary advantages to isobaric coordinates: (1) the pressure gradient force is simpler due to the absence of the density, and (2) the continuity equation does not include density and has no time derivatives.

3. Other Vertical Coordinates

Dynamic meteorology uses several other vertical coordinates for particular problems. Sigma coordinates, defined by the vertical coordinate $\sigma = p/p_s$, where p_s is the surface pressure, are useful when the height of the lower boundary changes, such as in models where the topography of the earth is taken into account. Isentropic coordinates, where the potential temperature θ is the vertical coordinate, have the advantage that for adiabatic motion, air parcels remain on a coordinate surface as they move about in the atmosphere. Furthermore, vertical motion in isentropic coordinates is directly proportional to the diabatic heating rate. Log-pressure coordinates combine many of the advantages of pressure coordinates with a vertical coordinate approximately equal to height. This coordinate is defined by $z^* = -H \ln p/p_{00}$, where p_{00} is a constant reference pressure, $H = RT_{00}/g$, and T_{00} is a constant reference temperature.

G. Vorticity, Circulation, and Potential Vorticity

1. Basic Aspects of Vorticity

Vorticity, a variable of fundamental importance in dynamic meteorology, is a measure of the rotation of a fluid and is defined as the curl of the velocity. Meteorologists consider both the absolute vorticity, which includes the rotation of the earth, and the relative vorticity, which is just the fluid rotation relative to the earth. Because the

velocity of a point on the rotating earth is given by $\Omega \times \mathbf{r}$, absolute and relative vorticities ζ_a and ζ , respectively, are defined by

$$\zeta_a = \nabla \times \mathbf{V} + 2\Omega \quad (34)$$

$$\zeta = \nabla \times \mathbf{V} \quad (35)$$

Much of dynamic meteorology uses only the vertical component of vorticity, commonly denoted by ζ , and the vertical component ζ_a of absolute vorticity. Thus

$$\zeta_a = \hat{\mathbf{k}} \cdot (\nabla \times \mathbf{V} + 2\Omega) = \hat{\mathbf{k}} \cdot (\nabla \times \mathbf{V}) + f \quad (36)$$

$$\zeta = \hat{\mathbf{k}} \cdot (\nabla \times \mathbf{V}) \quad (37)$$

so that $\zeta_a = \zeta + f$. The vertical component of vorticity is a useful quantity for the study of midlatitude synoptic storms. A primary reason for its usefulness is that any horizontal velocity can be decomposed into the sum of two vectors: one with a vertical component of vorticity but no horizontal divergence (the rotational or nondivergent wind), and the other with divergence but not vorticity (the irrotational wind). This is known as Helmholtz's theorem. For large-scale midlatitude weather patterns, the rotational wind is by far the larger, so that by studying the vertical component of vorticity, most of the flow is included. This fact has significant dynamical implications.

For synoptic-scale storms in the northern hemisphere, the air tends to flow cyclonically (counterclockwise) and the relative vorticity is positive. For large high-pressure systems, the air flows anticyclonically (clockwise) and the relative vorticity is negative. Because f is always positive, the absolute vorticity is almost always positive; only in rare situations of very strong anticyclonic flow does the absolute vorticity become negative.

Physical interpretation of vorticity is facilitated by application of Stokes's theorem. This theorem states that for an area A with normal vector $\hat{\mathbf{n}}$ enclosed by a curve or path P with contour element $d\mathbf{r}$,

$$\iint_A (\nabla \times \mathbf{V}) \cdot \hat{\mathbf{n}} dA = \oint_P \mathbf{V} \cdot d\mathbf{r} \quad (38)$$

For a circular horizontal area with radius r_0 it may then be shown that the area averaged vertical vorticity ζ_A is related to the average tangential velocity at the boundary v_T through the formula

$$\zeta_A = 2v_T/r_0. \quad (39)$$

Thus the area-average vorticity is twice the perimeter-average angular velocity. This argument shows that vorticity is related to the curvature of the flow; however, vorticity may also be present when there is wind shear: a gradient of wind speed perpendicular to the flow. For example, if the flow has only a u component, ζ becomes

$$\zeta = -\frac{1}{a \cos \phi} \frac{\partial u \cos \phi}{\partial \phi}. \quad (40)$$

Hence, vorticity may be present simply from a change in westerly wind speed in the north-south direction, such as frequently occurs in the middle and upper troposphere of midlatitudes. This source of vorticity is often important for the development and motion of storm systems.

2. Vorticity Equations

An equation for the time evolution of the vertical component of vorticity may be derived by taking the curl of Eq. (26a), applying several vector identities and taking the dot product of the result with $\hat{\mathbf{k}}$. Interpretation of the result is conveniently done in cartesian coordinates where this equation may be written

$$\begin{aligned} \frac{d}{dt}(\zeta + f) = & -(\zeta + f) \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) \\ & - \left(\frac{\partial w}{\partial x} \frac{\partial v}{\partial z} - \frac{\partial w}{\partial y} \frac{\partial u}{\partial z} \right) \\ & + \frac{1}{\rho^2} \left(\frac{\partial \rho}{\partial x} \frac{\partial p}{\partial y} - \frac{\partial \rho}{\partial y} \frac{\partial p}{\partial x} \right). \end{aligned} \quad (41)$$

This equation states that the rate of change of the vertical component of absolute vorticity equals the sum of the three terms on the right-hand side, called the divergence term, the tilting or twisting term, and the solenoidal term, respectively.

The divergence term is so called because $\partial u / \partial x + \partial v / \partial y$ is the horizontal divergence. Its role in generating vorticity can be seen by considering the analogy of a spinning ice skater. If the skater moves his arms toward his body, which represents convergence, his rate of spin increases. Conversely, if he moves his arms away from his body, which represents divergence, his rate of spin decreases. Thus convergence increases vorticity and divergence decreases it.

The twisting or tilting term represents vertical vorticity generated by tilting or twisting horizontal vorticity components into the vertical by a spatially varying vertical velocity field. For example, if there is shear vorticity in the x direction due to the y component of velocity increasing with height, a vertical velocity field that decreases in the x direction can tilt the horizontal vorticity into the vertical, thereby yielding a source of vertical vorticity.

[Figure 1](#) provides insight into the role of the solenoidal term in vorticity production. If the lines of constant pressure and constant density intersect as shown, the less dense air on the right experiences a greater acceleration due to the pressure gradient force than the heavier air on the left, giving rise to a velocity shear and positive vorticity. The

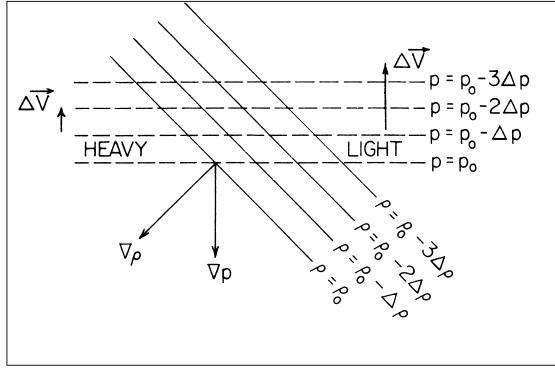


FIGURE 1 Vorticity generation by the solenoidal term.

solenoidal term may be written $(1/\rho^2)\hat{\mathbf{k}} \cdot (\nabla\rho \times \nabla p)$. If the surfaces of constant pressure and constant density do not coincide, the state of the fluid is called baroclinic and the baroclinic vector $\nabla\rho \times \nabla p \neq 0$. If the surfaces do coincide, then $\nabla\rho \times \nabla p = 0$ and the state of the fluid is called barotropic. For these reasons, a nonzero generation of vorticity by the solenoidal term is referred to as baroclinic production of vorticity.

In isobaric coordinates the vorticity equation takes a simpler form. Defining the vertical component of vorticity in pressure coordinates as $\zeta = \partial v / \partial x - \partial u / \partial y$ with derivatives now taken on constant pressure surfaces, the cartesian equation analogous to Eq. (41) is

$$\begin{aligned} \frac{d}{dt}(\zeta + f) &= -(\zeta + f) \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) \\ &\quad + \left(\frac{\partial \omega}{\partial y} \frac{\partial u}{\partial p} - \frac{\partial \omega}{\partial x} \frac{\partial v}{\partial p} \right), \end{aligned} \quad (42)$$

where, again, all horizontal derivatives hold pressure constant. Note that the total derivative is given by $d/dt = \partial/\partial t + u(\partial/\partial x) + v(\partial/\partial y) + \omega(\partial/\partial p)$, the twisting term involves derivatives with respect to pressure and horizontal gradients of ω , and there is no explicit baroclinic generation of isobaric vorticity by a solenoidal term.

3. Circulation and the Circulation Theorems

The relative circulation about a closed path P is defined by the right-hand side of Eq. (38).

$$C = \oint_P \mathbf{V} \cdot d\mathbf{r}. \quad (43)$$

Therefore, the circulation is the line integral about the contour P of the component of the velocity tangent to the contour and is another measure of the rotation of the fluid. Stokes's theorem provides the relationship to vorticity: the circulation divided by the area equals the average normal component of vorticity in the area.

The rate of change of relative circulation when P is a curve moving with the fluid may be derived by taking the total derivative of Eq. (43). After substitution from the momentum equation [Eq. (26a)] and some manipulation, it follows that

$$\frac{dC}{dt} = -2\Omega \frac{dA_n}{dt} + \iint_A \frac{\nabla\rho \times \nabla p}{\rho^2} \cdot \hat{\mathbf{n}} dA + \oint_P \mathbf{f}_r \cdot d\mathbf{r}, \quad (44)$$

where A_n is the projection of the area enclosed by P onto the plane perpendicular to Ω (the equatorial plane). This is known as the Bjerknes circulation theorem. The term involving dA_n/dt arises from the circulation due to the earth's velocity. It shows that the relative circulation around an area will decrease either when the area increases in size or when the area moves northward so that its projection onto the equatorial plane is larger. The second term on the right-hand side is analogous to the production of vorticity by the solenoidal term in the vorticity equation. The third term on the right-hand side is the production or destruction of circulation by friction.

The absolute circulation C_a is the sum of the relative circulation C and the circulation of the earth, which is just $2\Omega A_n$; therefore, the rate of change of the absolute circulation is

$$\frac{dC_a}{dt} = \iint_A \frac{\nabla\rho \times \nabla p}{\rho^2} \cdot \hat{\mathbf{n}} dA + \oint_P \mathbf{f}_r \cdot d\mathbf{r}. \quad (45)$$

This leads to Kelvin's circulation theorem, which states that for a barotropic fluid with no frictional forces acting, the absolute circulation is conserved following the motion.

4. Ertel Potential Vorticity

The vorticity equation describes how vorticity is changed by various properties of the flow. Only in very special circumstances would the vorticity be conserved following the flow. Kelvin's circulation theorem describes how an integral measure of vorticity is conserved but is valid only for barotropic flow and furthermore requires a knowledge of the time evolution of material surfaces. There does exist a quantity, referred to as the Ertel potential vorticity, that is conserved under more general conditions than either the vorticity or the circulation. It may be shown by combining the curl of the momentum equation [Eq. (26a)] with the continuity equation [Eq. (26c)] and the thermodynamic equation [Eq. (26b)] expressed in terms of potential temperature θ that

$$\frac{d}{dt} \left(\frac{\zeta_a}{\rho} \cdot \nabla \theta \right) = \frac{\zeta_a}{\rho} \cdot \nabla \frac{\theta q}{c_p T} + \nabla \theta \cdot \frac{\nabla \times \mathbf{f}_r}{\rho}. \quad (46)$$

The quantity $(\zeta_a/\rho) \cdot \nabla\theta$ is called the Ertel potential vorticity, after meteorologist H. Ertel, who derived it in 1942. It is a scalar quantity that may be viewed as the projection of the vorticity on the gradient of potential temperature. It is conserved if friction and diabatic heating can be neglected. The Ertel potential vorticity is an extremely significant and useful quantity. The constraints on its conservation are more straightforward and less restrictive than those for either vorticity or circulation. Because it embodies all the governing equations and combines into one quantity both dynamic and thermodynamic properties, it is viewed as one of the most general conservation statements that can be made in dynamic meteorology. It is a “potential” vorticity in the following sense: when the distance between two surfaces of constant θ increases, $|\nabla\theta|$ decreases, and ζ_a/ρ must increase proportionally if the potential vorticity is conserved. If the variation of ρ is small, this will be manifested primarily as an increase in the projection of ζ_a on the local θ surface. Thus, the packing together of surfaces of constant θ may be considered a reservoir of vorticity with the “potential” for creating vorticity as the surfaces move apart.

The conservation of Ertel potential vorticity has been applied by the detailed analysis of “isentropic potential vorticity” (IPV). Generally, IPV is computed hydrostatically and displayed on surfaces of constant potential temperature (θ -surfaces). The horizontal motion and temperature fields can be inferred from IPV, and vertical motion is obtained through the assumption that every parcel remains on its θ surface.

These developments follow theoretical advances built on vorticity theorems which are based on the conservation of both potential vorticity and potential temperature (i.e., entropy) on timescales for which friction, small-scale mixing, and diabatic heating can be ignored. IPV is a valuable atmospheric tracer which therefore shows the origin and predicted motion of atmospheric parcels; indeed, it is the *only* such dynamical (as opposed to chemical) tracer. Since heating/cooling and friction act to alter the IPV, it is clear that careful representation of these physical processes (i.e., “parameterization”) in models and theories is crucial.

II. FUNDAMENTAL SIMPLIFICATIONS AND APPROXIMATIONS

A. Scale Analysis

The motion of a rotating gaseous fluid expressed by Eq. (29) is very difficult to describe and predict. The circulation theorems and vorticity concepts, especially potential vorticity, allow for some basic insights into dynamic processes, but further understanding requires simplifica-

tion of the equations so that they describe only the essential aspects of a particular type of motion. To be consistent dynamically and mathematically, the equations need to be simplified in a systematic, not a haphazard, way.

Scale analysis is a method for estimating the magnitudes of the terms in the governing equations and provides a systematic way to simplify the equations for a given phenomenon. The method is as follows: characteristic scales (or sizes) for horizontal length, vertical depth, time, and horizontal and vertical velocity are used to nondimensionalize both the dependent and independent variables and, from them, the governing equations. The resulting equations are similar to the dimensional equations but differ in that they are written so that nondimensional parameters (which are products of the characteristic scales) appear multiplying each term. An implicit assumption is that the size of the characteristic scales has been chosen based on a particular type of motion under consideration so that the nondimensional variables have a magnitude of unity. Thus the magnitude of each term is determined by the magnitude or order of the nondimensional parameter which multiplies it. Typically, terms at least a factor of 10 smaller than the largest terms are then ignored (within appropriate conservation constraints). It needs to be emphasized that the results of scale analysis do not follow automatically from the governing equations. Rather, they follow from bringing a preconceived qualitative view, expressed through the choice of scale sizes, of the nature of the motion under consideration. For example, some variables are nondimensionalized by noting from observations that particular balances in the equations are valid.

The most important nondimensional parameters that appear in the course of performing scale analysis for many types of motion are

$$\text{Ro} = \frac{U}{fL} \quad (\text{Rossby number})$$

$$\varepsilon = \frac{L}{a} \quad (\text{horizontal size parameter})$$

$$\delta = \frac{D}{L} \quad (\text{aspect ratio})$$

$$\text{Re} = \frac{UL}{\nu} \quad (\text{Reynolds number})$$

$$E = \frac{\nu}{fD^2} \quad (\text{Ekman number}),$$

where U , L , and D are characteristic scales for horizontal velocity, horizontal length, and vertical depth, respectively, a is the earth’s radius, and ν is a coefficient of viscosity. Nondimensional parameters are not necessarily independent of each other; e.g., $E = \text{Ro}/(\delta^2 \text{Re})$. Typical values of L and U for several atmospheric phenomena are listed in Table I. The size of certain of these

TABLE I Typical Values of Length and Velocity Scales for Several Atmospheric Phenomena

Phenomena	Length scale L (m)	Velocity scale U (m/s)
Turbulence	$1\text{--}10^3$	$10^{-2}\text{--}10$
Tornado	$10^2\text{--}10^3$	10^2
Thunderstorm	10^4	$10\text{--}10^2$
Squall line	10^5	$10\text{--}10^2$
Hurricane	$10^5\text{--}10^6$	$10\text{--}10^2$
Midlatitude synoptic storm	10^6	$10\text{--}10^2$
Planetary-scale wave	10^7	10^2

nondimensional parameters (relative to unity) and their magnitudes relative to each other are exploited to justify the many simplified sets of equations used in dynamic meteorology. The following sections consider several of these simplified models of atmospheric motion.

B. Simple Wind Models

1. Geostrophic Wind

Observations indicate that, for many phenomena, the horizontal pressure gradient force is nearly in balance with the Coriolis force. The wind resulting from this balance is called the geostrophic wind. Rigorously, it is justified if the Ekman number $E \ll 1$ (allowing friction to be neglected), the Rossby number $\text{Ro} \ll 1$ (allowing the acceleration terms to be neglected), and $\varepsilon \ll 1$ (allowing the curvature terms to be neglected). (The notation \ll means much less than.) A scale analysis of the horizontal momentum equations in cartesian coordinates allows the geostrophic wind components to be written

$$v_g \equiv \frac{1}{\rho_s f_0} \frac{\partial p}{\partial x} \quad (47a)$$

$$u_g \equiv -\frac{1}{\rho_s f_0} \frac{\partial p}{\partial y}, \quad (47b)$$

where the subscript g denotes geostrophic velocity components, f_0 is the value of the Coriolis parameter at latitude ϕ_0 , and ρ_s is a reference density depending only on z . The geostrophic wind is determined from knowledge of the pressure distribution and density at a given height. In midlatitudes it usually approximates the true wind to within 10–15%. In low latitudes the Rossby number is no longer small (because f is smaller) and the geostrophic approximation is not generally valid. Finally, the geostrophic wind is horizontally nondivergent, that is, $\nabla_H \cdot \mathbf{V}_g = 0$, where $\mathbf{V}_g = u_g \hat{\mathbf{i}} + v_g \hat{\mathbf{j}}$ and ∇_H is the horizontal part of the ∇ operator.

In isobaric coordinates the geostrophic wind takes the form, using Eq. (33a,b),

$$v_g \equiv \frac{1}{f_0} \frac{\partial \Phi}{\partial x} \quad (48a)$$

$$u_g \equiv -\frac{1}{f_0} \frac{\partial \Phi}{\partial y}, \quad (48b)$$

where derivatives are taken holding pressure constant. In this formulation, the geostrophic wind depends on only the geopotential on a pressure surface. The density does not explicitly appear.

A vector form of geostrophic balance derived from Eq. (47a,b) is

$$f_0 \hat{\mathbf{k}} \times \mathbf{V}_g = \frac{1}{\rho_s} \nabla_H p. \quad (49)$$

On a horizontal map that displays the pressure fields as contour lines at a given height, Eqs. (47) and (49) indicate that the geostrophic wind is parallel to the pressure contours. Alternatively, the geostrophic wind is parallel to geopotential contours at a given pressure level, according to Eq. (48).

2. The Hydrostatic Approximation and the Thermal Wind

For many large-scale phenomena (e.g., midlatitude synoptic storms), the vertical pressure gradient is observed to be in very good balance with gravity. This balance yields the hydrostatic approximation. It is rigorously justified when the aspect ratio $\delta = D/L \ll 1$ (i.e., when the horizontal scale of the phenomena is significantly larger than the vertical scale). From Eq. (29c) the hydrostatic approximation is

$$\partial p / \partial z = -\rho g. \quad (50)$$

With this approximation it can be seen that the pressure at any height is equal to the weight of the unit cross section column of air above that point. Furthermore, if ρ is written in terms of temperature and pressure using the ideal gas law, Eq. (50) can be integrated to yield

$$\Delta z = \frac{R \bar{T}}{g} \ln\left(\frac{p_1}{p_2}\right), \quad (51)$$

where $\Delta z = z_2 - z_1 > 0$ is the thickness of the layer between pressure levels p_2 and p_1 ($p_2 < p_1$), and \bar{T} is the mean temperature (with respect to $\ln p$) in the layer. This hypsometric equation shows that thickness between two pressure surfaces is directly proportional to the mean temperature in the layer. For an isothermal atmosphere ($\bar{T} = \text{constant}$), the hydrostatic approximation implies that pressure and density vary with height as $\exp(-z/H)$, where $H = R \bar{T} / g$ is called the scale height. Thus, in the isothermal case, pressure and density decrease by a factor of e (≈ 2.718) over a vertical distance of one scale height.

In a hydrostatic atmosphere, the geostrophic wind must change with height if the temperature varies horizontally. This follows from the fact that the thickness will vary horizontally, causing the slope of the pressure surfaces to change with height. Thus the pressure gradient changes with height and from Eq. (47) the geostrophic wind changes with height. The thermal wind is defined as the vector difference between the geostrophic wind at two pressure levels. It may be shown that the thermal wind is proportional to the horizontal gradient of thickness, or, equivalently, the horizontal gradient of mean temperature. This is most easily demonstrated in isobaric coordinates. Using Eq. (48a,b) and the definition of geopotential, the thermal wind may be written

$$\mathbf{V}_T = \frac{g}{f_0} \hat{\mathbf{k}} \times \nabla_H(z_2 - z_1) \quad (52)$$

or, using Eq. (51),

$$\mathbf{V}_T = \frac{R}{f_0} \ln\left(\frac{p_1}{p_2}\right) \hat{\mathbf{k}} \times \nabla_H \bar{T}. \quad (53)$$

Strictly speaking, the thermal wind is a measure of the vertical shear of the geostrophic wind and not a real wind composed of air in motion. Yet it is useful to think of it as a velocity. From Eq. (52) the thermal wind is directed parallel to lines of constant thickness with the cold air (low thickness) on the left and is stronger in regions of greater mean temperature gradient.

There are many applications of the thermal wind concept in understanding the structure of the atmosphere. One such application is the jet stream—the region of maximum westerly winds located just below the tropopause at a latitude that varies from 30–60°N. Because the air in the troposphere is warm in the tropics and colder toward the poles, the thermal wind equation implies that the westerly component of geostrophic wind will increase with height. The westerly flow reaches a maximum near the tropopause because the horizontal temperature gradient in the stratosphere is generally reversed, with colder air in the tropics. Furthermore, the strongest tropospheric temperature gradient is concentrated in a narrow zone which varies with latitude, longitude, and season. Thus, the thermal wind, and hence the geostrophic wind shear, is most pronounced in this region, which has been called the polar front.

3. Gradient Wind

The geostrophic wind considers a balance between the horizontal pressure gradient and Coriolis forces. A less restrictive balance is one that includes the centripetal acceleration terms in the horizontal momentum equations. The balance that follows is not obtainable by a rigorous

scale analysis as the geostrophic wind is, but it yields a useful conceptual wind model. If vertical advection and curvature terms are neglected, the horizontal components of Eq. (26a) are

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = -\frac{1}{\rho} \frac{\partial p}{\partial x} + fv \quad (54a)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} = -\frac{1}{\rho} \frac{\partial p}{\partial y} - fu. \quad (54b)$$

The gradient wind can be obtained if these equations are rewritten in polar coordinates (r, φ) , where $r = \sqrt{x^2 + y^2}$ and $\varphi = \tan^{-1}(y/x)$. Defining tangential and radial velocities u_r and v_φ , respectively, the polar coordinate form of Eq. (54) is

$$\frac{\partial v_\varphi}{\partial t} + u_r \frac{\partial v_\varphi}{\partial r} + \frac{v_\varphi}{r} \frac{\partial v_\varphi}{\partial \varphi} + \frac{u_r v_\varphi}{r} + fu_r = -\frac{1}{\rho r} \frac{\partial p}{\partial \varphi} \quad (55a)$$

$$\frac{\partial u_r}{\partial t} + u_r \frac{\partial u_r}{\partial r} + \frac{v_\varphi}{r} \frac{\partial u_r}{\partial \varphi} - \frac{v_\varphi^2}{r} - fv_\varphi = -\frac{1}{\rho r} \frac{\partial p}{\partial r}, \quad (55b)$$

where $v_\varphi > 0$ corresponds to counterclockwise motion. Now if local and advective changes in the radial flow can be neglected in the radial wind equation [Eq. (55b)], letting $v_\varphi = V_{gr}$ we find

$$\frac{V_{gr}^2}{r} + f V_{gr} = \frac{1}{\rho} \frac{\partial p}{\partial r}. \quad (56)$$

The wind V_{gr} satisfying this equation is called the gradient wind. Although the flow is steady, it is curved and hence there is a centripetal acceleration. This acceleration is measured by the term V_{gr}^2/r and may be considered to define a centrifugal force per unit mass. Therefore, the gradient wind may be seen as a balance between centrifugal, Coriolis, and pressure gradient forces.

Geostrophic balance does not include the centrifugal force arising from curvature in the flow as gradient balance does. The relationship between the gradient and geostrophic winds is derived by noting that $(1/\rho f)(\partial p/\partial r)$ defines the geostrophic wind, denoted by subscript g, so that

$$\frac{V_g}{V_{gr}} = 1 + \frac{V_{gr}}{fr}. \quad (57)$$

Thus, the geostrophic wind overestimates the gradient wind for cyclones (where there is central low pressure and $V_{gr} > 0$) and underestimates it for anticyclones (where there is central high pressure and $V_{gr} < 0$).

As with the thermal wind, the gradient wind has many applications. One application is an explanation of why the pressure gradient is generally weaker near the center of an anticyclone than near the center of a cyclone. This can

be demonstrated by solving the quadratic Eq. (56) for V_{gr} , yielding

$$V_{\text{gr}} = -\frac{fr}{2} \pm \sqrt{\frac{f^2 r^2}{4} + \frac{r}{\rho} \frac{\partial p}{\partial r}}. \quad (58)$$

For physically meaningful solutions the quantity under the square root must be positive. Now for anticyclones $\partial p/\partial r < 0$ so that for a realizable gradient wind

$$\left| \frac{\partial p}{\partial r} \right| < \frac{\rho f^2 r}{4}. \quad (59)$$

For cyclones, $\partial p/\partial r > 0$, and the quantity under the square root is always positive, so that gradient wind balance places no restriction on the pressure gradient at the center of a cyclone.

C. Geometric Simplifications and Quasi-Geostrophic Theory

In the previous section, the geostrophic wind model was discussed. This wind is a balance between the Coriolis and pressure gradient forces and is an example of a diagnostic relation; that is, at any given time one variable is diagnosed from knowledge of other variables but there is no information about evolution in time. The gradient wind results from a less restrictive balance among the Coriolis, pressure gradient, and centrifugal forces. However, temporal tendencies are neglected so that gradient balance is also a diagnostic relation. Quasi-geostrophic theory further exploits the situations in which the Rossby number is small to arrive at a system of equations that is considerably simplified like the earlier balances but that, in contrast to them, allows for the time evolution of the flow. To take into account the time evolution, slight departures from geostrophic balance must be considered, hence the name quasi-geostrophic. This theory has become the cornerstone of modern dynamic meteorology, providing basic understanding of many dynamical processes.

Prior to developing quasi-geostrophic theory itself, it is useful to make some geometric simplifications. A phenomenon with horizontal scale much less than the radius of the earth does not sense the curvature of the earth. Synoptic storms (with length scale $L \approx 1000$ km) fall into this category because $L/a \ll 1$. Although curvature terms in the governing equations can be neglected on this basis, it is important to retain the latitudinal variation of the Coriolis parameter, which has important dynamical implications. Rigorously, the equations are written in a suitable conformal mapping, and then it is demonstrated that the spatial variation of the scale factors (which relate incremental curvilinear distances to increments in the coordinates) may be neglected when $L/a \ll 1$. The result is a system

of equations formally identical to a cartesian system formulated on a plane tangent to the earth at a reference latitude ϕ_0 with the exception that the Coriolis parameter retains its variation with latitude. The Coriolis parameter may be expanded in a Taylor series about latitude ϕ_0 and truncated after two terms to yield $f = f_0 + \beta y$, where $f_0 = 2\Omega \sin \phi_0$, $\beta = (\partial f / \partial y)_{\phi_0} = 2\Omega \times \cos \phi_0 / a$, and y measures the northward distance from ϕ_0 . This set of approximations, where the only effect of the earth's sphericity in an otherwise cartesian model is a linear variation of f with y , is called the beta-plane approximation.

The assumptions of quasi-geostrophic theory may now be outlined. First, in light of previous comments on isobaric coordinates, it is convenient to discuss quasi-geostrophic theory in this coordinate system. Second, instead of using the individual momentum equations on the beta-plane, the vorticity equation [Eq. (42)] is used instead. This is done because the divergence appears explicitly in the vorticity equation and hence the nongeostrophic part of the wind, which determines the time evolution, also explicitly appears, separate from the rotational part. Now the Rossby number represents the ratio of the relative vorticity to the Coriolis parameter. For midlatitude synoptic scales, the Rossby number is small so that ζ is small compared to f . Therefore, ζ is neglected compared to f in the divergence term. The divergence is the sum of two components, each of which individually has magnitude U/L . However, the divergence as a whole is observed to be much smaller than either of its two components, so that it scales as $(Ro) U/L$. The continuity equation then yields a scale for the vertical advection that enables it to be neglected compared to the horizontal advection. The twisting term may be similarly neglected. The wind may be replaced by its geostrophic value everywhere except in the divergence term, where it is the departures from geostrophic balance that are important. The vorticity ζ is replaced by its geostrophic value ζ_g . Finally, because the ratio of βy to f_0 scales as $L/a \ll 1$, f may be replaced by f_0 except where it is differentiated, in which case its derivative is given by the constant β . These approximations, together with the continuity equation on the beta-plane, yield the quasi-geostrophic vorticity equation

$$\frac{\partial \zeta_g}{\partial t} = -u_g \frac{\partial \zeta_g}{\partial x} - v_g \frac{\partial \zeta_g}{\partial y} + f_0 \frac{\partial \omega}{\partial p} - \beta v_g, \quad (60)$$

where

$$u_g = -\frac{\partial}{\partial y} \left(\frac{\Phi}{f_0} \right) \quad v_g = \frac{\partial}{\partial x} \left(\frac{\Phi}{f_0} \right),$$

and

$$\zeta_g = \nabla^2 \left(\frac{\Phi}{f_0} \right) = \frac{\partial^2}{\partial x^2} \left(\frac{\Phi}{f_0} \right) + \frac{\partial^2}{\partial y^2} \left(\frac{\Phi}{f_0} \right).$$

Quasi-geostrophic theory is an approximation to the complete set of governing equations. The quasi-geostrophic vorticity equation is one equation in the two dependent variables Φ and ω . A combination of the hydrostatic approximation and the thermodynamic equation completes the system, giving two equations in the two unknowns Φ and ω . The thermodynamic equation is approximated by replacing the horizontal velocity by its geostrophic value and considering the stability σ , defined next, and reference temperature T_s and density ρ_s to be functions of pressure only. Using the hydrostatic approximation of Eq. (33c) and neglecting the diabatic heating (which may be easily reintroduced if desired), the quasi-geostrophic thermodynamic equation is

$$\frac{\partial}{\partial t} \left(-\frac{\partial \Phi}{\partial p} \right) = -u_g \frac{\partial}{\partial x} \left(-\frac{\partial \Phi}{\partial p} \right) - v_g \frac{\partial}{\partial y} \left(-\frac{\partial \Phi}{\partial p} \right) + \sigma \omega, \quad (61)$$

where

$$\sigma = \left(-\frac{R}{p} \right) \left(\frac{\partial T}{\partial p} - \frac{RT}{pc_p} \right) = \frac{N^2}{\rho^2 g^2}.$$

Equations (60) and (61) constitute the quasi-geostrophic system. The rigorous justifications from scale analysis require that $E \ll 1$ (so that friction may be neglected), $\delta \ll 1$ (so that the hydrostatic approximation is valid), and $\varepsilon \ll 1$ (so that the beta-plane approximation is valid). Furthermore, it is required that $\text{Ro} \ll 1$. If each variable is expanded in terms of this small parameter, the balance of all terms of magnitude 1 in the resulting equations yields the geostrophic approximation. To obtain equations that account for the time evolution, the balance of terms of size Ro must be considered. This balance yields the quasi-geostrophic system.

The primary advantage of quasi-geostrophic theory is that it is simple enough to allow dynamical explanation of many large-scale phenomena. With this theory, the original system of five equations in five unknowns is reduced to one equation in terms of the geopotential. This may be seen by eliminating ω from Eqs. (60) and (61). The result is

$$\left(\frac{\partial}{\partial t} - \frac{1}{f_0} \frac{\partial \Phi}{\partial y} \frac{\partial}{\partial x} + \frac{1}{f_0} \frac{\partial \Phi}{\partial x} \frac{\partial}{\partial y} \right) q = 0, \quad (62)$$

where

$$q = \frac{1}{f_0} \frac{\partial^2 \Phi}{\partial x^2} + \frac{1}{f_0} \frac{\partial^2 \Phi}{\partial y^2} + f + f_0 \frac{\partial}{\partial p} \left(\frac{1}{\sigma} \frac{\partial \Phi}{\partial p} \right). \quad (63)$$

Thus, the core of quasi-geostrophic theory is that the quantity q is conserved (neglecting diabatic effects) following the geostrophic motion. This quantity q is the quasi-geostrophic analog to the Ertel potential vorticity and is called the *quasi-geostrophic potential vorticity*.

Equation (62) is called the quasi-geostrophic potential vorticity equation.

D. Linearization

The atmosphere is a turbulent fluid. Nonlinear advects cause the motions on one scale to interact and affect other scales of motion. Through the nonlinear interactions, energy cascades from one scale to another. Because it is impossible to obtain an accurate description of the atmosphere's state at all scales, a perfect weather forecast is impossible; the initial conditions for a time integration of the governing equations are never precisely known.

For the purpose of understanding the characteristics of the atmosphere's behavior at certain scales, it is often useful to isolate a particular scale and neglect the nonlinear interactions. Linearization enables the study of specific types of fluid perturbations, their propagation and dispersion characteristics, and the scale at which fluid instabilities are likely to be observed.

In this approximation, a basic state must be specified. Usually the basic state is chosen to be independent of longitude and time. All field variables are then written as the sum of the basic state part (denoted by an overbar) and a fluctuating component called the perturbation (denoted by a prime); for example,

$$\begin{aligned} u(x, y, z, t) &= \bar{u}(y, z) + u'(x, y, z, t) \\ v(x, y, z, t) &= \bar{v}(y, z) + v'(x, y, z, t). \end{aligned} \quad (64)$$

The governing equations are then written in terms of the known basic state and the unknown perturbations; for example, one advection term in the v -momentum equation is written

$$\begin{aligned} v \frac{\partial u}{\partial y} &= (\bar{v} + v') \frac{\partial}{\partial y} (\bar{u} + u') \\ &= \bar{v} \frac{\partial \bar{u}}{\partial y} + \bar{v} \frac{\partial u'}{\partial y} + v' \frac{\partial \bar{u}}{\partial y} + v' \frac{\partial u'}{\partial y}. \end{aligned} \quad (65)$$

If the governing equations are averaged with respect to x and the resulting mean equations are subtracted from the original equations, a set of perturbation equations results. The advection term in the perturbation v momentum equation becomes

$$v \frac{\partial u}{\partial y} - \bar{v} \frac{\partial \bar{u}}{\partial y} = \bar{v} \frac{\partial u'}{\partial y} + v' \frac{\partial \bar{u}}{\partial y} + \left(v' \frac{\partial u'}{\partial y} - \bar{v}' \frac{\partial \bar{u}'}{\partial y} \right). \quad (66)$$

The linearization assumption is then applied by neglecting all products of perturbation quantities (e.g., $v' \partial u' / \partial y$) relative to linear terms (e.g., $v' \partial \bar{u} / \partial y$). If the amplitudes of fluctuations from the basic state are very small, this step is equivalent to a series expansion in a small parameter proportional to the perturbation amplitude followed by

neglecting second-order quantities relative to first-order quantities in the first-order governing equations.

Because linear differential equations are generally much easier to solve than nonlinear equations, the resulting set is more tractable and leads to significant insight into the waves and instabilities in the atmosphere. Sometimes analytic solutions can be obtained if the basic state is sufficiently simple. This powerful technique is the mathematical cornerstone for the analysis of atmospheric motions.

E. Multilayer Models and the Shallow-Water Equations

1. Multilayer Models

In representing the vertical structure of the continuously stratified atmosphere, it is generally necessary to discretize the vertical dependence so that the semiinfinite atmosphere may be approximated by a finite number of parameters. Often finite differences at a prespecified number of levels are used to represent the vertical structure. Another method is to divide the atmosphere into a finite number of homogeneous layers, as shown in Fig. 2. This representation is often used in dynamic oceanography, which has many similarities to dynamic meteorology, essentially because the Rossby number is also small for large-scale motions in the ocean. Both disciplines may be considered subsets of geophysical fluid dynamics.

In a layered model, the upper boundary of the uppermost (N th) layer is considered to be a free surface with $p = 0$, corresponding to negligible mass content above layer N . Each layer is taken to be homogeneous ($\rho = \text{constant}$) and incompressible ($d\rho/dt = 0$). Assuming that the horizontal scales of interest are much greater than the depth of the fluid, the pressure is hydrostatic to a good approximation. The individual layers are immiscible, retaining their identity without mixing. The fluid will be assumed

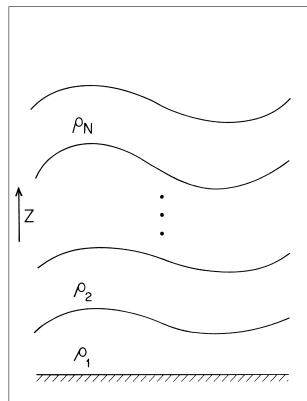


FIGURE 2 A multilayer fluid consisting of N homogeneous, immiscible layers.

to be stably stratified, with density decreasing upward: $0 < \rho_N < \dots < \rho_2 < \rho_1$. Each (n th) layer is capped by a free surface at $z = h_n(x, y, t)$.

2. Shallow-Water Equations

The special case of a single layer of fluid, which is governed by the so-called shallow-water equations, is now treated explicitly. The hydrostatic relation may be integrated from level z to the free surface $h_1 = h$, yielding

$$p(x, y, z, t) = p_s(x, y, t) + \rho g[h(x, y, t) - z]. \quad (67)$$

With negligible mass above $z = h$, the pressure at the free surface $p_s(x, y, t) = 0$. Because no fluid particles move through the free surface,

$$w(x, y, z = h, t) = dh/dt, \quad (68)$$

that is, a parcel on the surface remains on the surface forever. In this section we assume a flat lower boundary so that the vertical velocity at the lower boundary must vanish.

In the homogeneous fluid, the horizontal pressure gradient force is, from Eq. (67),

$$-\frac{1}{\rho} \nabla p = -g \nabla h, \quad (69)$$

where throughout this section ∇ indicates the horizontal gradient operator. This horizontal force is independent of a parcel's depth in the fluid. Writing the inviscid, horizontal momentum equation in cartesian geometry (with \mathbf{V} denoting the horizontal velocity),

$$\frac{\partial \mathbf{V}}{\partial t} = -u \frac{\partial \mathbf{V}}{\partial x} - v \frac{\partial \mathbf{V}}{\partial y} - w \frac{\partial \mathbf{V}}{\partial z} - f \hat{\mathbf{k}} \times \mathbf{V} - g \nabla h \quad (70)$$

it can be seen that if the horizontal flow is initially depth-independent ($\partial \mathbf{V} / \partial z = 0$), then all terms on the right are depth-independent. Thus $(\partial / \partial t)(\partial \mathbf{V} / \partial z) = 0$ and no vertical shear will develop. Hence it is assumed that the flow is independent of depth within the fluid for all time. Vertical advection of momentum may therefore be ignored, even though the vertical velocity is finite.

The continuity equation for an incompressible fluid may be written

$$\nabla \cdot \mathbf{V} + \partial w / \partial z = 0. \quad (71)$$

Because the horizontal flow is depth-independent, vertical integration from the lower boundary ($z = 0$) to the upper boundary ($z = h$) yields

$$w(h) - w(0) + h \nabla \cdot \mathbf{V} = 0. \quad (72)$$

With the kinematic boundary conditions on vertical velocity described above, this reduces to

$$dh/dt + h \nabla \cdot \mathbf{V} = 0, \quad (73)$$

or alternatively in flux form,

$$\partial h/\partial t = -\nabla \cdot h \mathbf{V}. \quad (74)$$

This equation states that the free surface raises/lowers only if there is mass convergence/divergence into the fluid column below.

The shallow water equations may be written compactly for future reference:

$$\frac{du}{dt} - fv = -g \frac{\partial h}{\partial x} \quad (75a)$$

$$\frac{dv}{dt} + fu = -g \frac{\partial h}{\partial y} \quad (75b)$$

$$\frac{dh}{dt} + h \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) = 0. \quad (75c)$$

Here only horizontal advection is included in the total time derivatives.

The relevance of these shallow-water equations for atmospheric motions is seen when the governing equations for the stratified atmosphere are linearized about an assumed basic state. If the basic-state wind field is assumed independent of height, the perturbation equations can generally be separated into an ordinary differential equation for the vertical structure of the variables, and a set of partial differential equations for the horizontal structure. The separation constant that appears in both sets is called the “equivalent depth” and is usually denoted by h_e . The equivalent depth is so-named because the horizontal structure equations appear very similar to the linearized equations for a shallow water fluid with mean fluid depth \bar{h} equal to h_e . The horizontal structure of atmospheric waves is essentially identical to the solutions of the shallow-water equations.

III. THEORY OF ATMOSPHERIC MOTION

A. Atmospheric Waves

Atmospheric motions are observed on nearly all scales of time and space. To understand these motions, it is necessary to determine which terms in the physical equations are most important in governing the behavior of a particular phenomenon on a particular time and space scale. If possible, the equations are scaled to isolate a single type or class of motions. Analysis of these reduced sets of equations provides insight into the fundamental dynamics of a given type of motion. Application of the linearization technique helps to isolate these scales.

Different wave types can be categorized according to the appropriate restoring mechanism. **Table II** summarizes

TABLE II Atmospheric Wave Types

Wave type	Restoring mechanism	Key parameter
A. Sound (acoustic)	Compressibility	c_s^2
B. Gravity	Buoyancy (stable stratification)	N^2
C. Inertial oscillation	Inertia (rotation)	f^2
D. Inertia-gravity		
Midlatitude	Inertia/buoyancy	$L_R = \sqrt{gh_e}/f_0$
Equatorial	Inertia/buoyancy	$L_R = (\sqrt{gh_e}/\beta_e)^{1/2}$
E. Rossby		
Midlatitude	Planetary vorticity gradient	$\beta = \beta_m = \frac{2\Omega}{a} \cos \phi$
Equatorial	Planetary vorticity gradient	$\beta = \beta_e = \frac{2\Omega}{a}$
F. Kelvin		
Midlatitude	Inertia/buoyancy	$L_R, v = 0$
Equatorial	Inertia/buoyancy	$L_R, v = 0, n = -1$
G. Mixed		
Rossby-gravity	Inertia/buoyancy and planetary vorticity gradient	$L_R, n = 0$

the restoring mechanisms that are discussed in this section. If a physical/dynamical constraint tends to return a displaced air parcel to its original position, the parcel will tend to oscillate about that position. In general, the parcel motion affects adjacent air parcels, so that the oscillation tends to propagate in space. As such a wave propagates, its associated energy may disperse and the energy density decrease. In general, for a given spatial scale the wave will propagate at a certain period τ and frequency $\sigma = 2\pi/\tau$ should be, which can be determined from the physical laws and which depend on the characteristics of the basic state.

Under certain dynamic and thermodynamic configurations of prescribed basic states, some scales of motion can draw energy from the basic state and increase in amplitude and energy. If the basic state is artificially held fixed, the perturbation will continue to grow. In such an instability, a parcel may or may not oscillate about a central location, but the energy of the disturbance will increase with time, being converted from the energy of the basic state. The growth rate and associated e -folding timescale (i.e., the time it takes for the disturbance to grow by a factor of e) are fundamental parameters of the instability. If energy is initially present at all scales, presumably the instability will grow to observable amplitude in the scales with the fastest growth rates. Of course, in nature the perturbation cannot grow indefinitely because the basic state must lose energy to the perturbation, thereby reducing and eventually eliminating the source of the instability.

In the following sections, the various waves and instabilities are isolated by considering simplified dynamical systems.

1. Sound Waves

Sound waves propagate acoustic energy in a compressible fluid by the alternating adiabatic compression and expansion of the fluid. The elastic nature of the fluid provides the restoring force for parcels to return to their undisturbed location; as air is compressed and its density changed ($\partial\rho/\partial t \neq 0$), the local pressure also changes ($\partial p/\partial t \neq 0$) because $\theta = (p/\rho R)(p_0/p)^\kappa$ is conserved for adiabatic motion of an ideal gas. The resultant pressure gradients cause air parcels to accelerate. Convergence/divergence couplets result, with corresponding density changes and thus further propagation of the disturbance.

Typical characteristic scales for sound waves are $\tau = 0.1$ s and $L = 10$ m. If a particle displacement scales as L , local accelerations scale as $L\tau^{-2} \approx 1000$ m/s². Clearly, gravity is negligible for such strong accelerations ($g \ll L\tau^{-2}$), and the Coriolis force can likewise be neglected on such short timescales ($f\tau \ll 1$). For a basic state, a uniform flow \bar{u} in any direction and locally uniform density $\bar{\rho}$, pressure \bar{p} , and temperature \bar{T} are assumed. Under these conditions the linearized perturbation equations become

$$\frac{D\mathbf{v}'}{Dt} = -\frac{1}{\bar{\rho}} \nabla p' \quad (76a)$$

$$\frac{D\rho'}{Dt} + \nabla \cdot \bar{\rho}\mathbf{v}' = 0 \quad (76b)$$

$$\frac{D}{Dt} \left(\frac{p'}{c_s^2 \bar{\rho}} - \frac{\rho'}{\bar{\rho}} \right) = 0, \quad (76c)$$

where $D/Dt = \partial/\partial t + \bar{u}(\partial/\partial x)$ represents the substantial time-derivative operator following the basic-state flow, $c_s^2 = \gamma \bar{p}/\bar{\rho} = \gamma R \bar{T}$ ($\gamma = c_p/c_v$) is a characteristic parameter of the ideal gas, and \mathbf{v}' is the three-dimensional perturbation velocity. Operating on the momentum equation with $\nabla \cdot \bar{\rho}$, substituting for $\nabla \cdot \bar{\rho}\mathbf{v}'$ from the continuity equation and for Dp'/Dt from the thermodynamic equation gives

$$D^2 p'/Dt^2 = c_s^2 \nabla^2 p'. \quad (77)$$

Assuming that all perturbation fields oscillate in time with frequency σ and in space with x , y , and z wavenumbers k , l , and m {i.e., all perturbation fields can be represented as a complex amplitude times $\exp[i(kx + ly + mz - \sigma t)]$ }, the following dispersion relation is obtained:

$$\sigma = k\bar{u} \pm c_s \sqrt{k^2 + l^2 + m^2}. \quad (78)$$

The term $k\bar{u}$ represents a Doppler shift in the direction of the mean flow \bar{u} . In a frame of reference moving with the mean wind, the phase front propagates uniformly in the direction of the wavenumber vector $\mathbf{k} = ik + jl + km$ with phase speed $c = c_s = \sqrt{\gamma R \bar{T}}$. This ideal sound wave is dynamically nondispersive, which means the phase speed is independent of the wavelength; however, the wave energy certainly disperses geometrically from a local source.

2. Buoyancy or Gravity Waves

Gravity waves propagate energy horizontally and vertically in a stably stratified fluid. Again, the fluid motions are assumed to be adiabatic, and gravity provides the restoring force. If a fluid parcel is displaced vertically, it will conserve its potential temperature under the adiabatic assumption. With stable stratification, potential temperature increases monotonically with height ($\partial\theta/\partial z > 0$). Thus a fluid parcel forced upward finds itself colder than its surrounding environment and therefore negatively buoyant; it tends to return to its original level. Likewise, a parcel forced downward is warmer than its environment and is therefore buoyed upward. The appropriate timescale for buoyancy oscillations is determined by the stratification $\tau \approx N^{-1}$, the inverse Brunt–Väisälä frequency. Typically, N is approximately 100 s⁻¹ in the troposphere and 200 s⁻¹ in the stratosphere.

In this section a basic state that is in uniform horizontal motion \bar{u} in the x direction with the basic-state thermodynamic variables vertically varying [$\bar{\rho}(z)$, $\bar{p}(z)$, $\bar{T}(z)$, $\bar{\theta}(z)$] and hydrostatically related is assumed. The density and pressure vary with a scale height $H = R\bar{T}/g$. Considered here are oscillations of smaller vertical scale D such that $D \leq 1$ km $\ll H$.

Because the timescale for gravity waves is three orders of magnitude greater than that characteristic of sound waves, the compressibility or elasticity of the air may be expected to be unimportant. Hence $d\rho/dt$ is neglected in the simplified continuity equation. Furthermore, because density varies little on the scale D , $\bar{\rho}$ is replaced by a constant ρ_0 . Finally, Poisson's equation for linearized perturbations can be written

$$\frac{\theta'}{\bar{\theta}} = -\frac{\rho'}{\bar{\rho}} + \frac{p'}{c_s^2 \bar{\rho}}. \quad (79)$$

The ratio of the pressure term to the density term scales as $gD/c_s^2 \approx 0.1$ for gravity waves. Therefore, a nondimensional perturbation variable buoyancy may be defined as

$$b' = \frac{\theta'}{\bar{\theta}} \approx -\frac{\rho'}{\rho_0}. \quad (80)$$

This set of approximations is often called the Boussinesq approximation. The resulting governing equations are written

$$\frac{D\mathbf{v}'}{Dt} = -\nabla \frac{p'}{\rho_0} \quad (81a)$$

$$\delta \frac{Dw'}{Dt} = -\frac{\partial}{\partial z} \left(\frac{p'}{\rho_0} \right) + gb' \quad (81b)$$

$$\nabla \cdot \mathbf{v}' + \frac{\partial w'}{\partial z} = 0 \quad (81c)$$

$$\frac{Db'}{Dt} + \frac{N^2}{g} w' = 0. \quad (81d)$$

The tracer δ is zero in the hydrostatic approximation and unity otherwise. Here ∇ is the horizontal gradient operator. These can be combined to obtain a single equation in w' :

$$\frac{D^2}{Dt^2} \left(\delta \nabla^2 + \frac{\partial^2}{\partial z^2} \right) w' + N^2 \nabla^2 w' = 0 \quad (82)$$

Assuming as before a normal-mode, plane-wave solution proportional to $\exp[i(kx + ly + mz - \sigma t)]$, the dispersion relation is

$$(\sigma - k\bar{u})^2 = N^2 \frac{k^2 + l^2}{\delta(k^2 + l^2) + m^2}. \quad (83)$$

From the dispersion relation, note that the Doppler shifted (or intrinsic) frequency $\hat{\sigma} = \sigma - k\bar{u}$ is always less than the Brunt–Väisälä frequency in magnitude. Thus the intrinsic period of a buoyancy oscillation exceeds $2\pi N^{-1} \approx 10$ min. Because the three-dimensional velocity is nondivergent [cf. Eq. (81c)], parcel displacements are located along phase fronts. If the basic state is at rest, a wave source oscillating at the Brunt–Väisälä frequency will cause waves with $m = 0$ and vertical displacements. As the source becomes more slowly oscillating, the displacements become more horizontal. In the limit of slowly oscillating sources with $\sigma^2 \ll N^2$, the displacements are primarily horizontal with $k^2 + l^2 \ll m^2$ and $w' \ll u', v'$. Horizontal length scales exceed the vertical length scale D so that the hydrostatic approximation may be made to good accuracy. In the hydrostatic limit ($\delta = 0$), the dispersion relation derived above becomes identical to the corresponding dispersion relation for a shallow-water fluid,

$$(\sigma - k\bar{u})^2 = gh_e(k^2 + l^2) \quad (84)$$

if the vertical wavenumber and the mean fluid depth h_e are related by

$$m^2 = N^2/gh_e. \quad (85)$$

In this limit, the horizontal structure of gravity waves in a shallow-water fluid layer and of buoyancy waves in a stratified ideal gas correspond exactly.

3. Effects of Rotation: Inertia–Gravity Waves and Rossby Waves

Thus far the earth's rotation has been ignored. As the frequency of the oscillation approaches f , explicit account must be taken of the Coriolis force. Thus the preceding analysis applies to periods of oscillation between about 10 min and few hours (depending on the local value of f).

In this section, a basic state at rest is assumed; alternatively, if a constant mean flow \bar{u} in the x direction were assumed, the frequency σ would be changed to a Doppler-shifted frequency $\sigma - k\bar{u}$ as in previous sections.

Furthermore, a stably stratified atmosphere is assumed. Under these conditions, waves propagate vertically with wavenumber m . The horizontal structure of large-scale ($D \ll L$) hydrostatic waves is then found from the shallow water equations with $gh_e = N^2/m^2$. The linearized form of the horizontal structure equations in Cartesian geometry is

$$\frac{\partial u'}{\partial t} - f(y)v' + \frac{\partial \Phi'}{\partial x} = 0 \quad (86a)$$

$$\frac{\partial v'}{\partial t} + f(y)u' + \frac{\partial \Phi'}{\partial y} = 0 \quad (86b)$$

$$\frac{\partial \Phi'}{\partial t} + gh_e \left(\frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y} \right) = 0. \quad (86c)$$

Here Φ' represents the temporal and horizontal structure of the geopotential fluctuations on pressure surfaces or pressure fluctuations on height surfaces. The Coriolis parameter is taken to be a function of y . This is in preparation for the beta-plane approximation (see also Section II.C) which allows the use of a cartesian coordinate system where the only effect of sphericity is a linear variation of the Coriolis parameter with y . Because all coefficients are independent of x and t , solutions with zonal wavenumber k and frequency σ proportional to $\exp[i(kx - \sigma t)]$ are considered. The three equations can then be combined into a single equation for the latitudinal structure $V(y)$ of v' , given by

$$\frac{d^2 V}{dy^2} + \left[-\frac{k}{\sigma} \left(\frac{df}{dy} \right) - k^2 + \frac{\sigma^2 - f^2(y)}{gh_e} \right] V = 0. \quad (87)$$

The quantity in brackets is designated $l^2(y)$ and represents an index of refraction that changes with latitude. In general, significant wave energy is located where $l^2 > 0$. Several particular cases involving different assumptions about $f(y)$ are now considered.

Case I. Nonrotating plane: $f = 0$. In this case the index of refraction is constant and the dispersion relation

$$\sigma^2 = gh_e(k^2 + l^2) \quad (88)$$

is just that pertaining to gravity waves from the previous section.

Case II. Midlatitude f -plane: $f = f_0 = \text{constant}$. The index of refraction is again constant. Frequency is related to horizontal wavenumbers by

$$\sigma^2 = f_0^2 + gh_e(k^2 + l^2) = \sigma_{\text{IG}}^2. \quad (89)$$

The Coriolis parameter f_0 places a lower bound on the frequency for these inertia-gravity waves at large horizontal scales. If the horizontal scales are large enough and the vertical scale (proportional to $\sqrt{gh_e}$) is small enough that pressure gradients are negligible, then σ approaches f_0 . This anticyclonic circulation is called an inertial oscillation. Fluid parcels orbit anticyclonically in an attempt to conserve their linear inertia in an absolute reference frame—but are constrained to reside on a horizontal surface. On the other hand, at small scales (large k, l), the gravity-wave characteristic dominates.

On the f -plane, a third solution can also be obtained: $\sigma = 0$ or steady motion. The horizontal momentum equations indicate that this steady but nontrivial flow is geostrophically balanced by pressure gradients. The integrated continuity equation demonstrates that the flow is nondivergent.

Case III. Midlatitude β -plane. If latitudinal excursions of parcel trajectories are great enough that a parcel senses a changing Coriolis parameter, steady nondivergent flow is no longer a solution. As the beta effect influences the dynamics, the exactly balanced solution transforms into a second class of waves that propagate to the west (relative to the mean flow). These oscillations can be studied by applying the midlatitude β -plane approximation, replacing $f(y)$ by the constant f_0 and df/dy by the constant β in Eq. (87). The resulting dispersion relation may be written

$$\sigma^2 - \frac{\beta k g h_e}{\sigma} = f_0^2 + gh_e(k^2 + l^2). \quad (90)$$

For each value of k, l there are now three different values of frequency σ , all real, which correspond to the three time derivatives in the shallow-water equations. Figure 3 gives the dispersion diagram for some representative parameters: the frequency σ , nondimensionalized by f_0 , is the ordinate; wavenumber k nondimensionalized by L_R^{-1} is the abscissa where $L_R = \sqrt{gh_e/f_0^2}$. (The length scale L_R , called the Rossby radius of deformation, is a fundamental parameter for rotating fluids subject to gravitational restoring forces. When the atmosphere is perturbed away from an equilibrium state, the ratio of L_R to the horizontal scale of the perturbation determines the character of the adjustment toward equilibrium.) In Fig. 3 pa-

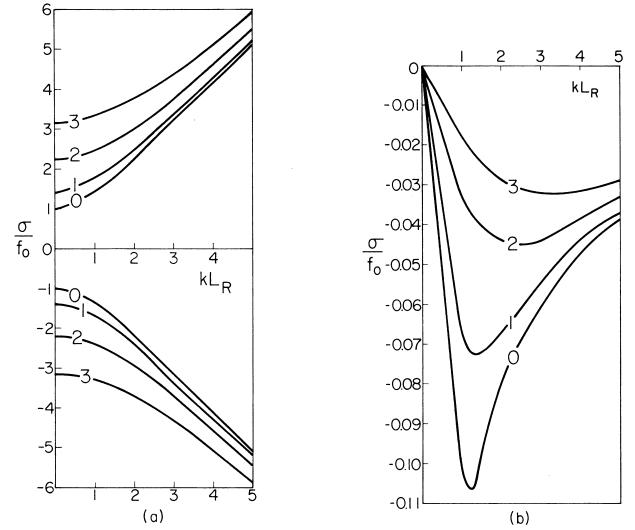


FIGURE 3 Dispersion diagram for the midlatitude β -plane: (a) inertia-gravity waves, (b) Rossby waves. (Note different scales along the ordinates.) The frequency σ and zonal wavenumber k are nondimensionalized by f_0 and the Rossby radius L_R , respectively. Curves are labeled by the index n corresponding to meridional wavenumbers $I_n = nL_R^{-1}$ for $n = 0, 1, 2, 3$. See text for additional details.

rameter values typical of midlatitude dynamics are used: $f_0 = 10^{-4} \text{ s}^{-1}$, $\beta = 2 \times 10^{-11} \text{ m}^{-1} \text{ s}^{-1}$, $gh_e = 10^4 \text{ m}^2 \text{ s}^{-2}$, $L_R = 1000 \text{ km}$. Frequencies satisfying the dispersion relation [Eq. (90)] for four different values of meridional wavenumber are displayed: $I_n = nL_R^{-1}$, where $n = 0, 1, 2, 3$.

From the dispersion diagram, note that the wave solutions separate clearly into two distinct classes:

1. High-frequency inertia-gravity waves. Two of the roots of the cubic dispersion relation have frequencies greater than f_0 . For these roots, $|\beta k g h_e / \sigma| \ll \sigma^2$ and a good first approximation is obtained by neglecting the term inversely proportional to σ , yielding

$$\sigma^2 \approx f_0^2 + gh_e(k^2 + l^2) = \sigma_{\text{IG}}^2 \quad (91)$$

or $(\sigma/f_0)^2 = 1 + (kL_R)^2 + (lL_R)^2$. These two roots are just the inertia-gravity waves of Case II, slightly modified by β so that the eastward-traveling waves ($\sigma < 0$) propagate with slightly higher frequency than the westward-traveling waves ($\sigma > 0$). All the inertia-gravity waves oscillate with frequency greater than f_0 .

2. Low-frequency Rossby waves. This second class of waves is limited to frequencies much less than f_0 in magnitude. For these oscillations, $\sigma^2 \ll |\beta k g h_e / \sigma|$ and the σ^2 term in Eq. (90) may be neglected to obtain a good first approximation to the dispersion relation given by

$$\sigma \approx -\frac{\beta k}{k^2 + l^2 + f_0^2/gh_e} = \sigma_R \quad (92)$$

or

$$\frac{\sigma}{f_0} \approx -\left(\frac{\beta L_R}{f_0}\right) \frac{(k L_R)}{(k L_R)^2 + (l L_R)^2 + 1}. \quad (93)$$

These Rossby waves are westward-traveling, with $\sigma < 0$. For the parameters chosen, $\beta L_R/f_0 = 0.2$, so that the maximum frequency is a factor of five smaller than f_0 . Equivalently, the period of oscillation exceeds $5(2\pi)/f_0 \approx 3.6$ days. This timescale closely fits the timescale for developing weather systems, so that Rossby waves may be expected to be relevant to weather forecasting.

Because $|\sigma| \ll f_0$, the horizontal accelerations in Rossby waves are much smaller than the Coriolis force and pressure-gradient terms. Indeed, this is the assumed balance in the quasi-geostrophic system. In the quasi-geostrophic formulation, the only propagating disturbances are Rossby waves with a dispersion relation essentially equivalent to Eq. (93); fast-moving inertia-gravity waves are scaled out of the equations (as are sound waves).

4. Tropical Waves

The preceding section showed that the midlatitude, hydrostatic waves are characterized by dynamic properties that depend crucially on the timescale of fluctuations relative to the Coriolis parameter, that is, on the nondimensional parameter σ/f_0 . Near the equator this scaling breaks down because $f = 2\Omega \sin \phi$ varies considerably, changing sign from northern to southern hemisphere. Although $f_0 = 0$ is appropriate for tropical waves, the ratio σ/f_0 then becomes meaningless and $L_R \rightarrow \infty$ at the equator.

This dilemma is solved by invoking the equatorial β -plane in the latitudinal structure equation [Eq. (87)]. Writing $f = \beta y$ (with $f_0 = 0$ implicitly), the structure equation becomes

$$\frac{d^2 V}{dy^2} + \left(-\frac{\beta k}{\sigma} - k^2 + \frac{\sigma^2 - \beta^2 y^2}{gh_e} \right) V = 0. \quad (94)$$

Here β and gh_e are the only parameters available to scale the frequency σ and the zonal wavenumber k . Dimensional analysis yields dynamical scales for length (L) and time (T):

$$L = (gh_e)^{1/4} \beta^{-1/2}$$

$$T = (gh_e)^{-1/4} \beta^{-1/2}.$$

Nondimensionalizing $k_* = kL$, $y_* = y/L$, $\sigma_* = \sigma T$, the structure equation becomes

$$\frac{d^2 V}{dy_*^2} + \left[\left(-\frac{k_*}{\sigma_*} - k_*^2 + \sigma_*^2 \right) - y_*^2 \right] V = 0. \quad (95)$$

This relatively simple governing equation has well understood solutions: parabolic cylinder functions. For so-

lutions required to be confined to the tropics, V can be written in a very simple closed form by letting the meridional domain expand to infinity. The result is

$$V_n(y_*) = c_n H_n(y_*) e^{-y_*^2/2}, \quad (96)$$

where c_n is an arbitrary constant and H_n is the n th-order Hermite polynomial; the accompanying dispersion relation is

$$-\frac{k_*}{\sigma_*} - k_*^2 + \sigma_*^2 = 2n + 1. \quad (97)$$

This cubic equation for σ_* is conspicuously similar to the midlatitude dispersion relation. Indeed, the dispersion diagram, shown in Fig. 4, has distinct similarities. Inertia-gravity waves are present in the high-frequency region and are approximately described by observing that if $|k_*/\sigma_*| \ll \sigma_*^2$ Eq. (97) yields

$$\sigma_*^2 \approx k_*^2 + 2n + 1, \quad \text{for } n = 1, 2, 3, \dots \quad (98)$$

Equatorial Rossby waves appear in the low-frequency limit $\sigma_*^2 \ll |k_*/\sigma_*|$:

$$\sigma_* \approx -\frac{k_*}{k_*^2 + 2n + 1}, \quad \text{for } n = 1, 2, 3, \dots \quad (99)$$

There are two waves that were not present in the midlatitude wave spectrum considered earlier. Both of these new waves extend from zero frequency to high frequency, unlike the midlatitude case, which contained a distinct gap

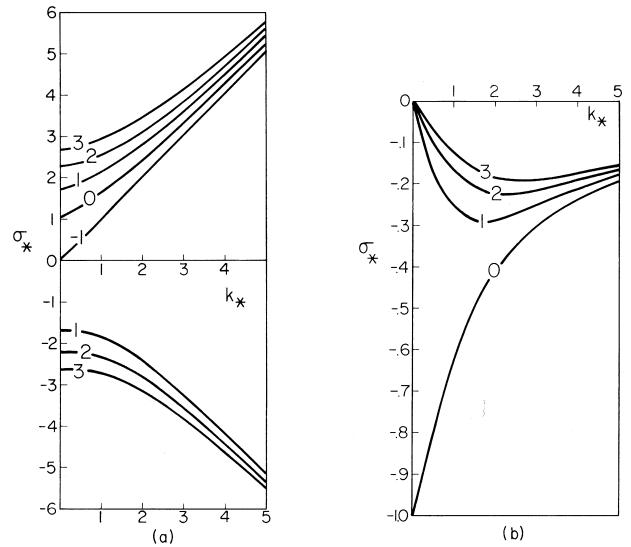


FIGURE 4 Dispersion diagram for the equatorial β -plane: (a) inertia-gravity waves and the Kelvin wave, (b) Rossby waves and the mixed Rossby-gravity wave. (Note different scales along the ordinates). The nondimensional frequency is σ_* and the nondimensional zonal wavenumber is k_* . Curves are labeled by the index n corresponding to different meridional structures. See text for additional details.

between the high-frequency and low-frequency waves. The $n = 0$ wave has Rossby type characteristics when it propagates westward at large wavenumber (k_* large, $\sigma_* < 0$) but joins the gravity-wave spectrum with high frequency when it propagates eastward with large wavenumber (k_* large, $\sigma_* > 0$). Thus it is called the mixed Rossby-gravity wave.

The second addition is the peculiar equatorial Kelvin wave. On the equatorial beta-plane, its meridional velocity is identically zero, satisfying the governing equation [Eq. (95)] trivially. Setting $v = 0$ in the shallow-water equations shows that the Kelvin wave is in geostrophic balance in one direction [from Eq. (86b)] but propagates zonally as a nondispersive pure gravity wave [from Eqs. (86a) and (86c)] with $\sigma_* = k_*$. This hybrid wave propagates only toward the east. (The corresponding westward-propagating wave has amplitude increasing away from the equator and thus does not satisfy boundary conditions.) The Kelvin wave is often denoted $n = -1$ because $\sigma_* = k_*$ is a solution to Eq. (97) with $n = -1$.

Vertically propagating Kelvin and mixed Rossby-gravity modes have been identified in the stratosphere; the evidence in the troposphere is less clear. Internal oceanic Kelvin waves are considered by many oceanographers to be the primary mechanism for the eastward transport of heat along the equator associated with the significant interannual phenomenon known as the El Niño/Southern oscillation (ENSO), which affects weather and climate globally.

Actually, Kelvin waves were originally named for mid-latitude oceanic waves that propagate along a coastline with vanishing velocity component normal to the coast. The classical Kelvin waves are in fact a special case of the inertial-gravity waves with the dispersion relation of Eq. (89). If x is the coordinate along the coast, y is directed toward the water, and the midlatitude f -plane assumption is made, $v' = 0$ is a solution to Eq. (86) if

$$\sigma^2 = gh_e k^2$$

and

$$0 = f_0^2 + gh_e l^2,$$

which clearly satisfies Eq. (89). The amplitude of the coastal Kelvin wave decays exponentially away from the coast with the horizontal scale L_R , the Rossby radius of deformation.

B. Atmospheric Instabilities

Just as geophysical waves can be characterized by a restoring force tending to return a displaced fluid parcel to an initial position, many instabilities can be described by a physical process tending to increase the displacement of

TABLE III Atmospheric Instabilities^a

Instability type	Local necessary condition for instability
Static instabilities	
Dry convective (B)	$N^2 < 0$
Moist convective (B)	$N_e^2 < 0$
Conditional instability (B)	Midtropospheric minimum in θ_e^* and low-level rising motion
Dynamic instabilities	
Wind shear (Kelvin–Helmholtz)	$Ri < \frac{1}{4}$
Inertial	
Horizontal (C)	$f\zeta_a < 0$
Vertical (D)	$Ri < 1$
Horizontal/vertical (D)	$f\bar{P} < 0$
Quasi-geostrophic	
Barotropic (E)	$\partial\zeta_a/\partial y \leq 0$
Baroclinic (E)	$\partial\bar{P}/\partial y \leq 0$ or $(\partial\bar{T}/\partial y)_{\text{surface}} \neq 0$

^a The instabilities are labeled so that they relate directly to the categories of wave types and restoring mechanisms in Table II. For example, E refers to an instability corresponding to a Rossby wave (category E of Table II). Kelvin–Helmholtz instability is not directly related to any wave type in Table II and so is unlabeled.

a parcel away from its initial equilibrium position. Several of the atmospheric instability mechanisms correspond to waves in which the restoring force is negative. In this section, instability mechanisms relevant to atmospheric motions on a broad spectrum of scales are examined. Table III summarizes these instabilities, particularly as they relate to the wave types and restoring mechanisms of Section III.A.

1. Static Instabilities

The term static instability refers to those motions that tend to amplify when the basic state is at rest (static). Because buoyancy is a key restoring force in a stratified fluid, static instability typically arises in situations in which the buoyancy force drives a parcel vertically away from its initial position. A simple analog is the heating of water on a stove: the heat source warms the bottom of the water pot, which then warms the water molecules very close to the bottom. Because warm water is generally less dense than cold water, any small random perturbation will cause the warm water to be displaced upward; by conservation of mass, the cold water above is displaced downward. This turbulent convective process transfers the heat through the entire fluid much more rapidly than molecular diffusion, and therefore is the dominant heat-transfer mechanism.

In the atmosphere, the same mechanism leads to static instability. However, the density of air changes substantially (about an order of magnitude in one vertical scale

height), whereas the density of the analogous water system changes by only a few percent. Formally, this compressibility is accounted for by using potential temperature θ rather than temperature T as the key parameter. In the water system, $\partial T/\partial z < 0$ corresponds to static instability. In the dry atmosphere (i.e., neglecting the effects of moisture), $\partial\theta/\partial z < 0$ is the condition for static instability. From the discussion on gravity waves, this implies that $N^2 < 0$ for static instability. Indeed, the dispersion relation for gravity waves then shows that $\sigma^2 < 0$, thereby implying that the frequency is pure imaginary; that is, $\sigma = i\sigma_i$ where σ_i is the imaginary part of the complex frequency σ . A pure imaginary frequency indicates the possibility of a disturbance with a secular change in amplitude, that is, growing or decaying with time depending on whether the disturbance gains energy from, or loses energy to, the basic state:

$$\exp(-i\sigma t) = \exp[-i(i\sigma_i)t] = \exp(\sigma_i t). \quad (100)$$

The unstable mode has $\sigma_i > 0$, and the stable mode has $\sigma_i < 0$. The most unstable mode is the one with the largest positive σ_i , and therefore the fastest growth rate or equivalently the shortest e -folding timescale ($1/\sigma_i$). Therefore, the dispersion relation of Eq. (83) indicates that static instability is likely to be manifested at smallest horizontal scale (i.e., largest horizontal wavenumber) and largest vertical scale (i.e., smallest vertical wavenumber). Of course, friction has been neglected; turbulence and molecular diffusion will place a limit on the instability on the smallest horizontal scales. Because the real part of the frequency, σ_r , is zero, static instability is stationary with respect to the mean wind and does not propagate horizontally. Dry static instability places a physical limit on the large-scale temperature lapse rate: if the lapse rate exceeds its dry adiabatic value of 10 K/km, overturning will occur spontaneously in the atmosphere.

This type of buoyancy instability is also manifested in convective boundary layers near the earth's surface. As the surface is heated by sunlight, the near-surface air warms and the lapse rate becomes unstable. Overturning occurs, mixing the air and forming what is called a mixed boundary layer. Furthermore, because water vapor may be present near the surface with mixing ratio as high as 3% by mass, the lesser mass of water-vapor molecules (molecular weight 18) relative to the mass of the air molecules (average molecular weight approximately 29) must be taken into account. Rather than adjust the gas constant R , meteorologists adjust the temperature and define the so-called virtual temperature and virtual potential temperature to account for the presence of water vapor.

If air parcels with typical moisture content (say, relative humidity greater than 10%) near the surface are raised, they cool adiabatically and the Clausius–Clapeyron equa-

tion (which shows how the vapor pressure of a saturated air parcel will change with temperature) indicates that they will eventually become saturated. The water vapor then condenses, releasing the latent heat of condensation and thereby warming the air parcel. For air at a given pressure level with specified temperature and relative humidity, a thermodynamic diagram (such as the skew $T - \log p$ diagram) gives the pressure level at which the parcel saturates, the so-called lifting condensation level (LCL). If the parcel is raised above the LCL, it will no longer conserve its potential temperature because of the latent heating. By applying the thermodynamic energy equation, meteorologists have derived another parameter that is approximately conserved as water is condensed in a cloud: the equivalent potential temperature θ_e . This is the potential temperature that a parcel of air would have if all of its moisture were condensed out and the resultant latent heat used to warm the parcel.

In this moist convective instability, the same derivation for static instability applies except that the θ_e profile is the relevant thermodynamic parameter, rather than the θ profile. Thus the condition for moist convective instability is $N_e^2 \equiv (g/\theta_e)\partial\theta_e/\partial z < 0$.

One final moist static instability is especially relevant for tropical meteorology: conditional instability. Near the surface in the tropics, θ_e has a relative maximum of roughly 350 K because of the high temperature and moisture content. In the midtroposphere, θ_e reaches a minimum of around 330 K. At higher levels, $\theta_e \approx \theta$ because the moisture content is negligible, and both θ and θ_e increase rapidly with height in the stratosphere. An additional important variable for determining conditional instability is the saturated equivalent potential temperature θ_e^* . This is the potential temperature of a hypothetically saturated atmosphere with the same temperature structure as the actual atmosphere. It has a similar vertical profile to θ_e , but is always larger and reaches a midtropospheric minimum of approximately 340 K. Figure 5 shows a typical tropical sounding for the hurricane season in the West Indies region. If a surface air parcel with equivalent potential temperature θ_e^P (superscript P meaning parcel) is forced to rise, θ_e^P , by definition, remains constant, even if it is lifted through its LCL where it becomes saturated. If it is lifted further upward to the point where $\theta_e^P = \theta_e^*$, then it becomes warmer than its environment and accelerates upward spontaneously as an unstable parcel. This critical point is known as the level of free convection (LFC). Thus conditional instability can be realized only if the following conditions are met: (1) there is a midtropospheric minimum in θ_e^* and (2) a circulation exists to force the parcel to its LFC. Conditional instability is responsible for the deep cumulonimbus convection that prevails in the tropics. Indeed, the entire tropical troposphere can be viewed

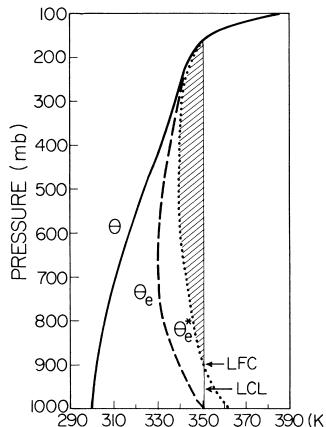


FIGURE 5 Vertical profiles of potential temperature θ , equivalent potential temperature θ_e , and the equivalent potential temperature θ_e^* of a hypothetically saturated atmosphere with the same temperature at each level. The thin, solid vertical line is the equivalent potential temperature of a surface air parcel forced to rise. The parcel's LCL (determined from temperature and moisture information not explicitly available from the diagram) and LFC are noted. The hatched area indicates the region of positive buoyancy, that is, where the equivalent potential temperature of the parcel is greater than θ_e^* . [Taken from the mean hurricane season sounding for the West Indies area, Jordan, C. L. (1958). *J. Meteorol.* **15**, 91–97.]

as a deep convective layer that is stirred by the cumulus transport of heat and moisture. In this scenario, the tropical tropopause is the maximum vertical extent of the mixing process.

2. Dynamic Instabilities

The term dynamic instability refers to those perturbations whose amplification depends inherently on the basic state being in motion relative to the reference frame of an observer who rotates with the earth.

As with static instabilities, the identification of appropriate temporal and spatial scales of maximum growth rate is often accomplished with a linear perturbation analysis in which dissipation is tentatively neglected. For basic states consisting of parallel flow in the x direction, a key element of many dynamic instabilities is the presence of a “critical surface” on which the speed of propagation of the instability (given by the real part of the phase speed c_r) equals the mean flow speed \bar{u} . By considering basic states that are specified and do not interact with the linear waves superimposed on them, the focus is placed on essentially linear instabilities. This restricts the applicability of the instability results to the initial stages of growth. As the initial disturbance grows to the stage when nonlinear effects cannot be neglected, the perturbations will begin to affect the mean flow, usually changing it in such a way as to reduce the source

of instability. These nonlinear instabilities are generally more difficult to understand and explain. Fortunately, linear instability studies have led to great insight into many phenomena.

a. Small-scale shear instabilities. Shear instability can develop if the flow speed varies in a direction perpendicular to unidirectional flow and if the molecular viscosity is small. The Richardson number, another key nondimensional parameter defined by

$$Ri = \frac{N^2}{(d\bar{u}/dz)^2} \quad (101)$$

measures the competition between the destabilizing influence of the wind shear and the stabilizing influence represented by a real Brunt–Väisälä frequency. For inviscid flow, shear instability develops when $Ri < \frac{1}{4}$. This criterion can be viewed heuristically as an energetic statement for the perturbations. If a perturbation can acquire kinetic energy from the environmental shear of the basic state faster than it loses potential energy by vertical displacements in stable stratification, then it can gain total energy from the basic state and amplify. With a large Richardson number, the wind shear is not strong enough for the displaced air parcels to gain the required kinetic energy. A small Richardson number indicates weak stratification; parcels can be displaced vertically without having to do much work against gravity, and displacements can amplify.

This instability mechanism can lead to turbulence. Kelvin–Helmholtz instability is a form commonly observed in the presence of a strongly sheared flow in the interior of a fluid. Turbulent surface layers can also develop by this mechanism.

b. Inertial instabilities. Intertial instabilities depend crucially on the rotation of the earth; the Coriolis parameter f plays a dominant role. The simplest formulation is to consider horizontal shear flow $\bar{u}(y)$ on an f -plane. As with the inertial oscillation, suppose that the scales and structure of the small-amplitude disturbance dictate that pressure gradients are unimportant for the dynamics of the linearized perturbations. Then the perturbation momentum equations can be written

$$\frac{\partial u'}{\partial t} + \bar{u}(y) \frac{\partial u'}{\partial x} - \left(f - \frac{\partial \bar{u}}{\partial y} \right) v' = 0 \quad (102a)$$

$$\frac{\partial v'}{\partial t} + \bar{u}(y) \frac{\partial v'}{\partial x} + fu' = 0. \quad (102b)$$

The variable v' can be eliminated to obtain a single equation in u' :

$$\left(\frac{\partial}{\partial t} + \bar{u} \frac{\partial}{\partial x} \right)^2 u' + \left[f \left(f - \frac{\partial \bar{u}}{\partial y} \right) \right] u' = 0. \quad (103)$$

The inertial parameter I^2 is now defined as

$$I^2 = f \left(f - \frac{\partial \bar{u}}{\partial y} \right) = f(f + \bar{\zeta}) = f \bar{\zeta}_a, \quad (104)$$

where $\bar{\zeta}_a$ is the vertical component of the basic-state absolute vorticity. The role of I^2 for horizontal displacements is similar to the role of N^2 for vertical displacements. The criterion for inertial instability is that $f \bar{\zeta}_a < 0$ somewhere in the fluid. For example, for symmetric perturbations (i.e., perturbations that do not vary in the down-wind direction x) in a basic flow with locally constant shear such that $\bar{\zeta}_a < 0$, the dispersion relation for Eq. (103), assuming solutions with time dependence $\exp(-i\sigma t)$, is

$$\sigma^2 = I^2. \quad (105)$$

Because I^2 is negative, σ is pure imaginary and the perturbations are unstable with e -folding time I^{-1} . The inertial instability develops in regions where the absolute vorticity $\bar{\zeta}_a$ has sign opposite to f . This derivation reduces to that for the stable inertial oscillation in the special case in which $\bar{u} = 0$ and $I^2 = f^2$.

When the mean zonal flow is more realistic, with both latitudinal and vertical shear [$\bar{u}(y, z)$], the basic-state potential vorticity \bar{P} replaces the absolute vorticity in determining instability criteria. A necessary condition for inertial instability is that $f \bar{P} < 0$ somewhere in the fluid. This criterion is easily related to the Richardson number if the basic-state flow is assumed geostrophic. From Section I.G.4, the potential vorticity can be written

$$\bar{P} = \frac{1}{\bar{\rho}} \bar{\zeta}_a \cdot \nabla \bar{\theta} = \frac{1}{\bar{\rho}} \left(\frac{\partial \bar{u}}{\partial z} \right) \left(\frac{\partial \bar{\theta}}{\partial y} \right) + (f + \bar{\zeta}) \frac{\partial \bar{\theta}}{\partial z}. \quad (106)$$

Using the approximate thermal wind relation

$$-f \frac{\partial \bar{u}}{\partial z} = g \frac{\partial \ln \bar{\theta}}{\partial y} \quad (107)$$

the instability criterion becomes

$$f \bar{P} = \frac{\bar{\theta}}{\bar{\rho} g} \left(f \frac{\partial \bar{u}}{\partial z} \right)^2 \left[\text{Ri} \frac{f + \bar{\zeta}}{f} - 1 \right] < 0. \quad (108)$$

In the case of horizontal shear only, $f \bar{P}$, is proportional to $f(f + \bar{\zeta})$ and the result of the previous derivation is obtained. For the case of vertical shear only, $f \bar{P}$ is proportional to $\text{Ri} - 1$ and the instability criterion becomes $\text{Ri} < 1$.

c. Barotropic and baroclinic instability. The final classes of instability considered here act on the synoptic and planetary scales, and therefore are relevant for large-scale weather phenomena. These instabilities result

from a change in sign of the basic-state potential vorticity gradient, rather than a sign change in the potential vorticity itself. They differ from inertial instabilities in that they lead only to asymmetric perturbations. Quasi-geostrophic theory is very useful for describing the dynamics of these large-scale phenomena.

By analogy with static instability and inertial instability, these instabilities may also be viewed as resulting from a restoring mechanism becoming negative somewhere in the atmosphere. For example, the restoring mechanism for a Rossby wave is β , which in a barotropic atmosphere with no basic-state wind shear is simply the gradient of potential vorticity. For a neutrally propagating Rossby wave, this potential vorticity gradient is positive. However, basic-state wind configurations exist where the potential vorticity gradient is negative, leading to a negative restoring mechanism.

Barotropic instability. Barotropic instability can result from a basic state with horizontal shear but no vertical shear of the zonal average wind. By the thermal wind relationship, the basic-state horizontal temperature gradient vanishes and temperature advection does not occur. Furthermore, it is assumed that the perturbed flow is purely horizontal, and therefore, by necessity, nondivergent in the horizontal plane with no vertical shear. In this situation the potential vorticity reduces to the absolute vorticity. H.-L. Kuo showed in 1949 that such a flow is susceptible to instability. A necessary condition for barotropic instability may be derived that states that somewhere in the fluid

$$\frac{d(f + \bar{\zeta})}{dy} = \beta - \frac{\partial^2 \bar{u}}{\partial y^2} = 0. \quad (109)$$

Limits on the unstable growth rate are obtained from a modified version of the so-called semicircle theorem, developed by researchers J. Miles and L. Howard in 1961. This significant global theorem states that the complex phase speed $c = c_r + i c_i$ (where c_i is the growth rate σ_i divided by the zonal wavenumber k) must satisfy the relationship

$$(c_r - \bar{u}_m)^2 + c_i^2 \leq \bar{u}_s^2 \left(1 + \frac{c_\beta}{\bar{u}_s} \right) \equiv U_R^2, \quad (110)$$

where

$$\bar{u}_m = \frac{(\bar{u}_{\max} + \bar{u}_{\min})}{2} \quad \text{a measure of the mean wind speed}$$

$$\bar{u}_s = \frac{(\bar{u}_{\max} - \bar{u}_{\min})}{2} \quad \text{a measure of the total wind shear}$$

$$c_\beta = \frac{\beta}{k^2 + \pi^2/4L^2} \quad \text{relative zonal phase speed of a Rossby wave with zonal wavenumber } k \text{ and meridional wave-number } \pi/2L.$$

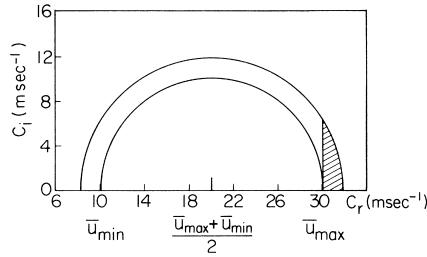


FIGURE 6 A graphical representation of the semicircle theorem for barotropic instability calculated for $\bar{u}_{\min} = 10 \text{ m s}^{-1}$, $\bar{u}_{\max} = 30 \text{ m s}^{-1}$, $\beta = 2 \times 10^{-11} \text{ m}^{-1} \text{ s}^{-1}$, $k = 2\pi/3000 \text{ km}$, $L = 1500 \text{ km}$. The inner and outer semicircles have radii \bar{u}_u and U_R , respectively.

The complex phase speed must lie inside a closed region of the complex phase speed plane with abscissa c_r and ordinate c_i as in Fig. 6. Because the upper half plane represents unstable perturbations, the complex phase speed c for an instability must be located within the semicircle with radius U_R and center at $(\bar{u}_m, 0)$. Part of the semicircle is cut off because it may be shown that $c_r < \bar{u}_{\max}$. It may be demonstrated that as wavenumber k approaches zero, the growth rate must also approach zero. Furthermore, with the exception of the small region within the semicircle to the left of \bar{u}_{\min} , barotropic instabilities always have a critical level y_c , where $c_r - \bar{u}(y_c) = 0$.

Baroclinic instability. Baroclinic instability is understood to be the dynamic cause for synoptic-scale, midlatitude storms. It is the result of a vertical shear in the basic-state zonal wind. Its elucidation by J. Charney in 1947 and E. Eady in 1949 provided the springboard for much of modern dynamic meteorology. Formally, baroclinic instability results can be derived by generalizing the derivation for barotropic instability to the case where $\bar{u} = \bar{u}(z)$. Physically the key difference between the two instabilities is the poleward advection of basic-state temperature by the perturbation meridional wind. The formal relationship is a consequence of the significance of the potential vorticity gradient for both instabilities. In baroclinic instability, the horizontal temperature gradient, which is proportional to vertical shear of the wind through the thermal wind relation, is the key source of a change in sign of the potential vorticity gradient. In fact, the conditions for instability can be satisfied even if the vorticity gradient is identically zero in the interior but a horizontal temperature gradient exists at the lower boundary (This is a generalization of the Eady problem in baroclinic instability theory). Because of the formal similarity to the barotropic instability problem, a semicircle theorem giving bounds for growth rate and phase speed is available for baroclinic instability as well. In baroclinic instability, the Rossby radius of deformation is the relevant horizontal scale of motion.

C. Dynamics of Midlatitude Storms

Quasi-geostrophic theory is the cornerstone of dynamic meteorology. It has proven especially useful in understanding the three-dimensional motion associated with midlatitude synoptic storms. The principal insights from quasi-geostrophic theory for these storms are contained in the so-called tendency and omega equations. These are addressed next.

1. Tendency Equation

The tendency equation is an equation describing the time rate of change of the geopotential Φ . It is actually a slightly different form of the potential vorticity equation [Eq. (62)] and is obtained by eliminating ω from Eqs. (60) and (61). The result may be written

$$\left(\nabla^2 + \frac{\partial}{\partial p} \left(\frac{f_0^2}{\sigma} \right) \frac{\partial}{\partial p} \right) \frac{\partial \Phi}{\partial t} = -f_0 \mathbf{V}_g \cdot \nabla \zeta_g - f_0 \mathbf{V}_g \cdot \nabla f - \frac{\partial}{\partial p} \left[-\mathbf{V}_g \cdot \nabla \left(-\frac{f_0^2}{\sigma} \frac{\partial \Phi}{\partial p} \right) \right]. \quad (111)$$

Throughout this section, ∇ refers to the horizontal part of the gradient operator.

Interpretation of the tendency equation is facilitated by considering a wavelike geopotential pattern. This is a convenient idealized way to view the wavy trough and ridge patterns that actually occur in the middle and upper troposphere. For such a wavelike disturbance with a typical vertical structure, it may be shown by Fourier analysis that the term on the left of Eq. (111) is proportional to $-\partial \Phi / \partial t$. The terms on the right-hand side represent the advection of relative vorticity, the advection of planetary vorticity, and the vertical change in horizontal temperature advection, respectively. (The last term is so-called because $-\partial \Phi / \partial p$ is proportional to temperature.) An idealized midlevel flow and surface cyclone pattern for a developing storm are shown in Fig. 7. The relative vorticity advection term has a maximum positive value to the east of the trough and a maximum negative value to the west. The planetary advection term is just the opposite. Both are zero at the trough and ridge axes. For synoptic-scale waves, the relative vorticity advection dominates, while for planetary-scale waves the planetary vorticity advection dominates. Where the absolute vorticity advection is positive, the tendency equation shows that the geopotential will decrease. Where the absolute vorticity advection is negative, the geopotential increases. Therefore, for the typical synoptic situation shown in Fig. 7, the vorticity advection causes height falls to the east of the trough and height rises to the west. This process moves the wave toward the east.

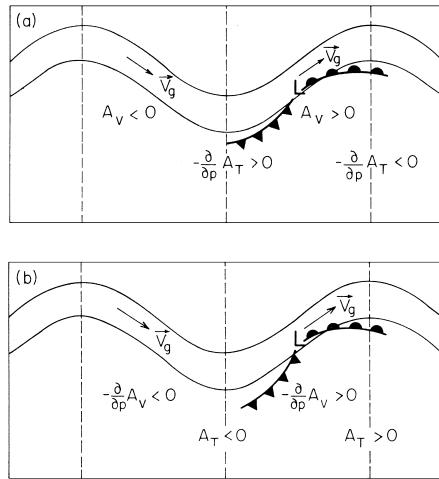


FIGURE 7 An idealized midlevel and surface cyclone pattern for a developing synoptic storm. The regions where the physical processes corresponding to terms in (a) the tendency equation and (b) the omega equation reach a maximum or minimum are shown. Vorticity advection is denoted by $A_v = -\mathbf{V}_g \cdot \nabla(\zeta_g + f)$ and thermal advection by $A_T = -(1/\sigma)\mathbf{V}_g \cdot \nabla(-\partial\Phi/\partial p)$. Trough and ridge axes are denoted by dashed lines.

It is important to note that the vorticity advection moves the geopotential pattern but cannot intensify it because the vorticity advection is zero at the trough and ridge axes. Temperature advection is the process that can amplify the disturbance. Because the contours of geopotential and temperature are more nearly aligned at midlevels, the temperature advection tends to decrease as height increases (or increase as pressure increases). The rate at which temperature advection decreases with height has a maximum positive value at the ridge axis (due to strong warm-air advection associated with the warm front) and a maximum negative value at the trough axis (due to strong cold-air advection associated with the cold front). Where there is warm advection decreasing with height, the tendency equation implies a geopotential rise. Similarly, cold advection decreasing with height implies a geopotential fall. Therefore, in the typical synoptic situation, the temperature advection causes height rises at the ridge axis and height falls at the trough axis.

In summary, the tendency equation shows that geopotential rises/falls are proportional to negative/positive vorticity advection plus the rate of decrease with height of cold/warm advection.

2. Omega Equation

The diagnostic omega equation relates the vertical velocity to the geopotential. Because precipitation is associated with large-scale upward motion, the omega equation is useful in determining those regions where precipitation

may occur. The omega equation is obtained by eliminating the time derivatives from the quasi-geostrophic system of Eqs. (60) and (61), giving

$$\left(\nabla^2 + \frac{f_0^2}{\sigma} \frac{\partial^2}{\partial p^2} \right) \omega = -\frac{f_0}{\sigma} \frac{\partial}{\partial p} [-\mathbf{V}_g \cdot (\zeta_g + f)] - \nabla^2 \left[-\frac{1}{\sigma} \mathbf{V}_g \cdot \nabla \left(-\frac{\partial \Phi}{\partial p} \right) \right]. \quad (112)$$

The omega equation may be interpreted in a manner similar to the tendency equation. If ω has a wavelike pattern with typical vertical structure, then the term on the left-hand side is proportional to $-\omega$. The terms on the right-hand side represent the vertical change in absolute vorticity advection and the horizontal Laplacian of temperature advection. Because the geopotential and vorticity patterns are more closely aligned at lower levels, the vorticity advection tends to increase with height (or decrease with pressure). The rate at which the absolute vorticity advection increases with height reaches a maximum positive value east of the trough and a maximum negative value west of the trough. Where there is positive vorticity advection increasing with height, the omega equation implies that $\omega < 0$ or upward motion. Where there is negative vorticity advection increasing with height, there is downward motion. Therefore, in the typical developing synoptic storm shown in Fig. 7, the vorticity advection causes upward motion east of the trough and downward motion west of the trough.

The horizontal Laplacian of temperature advection may be shown to be proportional to the temperature advection itself. There is maximum warm-air advection at the ridge axis associated with the warm front and maximum cold-air advection at the trough axis associated with the cold front. If there is warm-air advection, the omega equation implies $\omega < 0$ or upward motion. For cold-air advection, downward motion is implied. Therefore, in the typical synoptic situation, temperature advection produces upward motion at the ridge axis and downward motion at the trough axis.

In summary, the omega equation shows that rising/sinking motion is proportional to the rate of increase with height of positive/negative vorticity advection plus warm/cold advection. A disadvantage of the omega equation expressed in the form of Eq. (112) is that the vorticity advection and the thermal advection are not independent and tend to cancel one another.

This disadvantage can be overcome by combining the right-hand side of Eq. (112) into one term given by the divergence of a vector field. This so-called **Q**-vector technique is simply demonstrated by neglecting in the Coriolis force the generally small term involving both β and the

ageostrophic wind \mathbf{V}_a . We define \mathbf{Q} as the frontogenesis vector, which is proportional to the Lagrangian rate at which the temperature gradient ∇T is changed by the geostrophic wind \mathbf{V}_g :

$$\mathbf{Q} \equiv -\left(\frac{R}{p}\right) \nabla \mathbf{V}_g \cdot \nabla T.$$

It can be shown that the geostrophic flow by itself tends to destroy the thermal wind balance between geostrophic vertical wind shear and horizontal temperature gradient. A secondary circulation, involving both vertical motion and ageostrophic wind divergence, is *required* to maintain thermal wind balance. Specifically, by differentiating the approximated momentum equation with respect to pressure, taking the horizontal gradient of the thermodynamic equation, and applying the thermal wind relation, the right-hand side of Eq. (112) can be rewritten:

$$-2\nabla \cdot \mathbf{Q} + \left(\frac{\beta f_0}{\sigma}\right) \frac{\partial v_g}{\partial y}$$

Often in synoptic situations, the last term can be safely ignored.

Rising motion occurs where the vectors converge because the “dynamically modified” three-dimensional Laplacian of ω (which is qualitatively similar to the upward motion field) equals twice the convergence of \mathbf{Q} . It is common practice to plot \mathbf{Q} and $-\nabla \cdot \mathbf{Q}$ to aid in the analysis of vertical motion. Knowledge of large-scale vertical motion, along with relative humidity, is essential for forecasting and diagnosing areas of stratiform precipitation.

3. Secondary Circulations

The vertical motion pattern inferred from the omega equation and the divergence associated with it is referred to as a secondary circulation. This secondary circulation is extremely important as it acts to continuously adjust the wind and temperature fields toward geostrophic and hydrostatic balance (or, equivalently, thermal wind balance). For example, east of the trough depicted in Fig. 7, there is positive midlevel vorticity advection. Thermal wind balance imposes a relationship between the vertical change in geostrophic vorticity and the Laplacian of temperature such that a temperature decrease must accompany the increases in midlevel vorticity. This temperature decrease is achieved by the adiabatic cooling associated with upward motion east of the trough as deduced from the omega equation. Similarly, at the ridge axis, there is warm-air advection. The thermal wind constraint implies that the midlevel vorticity must decrease. This is accomplished by the midlevel divergence that occurs in response to the upward motion at the ridge axis as deduced from the omega equation. These examples show that the secondary circu-

lation acts to ensure that the simultaneous changes in temperature and vorticity tend toward a thermal wind balance.

D. Eliassen–Palm Flux

As stated previously, linearization is the mathematical cornerstone in the analysis of atmospheric motion. In addition to making the mathematical analysis more tractable, linearization also leads to more insight into the perturbation (or eddy) motions of the atmosphere. One way in which this insight is manifested is in the use of the Eliassen–Palm (EP) flux, named after the two researchers who first derived it in 1961. In this analysis, the effect of eddy motions is discussed in terms of the zonally averaged eddy fluxes such as the average eddy northward temperature flux $\overline{v' T'}$. In the context of quasi-geostrophic theory, the EP flux may be considered to be a two-dimensional vector in the meridional (latitude–height) plane that measures the zonally averaged northward eddy fluxes of momentum and temperature. Eliassen and Palm derived a theorem that states that for perturbations that have constant amplitude and are conservative (no friction or diabatic heating), the divergence of the EP flux (in the meridional plane) is zero. The significance of EP fluxes is that they are useful as a diagnostic tool for disturbances on a zonal flow. This happens for two reasons. First, they are a measure of “wave activity” in that regions of nonzero EP flux divergence show where the waves are not steady or not conservative. Second, the EP flux divergence is directly related to the effect that the eddies have on the basic state. A quasi-geostrophic exposition of these ideas follows.

If the geopotential Φ is written as the sum of a basic-state part $\bar{\Phi}(y, p, t)$ and a perturbation part $\Phi'(x, y, p, t)$, where the overbar refers to an x -average, the averaged quasi-geostrophic potential vorticity equation [Eq. (62)] may be written (dropping the subscript g on the winds)

$$\frac{\partial \bar{q}}{\partial t} = -\frac{\partial}{\partial y} (\overline{v' q'}) + \bar{D}, \quad (113)$$

where

$$\begin{aligned} \bar{q} &= -\frac{\partial \bar{u}}{\partial y} + f + f_0 \frac{\partial}{\partial p} \left(\frac{1}{\sigma} \frac{\partial \bar{\Phi}}{\partial p} \right) \\ q' &= \frac{\partial v'}{\partial x} - \frac{\partial u'}{\partial y} + f_0 \frac{\partial}{\partial p} \left(\frac{1}{\sigma} \frac{\partial \Phi'}{\partial p} \right) \end{aligned} \quad (114)$$

and \bar{D} is a term representing the averaged nonconservative forces. It may be shown that

$$\overline{v' q'} = \nabla \cdot \mathbf{F} = \frac{\partial}{\partial x} (-\overline{u' v'}) + \frac{\partial}{\partial p} \left(\frac{f_0 \overline{v' \partial \Phi' / \partial p}}{\sigma} \right), \quad (115)$$

where \mathbf{F} is a vector defined as

$$\mathbf{F} = -\overline{u'v'}\hat{\mathbf{j}} + f_0 \frac{\overline{v'\partial\Phi'/\partial p}}{\sigma} \hat{\mathbf{k}} \quad (116)$$

and ∇ in this context refers to divergence in the yp -plane. Therefore,

$$\frac{\partial \bar{q}}{\partial t} = -\frac{\partial}{\partial y}(\nabla \cdot \mathbf{F}) + \bar{D}. \quad (117)$$

The vector \mathbf{F} is the quasi-geostrophic EP flux. Its components are the zonally averaged northward eddy flux of zonal momentum and the zonally averaged northward eddy flux of temperature (because $\partial\Phi'/\partial p$ is proportional to temperature through the hydrostatic approximation). Equation (117) shows that the EP flux divergence, defined by Eq. (115), induces changes in the basic state.

The perturbation potential vorticity equation may be written

$$\frac{\partial q'}{\partial t} + \bar{u} \frac{\partial q'}{\partial x} + v' \frac{\partial \bar{q}}{\partial y} = D', \quad (118)$$

where D' represents the perturbation nonconservative forces. Multiplying by $q' / (\partial \bar{q} / \partial y)$ and averaging yields, assuming $\partial \bar{q} / \partial y \neq 0$,

$$\frac{\partial}{\partial t} \left(\frac{\overline{q'^2}}{2 \overline{\partial \bar{q} / \partial y}} \right) + \nabla \cdot \mathbf{F} = \frac{\overline{q'D'}}{\overline{\partial \bar{q} / \partial y}}. \quad (119)$$

This result may be thought of as a generalized Eliassen-Palm theorem because it shows the relationship between EP flux divergence and nonsteady, nonconservative perturbations. The term in parentheses on the left of Eq. (119) is a measure of wave activity, so that Eq. (119) clearly shows the relationship between EP fluxes and wave activity.

In summary, the EP flux is a vector in the meridional plane whose components, in quasi-geostrophic theory, measure the zonally averaged northward eddy fluxes of momentum and temperature. Maps of the EP flux and its divergence may be prepared from observations or the results of model calculations. Such maps may be used to infer where and to what extent the eddy motions are acting to change the mean flow and where transient, nonconservative eddy motions are present.

E. The General Circulation and Atmospheric Energetics

The general circulation may be thought of as all the motions that characterize the large-scale structure of the atmosphere. This large-scale structure evolves in response to latitudinally dependent energy input from the sun, characteristics of the underlying surface, small-scale effects, and the ever-present growth and decay of transient weather systems.

Conceptually, this evolution occurs in the following way. The atmosphere, land surface, and ocean surface

are heated by absorption of shortwave solar radiation and cooled by longwave infrared radiation loss to space. The difference between the absorbed shortwave energy and the radiated longwave energy is the radiation imbalance. At the earth's surface (land and ocean), this imbalance is positive, thereby implying a radiative excess. Furthermore, because of the strong latitudinal variation of the incoming solar radiation, this radiation balance has a strong latitudinal dependence with a maximum at the tropics and a minimum at the poles. In the atmosphere, the radiation balance is negative, implying a radiation deficit. In contrast to the earth's surface, this deficit is, to a rough approximation, constant with latitude. As a consequence of the surplus at the surface and the deficit in the atmosphere, a transfer of heat from the surface to the atmosphere must take place if neither is to cool or warm indefinitely. However, equatorward of about 40° the surface surplus exceeds the atmospheric deficit, while poleward of 40° the atmospheric deficit exceeds the surface surplus. Thus, the earth-atmospheric system as a whole experiences a radiative surplus equatorward of 40° and a radiative deficit poleward of that latitude. Thus a poleward transport of heat is required to prevent these regions from cooling or warming indefinitely. In summary, a two-way transport is needed: from the surface to the atmosphere and from the equator to the pole. The general circulation of the atmosphere (as well as of the ocean) results from these transport requirements.

It is convenient to discuss the general circulation in terms of its energetics. Energy is a useful concept whose application often leads to significant insight into a given phenomenon. Energy may be classified into several types. The familiar kinetic energy is the energy associated with motion and is equal to half the square of the velocity multiplied by the mass. The gravitational potential energy is the energy associated with the vertical position of an air parcel relative to mean sea level. The internal energy, discussed in the derivation of the first law of thermodynamics, may be shown to be proportional to the potential energy when integrated over the entire depth of the hydrostatic atmosphere. Therefore, the sum of the two may be referred to as a total potential energy. But this total potential energy is not the most relevant energy quantity for the atmosphere, because not all of it is available for conversion to kinetic energy, according to the following argument. Consider a stably stratified atmosphere at rest with no horizontal variation of pressure and temperature and hence no horizontal variation of potential temperature. There is no means to generate motion in such a state, and hence none of the potential energy present in that configuration can be converted to kinetic energy. This is a minimum or ground state of potential energy. If this state is disturbed so that potential temperature surfaces are no longer horizontal,

potential energy becomes available for conversion to kinetic energy through an adiabatic adjustment toward the ground state. Thus the available potential energy is defined as the difference between the actual total potential energy and the total potential energy of the ground state. Observations indicate that only about 0.5% of the total potential energy is available for conversion to kinetic energy. This suggests that the atmosphere is rather inefficient, yet it is this small amount of available potential energy that is associated with most of the atmosphere's significant weather.

Quasi-geostrophic theory is a convenient framework in which to discuss the energetics of the general circulation. It is restrictive in several ways, most notably the inability to correctly describe tropical motions, yet it is capable of giving qualitative insight into the dynamics of the general circulation. The starting point is the quasi-geostrophic vorticity and thermodynamic equations with diabatic heating Q and frictional dissipation ε included. Instead of Φ , the geostrophic streamfunction $\psi = \Phi/f_0$ is used as the dependent variable. Consistent with the framework of a zonal mean and superimposed eddies that was found useful in understanding waves, instabilities, and the wave and mean flow properties and interactions inherent in EP fluxes, ψ is separated into a zonally averaged part $\bar{\psi}$ and a deviation ψ' representing the eddy motion. The flow is assumed to occur in a zonal channel on the β -plane. Boundary conditions are applied representing the constraints of no meridional velocity at the north and south boundaries, no vertical velocity at the top of the model, and a vertical velocity near the lower boundary proportional to frictional convergence in the planetary boundary layer.

Equations for the mean and eddy kinetic and available potential energies may be derived from the vorticity and thermodynamic energy equations. The results are summarized as:

$$\begin{aligned}\frac{d\bar{K}}{dt} &= \{K' \cdot \bar{K}\} + \{\bar{P} \cdot \bar{K}\} - \bar{\varepsilon}, \\ \frac{dK'}{dt} &= -\{K' \cdot \bar{K}\} + \{P' \cdot K'\} - \varepsilon', \\ \frac{d\bar{P}}{dt} &= -\{\bar{P} \cdot P'\} - \{\bar{P} \cdot \bar{K}\} + \bar{G}, \\ \frac{dP'}{dt} &= \{\bar{P} \cdot P'\} - \{P' \cdot K'\} + G',\end{aligned}$$

where the mean and eddy kinetic energies \bar{K} and K' , the mean and eddy available potential energies \bar{P} and P' , the conversion terms from the first type of energy in the brackets to the second, and mean and eddy generation terms are given by

$$\begin{aligned}\bar{K} &= \int_0^{p_{00}} \int_A \frac{(\nabla \bar{\psi})^2}{2} dA \frac{dp}{g} \\ \bar{P} &= \int_0^{p_{00}} \int_A \frac{f_0^2}{2\sigma} \left(\frac{\partial \bar{\psi}}{\partial p} \right)^2 dA \frac{dp}{g} \\ K' &= \int_0^{p_{00}} \int_A \frac{(\nabla \psi')^2}{2} dA \frac{dp}{g} \\ P' &= \int_0^{p_{00}} \int_A \frac{f_0^2}{2\sigma} \left(\frac{\partial \psi'}{\partial p} \right)^2 dA \frac{dp}{g} \\ \{K' \cdot \bar{K}\} &= - \int_0^{p_{00}} \int_A \bar{\psi} \frac{\partial^2}{\partial y^2} (\bar{u}' \bar{v}') dA \frac{dp}{g} \\ \{\bar{P} \cdot \bar{K}\} &= \int_0^{p_{00}} \int_A f_0 \bar{\omega} \frac{\partial \bar{\psi}}{\partial p} dA \frac{dp}{g} \\ \{P' \cdot K'\} &= \int_0^{p_{00}} \int_A f_0 \omega' \frac{\partial \psi'}{\partial p} dA \frac{dp}{g} \\ \{\bar{P} \cdot P'\} &= \int_0^{p_{00}} \int_A \frac{f_0^2}{\sigma} \frac{\partial \bar{\psi}}{\partial p} \frac{\partial}{\partial y} \left(\bar{v}' \frac{\partial \psi'}{\partial p} \right) dA \frac{dp}{g} \\ \bar{G} &= - \int_0^{p_{00}} \int_A \frac{f_0^2}{\sigma} \bar{Q} \frac{\partial \bar{\psi}}{\partial p} dA \frac{dp}{g} \\ G' &= - \int_0^{p_{00}} \int_A \frac{f_0^2}{\sigma} Q' \frac{\partial \psi'}{\partial p} dA \frac{dp}{g}\end{aligned}$$

and $\bar{\varepsilon}$ and ε' represent mean and eddy dissipation, respectively. The energy quantities have been integrated over the volume of the channel; dA is an incremental horizontal area. The gradient operator operates only in the horizontal and subscripts on the geostrophic winds have been dropped.

The kinetic energies involve the quantity $(\nabla \psi)^2/2$, which is equivalent geostrophically to the more familiar $(u^2 + v^2)/2$. In quasi-geostrophic theory, the vertical velocity and the divergent component of the horizontal velocity do not appear in the kinetic energy. The quantity $(-\partial \psi / \partial p)$ appears in several terms and through the hydrostatic equation is proportional to the temperature deviation from the horizontal mean. Thus the available potential energy in quasi-geostrophic theory is proportional to the square of the temperature deviation.

Bearing these facts in mind, the energetics of the general circulation may now be interpreted (in the quasi-geostrophic framework) as follows. Observations indicate that the directions of energy conversions are as shown in the "4-box" diagram of Fig. 8. Mean APE (available

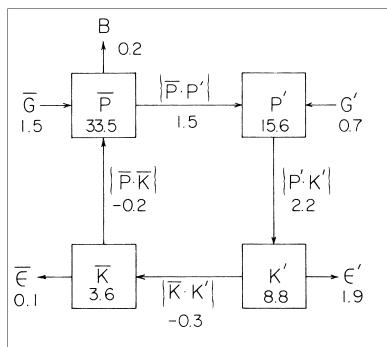


FIGURE 8 The energetics of the general circulation. Values are annual means for the northern hemisphere [taken from Oort, A. H., and Peixoto, J. P. (1974). *J. Geophys. Res.* **79**, 2705–2719]. Term B refers to a flux into the southern hemisphere. Values in the squares have units of 10^5 J/m^2 , while numbers next to the arrows are in units of watts per square meter.

potential energy) is generated if the mean diabatic heating is positive where the mean temperature is high, and negative where the mean temperature is low. This is actually the case because the radiation balance of the earth-atmosphere system has a surplus in the warm tropics and a deficit in the cool polar regions. Potential energy increases in association with the increase in the pole-to-equator temperature gradient. Eventually, baroclinically unstable eddies develop and mean APE is converted to eddy APE by the northward transport of warm air and the southward transport of cold air. At the same time, the vertical motions in the eddies, specifically the rising of warm air and the sinking of cold air, convert the eddy APE to eddy KE (kinetic energy). In this way, heat is transported both upward and poleward so that the observed energetic equilibrium is realized. The eddy KE is converted to mean KE when the trough and ridge patterns tilt from southwest to northeast. With this configuration, u' is large when $v' > 0$ and u' is small when $v' < 0$, so that in the average, zonal momentum is transported poleward by the eddies. This transport helps to maintain the westerlies against frictional loss. Finally, mean KE can be converted into mean APE if the mean motion is upward where it is cold and downward where

it is warm. This is the case in midlatitudes (although the opposite occurs in the tropics). Note also that both kinetic energies are dissipated through friction and that eddy APE may also be generated through diabatic effects.

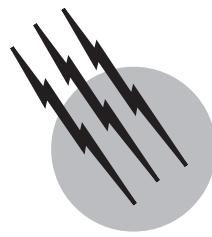
In conclusion, this energetics analysis of the general circulation has illustrated the fundamental manner in which the midlatitude atmosphere responds to the sun's radiative forcing. The fundamental conservative principles of energy, mass, and potential vorticity have been combined with the quasi-geostrophic approximation and the theoretical concept of baroclinic instability to yield an explanation of many observed atmospheric phenomena.

SEE ALSO THE FOLLOWING ARTICLES

ATMOSPHERIC DIFFUSION MODELING • ATMOSPHERIC TURBULENCE • CLIMATOLOGY • CLOUD PHYSICS • GRAVITATIONAL WAVE PHYSICS • METEOROLOGY, DYNAMIC (STRATOSPHERE) • PLANETARY WAVES (ATMOSPHERIC DYNAMICS) • TROPOSPHERIC CHEMISTRY • WEATHER PREDICTION, NUMERICAL

BIBLIOGRAPHY

- Dutton, J. A. (1976). "The Ceaseless Wind," McGraw-Hill, New York.
 Gill, A. E. (1982). "Atmosphere-Ocean Dynamics," Academic Press, New York.
 Haltiner, G. J., and Williams, R. T. (1980). "Numerical Prediction and Dynamic Meteorology," Wiley, New York.
 Holton, J. R. (1992). "An Introduction to Dynamic Meteorology," 3rd ed. Academic Press, New York.
 Jordan, C. L. (1958). Mean soundings for the West Indies area. *J. Meteorol.* **15**, 91–97.
 Lindzen, R. S. (1991). "Dynamics in Atmospheric Physics," Cambridge Univ. Press, Cambridge.
 Oort, A. H., and Peixoto, J. P. (1974). The annual cycle of the energetics of the atmosphere on a planetary scale. *J. Geophys. Res.* **79**, 2705–2719.
 Pedlosky, J. (1979). "Geophysical Fluid Dynamics," Springer-Verlag, New York.
 Wallace, J. M., and Hobbs, P. V. (1977). "Atmospheric Sciences: An Introductory Survey," Academic Press, New York.



Nitrogen Cycle, Atmospheric

Dan Jaffe

University of Washington-Bothell

- I. Key Atmospheric Nitrogen Compounds
- II. Role of Nitrogen in the Modern-Day Atmosphere

GLOSSARY

Acid precipitation Over most of the earth, rainwater is weakly acidic (pH of about 5.6) due to the presence of the weakly acidic compound, CO₂, in the earth's atmosphere. The term *acid precipitation* is generally reserved for rainwater that is more acidic than this due to the presence of other acids associated with anthropogenic emissions. The most common acidifying compounds are sulfuric and nitric acid from sulfur dioxide and nitrogen oxide emissions.

Ammonification Breaking down of organic nitrogen compounds into NH₃ or NH₄⁺.

Denitrification The process by which certain microbes utilize the NO₃⁻ ion as an oxidant and release gaseous N₂ or N₂O to the atmosphere.

Dry deposition Removal of a gas or aerosol compound by reaction, absorption, adsorption, or dissolution onto a wet or dry surface (e.g., a lake or grass).

Eutrophication Addition of excess nutrients, such as nitrates or phosphates, to a water body which causes excessive plant growth and eventually oxygen deficits.

Lifetime The time it takes for a compound to decrease to 37% of its original value, assuming a first-order loss process. Note that, for most species, the lifetime depends on local conditions, so that a global average lifetime will not reflect the true lifetime in all circumstances.

Mixing ratio Unitless ratio that describes the number of molecules of the target compound mixed into the total number of molecules of air; hence, 1 part per million (ppm) refers to 1 target molecule mixed into 1 million molecules of air. Alternatively, a mixing ratio could be defined using volumes, as a result of the ideal gas law.

Molecular density A unit of concentration given by molecules/cm³.

Nitrogen fixation The process by which N₂ in the atmosphere reacts to form *any* nitrogen compound. Biological nitrogen fixation is the enzyme-catalyzed reduction of N₂ to NH₃, NH₄⁺, or any organic nitrogen compound.

Photochemical smog Photochemical smog is the mixture of photochemically generated compounds, especially ozone, which forms in many urban areas from the action of sunlight on nitrogen oxides and hydrocarbons. Photochemical smog can contain ozone mixing ratios of several hundred parts per billion, a level known to cause significant impairments in lung function.

Wet deposition Removal of a chemical compound via precipitation.

NITROGEN GAS (N₂) is the most abundant compound in the earth's atmosphere (78.1% by volume). Uptake of N₂ by nitrogen-fixing organisms is the primary natural source of nitrogen for terrestrial and marine ecosystems; however, from a chemical kinetics point of view, reactions

of N₂ are too slow to be important to the chemistry of the bulk of the atmosphere. Instead, a number of trace atmospheric nitrogen compounds play key roles in regulating the oxidative and acid-base chemistry of the lower atmosphere (troposphere and stratosphere). Most of these nitrogen compounds have probably been present in the earth's atmosphere for millions of years; however, human activities, especially fossil fuel combustion, biomass burning, and agriculture, have significantly increased the concentration of these compounds in the earth's atmosphere. Collectively, these nitrogen compounds cause or contribute to a number of environmental problems, including urban photochemical smog, acid precipitation, watershed eutrophication, and climate change and may also contribute to stratospheric ozone depletion.

I. KEY ATMOSPHERIC NITROGEN COMPOUNDS

While several dozen trace nitrogen compounds have been found in the earth's atmosphere, in this section we will focus on four categories of nitrogen compounds and give an overview of their respective chemistry, sources, and sinks. In a later section, we present a quantitative budget of the sources and sinks for these compounds. The four categories of nitrogen compounds are

1. N₂ and N₂O
2. NH₃ and NH₄⁺ compounds
3. The NOx/NOy family (NO, NO₂, HNO₃, etc.)
4. Organic N and aerosol nitrogen compounds

A. N₂ and N₂O

Dinitrogen or simply nitrogen gas, N₂, is the most abundant gas in the atmosphere at 78.1% by volume. Nitrogen is a colorless gas at room temperature. While N₂ is generally quite inert chemically, it is utilized by nitrogen-fixing organisms and this is the primary natural nitrogen input to terrestrial and marine ecosystems. Biological nitrogen fixation refers to the natural process that is performed by a variety of bacteria and algae, both symbiotic and free living. Industrial nitrogen fixation refers to the industrial production of NH₃ and nitrates from N₂, mainly for fertilizers. In addition, nitrogen-fixing organisms are found on the roots of many leguminous plants (clover, soybeans, chickpeas, etc.) and have been used agriculturally as a means of replenishing soil nitrogen ("green manures").

The natural flux of N₂ from the atmosphere to oceans and terrestrial ecosystems is balanced by a nearly equal source of N₂ to the atmosphere in a process called denitrification. In denitrification, certain microbes utilize aqueous

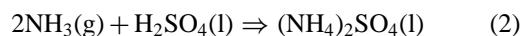
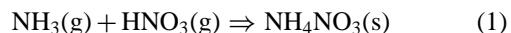
NO₃⁻ ion as an oxidant in low-oxygen environments. Thus, the combined impacts of natural biological nitrogen fixation and denitrification result in a nearly balanced cycle with respect to atmospheric N₂. Denitrification releases a smaller amount of N₂O, as well.

Nitrous oxide, or N₂O, is a colorless, fairly unreactive gas at room temperature. It is the second most abundant nitrogen compound in the atmosphere with a current (year 2000) average global mixing ratio of 316 parts per billion by volume (ppbv). Because of its low reactivity and low water solubility, N₂O has a lifetime of about 100 years in the atmosphere and is well mixed throughout the global troposphere. Having no significant sinks in the troposphere, N₂O eventually gets transported to the stratosphere where it can react or photolyze. The products of this decomposition play key roles in the chemistry of the stratosphere.

In denitrification, 80 to 100% of the nitrogen released is in the form of N₂, with the remainder being N₂O, although under certain environmental conditions nitrous oxide can become a major end product. While denitrification is a natural process, there is a significant increase in denitrification and N₂O release following the application of nitrogenous fertilizers, which is certainly a contributor to the global increase in N₂O concentrations being observed. N₂O is also produced from biomass burning. Because of these additional anthropogenic contributions, N₂O sources are now larger than the sinks and N₂O is accumulating in the atmosphere. Because N₂O is a greenhouse gas and plays a significant role in stratospheric ozone depletion, understanding the contributions from agricultural releases and biomass burning of N₂O and, if possible, reducing these fluxes are ongoing areas of investigation.

B. NH₃/NH₄⁺

Ammonia is a colorless gas with a very pungent odor and is toxic at high concentrations. It is a strong base, forms hydrogen bonds, is highly soluble in water, and is fairly reactive. Because of this reactivity and water solubility, atmospheric ammonia is short lived and has a very inhomogeneous distribution in the atmosphere: high concentrations near sources and extremely low concentrations in remote regions. In the atmosphere, ammonia is one of the few "reduced" compounds and also one of a small number of gaseous bases. For this reason, it will react with acids either in the gas or aqueous phase to produce an ammonium salt, such as:



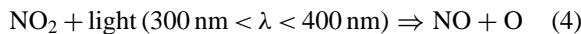
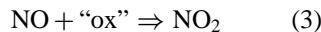
These reactions reduce the level of acidity, or increase the pH, in rainwater. Under most conditions, NH_4^+ ion (solid, liquid or in an aerosol) will be present in greater amounts than gaseous NH_3 . In addition, ammonia is produced in large amounts industrially, principally for use in fertilizers or explosives.

C. NO, NO_2 , and HNO_3

This group of compounds is characterized by relatively short atmospheric lifetimes (hours to days). They are best thought of as a group in that NO is the primary source and HNO_3 is the primary sink. These compounds are responsible for formation of photochemical smog and contribute to acid precipitation. Two subgroups within this broader class are defined. NO_x is defined as $\text{NO} + \text{NO}_2$ and NO_y refers to the sum of all reactive nitrogen oxides. NO_y specifically excludes less reactive compounds such as N_2O .

Nitric oxide (NO) is a colorless gas at room temperature. A small amount of NO is produced by microbial action in natural soils, but much larger amounts are produced during fossil fuel combustion. In combustion, NO is produced by the oxidation of nitrogen in the fuel and through the high temperature reaction of N_2 with O_2 . The kinetics and thermodynamics of this reaction are strongly temperature dependent, such that only at temperatures exceeding about 1000°C are significant quantities of NO produced.

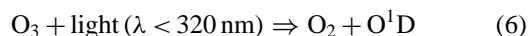
Nitrogen dioxide (NO_2) is a brown/yellow gas at room temperature due to its light absorption in the blue spectral region. Nitrogen dioxide has a very irritating odor and is fairly toxic. It is produced by the oxidation of NO, but NO_2 also photolyzes back to NO:



In reaction (3), "ox" refers to a number of known oxidants including O_3 , HO_2 , or RO_2 . As a result of reactions (3) and (4), a steady state between the NO and NO_2 is quickly established which depends on the amount of oxidants and available light. At night, NO essentially disappears away from direct sources, to reappear again the next morning. For this reason, sources or sinks of either compound represent a source or sink of both. The sum of the two compounds, NO plus NO_2 , is referred to as NO_x . The principal NO_x removal is the reaction:



and this results in a NO_x lifetime of about one day under typical conditions. The OH radical is an important atmospheric oxidant, which is produced from the photolysis of ozone and subsequent reaction with water vapor:



O^1D refers to an oxygen atom in an electronically excited state, in contrast to a ground-state atom. Since NO is formed from virtually all combustion processes, NO and NO_2 are found in much higher concentrations in urban areas compared with non-urban areas. In the presence of NO_x , hydrocarbons, and sunlight, ozone is produced and is a major contributor to urban photochemical smog in many large cities (see Section II.C.1).

Gaseous nitric acid, or HNO_3 , is highly water soluble and is a strong acid. As shown above, it is principally formed from the reaction of NO_2 with OH. Its principal sinks are wet deposition (contributing to acid rain) and dry deposition. Because it is the terminal step in the NO_x/NO_y cycle and because its removal contributes significantly to acidic precipitation, HNO_3 plays an important role in atmospheric chemistry. Nitric acid is also produced in large amounts as a feedstock to manufacture fertilizers and explosives.

A number of other NO_y compounds are involved in nitrogen oxide cycling, including NO_3 , N_2O_5 , HONO, and HNO_4 . For a description of the sources, sinks, and significance of these compounds, the reader should consult one of the more specialized references given in the bibliography.

D. Organic Nitrogen and Aerosol Nitrogen Compounds

A wide variety of organic nitrogen compounds have been identified in the atmosphere, including amines, amides, alkyl nitrates, alkyl nitrites, nitrosamines, nitroarenes, and peroxyacetyl nitrates. This last category includes one of the most well-studied organic nitrogen compounds, peroxyacetyl nitrate (PAN), which is an important lung and eye irritant formed in photochemical smog. Some of these compounds are directly emitted into the atmosphere from industrial emissions. This includes, for example, emissions of amines from sewage treatment or waste incineration and emissions of nitroarenes from fossil fuel combustion. Natural sources of amines are also common—for example, from decaying organisms.

Secondary products, formed from *in situ* atmospheric photochemical reactions, include the alkyl nitrates, peroxyacetyl nitrate, and many other compounds. In general, the contribution of these species is usually fairly modest (a few percents) compared to the inorganic species already mentioned. However, because some of these compounds have well-documented health implications (e.g., PAN, nitrosamines, nitroarenes) and because some of these can play roles in regulating the chemistry of ozone (e.g., PAN),

the importance of these compounds is greater than their modest abundance would suggest.

Aerosols play a key role in the atmospheric cycles of many elements. Quantitatively, a relatively small fraction of total atmospheric nitrogen is found in aerosol particles. This is because most nitrogen compounds are relatively volatile and therefore are found preferentially in the gas phase. On the other hand, aerosols can be important to the overall cycling in certain respects. Nitrogen species that are found in an aerosol form include the NO_3^- ion, as well as many of the higher molecular weight or polar organic nitrogen compounds mentioned above. Aerosol NO_3^- can form either by uptake of gaseous nitric acid on a pre-existing aerosol particle or by formation of NO_3^- directly on the surface of the particle.

Observations show that NO_3^- is found in greater amounts on larger aerosol particles, those with diameters greater than about 2 micrometers (μm). These larger particles are commonly composed of alkaline, crustal material, which will retain a volatile but acidic gas such as HNO_3 . Smaller aerosol particles (e.g., those with diameters less than 2 μm) are usually acidic and so tend to retain less HNO_3 . In urban areas, high concentrations of aerosol NO_3^- can be found on both large and fine aerosol

particles, presumably due to the high levels of precursor NO_x , N_2O_5 , and HNO_3 . In both urban and remote areas, gaseous HNO_3 concentrations are generally greater than aerosol NO_3^- concentrations by at least a factor of 3.

Other aerosol nitrogen compounds include high molecular weight polycyclic organic compounds, such as 1-nitropyrene which is a known animal carcinogen and a possible human carcinogen. This, along with many other polycyclic organic compounds, is emitted in significant quantities in fuel combustion, especially diesel exhaust. The polycyclic compounds are typically found associated with graphitic carbon particles or soot.

II. ROLE OF NITROGEN IN THE MODERN-DAY ATMOSPHERE

A. Natural and Anthropogenic Fluxes of Key Constituents

Figure 1 gives the sum of all sources and sinks for the key atmospheric nitrogen compounds, as well as the contribution from human activities. While it must be recognized that the values given in Fig. 1 have significant uncertainties (in some cases, as much as 50%), the broad picture of the

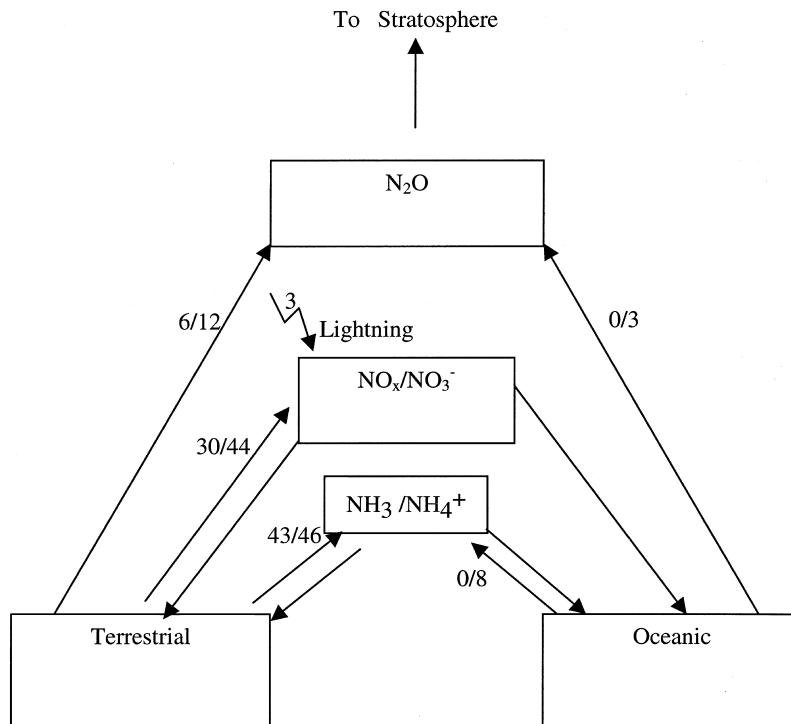


FIGURE 1 Sources and sinks for key atmospheric nitrogen compounds. The numbers give the flux in tgN per year. The second number gives the total for all sources (natural plus anthropogenic) and the first number shows only the anthropogenic flux. For the short-lived species, NO_x and NH_3 , the sources and sinks will be very nearly in balance. For N_2O , the sources are greater than the sinks; therefore, about 4 tgN is accumulating in the atmosphere each year.

atmospheric nitrogen cycle is fairly well understood. For a detailed discussion of how these numbers are obtained and the range of uncertainties of specific fluxes, the reader is referred to some of the specialized references given in the bibliography.

For most of the trace nitrogen compounds, the sources have increased significantly as a result of human activities. For some classes of compounds, such as NO_x and NH₃, humans are the dominant source. For N₂O, the human contribution is approximately 40% of the total. Nonetheless, because N₂O is long lived, the human sources result in an imbalance in the global source–sink relationship and an ongoing increase in the tropospheric burden.

With respect to nitrogen fixation, the anthropogenic contribution (fertilizer production and planting of legumes, clover, and other N-fixing plants) approximately equals the natural biological contribution. Production and use of industrial nitrogen fertilizers is a late 20th-century phenomenon. Industrially produced nitrogen fertilizers come in many forms, including NH₃, (NH₄)₂SO₄, NH₄NO₃, (NH₄)₃PO₄, and urea. Starting around 1950, when production was less than 5 tgN/year, it has increased to a current value of around 80 tgN/year (1 teragram N is 10¹² grams of nitrogen). This substantial enhancement to the global nitrogen cycle explains how we are able to feed the 6 billion people the planet now holds but is also responsible for large-scale changes in the distribution of nitrogen on the earth.

Natural sources of NO, principally microbial processes in soil and lightning, are relatively modest; however NO sources have increased dramatically as a result of human activities, principally fossil fuel combustion. As mentioned above, once released into the atmosphere NO will reach a steady state with NO₂ within minutes, so that sources of NO can be thought of as sources of NO_x. Globally, NO_x has a lifetime of ~1 day due to the reaction of NO₂ to form HNO₃. The NO_x lifetime is shorter in the atmospheric boundary layer and somewhat longer in the free troposphere. HNO₃ is then removed from the atmosphere by either wet or dry deposition. The HNO₃ removed by wet deposition is a significant contributor to acid precipitation in areas that are downwind of major NO_x sources. Thus, as a result of the large increase in NO_x emissions, there has been a substantial increase in NO₃⁻ deposition in many areas.

Volatilization of NH₃ from decaying organic matter and animal feces is a natural process; however human intervention has resulted in substantial changes in the natural cycle. While the actual fluxes are highly uncertain, due to the nature of the sources, it is clear that domestic animals are large NH₃ sources, greater than natural sources. In addition, biomass burning and fertilizer usage are also significant sources. Once released, NH₃ can be removed

by wet or dry deposition, but removal by wet deposition is quantitatively more significant. Because NH₃ is one of the few basic compounds in the atmosphere, it plays an important role in buffering the acidity of other acids (e.g., HNO₃ and H₂SO₄). Thus, it is quite common to find the compound (NH₄)₂SO₄ in atmospheric aerosols, which indicates that NH₃ has completely neutralized the strong acid H₂SO₄. In aqueous solutions or rainwater, NH₃ forms the NH₄⁺ ion, which will increase the pH, or reduce the acidity, of the solution.

B. Concentrations of Major Constituents in the Modern-Day Atmosphere

Table I shows the global average lifetime, global average mixing ratio or the range of observed mixing ratios, and an estimate of the global burden for some of the key nitrogen species. Long-lived compounds, such as N₂ and N₂O, have very uniform distributions throughout the globe. This makes it relatively easy to determine their total abundance and average global concentration based on a relatively small number of observations. However, short-lived species, such as NO_x and NH₃, have highly inhomogeneous distributions, which makes it much more difficult to determine their global abundance. For example, NO_x mixing ratios vary over 5 orders of magnitude between urban and remote regions. In addition, for both NO_x and NH₃, the lifetime is dependent on local conditions. For NO_x, the lifetime in urban air due to reaction with OH is about half a day, whereas in the upper troposphere it can be many days, depending on the OH molecular density. So for both NO_x and NH₃, the global burdens have high uncertainties, on the order of ±100%.

For N₂O there is good evidence that modern-day mixing ratios are higher than pre-industrial values by ~12–15%. This is based on direct atmospheric observations over the past two decades and ice core data extending back several thousand years. Figure 2 shows recent atmospheric data, collected at four background stations by the

TABLE I Mixing Ratios and Burdens of Atmospheric Nitrogen Compounds

Compound	Lifetime	Range of observed mixing ratios or global mean value	Global burden (tgN)
N ₂	Millions of years	78.08% (global mean)	3.9 × 10 ⁹
N ₂ O	~100 years	316 ppbv (global mean)	1600
NH ₃	~3 days	0.05–50 ppbv	~0.4
NO _x	~1 day	.005–200 ppbv	~0.2

Note: The N₂O mixing ratio is rising by 0.7 ppbv per year. The values given in the table are for the year 2000.

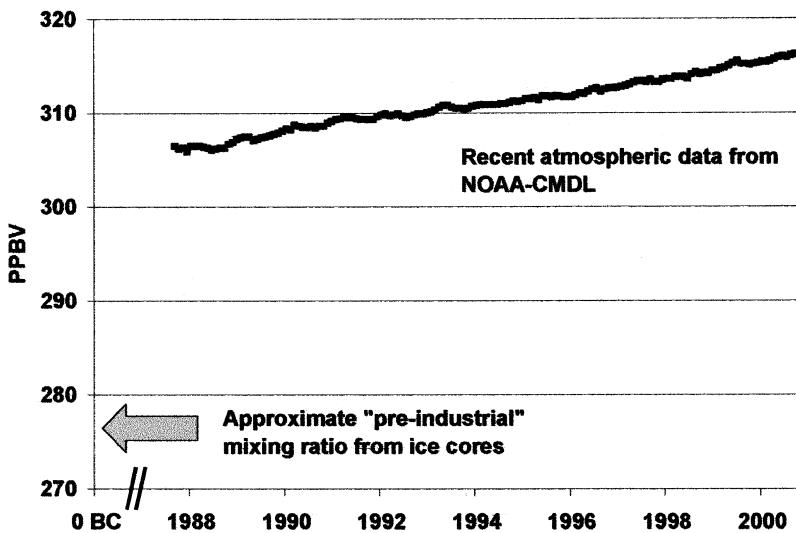


FIGURE 2 Global average N_2O mixing ratio (ppbv) from recent atmospheric data and estimated from ice core data. The NOAA-CMDL data are based on observations made at four background stations.

National Oceanographic and Atmospheric Administration's Climate Monitoring and Diagnostics Laboratory (NOAA-CMDL). Also shown is an estimate of the pre-industrial mixing ratio of N_2O , as determined by ice core data. For NH_3 and NO_x , the global burden has certainly increased significantly as a result of human sources; however, due to the highly variable nature of these reactive gases, it is difficult to quantify the exact amounts.

C. Significance and Consequences

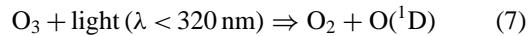
Atmospheric nitrogen cycling plays an important role in a number of key environmental processes, including atmospheric photochemistry and smog, acid rain, climate, stratospheric chemistry, and eutrophication.

1. Photochemistry and Smog

Photochemistry plays a significant role in the atmospheric cycling of nitrogen oxides and, conversely, nitrogen oxides significantly impact these photochemical processes. This is because photochemical processes—for example, reaction (4), above—cycle nitrogen oxides between the various species. Nitrogen oxides that can be photolyzed include NO_2 , NO_3 , HNO_3 , and ClONO_2 , as well as a number of other species. At the same time, the presence of nitrogen oxides at high concentrations dramatically changes the nature of the photochemical processes, resulting in an increase in the oxidizing power of the atmosphere through the formation of ozone.

One of the key atmospheric species involved in photochemical transformations is the OH radical. This gaseous

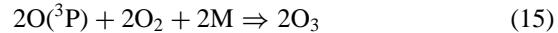
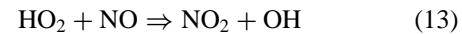
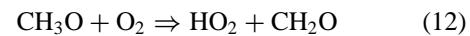
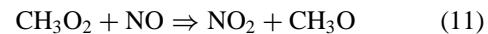
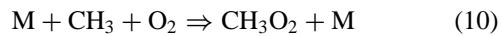
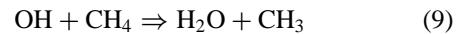
radical species is the primary oxidant in the atmosphere. The first step in its formation is the photolysis of O_3 :



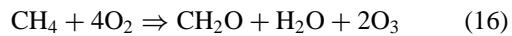
$\text{O}^{(1)\text{D}}$ is an electronically excited oxygen atom. It can decay back to a ground-state oxygen atom (O^3P , which will regenerate an ozone molecule), or else it can react with water to produce two OH radicals:



The OH radical is the primary oxidizer in the atmosphere, oxidizing CO to CO_2 and CH_4 and higher hydrocarbons to CH_2O , CO, and eventually CO_2 . It is also responsible for the conversion of NO_2 to HNO_3 via reaction (5). In the presence of NO_x , OH is responsible for the oxidation of CH_4 to form O_3 via the following sequence of reactions:



where $\text{M} = \text{N}_2$ or some other inert species. The net reaction is



In this sequence, NO_x and OH act as catalysts. Eventually, NO_x gets removed via reaction (5), which ends the cycle.

While there are many other possible reaction sequences, considering the large number of hydrocarbon reactants present in the atmosphere, the sequence in all cases is similar to that shown in reactions (9) through (16). These reactions are important in a cycle that oxidizes CO and hydrocarbons and produces ozone, in the presence of NOx ($\text{NO} + \text{NO}_2$). While there are many sources of hydrocarbons, both natural and anthropogenic, fossil fuel burning is the main source of NOx. Thus, in urban areas that receive ample sunshine, all of the ingredients for photochemical smog and ozone production are present. In photochemical smog, ozone can build up to unhealthy levels of several hundred parts per billion as a result of these reactions. Both hydrocarbons and nitrogen oxides are required for O_3 formation to occur; however, in most instances the nitrogen oxides limit the process.

While O_3 is not the only secondary product with health implications formed in photochemical smog, it is one of the key products and one that can be readily measured. For these reasons, O_3 mixing ratios are monitored in many urban areas around the world and most countries have established health standards. In the United States, the O_3 standard is set at 120 ppb for a 1-hour averaging period. There is evidence that a stricter standard is needed to protect public health. The U.S. Environmental Protection Agency has proposed a new standard of 80 ppb for an 8-hour averaging period, but at present this has not been implemented. Unfortunately, these standards are exceeded on a routine basis in many urban areas, principally regions with high solar availability, high traffic densities, and unfavorable meteorology or geography, which lead to poor dispersion conditions. Some of the worst regions of the world for photochemical smog include Los Angeles, Houston, Mexico City, and Beijing.

Due to the large number of hydrocarbons in the atmosphere, the number of possible reactions occurring in photochemical smog is in the thousands. Because of the complexity of the photochemistry, sources, and meteorology, computer models have been used as a primary tool to identify the best strategies to reduce photochemical smog. These models incorporate the most important reactions and their rate constants and sources, as well as meteorological and geographical data. Unfortunately, the complexity makes developing strategies a very challenging task. Past strategies for dealing with photochemical smog have focused mainly on reducing hydrocarbon emissions; however, this approach has met with very limited success. It now seems clear that reductions in both nitrogen oxides and hydrocarbon emissions are required to reduce urban photochemical smog. In Los Angeles, for example, substantial reductions in O_3 concentrations have been achieved over the past decade as a result of vigorous efforts to reduce both precursors.

Ozone also causes significant damage to vegetation. In some regions where intensive industrial emissions and agriculture coexist, there is the possibility for substantial impacts on food production due to ozone. This is because ozone damage to crops can occur at mixing ratios as low as 60 ppb. Also, the application of nitrogen fertilizers will increase local NO emissions due to denitrification. Thus, to some extent there is a “self-limiting” effect from adding additional fertilizer in that the increased NOx emissions will result in decreased crop yields due to ozone damage.

To summarize this process:

1. The oxidation of CO, CH_4 , and all other hydrocarbons is primarily driven by sunlight via the OH radical.
2. In the presence of sufficient NOx (roughly 30 parts per trillion), this oxidation produces ozone.
3. As a result of the rapid reaction between NO and O_3 (reaction (3), above), O_3 mixing ratios are often suppressed in urban centers and reach their highest values in suburban areas downwind of the main source region.
4. Most NOx gets removed as HNO_3 . The NOx lifetime is on the order of half a day in urban areas, somewhat longer in the free troposphere.

2. Acid Rain

Acid precipitation, or acid rain, can cause significant impacts on freshwater, coastal, and forested ecosystems. Both NO_3^- (from NOx emissions) and SO_4^{2-} (from SO_2 emissions) contribute significantly to acid rain. The relative ratio of $\text{SO}_4^{2-}/\text{NO}_3^-$ in precipitation will be substantially determined by the regional emissions of SO_2 and NOx. In regions that get most of their energy from coal and other high-sulfur fuels, there will be significant emissions of SO_2 unless scrubber technology is employed. Due to declining emissions of SO_2 in developed regions of the world and increasing NOx emissions, from automobiles, the relative contribution of NO_3^- is changing, with NO_3^- contributing an increasing fraction of the acidity in acid rain. In ice cores collected in remote regions of the northern hemisphere, SO_4^{2-} and NO_3^- concentrations have increased significantly in the past 50 years, reflecting the large increase in source emissions due to anthropogenic sources.

3. N_2O , Tropospheric O_3 , and N Fertilization

Both N_2O and O_3 are radiatively active in the infrared; thus, they contribute to the overall planetary warming caused by the “greenhouse effect.” Increases in the global concentrations of these two gases will contribute to an increase in radiative forcing, or heating, of the earth-atmosphere system. Because of its long lifetime, the global

distribution of N₂O and its steadily increasing concentrations are well documented and reasonably well understood (see Fig. 2), with humans contributing about 6 tgN of the 15 tgN total flux per year.

The global budget of ozone, on the other hand, is not nearly as well understood, due to the fact that it is a secondary product formed from a complex sequence of reactions. Also, because ozone's lifetime is much shorter than N₂O, on the order of weeks to a few months, it has a global distribution that is much less uniform than N₂O. There is good evidence that surface ozone throughout the northern hemisphere (not just in urban areas) has increased significantly as a result of the widespread use of fossil fuels. However, radiative forcing in the infrared is most important in the colder regions of the atmosphere, between about 10 and 15 km high. This is a region of the atmosphere that has highly variable O₃ mixing ratios; therefore, it is not certain how anthropogenic activities impact O₃ in this region.

Using global models it is possible to quantify the expected changes in radiative forcing due to changes in O₃; however, these calculated values have fairly large uncertainties associated with them. Based on the concentrations changes from pre-industrial times (1850) to 1992, CO₂, tropospheric O₃, and N₂O contribute 1.56, 0.4, and 0.14 watts/m², respectively, to global average radiative forcing. The uncertainty in the O₃ forcing is about 100%, whereas for CO₂ and N₂O it is much smaller. Based on the expected trends in energy consumption and agricultural emissions, all of these will likely continue to increase in the coming decades.

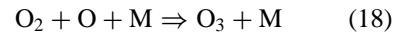
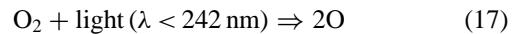
Nitrogen oxides also contribute to climate change in another, very different manner. This is a result of the fact that nitrogen is a limiting nutrient in many ecosystems. Additions of fixed nitrogen can significantly increase plant growth. This is known as nitrogen fertilization. In regions where anthropogenic nitrogen is deposited two results are possible: reduced plant growth due to acid precipitation or increased growth due to nitrogen fertilization. Typically, the initial nitrogen deposition will stimulate growth; however, at high N deposition a plant can become nitrogen saturated and no longer respond to additional nitrogen inputs. Nitrogen fertilization that causes increased plant growth results in increased uptake of atmospheric CO₂, increased global biomass, and a slowing in the growth of atmospheric CO₂, which then change the radiative forcing for CO₂, which as explained in the previous discussion is the largest contributor to the greenhouse effect.

One of the problems with understanding changes in the global carbon cycle has been the problem of missing sinks: The well-quantified anthropogenic sources of CO₂ are larger than the amount that can be accounted for in the

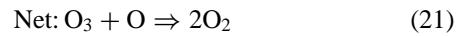
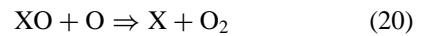
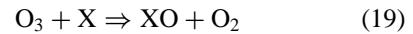
steadily increasing amount showing up in the atmosphere. The oceans are absorbing some of this missing CO₂, but it is unlikely that the oceans are adsorbing the entire missing sink. It now appears that increasing plant growth in the northern hemisphere, due to nitrogen fertilization, can account for the rest of this missing carbon, and this has the effect of slowing the buildup of CO₂. However, it is unclear how long this process can continue. Should the northern forests switch over from nitrogen fertilization to nitrogen saturation, or if other nutrients become limiting in these ecosystems, then the rate of rise of atmospheric CO₂ would increase. Overall, this uncertainty is an important limitation on our ability to predict future concentrations of atmospheric CO₂.

4. Stratospheric Chemistry

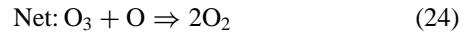
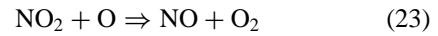
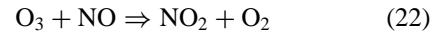
Nitrogen oxides also play a significant role in regulating the chemistry of the stratosphere. In the stratosphere, ozone is formed by:



This ozone production sequence is balanced by various catalytic destruction sequences:



where X can be any of the radical species NO, H, or Cl. For example, substituting NO for X yields:



The NOx reaction sequence is responsible for approximately half of the normal O₃ removal in the upper stratosphere, but much less in the lower stratosphere. Since this is a natural steady-state process, this is not the same as a long-term O₃ loss. The principal source of NO to the stratosphere is the slow upward diffusion of tropospheric N₂O and its subsequent reaction with O atoms:



Studies on stratospheric chemistry in the early 1970s suggested that direct emissions of nitrogen oxides from supersonic aircraft would result in a significant decline in stratospheric O₃. Partly as a result of this concern, only a small fleet of supersonic aircraft was built.

More recently, the widespread use of chlorofluorocarbons (CFCs) has resulted in significant perturbations to stratospheric O₃. Data from Antarctica in the 1980s revealed the existence of an Antarctic “ozone hole” each spring. Aircraft observations made within the Antarctic vortex determined that the O₃ depletion was occurring as a result of chlorine reactions due to fragmentation of CFCs. This process is accelerated in the extremely cold vortex that forms over Antarctica each winter. Further studies confirmed that the nitrogen oxides, which would normally remove Cl fragments, are tied up in ice particles in polar stratospheric clouds, which form only at very low temperatures. Once the nitrogen oxides are “frozen out,” the Cl fragments can go on to catalytically destroy ozone. Although the chemistry is complex, it is clear that N₂O and NO_x in the stratosphere play a critical role in the chemistry of ozone.

What is difficult to predict is how stratospheric chemistry will change as a result of continued increases in the concentration of atmospheric N₂O at a time when CFCs and total atmospheric Cl concentrations are declining due to the Montreal Protocols, an international agreement to ban CFC use. An additional area of concern with respect to stratospheric ozone is possible direct emissions of NO_x into the stratosphere by high-flying supersonic aircraft, which are once again being proposed. However, despite a substantial research effort to understand stratospheric chemistry, the question is complicated by the changing levels of stratospheric chlorine due to, first, the accumulation of tropospheric CFCs, followed by a rapid decline in CFC emissions as a result of the Montreal Protocol. To quote from the executive summary of the 1994 WMO/UN scientific assessment of ozone depletion ([WMO, 1995](#)):

Atmospheric effects of supersonic aircraft depend on the number of aircraft, the altitude of operation, the exhaust emissions, and the background chlorine and aerosol loadings. Projected fleets of supersonic transports would lead to significant changes in trace-species concentrations, especially in the North-Atlantic flight corridor. Two-dimensional model calculations of the impact of a projected fleet (500 aircraft, each emitting 15 grams of NO_x per kilogram of fuel burned at Mach 2.4) in a stratosphere with a chlorine loading of 3.7 ppb, imply additional (i.e., beyond those from halocarbon losses) annual-average ozone column decreases of 0.3–1.8% for the Northern hemisphere. There are, however, important uncertainties in these model results, especially in the stratosphere below 25 km. The same models fail to reproduce the observed ozone trends in the stratosphere below 25 km between 1980 and 1990. Thus, these models may not be properly including mechanisms that are important in this crucial altitude range.

5. Eutrophication

Eutrophication occurs when NO₃⁻ (or other nutrients such as PO₄³⁻) accumulates in lakes, ponds, or estuaries from agricultural runoff, sewage, detergents, or NO₃⁻ deposited with rainwater. These nutrients will accelerate plant growth, often leading to algal blooms, oxygen depletion, and sometimes fish die-offs. Eutrophication can also lead to substantial impacts on the overall aquatic ecosystem. Impacts due to eutrophication are well known in the Baltic Sea, Black Sea, Chesapeake Bay, and other regions.

D. The Future?

As the global population and energy use continue to rise, fertilizer use and fossil fuel combustion will certainly continue to grow. This is especially true in developing regions of the world as they attempt to achieve standards of living that are similar to those in the more developed and affluent regions. One researcher has estimated that global fertilizer use will increase by 70% and fossil fuel emissions of NO_x will increase by 115% by the year 2020. While the exact rate of change in nitrogen emissions is not certain, there is good reason to believe that emissions of many of these compounds will continue to increase in the coming decades. Thus, it seems likely that changes in the global nitrogen cycle will continue into the 21st century.

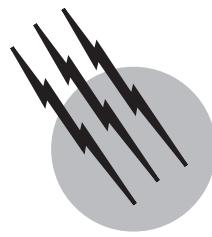
SEE ALSO THE FOLLOWING ARTICLES

BIOINORGANIC CHEMISTRY • CARBON CYCLE • CLIMATOLOGY • GREENHOUSE EFFECT AND CLIMATE DATA • NITROGEN CYCLE, BIOLOGICAL • OZONE MEASUREMENTS AND TRENDS • PLANETARY ATMOSPHERES • TRANSPORT AND FATE OF CHEMICALS IN THE ENVIRONMENT • TROPOSPHERIC CHEMISTRY

BIBLIOGRAPHY

- Brasseur, G. P., Orlando, J. J., and Tyndall, G. S. (1999). “Atmospheric Chemistry and Global Change,” Oxford University Press, New York.
- Crutzen, P. J., Lawrence, M. G., and Poschl, U. (1999). “On the background photochemistry of tropospheric ozone,” *Tellus* **51**, 123–146.
- Finlayson-Pitts, B. J., and Pitts J. N. Jr. (2000). “Chemistry of the Upper and Lower Atmosphere,” Academic Press, New York.
- Galloway, J. N. (1999). “Nitrogen fixation: anthropogenic enhancement-environmental response,” *Global Biogeochem. Cycles* **9**, 235–252.
- Holland, E. A., Dentener, F. J., Braswell B. H., and Sulzman, J. M. (1999). “Contemporary and pre-industrial global reactive nitrogen budgets,” *Biogeochemistry* **46**, 7–43.

- Horowitz, L. W., and Jacob, D. J. (1999). "Global impacts of fossil fuel combustion on atmospheric NO_x," *J. Geophys. Res.* **104**, 823–840.
- Houghton, J. T., Meira Filho, L. G., Callander, B. A., Harris, N., Kattenberg, A., and Maskell, K., eds. (1995). "Climate Change 1995: The Science of Climate Change (IPCC-1995)," Cambridge University Press, Cambridge, U.K.
- Jaffe, D. A. (2000). "The nitrogen cycle." In "Earth System Science" (M. Jacobson, R. J. Charlson, H. Rodhe, and G. H. Orians, eds.), pp. 322–342, Academic Press, San Diego, CA.
- Lents, J. M., and Kelly, W. J. (1993). "Clearing the air in Los Angeles," *Sci. Amer. Oct.*, 32–39.
- Matson, P. A., Naylor, R., and Ortiz-Monasterio, I. (1998). "Integration of environmental, agronomic and economic aspects of fertilizer management," *Science* **280**, 112–114.
- Smil, V. (1997). "Global population and the nitrogen cycle," *Sci. Amer. July*, 76–81.
- WMO (1995). "Scientific Assessment of Ozone Depletion: 1994," World Meteorological Organization Global Ozone Research and Monitoring Project, Report No. 37, Geneva.



Nitrogen Cycle, Biological

Elisabeth A. Holland

Max-Planck-Institut für Biogeochemistry and National Center for Atmospheric Research

Antje M. Weitz

Max-Planck-Institut für Biogeochemistry

- I. Introduction to the Nitrogen Cycle
- II. Reservoirs and Fluxes of the Biological Nitrogen Cycle
- III. Nitrogen Transformation Processes
- IV. Links between the Nitrogen and Carbon Cycles

GLOSSARY

Ammonification Mineralization of organic compounds by microbes, formation of ammonium (NH_4^+).

Biogeochemical processes Chemical and physical reaction processes involving abiotic and biotic components of ecosystems.

Denitrification The reduction of oxidized nitrogen compounds (nitrate NO_3^- , nitrite NO_2^-) by denitrifying (anaerobic) bacteria. Nitrous oxide is an intermediate gaseous denitrification product. The final end product is molecular nitrogen (N_2). Gases are either consumed by organisms or emitted to the atmosphere.

Mineralization The conversion of organic nitrogen to its mineral forms (NH_4^+ or NO_3^-) during decomposition. Gross N mineralization is the gross rate of the process, which can be determined using ^{15}N dilution. Net N mineralization is gross N mineralization minus N immobilization or uptake by microbes.

Nitrification The oxidation of reduced nitrogen compounds (ammonia NH_3 , NH_4^+) to NO_2^- and NO_3^- by nitrifying organisms. Nitrate is a plant available nitro-

gen form. By-products of nitrification are gaseous NO and N_2O . Gases are either consumed by organisms or emit to the atmosphere.

Nitrogen fixation (N-fixation) Conversion of nonreactive, atmospheric molecular nitrogen (N_2) into biochemically available forms, NH_4^+ and amino acids.

I. INTRODUCTION TO THE NITROGEN CYCLE

The *nitrogen cycle* refers to the internal cycling of N within an ecosystem, as well as the transformation of nitrogen at the planetary scale. The ecosystem may be terrestrial or hydrological (lakes, rivers, coastal, or open oceans). The biological processes which define the internal biological nitrogen cycle include the conversion of stable N_2 in the atmosphere to organic nitrogen, mineralization of nitrogen to ammonium, oxidation to nitrate, plant uptake of N, herbivore consumption of plants, volatilization of N through trace gas formation, and reduction of nitrogen back to gaseous N_2 (see bottom of Fig. 1). The *global nitrogen cycle* encompasses the N cycle of

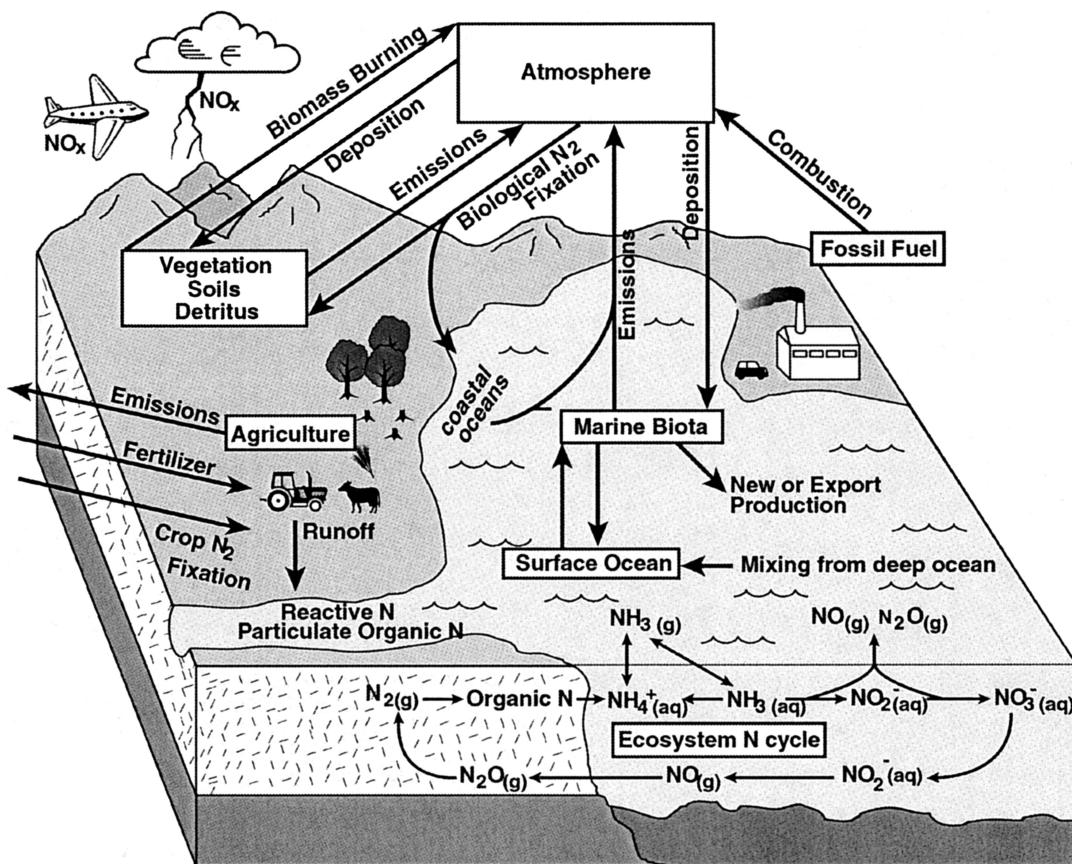


FIGURE 1 The global nitrogen cycle includes the internal transformations of nitrogen, mineralization, nitrification, denitrification, plant uptake, and N trace gas production which occur in soils and water. It also includes the inputs of nitrogen through nitrogen deposition, nitrogen fixation, and fertilizer application. The transformation of the modern global nitrogen cycle compared to the pre-industrial nitrogen cycle has been greater than for any other biogeochemical cycle. There are strong interactions with the carbon cycle, atmospheric chemistry, aerosol formation, and radiative forcing. The units most often used to describe the global nitrogen cycle are $\text{Tg} = 10^{12} \text{ g}$.

the earth and includes the transfer of mass among the earth's biogeochemical reservoirs: oceans, lakes, rivers, groundwater, the atmosphere, terrestrial biosphere, and geosphere (Fig. 1). Increases in the transfers of N among all of these compartments and the accompanying environmental changes have led to increasing concern about human alteration of the global N cycle for the following reasons:

- Pollution of streams, rivers, lakes, groundwater, and oceans with excess nitrogen
- Increased concentrations of nitrous oxide, a long-lived greenhouse gas involved in the destruction of the stratospheric ozone layer
- Increased concentrations of oxidized nitrogen that drive the formation of photochemical smog and tropospheric ozone in urban and rural areas
- Loss of base cations and soil nutrients such as calcium and potassium

- Acidification of soils, streams, and lakes through deposition of nitric acid
- Increased carbon uptake in terrestrial, aquatic, and marine ecosystems
- Accelerated losses of biological diversity
- Changes in the composition and functioning of estuarine and near-shore ecosystems that contribute to long-term declines in coastal marine fisheries

Both the global and internal nitrogen cycles are biologically regulated, but the transformations are the result of biotic and abiotic interactions. Electron transfers are at the center of the transformations which make the N cycle. Nitrogen loses and gains electrons during abiotic and biotic chemical transformation. The oxidation state of nitrogen can range from -3 to $+5$, an 8 electron difference. Knowing the oxidation state of nitrogen throughout its transformations helps to understand the role of the N-containing compounds in different environments.

Species	Oxidation state	Common name
NH ₃	-3	Ammonia
NH ₄ ⁺	-3	Ammonium
NH _x	-3	NH ₃ + NH ₄ ⁺
N ₂	0	Molecular nitrogen
N ₂ O	+1/2	Nitrous oxide
NO	+1	Nitric oxide
NO ₂ ⁻	+3	Nitrogen dioxide
HNO ₃ ⁻	+4	Nitric acid
NO ₃ ⁻	+5	Nitrate

II. RESERVOIRS AND FLUXES OF THE BIOLOGICAL NITROGEN CYCLE

General cycle—The biological N cycle embraces reservoirs and fluxes of those nitrogen compounds in water, land, and atmosphere that are available to organisms. Since not all nitrogen is biologically available, only a part of the total N participates in the biological nitrogen cycle. Biologically available nitrogen is bound in living and dead organic matter, bound to mineral compounds, dissolved in solutions, or present as oxidized and reduced gaseous nitrogen species.

Reservoir: atmosphere—Globally the atmosphere represents the largest nitrogen reservoir (3.9×10^{21} g N), with molecular nitrogen (N₂) accounting for about 79% of the air. N₂ molecules are, for the most part, excluded from the biological nitrogen cycle because of the stable triple bond between the atoms. A relatively small amount of N₂ undergoes N-fixation processes and enters the biological nitrogen cycle.

Reservoir: biosphere—The biosphere represents the second largest nitrogen reservoir after the atmosphere. Estimated global nitrogen content is 4100 Tg in vegetation, is 5000 Tg in detritus or dead organic matter, and is 95,000 Tg in soils. Oceans contain an estimated 2000×10^2 Tg of dissolved organic nitrogen. The nitrogen and carbon cycles are closely linked since N is essential for the functioning and growth of organisms. The transformation of N₂ into ammonium and amino acids via nitrogen fixation was one of the key building blocks for the development of life on earth. This nitrogen could then be further transformed by other paths and microbes. Ammonification and nitrification are the primary source for ammonium (NH₄⁺) and nitrate (NO₃⁻), which are both biologically available nitrogen species. Denitrification, however, represents a biological N-transformation process turning biologically available N oxides into unavailable N₂, providing closure of the entire N cycle. Biologically available nitrogen is a limiting factor for the productivity of terrestrial and aquatic ecosystems.

Reservoir: pedosphere (soils and rock)—Biologically available soil nitrogen is mainly stored in organic matter, bound reversibly to mineral soil particles, or dissolved in soil solution. The reservoir of soil inorganic nitrogen is relatively small compared to the amount of N cycling rapidly through the biosphere. Turnover times, and thus quantity and quality of soil organic matter, are major controls for the productivity of terrestrial ecosystems. The quality of organic material relates to its availability for biological processes upon decay, thus its aptness to decomposition and mineralization. The C to N ratio of organic material is the key regulator for mineralization and immobilization of organic compounds. Mineralized nitrogen compounds may undergo biological nitrogen transformation processes (ammonification, nitrification, denitrification) completing the biological N cycle or returning to the atmosphere as N₂. Some nitrogen is stored in geologic material, mainly in rocks derived from ocean sediments, restricted nitrate deposits, and hydrothermally altered basaltic rocks.

Reservoir: hydrosphere—The nitrogen load of rivers and groundwater results from leaching from soils. Coastal oceans receive nitrogen from river discharge, about 40 Tg N per year (Galloway *et al.*, 1995). Fixed nitrogen may enter the aquatic food chain. Dead organic material may sink to the deep sea and the ocean floor, the only mechanisms for nitrogen storage in marine systems. N₂ produced in anaerobic water or sludge through denitrification may emit to the atmosphere.

Abiotic and biotic N-fixation—In the atmosphere lightning and fire may trigger abiotic N-fixation reactions through temporarily very high pressure and temperature. Lightning is estimated to fix about 13 Tg N (Price, Penner, and Prather, 1997) per year into biologically available nitrogen forms. Biological nitrogen fixation occurs naturally in soils and oceans through biotic chemical reactions processes mediated by free and symbiotic living microbes. Estimations for biological N-fixation in aquatic ecosystems range from 30 to >300 Tg N (Galloway *et al.*, 1995) per year. Potential nitrogen fixation by natural ecosystems is 100–290 Tg N (Cleveland *et al.*, 1999). Human activity through agricultural practice increased the biological N-fixation. Cultivation of legumes enhanced N-fixation in terrestrial agriculture; in aquatic cultures (e.g., flooded rice fields) N availability increased through cultivation of N-fixing algae. Symbiotic N-fixation by crops is estimated to be 20 Tg N per year (Smil, 1999). Significant uncertainty surrounds all of these estimates.

Human-induced N-fixation—For about 60 years mankind has used industrial processes (Haber-Bosch) to form ammonia through the combination of hydrogen and atmospheric nitrogen under high pressure and temperature. Globally, industrial N-fixation through fertilizer

production amounts to about 80 Tg per year (Smil, 1999). Globally, nitrogen fixation by fossil fuel combustion is estimated to be more than 20–25 Tg N per year (Holland *et al.*, 1999).

Atmosphere–biosphere N-exchange—Direct exchange of N between the atmosphere and the biosphere occurs only through reactive nitrogen species [e.g., nitrogen dioxide (NO_2), ammonia (NH_3)]. The major exchange pathways are wet and dry deposition of atmospheric gaseous and particulate nitrogen compounds, diffusive gas exchange, and emission of biologically produced gaseous nitrogen species from soil and water surfaces.

Biosphere N-exchange to other reservoirs—Microbes are mediating biological nitrogen conversion processes in soils and water. These biological processes are controlled by N availability, temperature, moisture, redox, and pH conditions. Microbes produce ammonia (NH_3) and ammonium through ammonification of organic compounds. Volatile NH_3 may be emitted to the atmosphere, transformed and transported in the air, and deposited at any distance from the point of emission. NH_4^+ may become incorporated into organic compounds through ammonia assimilation and assimilatory nitrate reduction. Nitrifying bacteria oxidize reduced nitrogen (NH_3 , NH_4^+) to nitrite (NO_2^-) and nitrate, which is available for plant uptake. Dissolved in soil pore water, nitrate and organic nitrogen (DON) are spatially mobile through passive transport (leaching) with draining soil water into groundwater, streams, rivers, and coastal oceans.

Anthropogenic effect on N fluxes—Naturally, atmospheric concentration and fluxes of reactive nitrogen species are subject to considerable diurnal, seasonal, and spatial variability. Human activity, mainly through fossil fuel combustion, increases the atmospheric reactive nitrogen reservoir and alters the natural concentrations and fluxes at both temporal and spatial scales. Wet and dry deposition of atmospheric reactive nitrogen may enhance ecosystem productivity. Excessive deposition may result in acidification and nitrogen saturation of ecosystems, introducing a negative feedback reaction on ecosystem productivity. Saturation effects are, e.g., reported for forest systems in Europe and the northeastern United States. Human land use practice increased nitrogen loads in rivers, estuaries, and coastal oceans. Land use changes mobilize nitrogen and increase the potential for nitrogen loss. About 20 Tg N per year N loss in terrestrial ecosystems is attributed to forest conversion, and about 10 Tg N per year is attributed to wetland drainage.

Fertilization increased agricultural productivity (green revolution) through enhanced biomass production, though commonly only part of the added fertilizer N gets incorporated into the cultivated crop. Fertilizer is partly

consumed by soil microbes and partly stays in the soil available N-pool and may be transferred from the rooting zone to groundwater and rivers by leaching with draining soil water. At the community scale, plant species shifts may be induced by N additions, leading to species distribution changes which often favor weeds over native plants, which may reduce species diversity and later ecosystem function (Vitousek *et al.*, 1997). In oceans, external nutrient sources may reduce nitrogen fixation because of competition with other phytoplankton (Karl *et al.*, 1997). In more nutrient-rich coastal regions, nutrient eutrophication has been linked to the rise in harmful algal blooms, hypoxia, and shifts at higher trophic levels away from more traditional crustacean-dominated food webs that support many fisheries (Carpenter *et al.*, 1998).

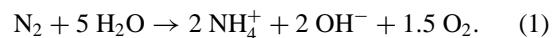
N fluxes to the atmosphere—By-products of nitrification processes are nitric oxide (NO) and nitrous oxide (N_2O) gas which may emit to the atmosphere. With progressing oxygen depletion in the anaerobic environment, denitrifying microbes transform NO_3^- into N_2O (incomplete denitrification) or N_2 (complete denitrification). Thus, in the anaerobic environment, complete denitrification removes nitrogen from the biological cycle. Organically bound nitrogen may be converted to N_2 by fire (pyrodenitrification). Biomass burning is estimated to return about 5–33 Tg N per year to the atmosphere by volatizing fixed nitrogen.

III. NITROGEN TRANSFORMATION PROCESSES

A. Fixation of Atmospheric N_2

Molecular nitrogen (N_2) is a relatively inert gas due to the strong triple bond between the two nitrogen atoms. The biological N cycle is largely restricted to molecules in which nitrogen is bound less strongly to elements other than nitrogen. Atmospheric N_2 may enter the biological nitrogen cycle through biotic or abiotic nitrogen fixation processes.

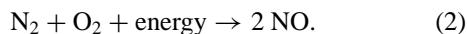
Microbial (biotic) fixation of atmospheric N_2 is the most important natural nitrogen source for the biosphere. Nitrogen-fixing microorganisms in terrestrial and oceanic ecosystems produce ammonia following the basic reaction:



Nitrogenase is the key enzyme for biotic nitrogen fixation, acting as the catalyst in the biochemical break up of the N_2 triple bond. Free-living autotroph bacteria in soils (e.g., *Acotobacter* in aerobic soils, *Clostridium* in anaerobic environments), blue-green algae (*Cyanobacteria*), and

bacteria living in a symbiosis with plants (e.g., *Rhizobium* with legumes) are able to fix atmospheric N₂ and incorporate N into bacterial biomass. Fixed nitrogen becomes available for other organisms after the death and decomposition of microbes. In symbiotic relations, the host plant may take advantage of the nitrogen fixed by the bacteria, while the microbes benefit from the nutrients assimilated by plant roots.

Abiotic fixation of molecular nitrogen requires a very high energy input to split the triple bond of N₂. In the presence of extremely high temperatures (e.g., during lightning strikes), N₂ molecules dissociate. A fraction of the dissociated nitrogen gets oxidized to form nitric oxide (NO) gas:



The primary mankind-induced abiotic nitrogen fixation processes are fossil fuel burning and fertilizer production (Haber-Bosch process). Industrial activity throughout the past 150 years affected the global biological nitrogen cycle noticeably. Human activity significantly altered reservoir sizes (e.g., atmospheric nitrogen oxide concentrations, soil nitrogen loads) and nitrogen fluxes (e.g., emission of nitrogen oxides from soils) of the global nitrogen cycle.

B. Atmospheric Deposition of Reactive Nitrogen Gases

What is reactive nitrogen deposition? The definition is most easily split into its wet and dry components (<http://nadp.sws.uiuc.edu/>; <http://www.epa.gov/castnet/>). Wet deposition of nitrogen includes deposition of ammonia, ammonium, nitrate, and nitric acid and a range of organic reduced and oxidized compounds in precipitation. The timing of nitrogen deposition events is driven by the episodic nature of precipitation and is likely important for coastal and open ocean ecosystems (Paerl, 1997). Dry deposition of nitrogen is the removal of N-containing gases by deposition onto a surface, e.g., leaves, soil, and open water. The N compounds which are exchanged via dry deposition include NH₃ (g), NO₂ (g), HNO₃ (g), particulate NH₄⁺ and NO₃⁻, and organic compounds including PAN, organic acids, nitrogen-containing aerosols, and simple amino acids. Much of our understanding of the spatial patterns and magnitude of N deposition is restricted to the well-sampled areas of the United States and Europe (<http://nadp.sws.uiuc.edu/>; <http://www.epa.gov/castnet/>; <http://www.emep.int/>). Both the measurements and the models are incomplete. The uncertainty associated with wet deposition measurements is much less than the substantial uncertainty associated with measurements/estimates of dry deposition.

C. Ammonia Assimilation and Plant Nitrogen Uptake

Autotroph organisms assimilate inorganic nitrate (NO₃⁻) ions into their body substances after conversion of NO₃⁻ into ammonium. The combined process of nitrate reduction and ammonia assimilation is referred to as the assimilatory nitrate reduction. Immobilization of nitrogen into organic N reduces the probability of nitrogen loss from the ecosystem.

Inorganic nitrogen (NO₂⁻ and NO₃⁻) molecules may enter the biological N cycle through plant uptake via roots or leaves. Plants assimilate inorganic nitrate dissolved in soil pore water or bound exchangeably to soil particles with the water sucked through the root tissue into the plant internal transport flow. Stomata dynamics control the transpiration flow and, thus indirectly, the root nutrient uptake. Agronomists use the NO₃⁻ concentration of the plant sap flow to evaluate the plant nitrogen supply. Stomata conductivity controls the uptake of gaseous nitrogen (NO₂) from the atmosphere into leaves through passive diffusive transport. Atmospheric NO₂ together with carbon dioxide (CO₂) diffuses through the stomata opening. Inside the leaf, NO₂ dissolves into the intercellular water of the stomata tissue and gets transformed to NO₂⁻ or NO₃⁻. In plant cells, inorganic nitrogen may be assimilated into the biological nitrogen cycle through direct incorporation into organic compounds or after reduction by the enzyme nitrate reductase. Isotope studies suggest that assimilated atmospheric nitrogen may be allocated in any growing part of the plant. Direct uptake of nitrogen deposited from the atmosphere onto above-ground plant surfaces (cuticula, bark) is of minor importance for the nitrogen supply of plants.

Until recently, it was thought that all N taken up by plants was taken up as a mineral form (NH₄⁺ and NO₃⁻) through their roots or as a gas through leaves and stomata (NH₃, NO, NO₂, or HNO₃). There is an accumulating body of evidence to suggest that plant roots are capable of taking up relatively simply amino acids directly, thus bypassing N mineralization (Näsholm *et al.*, 1999; Schimel and Chapin, 1996). This pathway is particularly important for boreal and tundra plants. Plant associations with mycorrhizal fungi may also play an important role in the nitrogen nutrition of plants through increasing surface area available for absorption and the production of proteases.

D. Mineralization of Nitrogen

Nitrogen is incorporated into organic substances or is available as inorganic NH₄⁺ and NO₃⁻ molecules. Decomposition and mineralization mobilizes nitrogen that is fixed in organic matter; the produced nitrogen species

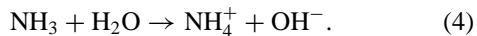
may complete the biological N cycle or return to the atmospheric N₂ pool. Microbial communities perform mineralization processes in both terrestrial and aquatic ecosystems.

Decomposition is primarily a biological process. In terrestrial ecosystems, it may be supported by physical processes, e.g., freezing breaks structural elements in dead organic matter. Organisms consume and transform organic material to gain energy and substance, with N mineralization and immobilization processes occurring simultaneously. Consequently, gross mineralization exceeds net mineralization due to immobilization of mineralized nitrogen into newly formed organic substances. The carbon to nitrogen (C:N) ratio of the decomposing organic matter largely determines transformation processes. Net immobilization occurs at C:N ratios >25, and net mineralization occurs at C:N ratios <25. High carbon content is related to structural, supportive elements in organic matter, which leads to higher resistance against decomposition and mineralization. Organic matter with a low C:N ratio decomposes easily. For example, microbial biomass with a C:N ratio of about 4–5 is rapidly mineralized given appropriate environmental conditions. Biological processes follow optimum functions for temperature and moisture and thus are subject to seasonal and general climatic influences, e.g., mineralization is slow in cold and dry environments and high in warm and moist climates. Net mineralization is the result of many interacting factors, mainly microbial biomass, C:N ratio, and the environmental conditions through mineralization processes.

A variety of soil microbial species are able to mineralize organic nitrogen through ammonification:



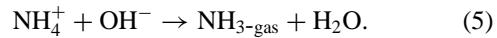
Ammonia dissolves rapidly in water, forming the easily bioavailable ammonium (NH₄⁺) molecules



Biologically available inorganic nitrogen (NH₄⁺, NO₃⁻) in soils is either dissolved in pore water or adsorbed reversibly to charged locations on surfaces of mineral particles. Commonly low anion exchange capacity of soils leaves NO₃⁻ molecules in solution. Leaching with draining soil water may deprive the soil nitrogen reservoir. Leaching loss is correlated to the seasonal dynamics of nitrogen mineralization, soil water content, soil properties, and agricultural management. Nitrate loss is generally smaller from heavier textured soils (loam and clay) with higher cation exchange capacity compared to light textured sandy soils. Natural ecosystems show low leaching losses since growing vegetation utilizes soil available nitrate. High leaching loss rates are known, especially from fertilized agricultural soils. Appropriate agricul-

tural management practice controls leaching loss from soils.

High concentration of dissolved NH₄⁺ and OH⁻ ions in recently fertilized or limed soils may cause the formation of gaseous ammonia (NH₃):

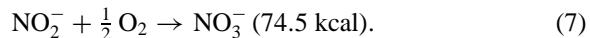
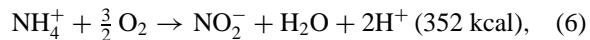


Ammonia molecules diffuse to the atmosphere (ammonia volatilization), get transported with wind, and, finally, deposit as NH₃ or NH₄⁺. Volatilized NH₃ is lost for the biological nitrogen cycle at the emission location, though considering larger scales it re-enters the biological N cycle after a time lag due to migration.

Ammonium molecules may be bound nonexchangeably into the mineral structure of silica, substituting similar sized potassium atoms. The amount of fixed nitrogen in the lithosphere is estimated to be about 50 times the amount of nitrogen in the atmosphere. However, this reservoir is not effective for the biological nitrogen cycle since nitrogen release by weathering is a negligently slow process.

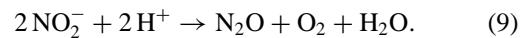
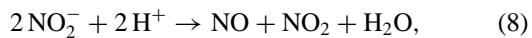
E. Nitrification and Denitrification

In aerobic soils, autotroph nitrifying bacteria gain energy for their metabolic processes by catalyzing the exotherm oxidation reaction of NH₄⁺ to NO₂⁻ (*Nitrosomas*) and of NO₂⁻ to NO₃⁻ (*Nitrobacter*):



The nitrification reactions change the nitrogen oxidation state from -3 in the reduced form (NH₄⁺) to +5 in the most oxidized form (NO₃⁻). Nitrifying bacteria are ubiquitous in soils. Some heterotrophic organisms are also capable of nitrification. Nitrification rates are controlled by nitrogen availability, temperature, moisture, pH, and redox state (oxygen availability). Temporal dynamics (diurnal and seasonal) of these variables affect nitrification. For example, in temperate zone soils nitrification rates are smaller during winter compared to summer (temperature is the major control), and in semi-arid areas rates slow down with start of the dry season (soil moisture is the major control). Nitrate availability is the major control in humid tropical soils.

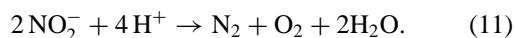
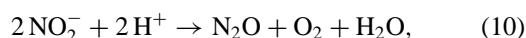
A fraction of the nitrate entering the nitrification process gets oxidized to gaseous forms (NO and N₂O):



Nitrogen availability, temperature, and soil moisture are the key controls for gaseous loss. Oxygen availability in soils controls the partitioning between the nitrification

by-products NO, NO₂, and N₂O. In the aerobic environment microbes predominantly produce NO. With increasing soil moisture content, the oxygen availability decreases and gaseous nitrogen production shifts towards N₂O. Increasing water content affects soil trace gas fluxes, because gas molecules diffuse through the air-filled pore space in soils. In structured, aggregated soils, aerobic and anaerobic microspots may occur in direct vicinity, resulting in simultaneous production of NO and N₂O gas.

In the aerobic soil environment facultative anaerobic bacteria (e.g., *Pseudomonas*) gain energy through heterotrophic respiration. However, in wet, oxygen limited soils, the same bacteria perform denitrification reactions to gain energy, using oxygen bound in NO₃⁻ and NO₂⁻ as electron acceptors:



A major control for the occurrence of either reaction is the oxygen availability or redox condition in soils. Nitrogen reduction increases with increasing oxygen limitation. Thus, with decreasing aeration, denitrification shifts more and more towards N₂ as the final product. Globally, denitrification returns 54–115 Tg N to the atmosphere as NO + N₂O + N₂.

Nitrogen emitted from soils in gaseous form is largely lost to the biosphere. Fertilizer nitrogen lost from soils in gaseous form (NO and N₂O) reduces the fertilizer efficiency. Effects of agricultural management and fertilization on soil trace gas fluxes also became important research questions in the context of global atmospheric warming and pollution.

IV. LINKS BETWEEN THE NITROGEN AND CARBON CYCLES

Biota and biogeochemical processes closely link the nitrogen and the carbon cycles. The ecosystem losses of nitrogen are determined by the coupling of N release and uptake. Because biospheric carbon uptake is N limited, the addition of nitrogen through nitrogen deposition (or fertilization) can stimulate carbon uptake and storage. The deposited nitrogen, which is immobilized by microbes to form soil organic matter, is stored at a C:N ratio close to that of microbial biomass, between 4 and 15. Nitrogen assimilated by plants is stored at a much wider C:N ratio between 30 and 90 for leaves and a C:N ratio of 150–300 for wood (Townsend *et al.*, 1996). Microbial N has a residence time of less than a year compared to the nitrogen in wood, which can have a residence time of decades to centuries. Decomposition of organic material which starts at a C:N ratio of ~25 requires immobilization of N by mi-

crobes. The quantity of nitrogen required depends on the microbial growth efficiency. Microbial N immobilization is the mechanism by which microbes compete with plants for available nitrogen. Below a C:N ratio of ~25, there is net nitrogen mineralization, releasing nitrogen into the “bioavailable” nitrogen pool for plant or further microbial uptake. Gross nitrogen mineralization rates in soils are controlled by the physical structure and chemical quality of decomposing organic material, soil temperature, moisture, pH, and redox conditions. As with most biological reactions, there is a temperature optimum for mineralization. Soil moisture also regulates mineralization. In dry soils mineralization increases with increasing moisture content. Above field capacity, reduced oxygen availability in increasingly wet soils slows mineralization rates down.

In marine ecosystems, bioavailable nitrogen is one of the key limiting nutrients in the upper ocean along with phosphorus, silicon, and iron. As with terrestrial ecosystems, atmospheric nitrogen deposition serves as a “new” nutrient source to the ecosystem, and increased inputs typically lead to enhanced primary production and net community export of organic matter either through sinking particles or through advection of dissolved organic matter. There are parallels between the carbon and nitrogen links in terrestrial and marine ecosystems. The vertical partition of carbon in the ocean and net air-sea CO₂ flux are, in turn, strongly modulated by the export flux, highlighting the close linkage of the marine nitrogen and carbon cycles. However, the C:N ratio is operated in a much more restricted range, much closer to that of microbial biomass. Atmospheric N deposition decouples the macronutrient input from ocean physics (which also tends to supply high metabolic CO₂ water to the surface layer) and therefore can alter marine biogeochemistry disproportionately.

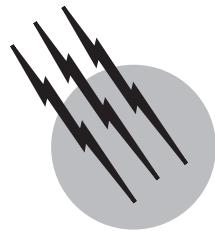
SEE ALSO THE FOLLOWING ARTICLES

BIOINORGANIC CHEMISTRY • CARBON CYCLE • GREENHOUSE EFFECT AND CLIMATE DATA • GREENHOUSE WARMING RESEARCH • NITROGEN CYCLE, ATMOSPHERIC • OCEAN-ATMOSPHERIC EXCHANGE • POLLUTION, ENVIRONMENTAL • SOIL AND GROUNDWATER POLLUTION • TRANSPORT AND FATE OF CHEMICALS IN THE ENVIRONMENT

BIBLIOGRAPHY

- Aber, J., McDowell, W., Nadelhoffer, K., Bernston, M. A. G., Kamakea, M., McNulty, S., Currie, W., Rustad, L., and Fernandez, I. (1998). “Nitrogen saturation in temperate forest ecosystems,” *Bioscience* **11**, 921–934.
 Carpenter, S. R., Caraco, N. R., Correll, D. L., Howarth, R. W., Sharpley, A. N., and Smith, V. H. (1998). “Nonpoint pollution of surface waters with phosphorus and nitrogen,” *Ecol. Appl.* **8**, 559–568.

- Cleveland, C. C., Townsend, A. R., Schimel, D. S., Fisher, H., Howarth, R. W., Hedin, L. O., Perakis, S. S., Latty, E. F., Von Fischer, J. C., Elseroad, A., and Wasson, M. F. (1999). "Global patterns of terrestrial biological nitrogen (N₂) fixation in natural ecosystems," *Global Biogeochem. Cycles* **13**, 623–645.
- Galloway, J. N., Schlesinger, W. H., Levy, H., II, Michaels, A., and Schnoor, J. L. (1995). "Nitrogen fixation: Anthropogenic enhancement-environmental response," *Global Biogeochem. Cycles* **9**, 235–252.
- Holland, E. A., Dentener, F. J., Braswell, B. H., and Sulzman, J. M. (1999). "Contemporary and pre-industrial global reactive nitrogen budgets," *Biogeochemistry* **46**, 7–43.
- Howarth, R. (1996). "Nitrogen Cycling in the North Atlantic Ocean and Its Watersheds" (R. Howarth, ed.), Kluwer Academic, Boston.
- Karl, D. M., Letelier, R., Tupas, L., Dore, J., Christian, J., and Hebel, D. (1997). "The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean," *Nature* **388**, 533–538.
- Näsholm, T., Ekblad, A., Nordin, A., Giesler, R. S., Höglberg, M., and Höglberg, P. (1999). "Boreal forest plants take up organic nitrogen," *Nature* **392**, 914–916.
- Paerl, J. W. (1997). "Coastal eutrophication and harmful algal blooms: Importance of atmospheric deposition and groundwater as "new" nitrogen and other nutrient sources," *Limnol. Oceanogr.* **42**,
- Price, C. J., Penner, J., and Prather, M. (1997a). "NO_x from lightning 1. Global distribution based on lightning physics," *J. Geophys. Res.* **102**, 5929–5941.
- Schimel, J. S., and Chapin, F. S., III (1996). "Tundra plant uptake of amino acid and NH₄⁺ nitrogen in situ: Plants compete well for amino acid N," *Ecology* **77**, 2142–2147.
- Smil, V. (1999). "Nitrogen in crop production: An account of global flows," *Global Biogeochem. Cycles* **13**, 647–662.
- Townsend, A. R. (1999). New perspectives on nitrogen cycling in the temperate and tropical americas. In "International SCOPE Nitrogen Project," Kluwer Academic, Dordrecht/Norwell, MA.
- Townsend, A. R., Braswell, B. H., Holland, E. A., and Penner, J. E. (1996). "Spatial and temporal patterns in potential terrestrial carbon storage resulting from deposition of fossil fuel derived nitrogen," *Ecol. Appl.* **6**, 806–814.
- Vitousek, P. M., Aber, J. D., Howarth, R. W., Likens, G. E., Matson, P. A., Schindler, D. W., Schlesinger, W. H., and Tilman, D. G. (1997). Human alteration of the global nitrogen cycle: Sources and consequences," *Ecol. Appl.* **7**, 737–750.



Ozone Measurements and Trends (Troposphere)

Johannes Staehelin

Swiss Federal Institute of Technology

- I. Introduction
- II. Ozone Production in the Troposphere
- III. The Tropospheric Ozone Cycle and Ozone Distribution
- IV. Tropospheric Ozone Measurements
- V. Trend Determination
- VI. Surface Ozone Trends in North America (Including Mexico)
- VII. Surface Ozone Trends in Europe
- VIII. Surface Ozone Trends in Asia
- IX. Surface Ozone Trends in the Tropics and Southern Hemisphere
- X. Ozone Trends in the Free Troposphere
- XI. Summary and Conclusions

GLOSSARY

NO_x Nitrogen oxides including NO and NO₂.

NO_x sensitivity Chemical environment, in which the ozone production increases with the concentration of nitrogen oxides. NO_x sensitivity usually occurs at low NO_x concentrations and therefore at rural and remote sites.

Ozone precursors Chemical compounds which lead to the formation of ozone and other photooxidants in the

troposphere including nitrogen oxides, reactive organic gases, such as hydrocarbons (methane and others) and carbon monoxide.

Photo smog Air pollution caused by ozone and other oxidizing compounds, which are formed in ambient air from the ozone precursors under the influence of sunlight.

Planetary boundary layer Air layer close to the earth's surface which has a typical vertical extension of approximately 1 km.

Summer smog Same as photo smog.

Temperature inversion Increase of temperature with height leading to restricted vertical mixing of the pollutants emitted from ground.

Tropopause The tropopause separates the troposphere from the stratosphere. It is usually defined by the vertical temperature gradient. The annual mean value of the tropopause altitude at midlatitudes is approximately at 10 km amsl.

Tropospheric ozone Ozone in the troposphere, which extends from the earth's surface to the tropopause.

VOC Volatile Organic Compound, including volatile hydrocarbons (e.g., methane or toluene).

VOC sensitivity Chemical environment, in which ozone production increases with the concentrations of VOCs. VOC sensitivity usually occurs in polluted cities.

SCHÖNBEIN DISCOVERED ozone (O_3) in the year 1840 and he showed in 1845 that ozone is a natural trace component of the atmosphere surrounding our planet. Today we know that ozone concentrations are highest in the stratosphere (see Fig. 1) where ozone is formed from the photolysis of molecular oxygen (O_2). Ozone strongly absorbs the UV-light and protects the earth's surface from the detrimental part of the sunlight in the UV-B and UV-C region. Until World War II it was believed that ozone in the troposphere entirely originates from the mixing down from the stratosphere. In the 1940s, large ozone concentrations have been measured near the earth's surface in the Los Angeles basin in Southern California, and it has been discovered that ozone can be formed in the polluted planetary boundary layer from primary air pollutants including nitrogen oxides (NO and NO_2 : NO_x), Volatile

Organic Compounds (VOC) and carbon monoxide (CO) by photochemical conversions which are driven by sunlight in the UV-A and UV-B region. Besides ozone a large number of other oxidizing compounds such as peroxy-acetyl nitrates (PANs), carbonyls (aldehydes and ketones), and nitric acid (HNO_3) are also formed by these photochemical reactions. This type of air pollution is called photo smog or summer smog. In the 1970s and 1980s, it became more and more clear that photooxidants are a major air pollution problem surrounding many large cities in the industrialized world.

I. INTRODUCTION

Ozone is a strong oxidant and it can be detrimental to human health at elevated concentrations in the ambient air. Adverse health effects, such as impaired lung function, occur in the population when exposed for several hours at ozone concentrations of 120 ppb. Sensitive people show health effects at substantially lower concentrations. The guideline of the World Health Organization (WHO) for protection of human health effects of ozone in ambient air is approximately 60 ppb for 8 hr. It has been reported that the elevated concentration of ozone for over 70 million people living in the United States still exceeded the National Ambient Air Quality standard (80 ppb for 8 hr) in 1998 despite considerable regulatory efforts to reduce air pollution.

Ozone is a stress factor for plants and present ozone concentrations significantly diminish agricultural crop yields in many areas over the world. When plants are grown in an atmosphere with high ozone concentrations, visible signs of injury may appear within a few hours or days, or in absence of these signs of injury, leaf discoloration appears within a few hours. The response of plants on ozone stress strongly depends not only on the species but also on other stress factors. In open-top chambers, plants are grown in the field under the exposure of different ozone concentrations. Such studies are used to quantify agricultural crop yield losses caused by ozone. Based on ambient air concentrations in the years 1981–1983 it was estimated that annual agricultural production prices in the United States would have been higher by 1900 million dollars if ozone concentrations would have been lower by 25%, whereas a reduction in ozone concentrations of 40% would have resulted in a benefit of agricultural production of 3000 million dollars.

Figure 1 shows that ozone in the stratosphere at northern mid-latitudes has decreased during the last decades, which is mainly caused by ozone depletion of halogen containing volatile gases such as chlorofluorocarbons. During the same time ozone in the troposphere has increased due

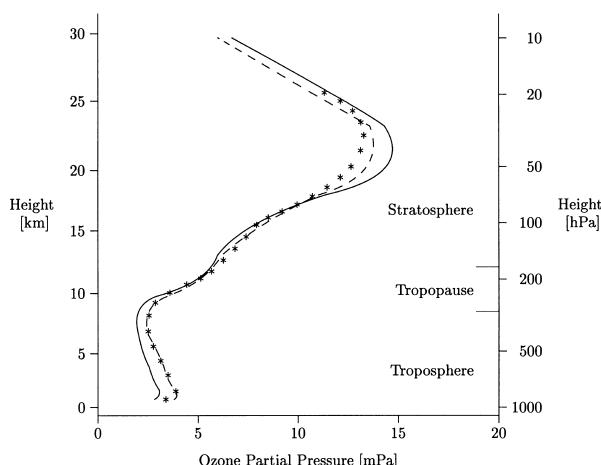


FIGURE 1 Averaged vertical ozone profiles of selected years measured over Payerne, Switzerland. The ozone concentrations are given in partial pressure (nanobar).

to tropospheric ozone formation. Anthropogenic ozone formation not only increases ozone concentrations downwind of urban areas but also affects ozone concentrations in the entire troposphere. Stratospheric ozone depletion increased the intensity of the UV-B light in the troposphere, which influences tropospheric ozone production. This effect strongly depends on nitrogen oxide concentration. Enhanced ozone production occurs if nitrogen concentrations are low.

The photolysis of ozone leads to OH radicals which strongly determine the oxidation capacity of the troposphere. The oxidation capacity regulates the atmospheric residence time of many gaseous species including greenhouse gases such as methane. Ozone strongly absorbs the infrared radiation and ozone is an important greenhouse gas like water vapor and carbon dioxide. Changes in ozone concentrations therefore contribute to the anthropogenic disturbances of the radiative balance of the earth which is usually quantified by the quantity of the radiative forcing which describes the driving force of changes in climate. Ozone has a rather unique property as its greenhouse gas strength depends on altitude. It has been shown that the greenhouse warming of ozone is strongest close to the tropopause at mid-latitudes. It has been estimated that the increase in tropospheric ozone contributed by approximately 25% of the amount of the carbon dioxide increase to the change in global mean radiative forcing since preindustrial time. Therefore, trends of ozone not only in the surface layer and in the stratosphere but also near the tropopause are important issues in atmospheric science.

Section II summarizes the chemical theory most important for the discussion of ozone in the troposphere, and typical tropospheric ozone concentrations are presented in Section III. Section IV contains a short description of the methods which are currently used to measure tropospheric ozone and problems and methods of tropospheric ozone trend determination are discussed in Section V. The present knowledge of trends of surface ozone in different parts of the world is summarized in Sections VI–IX and ozone trends of the free troposphere are summarized in Section X. Section XI includes the summary and some conclusions.

II. OZONE PRODUCTION IN THE TROPOSPHERE

Ozone is formed in the troposphere from its precursors by a complex series of chemical reactions which take place under the influence of solar light. Surface ozone concentrations often reach maximal concentrations between 30 and 70 km downwind of large emission sources depend-

ing on meteorological conditions such as wind speed and height of the mixing layer and the precursor concentration mix. **Table I** shows examples of maximal ozone concentrations measured in such urban air plumes. Very high-surface ozone concentrations were reported from many large cities all over the world in which road traffic is commonly used for transport. Maximal ozone concentrations in the urban plumes depend on meteorological conditions, the size of the city (compare, e.g., maximal ozone concentrations reported from different cities in the United States), and the air pollution abatement legislation which influences the strength of the anthropogenic precursor emissions.

To our knowledge 680 ppb was the largest surface ozone concentration measured by a reliable technique (see **Table I**). It was reported in the surroundings of Los Angeles in 1973. Also in the middle of the 1970s very large ozone concentrations, up to 310 ppb, were measured in the surroundings of Tokyo. It appears, that the largest ozone concentrations presently occur in the surroundings of Mexico city, where the highest concentrations have been measured in 1992. High ozone concentrations downwind of very large cities which have been named “Mega-cities” are a concern for human health. The problem could become particularly severe in the future in “Mega-cities” in third-world countries where the agglomerations are predicted

TABLE I Maximal Ozone Concentrations in Urban Plumes

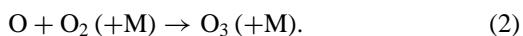
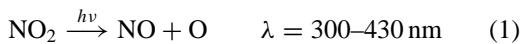
Site Continent	City	Maximal O ₃ concentration, ppb	Date
Northern America	Los Angeles	680	Summer 1973
		454	Oct. 13, 1978,
		330	1990
	New York	310	June 10, 1974
		260	Sept. 8, 1975
		189	Aug. 14, 1978
		140	Aug. 15, 1977
Central America	Mexico City	477	1992
Southern America	Sao Paulo ^a	<200	
Asia	Seoul	322	July 23, 1994
	Tokyo (Yokohama)	310	July 15, 1975
Europe	Milan	200	May 13, 1998
	Athens ^b	ca. 200	Sept. 9, 1994
	London	174	July 7, 1984
	Berlin ^b	150	July 26, 1994
	Vienna	139	August 8, 1986

^a Exceedance of WMO value (100 ppb hourly mean value) by more than a factor of 2.

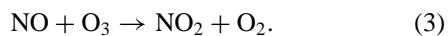
^b Measurements from airplanes in the planetary boundary layer.

to grow rapidly and the financial resources may not be available for appropriate air pollution abatement.

The precursors of tropospheric ozone formation include nitrogen oxides (NO_x), volatile organic gases (VOC) (e.g., toluene ($C_7\text{H}_8$), methane (CH_4)), and carbon monoxide (CO). Anthropogenic emissions of NO_x (mainly as NO) originate from high-temperature combustion of fossil fuels (such as automobile exhaust), whereas VOCs originate from fossil combustion and solvent evaporation. Biomass burning leads to significant anthropogenic emissions of ozone precursors particularly in the tropics. Here only the reactions most important for tropospheric ozone trends are briefly summarized. Photolysis of nitrogen dioxide is the source of oxygen atoms, which produce ozone in the troposphere ($h\nu$ describes the influence of (solar) light):



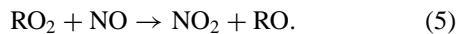
M is an unspecified molecule, which is not converted in the chemical reaction but required to remove the excess energy of the reaction. NO reacts with O_3 forming NO_2



Reactions (1) to (3) form an equilibrium which is called the photostationary state. It depends on the intensity of solar radiation. The photostationary state does not lead to a net ozone production. The ozone destruction by NO is important close to large emission sources such as busy streets, in particular in the nights (see Section III). O_x has been introduced to account for the influence of the reaction of NO with O_3 , because O_x is conserved by the “titration” of O_3 by NO (reaction (3)):



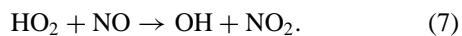
During the day NO is converted to NO_2 by photochemically produced peroxy radicals (RO_2 , where R means an organic entity) which leads to the net tropospheric ozone production by the subsequent reactions (1) and (2)



Oxiradicals (RO) react further to form hydroperoxy radicals (HO_2) and carbonyls (CARB):



HO_2 also oxidizes NO:

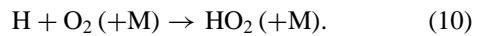
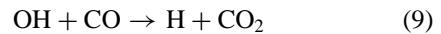


Organic peroxy radicals (see reaction (5)) originate from the oxidation of VOCs (Volatile Organic Compounds) with OH radicals. The first reaction of the sequence (reaction 8) includes either an addition of OH at a double bond of an aromatic hydrocarbon or an alkene or an

H-subtraction from a saturated C–H bond. These radicals quickly add molecular oxygen:



The reaction of carbon monoxide (CO) with OH radicals leads to the formation of hydroperoxy radicals:

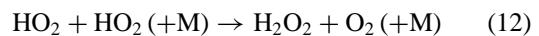


The main source of OH radicals in the troposphere is the photolysis of O_3 and carbonyls. The reaction sequences (8), (5), (6), and (7), respectively, (9), (10), and (7) form radical chains, in which the reactive OH radical is efficiently recycled. The different oxygen containing radicals RO_x (HO , HO_2 , RO_2 , and RO) are quickly converted because of their high reactivities. Tropospheric ozone is produced by this cycle because reactions (5) and (7) are included in this reaction sequence.

Termination reactions regulate the concentrations of the RO_x radicals. The main sink of RO_x and NO_x at high NO_x concentrations, such as in urban areas, is the reaction forming nitric acid:



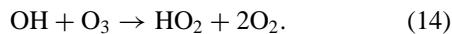
If NO_2 concentrations are low, the dominant termination reactions lead to the formation of hydrogen peroxide (H_2O_2) and organic peroxides (ROOH):



close to large emission sources ozone formation is usually limited by the concentrations of VOCs. The increase of nitrogen oxide concentrations enhances the rate of the radical termination by HNO_3 production (reaction (11)) leading to reduced local ozone production. Therefore, NO_x reduction without simultaneous reduction of VOCs leads to enhanced local ozone production. This regime is called VOC sensitive. Nitrogen oxides have a rather short lifetime in the polluted planetary boundary layer because reaction (11) is fast. The half-life of NO_x is in the order of a few hours under favorable photochemical conditions such as high solar irradiance in the summertime. Because of this rather fast removal of NO_x , O_3 formation is usually limited by the nitrogen oxide concentrations in rural sites more distant from the emissions sources. This regime is called NO_x sensitive. In rural sites VOCs such as isoprene and terpenes are emitted by biogenic sources (mainly trees). This further accentuates the need for NO_x control to reduce ozone formation in rural areas.

The concentrations of the ozone precursors are usually much smaller in the free troposphere than in the polluted

planetary boundary layer. Important sources of NO_x in the free and upper troposphere include (besides the transport from the planetary boundary layer) production in lightning strokes, mixing down from the stratosphere (where it is formed by photolysis of N₂O) and air traffic emissions. In the free troposphere, carbon monoxide and methane usually dominate the conversion of OH into HO₂ radicals. At very low NO_x concentrations (NO below 30 ppt) peroxy radicals (HO₂) no longer predominantly react with NO, to form NO₂, (reaction (7)) but with O₃:



This leads to a reaction chain in which O₃ is destroyed and not formed as under usual conditions in the troposphere. This type of chemistry is important in very clean environments such as in the unpolluted marine boundary layer in the southern hemisphere. Ozone also strongly regulates the oxidant capacity of the troposphere, because photolysis of O₃ is the most dominant tropospheric source of OH radicals. These radicals are most important for the chemical degradation of the gaseous reduced compounds which include important greenhouse gases such as, e.g., methane.

III. THE TROPOSPHERIC OZONE CYCLE AND OZONE DISTRIBUTION

The tropospheric ozone budgets include four terms. Their magnitudes (given in unit of mass (teragrams (tg)) per year) are roughly known from numerical simulations:

- The photochemical ozone production (from anthropogenic and natural precursors) of the entire troposphere (the respective processes are described in the last section) is believed to amount to approximately 3500–4000 tg O₃/yr.
- Ozone is also destroyed by the tropospheric photochemistry (see last section) leading to an annual ozone loss of approximately 3000–4000 tg yr⁻¹.
- Ozone is produced in the stratosphere where its concentrations are much higher than in the troposphere (see Fig. 1). The mixing down of air from the stratosphere into the troposphere leads to an important source of ozone in the troposphere (approximately 400–850 tg yr⁻¹).
- Ozone is destroyed at the earth's surface leading to the ozone removal which is also called dry deposition. Its magnitude is estimated to lead to an ozone loss of approximately 500–1200 tg yr⁻¹.

Photochemical ozone formation needs sunlight and its rate is many times faster in summer than in winter. However, photochemical lifetime also depends on the intensity of solar light and is therefore much larger in winter than

in summer. O₃ is therefore transported over much larger distances in winter than in summer.

The four terms of the ozone budget including seasonal variations and interactions may explain the ozone concentrations given as climatology in Fig. 2, which is mainly based on a large data set of ozone sonde measurements (see Section IV) averaged over the time period of 1980 to 1993. Ozone monthly mean concentrations show large vertical concentration gradients at the tropopause altitude, which lies at much higher altitudes in the tropics than in the extratropical regions. Ozone concentrations are in the order of 100 ppb at the tropopause. Ozone concentrations often start to increase within 1 km below the tropopause which is usually defined by the vertical temperature gradient. The difference between the vertical gradients of ozone and temperature is thought to be caused by small-scale mixing processes taking place close to the tropopause.

The latitudinal transect at a longitude of 170°W (see Fig. 2) indicates that zonal mean ozone concentrations in the middle troposphere exhibit almost no gradient from 30 to 75° in the winter hemispheres, whereas in the summer ozone mean values differ by 10 ppb, with the highest values ranging from 37–50°N. Ozone gradients are larger during the summer than in winter because of the faster photochemical ozone production and its shorter lifetime. In the middle troposphere winter values of ozone at northern mid-latitudes are 5–10 ppb higher than those in the southern mid-latitude troposphere, and this difference amounts to 20–30 ppb in the summer. Also surface ozone measurements at background sites indicate highest concentrations in northern mid-latitudes (see Table II). This difference is probably caused by the hemispheric differences in the ozone precursor strengths. Anthropogenic emissions of ozone precursors are largest in the northern mid-latitudes, but also the lightning source of NO_x is larger in the northern than in the southern hemisphere, because lightning occurs more frequently over continents than over the ocean and the area of continents is much larger in the northern hemisphere.

At northern mid-latitudes the seasonal cycle in the middle and upper part of the troposphere peaks in spring or summer (see Fig. 3, concentrations at 800 hPa (corresponding to an altitude of roughly 2 km amsl.) and 500 hPa (at approx. 5 km amsl.))—note that ozone at 300 (at approx. 8 km amsl.) and 200 hPa (approx. 12 km amsl.) is strongly affected by ozone of the lower stratosphere and the seasonal variation of the tropopause altitude). The spring maximum (see Fig. 3, left panel) is attributed to the interaction of the seasonal variation of the flux of ozone from the stratosphere, which peaks in spring, and the seasonal changes in the tropospheric ozone photochemistry. Ozone concentrations at the Canadian sites (located at 53–75°N) are lower than at the European sonde stations

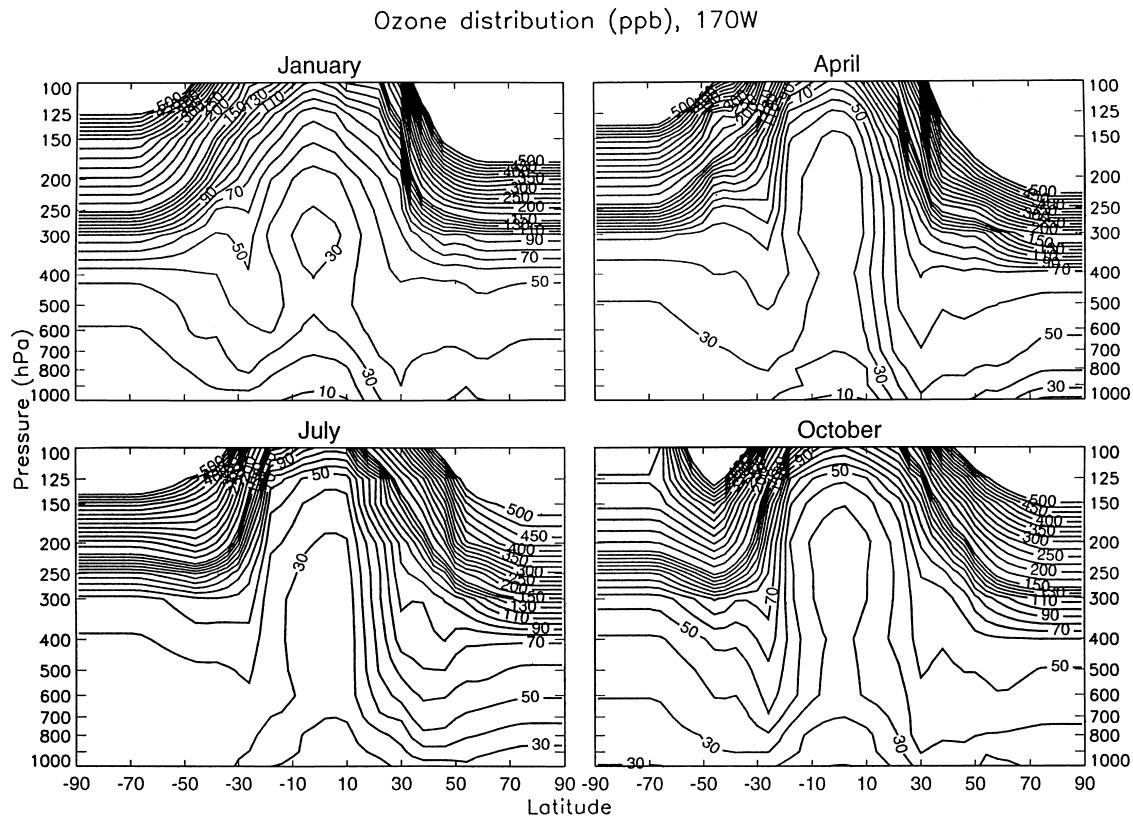


FIGURE 2 Seasonal variations of ozone as function of pressure (altitude) for 170°W , in volume mixing ratio (ppb). Contours are given every 10 to 150 ppb, and every 25 to 500 ppb. This climatology is mainly based on ozone sonde measurements of the years 1980 to 1993. (From Logan, J. A. (1999). *J. Geophys. Res.* **104**, 16, 115–149.)

($47\text{--}79^{\circ}\text{N}$), in particular at 800 hPa. The Canadian measurements show the largest ozone concentrations in the spring, while the European measurements indicate a broad summer maximum which can be explained by enhanced photochemical ozone production, from anthropogenic ozone precursors, which is fastest in the summer because of the maximum solar irradiation. The comparison of ozone measurements of in the United States and Europe is somewhat restricted because of the different types of sondes (see Section IV). Nevertheless, median ozone concentrations at tropopause are similar in North America and Europe. The Japanese ozone sonde measurements show a summer maximum at Sapporo (43°N) and a summer minimum at Naha (26°N). The summer minimum is caused by the summer monsoon, which transports low ozone containing air from the tropical Pacific by a southwesterly flow. During winter the prevailing flow at the Japanese sites is usually from the Asian continent. The residuals from ozone satellite measurements confirm the highest values of tropospheric ozone downwind of Asia, North America, and Europe in the summer.

Tropospheric ozone in the tropics shows large longitudinal and latitudinal gradients which are caused by dy-

namics and biomass burning, which injects large amounts of ozone precursors into the troposphere. The ozone concentrations are larger over the Atlantic than over the western Pacific from 800 hPa to the tropopause, possibly a consequence of biomass burning. Ozone concentrations are similar in May and June when the biomass burning is smallest over the tropics. In southern mid-latitudes free tropospheric ozone concentrations usually peak around southern hemispheric spring. Evidence for an influence of biomass burning on southern mid-latitude ozone was presented.

Surface ozone concentrations in the United States are generally lower along the west coast than along the east coast, which probably reflects the greater emission density in the east and the prevailing westerly winds leading to an increase in ozone background concentrations. In Europe, north of the Alps, a regional increase in surface ozone concentrations from the northwest to the southeast was found. Background ozone concentrations at mountain peaks usually increase with altitude (see Table II, European sites), which probably reflects the decreasing efficiency of the dry deposition with height. The surface afternoon monthly mean values from rural and remote continental sites of

TABLE II Ozone Concentrations in the Late 1990s and Long-Term Ozone Trends of Selected Rural, Remote, or Alpine Stations

Station	Latitude/longitude/ elevation, in m amsl.	Present-day concentration, in ppb	Period of measurements	Annual trend (%/yr)
<i>Arctica</i>				
Zeppelin Mountain, Spitzbergen	79°N/12°E/474	32		
Barrow, Alaska	71°N/157°W/11	27	1973–1997	0.30 ± 0.22
			1974–1995	0.18 ± 0.24
			1980–1997	−0.06 ± 0.34
Areskutan, Sweden	63.4°N/13°E/1240	37		
Karvatan, Norway	62°N/8°E/210	31 ^a	1988–1994	
Voss, Norway	60°N/6°E/500	30.4 ^a	1988–1994	
Birkenes, Norway	58°N/6°E/190	33 ^a	1988–1994	
<i>Northern mid-latitudes</i>				
Europe				
Mace Haed	53°N/9°W/10	35 ^b	1990–1994	No trend
Zugspitze, Germany	47°N/11°E/2937	49	1978–1995	1.48 ± 0.51 ^c
Sonnblick, Austria	47°N/13°E/3106	47.5		
Arosa, Switzerland	47°N/9°E/1880	47	1996–1997	No trend
		35 ^d		since 1989
Jungfraujoch Observatory, Switzerland	46°N/7°E/3580	51	1996–1997	No trend
Pic du Midi, France	43°N/0°E/2877	49	1982–1994	since 1988
Northern America				
Whiteface Mountain	44°N/		1974–1995	0.45 ± 0.22
Japan				
Oki	36°N/133°E/90	40 ^e	1994–1996	
<i>Tropics</i>				
Izana, Canary Islands	28°N/16°W/2360	45	1987–1995	−0.22 ± 0.4 ^c
Mauna Loa, Hawaii	20°N/156°W/3397	41	1973–1997	0.36 ± 0.11 ^c
			1980–1997	0.01 ± 0.28 ^c
Barbados	13°N/59°W/45	20	1989–1997	
Samoa	14°S/171°W/82	13	1976–1994	−0.26 ± 0.38
<i>Southern mid-latitudes</i>				
Cape Point, South Africa	34°S/18°E/75	22	1983–1995	0.53 ± 0.34
Cape Grim, Australia	41°S/145°E/94	25	1982–1995	0.18 ± 0.14
<i>Antarctica</i>				
Syowa, Antarctica	69°S/40°E/21	20	1989–1997	
South Pole, Antarctica	90°S/2835	26	1975–1997	−0.70 ± 0.17
			1980–1997	−0.83 ± 0.24

^a Estimated background.^b Atlantic air.^c Only downslope condition.^d Estimated Atlantic background.^e Asia continental background.

the United States and Europe show typically a spring or a broad summer maximum with maximal monthly mean values up to 75 ppb (see Fig. 4 and 7). The summer maximum provides evidence that the polluted continental planetary boundary layer is a net photochemical ozone source in summer. In Alaska and at other north polar sites ozone

concentrations are at a minimum in March to May, which was attributed to ozone destruction by bromine species which occur over the ice shield of the Arctic.

Surface ozone at remote marine sites in southern mid-latitudes shows a different annual cycle than that at northern mid-latitudes with a maximum in winter

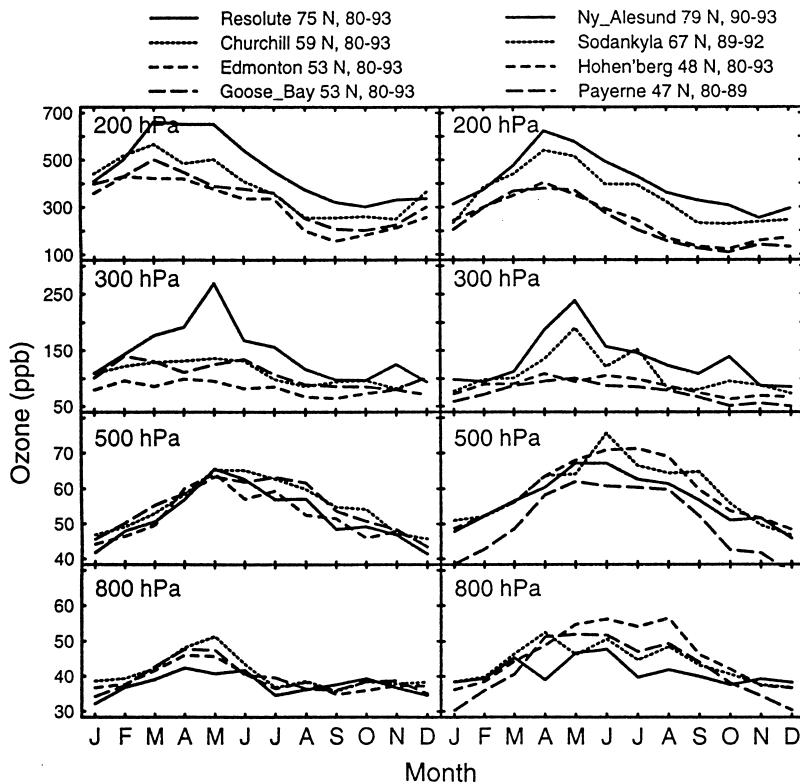


FIGURE 3 Seasonal variation of ozone at different altitudes (given in pressure levels, for conversion to altitude, see Fig. 1) from ozone sonde measurements of Canada (left) and Europe (right) of the period of 1980 to 1993. (From Logan, J. A. (1999). *J. Geophys. Res.* **104**, 16, 115–16, 149.)

(August) and a minimum in summer. At these sites NO_x concentrations are expected to be very low (less than 30 ppt) under which condition the photochemical destruction of ozone plays a dominant role (see Section II).

Ozone concentration in the planetary boundary layer strongly varies with space and time. The averaged diurnal cycle in polluted areas usually shows maximal ozone concentrations sometimes in the afternoon depending on the distance from the source region of the photooxidant precursors. At night ozone at surface decreases below the inversion layer because of the ozone depletion by dry deposition. Mixing from the free troposphere and photochemical formation leads to increasing ozone concentrations during the morning and the early afternoon. Close to large emissions sources, e.g., in cities or close to busy streets ozone concentrations are reduced because of the chemical reaction of O_3 with NO (reaction (3)). In particular during the night O_3 concentrations can become zero when no ozone is produced by the photolysis of NO_2 (reaction (2)) and when the nocturnal inversion prevents dilution of NO emissions, leading to large amplitudes in the diurnal variation of the ozone concentration. At rural sites the transport can strongly influence ozone concentrations. The amplitude of the diurnal variation at continental sites decreases with the distance to large emissions sources

and with altitude. Small amplitudes in the diurnal variations are found at high mountain stations such as at the Jungfraujoch in Switzerland.

Ozone concentrations are generally strongly dependent on meteorological conditions. Surface ozone in polluted regions strongly depends on the temperature which is the most suitable variable to describe the influence of meteorological conditions. Ozone concentrations near the surface are usually highest at high temperature because high temperatures are reached at mid-latitudes during high-pressure conditions in which the dilution of the primary pollutants and ozone is small. Furthermore, high temperature implies strong solar irradiation which accelerates photooxidant formation.

IV. TROPOSPHERIC OZONE MEASUREMENTS

Ozone in ambient air is presently measured at many sites over the world. In North America approximately 4300 air monitoring stations are currently in operation, which typically measure ozone and some primary air pollutants such as nitrogen oxides (NO_x), sulfur dioxide (SO_2), and carbon monoxide (CO). They are mainly in and near

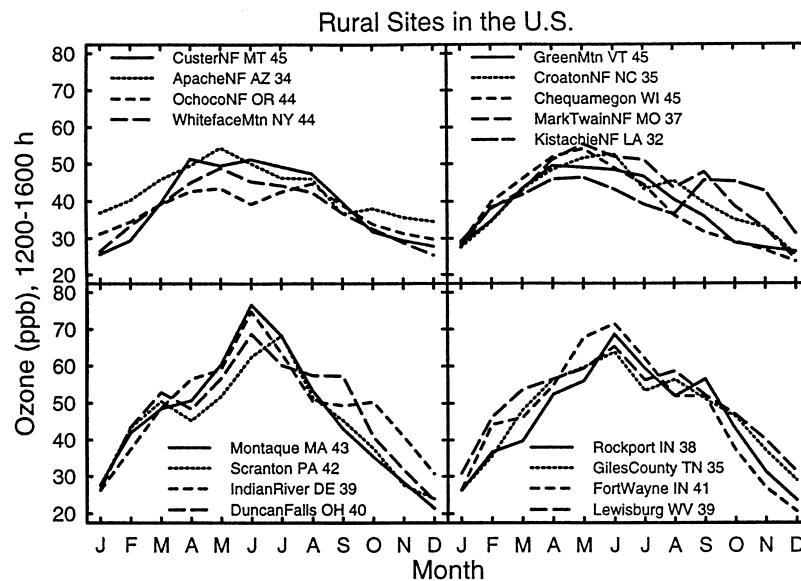


FIGURE 4 Seasonal variation of surface ozone of selected (rural or remote) sites from the United States (From Logan, J. A. (1999). *J. Geophys. Res.* **104**, 16, 115–149.)

highly populated centers of the United States, Canada, and Mexico. Also in other parts of the world ozone concentration is monitored at a large number of sites in national air quality programs. The program EMEP (Cooperative program for monitoring and evaluating the long-range transmission of air pollutants in Europe) is designed to study the trans-border air pollution in Europe. Therefore, the measurement sites of this program are located in rural areas in order to minimize the influence of large nearby primary emission sources. The measurements of the network of the Global Atmosphere Watch program (GAW) run under the auspices of the World Meteorological Organization (WMO). The sites are selected to reflect as much as possible trends in the hemisphere and the global background.

The documentation of ozone trends is a difficult task because the long-term changes in ozone concentrations are influenced by many factors including natural variability, etc. (see Section III) and anthropogenic long-term changes are usually slow. For reliable trend determination the data quality requirements of the measurements are high. Long-term trend analysis of tropospheric ozone is strongly restricted by the lack of measurements in the time before regular monitoring programs began (in the United States, most states started air pollutant monitoring in 1973, whereas most national monitoring programs in Europe started in the 1980s). All monitoring networks must include a data quality assurance program. The Global Atmosphere Watch program (GAW) of the World Meteorological Organization (WMO) runs a data quality assurance program for ozone in which surface ozone instruments and

sensors used to measure ozone profiles by light balloons are included.

Here we divide the methods of ozone measurements into the groups of *in-situ* and remote sensing methods and measurements from (satellites).

A. *In-Situ* Measurements of Ozone

1. UV-Absorption

Ambient air is sucked through a cell, where the ozone concentration is measured by its UV-absorption at 253.7 nm. At this wavelength the ozone absorption coefficient is close to its maximum. The UV light is produced by a Hg resonance lamp at 253.7 nm and the intensity of the light is measured by a photomultiplier. The ozone concentration is determined from the absorbance making use of the law of Lambert Beer:

$$[\text{O}_3] = \frac{1}{kl} \frac{T}{T_0} \frac{p_0}{p} 10^9 \log \frac{I_0}{I}, \quad (15)$$

where: $[\text{O}_3]$, concentration of ozone in ppb; T , temperature in K; T_0 , standard temperature (273 K); p , pressure in hPa; p_0 , standard pressure (1013.25 hPa); I_0 , intensity of air containing no O_3 ; I , intensity of the air when O_3 is present; k , Absorption cross section; l , length of the cell. Substances other than O_3 also absorb UV light at 250 nm in ambient air. A scrubber is used to measure the zero signal of the cell, and the absorbance of air with and without ozone is alternatively measured. The ozone concentration is determined from the difference of the signals of air with and without ozone.

TABLE III Projects in Which Commercial Aircrafts are Used for Air Sampling at Cruise Altitude

Acronym and name of the project	Measured species	Type of airplane (airliner)	Period of measurements
GASP (Global Atmospheric Sampling Program)	O ₃ , CO	B-747	1970s
MOZAIC (Measurements of Ozone and water vapor by AIRBUS In-service airCraft)	O ₃ , H ₂ O	7 AIRBUS 340 (Air France, Austrian Airlines, Lufthansa, Sabena)	Since 1994
NOXAR (Nitrogen Oxides and Ozone along Air Routes)	O ₃ , NO, NO _x	B-747 (Swissair)	May 1995–May 1996 Aug.–Nov. 1997
CARIBIC (Civil Aircraft for Remote Sensing and <i>In-situ</i> Measurements in Troposphere and Lower Stratosphere Based on the Instrumentation Container Concept)	O ₃ , CO, CH ₄ Aerosol part., Various organic trace gases	B-767 (LTU)	Since late 1990s
JAL	CO ₂ , CO, CH ₄ , N ₂ O	B-747 (JAL)	Since 1993

This technique is presently used in most air quality monitoring programs. It can provide reliable ozone measurements in ambient air. Contaminated instruments can exhibit a negative interference by water vapor yielding a reduction of the signal up to 10% at a relative humidity of 100%. Data quality of the instruments needs to be checked at regular intervals. In well-calibrated networks the readings of the instruments operated at the stations are compared with a standard instrument within a few months. In GAW the standard instrument of the Swiss Laboratories for Material Testing and Research (EMPA, Dübendorf, Switzerland) is compared with the global and regional station instruments operated in the European and African networks. Table II includes a list of sites which have provided reliable surface ozone measurements at rural, Alpine, or remote stations all over the world.

Changes in the accepted calibration methodology caused severe problems in the homogeneity of monitoring measurements in the United States. Before the late 1970s the methodology for calibration of the ozone sensors was based on the KI-method (see following); thereafter the UV-absorption technique was used. Comparisons of the two methods showed that the KI-method produced higher readings (by 3.0–3.5%) than the UV-method, but the adjustment factors varied, probably caused by interference by the KI-method (see following).

The UV-method is often used to make O₃ measurements from airplanes. However, the time resolution is comparatively slow because of the need of alternative measurements of air with and without ozone. The response time of commercially available instruments is usually in the order of 10 to 20 sec. Response times as short as 1 sec can be achieved by this method, which is important for particular applications such as ozone measurements from air planes.

Permanently installed ozone instruments on board commercial airplanes can be used for measurements with large spatial and temporal coverage as first demonstrated in the

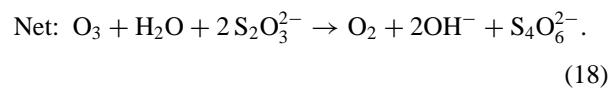
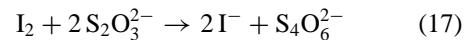
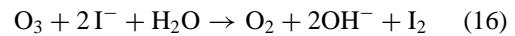
program GASP in the 1970s. Most of the measurements are performed at cruise altitude, which is approximately 10 to 12 km above ground and therefore in the upper part of the troposphere in the tropics and close to the mean tropopause altitude at mid-latitudes. Three further projects provided extended ozone measurements at cruise altitude since the second half of the 1990s (see Table III). Measurements from commercial airplanes might be used in the future for quasi-global upper tropospheric ozone trend analysis if the programs continue. They may play an important role in the future similar to the role played by satellite measurements in stratospheric ozone research.

2. Chemiluminescence

Ozone reacts very fast with NO (reaction (3)) which produces a chemiluminescence signal. This reaction can be used as a basis for a very sensitive and specific ozone instrument. Also the chemiluminescence reaction between ozone and unsaturated hydrocarbons (e.g., ethylene) and other compounds has been used in instruments to measure ozone in tropospheric air. These methods allow one to measure ozone with a very high time resolution which is important in many airplane measurements in research projects.

3. Measurements by the KI-Method

Ambient air is sucked through an aqueous solution which contains potassium iodide. O₃ is measured by the oxidation of I[−] to I₂ in aqueous solution which is stabilized by its reaction with thiosulfate (S₂O₃^{2−}) (Eq. (17)).



The amount of $S_2O_3^{2-}$ consumed is proportional to the O_3 present in ambient air. In the standard method, $S_2O_3^{2-}$ is measured by back-titration with I_2 using starch as indicator, but also electrochemical detection was employed. Arsenite (AsO_3^-) also can be used instead of thiosulfate which was used in the extended ozone measurements performed from 1876 to 1911 at the Observatoire de Montsouris close to Paris (France). The KI-method was widely used in the 1940s and 1950s.

The KI-method suffers from several interferences, particularly by sulfur dioxide (SO_2). Ozone measurements by the KI-method from urban sites must be regarded with some caution because sulfur dioxide is a major air pollutant produced by the combustion of fossil fuels (e.g., of coal). At some stations a CrO_3 filter was introduced to remove the bias of SO_2 , causing possible problems in the homogeneity of the long-term series. Measurements of rural and Alpine sites are much less influenced by such problems. Because of these interferences present monitoring measurements are usually based on the UV-method (see earlier).

The KI-method is still used in O_3 profile measurements made from light balloons, in which the UV-method is not practicable. Such balloons which are used for routine meteorological measurements (temperature, wind speed, wind direction, and humidity) reach an altitude of approximately 30 km where they burst. Thereafter the equipment flies back to the earth's surface by parachute. The ambient air is pumped through a sensor which is based on the electrochemical detection of the reaction of O_3 with KI providing profiles with high vertical resolution. The presently used sensors can be divided into two main classes depending on their design. In the Brewer–Mast (BM) sonde the two electrodes are located in the same cell with a silver anode and a platinum cathode immersed in an alkaline potassium iodide solution. The ECC-type sensors (Electro Chemical sonde or carbon–iodine ozone sonde) contain two half-cells, both containing a platinum mesh serving as electrode. The two chambers are linked together with an ion bridge. The sondes used in Japan (KC) are based on a modified version of the carbon–iodide ozone sensor.

Experience showed that the performance of Brewer–Mast sondes critically depends on the prelaunch procedure of the sondes, whereas the ECC sondes need less sophisticated control before launch. The integrated ozone amount measured by the sonde can be compared with the total ozone amount measured by sun photometry (usually performed by Dobson spectrophotometers). This requires the extrapolation of the O_3 profile above the burst level of the balloons (residual ozone). This comparison can be used to check the quality of the ozone profile of the sonde. The factor obtained by the sum of the ozone measured by the sonde plus the residual ozone divided

by the measured total ozone amount is called the correction factor (CF). In the operational procedure of WMO the sonde profile is linearly scaled by CF. It has been standard practice to use the mixing ratio at or near the top of the profile (before the burst of the sonde) to estimate the residual ozone. Long-term trend determination of free tropospheric ozone covering the time since the 1970s is restricted to a few ozone sonde stations (see [Table IV](#)). Unfortunately they are mainly located in the northern mid-latitudes. The reliable measurement of tropospheric ozone by the electrochemical sondes is difficult because of the low concentrations. Recent comparisons with extended O_3 measurements from airplanes (program MOZAIC, see [Table III](#)) showed encouraging data quality of averaged ozone sonde measurements of several stations during the late 1990s. However, the evaluation of the reliability of the earlier ozone sonde measurements remains a challenging task. Nevertheless, the growing network of ozone sonde measurements is still the most important archive for free tropospheric long-term trend determination and one of the most important sources of information for tropospheric ozone research in general.

4. Ozone Measurements with Schönbein Papers

Schönbein, who discovered ozone, used strips containing a solution of potassium iodide and starch which develop a bluish color when exposed to ozone in ambient air. Measurements were recorded in a scale from 0 to 10 according to the change of color. There are several unresolved problems with this method, including its sensitivity to humidity and the effect of decoloration. Unfortunately, the exact recipe of the preparation of the papers is not known. Already in the 19th century the method was criticized. It seems likely that the method is only semi-quantitative. We therefore regard these measurements as rather unsuitable to determine reliable ozone concentrations of the 19th century before the problems of the method are solved properly. These measurements are therefore not used in this article.

B. Remote Sensing Measurements from the Earth's Surface

1. Open Path UV-Absorption Spectrophotometry (Including DOAS)

In the open path absorption method ozone in ambient air is measured by its absorption in the UV. A UV-source with known intensities at different wavelengths provides the source of the UV-light and a receiver allows one to measure the decrease of the light at the respective wavelengths when the light beam passes through the air. The

TABLE IV Long-Term Ozone Sonde Records and LIDAR Measurements

Name of the station	Lat./long.	Type	Period or start of the measurements	Average number of ascents per month
Ozone sonde records^a				
Northern America				
Resolute, Canada	75°N/95°W	BM ECC	Jan. 1966–Nov. 1979 Since Dec. 1979	Jan. 70–Dec. 96:3.3
Curchill, Canada	59°N/94°W	BM ECC	Oct. 1973–Aug. 1979 Since Sept. 1979	Oct. 93–Dec. 96:3.3
Goose Bay, Canada	53°N/60°W	BM ECC	June 1969–Aug. 1980 Since Sept. 1980	Jan. 70–Dec. 96:3.7
(Alert, Canada		ECC	Start: 1987)	
Edmonton, Canada	53°N/114°W	BM ECC	Oct. 1972–Aug. 1979 Since Sept. 1979	Jan. 73–Dec. 96:3.5
Boulder, CO	40°N/105°W	(BM ECC	August 1963–July 1966) Since March 1979	Jan. 80–Dec. 84:1.5 Jan. 85–Dec. 96:3.8
Wallop Island, USA	38°N/76°W	ECC	May 1970	May 70–Apr. 95:2.5
Europe				
(Lindenberg, Germany	52°N/14°E	GDR	Jan. 1975–Dec. 1982)	
Uccle, Belgium	51°N/4°E	BM	Jan. 1969	Jan. 70–June 89:9.1 July 89–Dec. 96:11.0
Hohenpeissenberg, Germany	48°N/11°E	BM	Nov. 1966	Jan. 70–Dec. 77:3. Jan. 78–Dec. 96:9.9
Payerne, Switzerland	47°N/7°E	BM	Nov. 1966	Jan. 70–Dec. 75:8.2 Jan. 76–Dec. 96:10.9
(Biscarrosse, France	44°N/1°W	BM	March 1976–Dec. 1982)	
(Haute de la Provence, France	44°N/6°W	ECC	1989–present	1 per week)
Asia				
Sapporo, Japan	43°N/141°E	KC	Dec. 1968	Jan. 70–Dec. 74:2.6 Jan. 75–June 89:1.0 July 89–Dec. 96:3.2
Tskuba (Tateno), Japan	36°N/140°E	KC	Nov. 1968	Jan. 70–Dec. 74:2.4 Jan. 75–June 89:1.7 July 89–Dec. 96:4.2
Kagoshima, Japan	32°N/131°E	KC	Jan. 1969	Jan. 70–Dec. 74:2.3 Jan. 75–June 89:1.0 July 89–Dec. 96:3.0
Tropics				
(Naha, Japan	26°N/127°E	KC	Start: Sept. 1989)	
Hilo, USA	20°N/155°W	ECC	Dec. 1982	Sept. 82–Dec. 96:3.3
Southern midlatitudes				
Lauder, New Zealand	45°S/170°E	ECC	Aug. 1986	Aug. 86–Dec. 96:5.7
Antarctica				
Syowa	69°S/40°E/21	KC	1966	
LIDAR series				
Europe				
Hohenpeissenberg, Germany	47°N/11°E		1987	
Haute de la Provence, France	44°N/6°E		1986	
Asia				
Tskuba, Japan	36°N/140°E		1988	
Northern America				
Table Mountain Mesa, USA	34°N/118°W		1988	
Tropics				
Mauna Loa, Hawaii, USA	20°N/156°W		1992	
Southern mid-latitudes				
Lauder, New Zealand	45°S/170°W 7		1992	

^a The stations in brackets have not been included in the long-term trend analysis of SPARC, mainly because of insufficient length of the record. BM: Brewer Mast type; ECC: Electrochemical sonde; GDR: Brewer type, used in the former GDR; KC: Japanese sondes.

distance between the light source and the receiver varies between a few hundred meters to several kilometers. Measurements have to be performed at different wavelengths in order to attribute the absorption to ozone since other substances including aerosol particles also absorb UV-light in

the atmosphere. This method was used for reliable surface ozone measurements since 1918. At Arosa (Switzerland) ozone near the surface was measured in the 1930s by using measurements at 8 to 18 wavelengths which extended from 248 to 303 nm. However, such measurements could

only be made during the nights and the determination of the ozone concentration was rather time consuming.

The ozone spectrum exhibits distinct fine structures with several maxima and minima in the UV (260 to 350 nm) and in the visible which can be used to measure ozone by differential optical absorption spectroscopy (DOAS). The light source is a Xenon high-pressure lamp which emits light from 200 to 700 nm. A fast rotating disk allows one to measure the received light by a photomultiplier with very high resolution in time and wavelength. The high time resolution is required to eliminate the influence of changes in the trace gas concentrations caused by turbulence, whereas the high optical resolution (less than half of 1 nm) allows one to identify and to quantify the concentrations of the trace gases. Ozone measurements are performed, e.g., in the wavelength region around 335 or 425 nm and the high wavelength resolution allows one to separate the absorption of ozone from other compounds in ambient air which include SO₂, NO, and NO₂. The averaged ozone concentration over the light pass is calculated based on its absorption spectrum.

DOAS instruments are commercially available. In the absence of large spatial gradients the results of DOAS measurements (averages over the light path) yield equivalent results to the *in-situ* UV-method.

2. Ozone Profile Measurements by LIDAR

In Differential Absorption (DIAL) LIDAR (light detection and ranging) a laser source is used to emit pulsed light from the ground at selected wavelengths which is reflected in the atmosphere by aerosol particles and gases. For atmospheric ozone measurements one wavelength of the emitted light is strongly absorbed by atmospheric ozone, while the second wavelength is only weakly absorbed. The O₃ profile can be determined from the time delay between emission and return of the pulsed light and the ozone amount from the differential absorption between the two wavelengths. Different types of LIDAR instruments were developed. Some are more suitable to measure tropospheric ozone than stratospheric ozone. Most challenging are accurate ozone profile measurements in the planetary boundary layer, in particular, in polluted areas, where aerosol particles interfere with ozone measurements. Upper tropospheric ozone LIDAR measurements are disturbed by clouds. The first continuous LIDAR series of upper tropospheric ozone started in the second half of the 1980s (see [Table IV](#)). They complement the ozone sonde measurements.

C. Ozone Measurements from Satellites

Measurements from satellite instruments have allowed one to monitor the global ozone shield since the late 1970s.

They are widely used for column and stratospheric ozone trend determination. Tropospheric ozone measurements from space are more difficult since the light from space has to pass the ozone layer (see [Fig. 1](#)) before the much lower ozone concentrations in the troposphere can be measured.

1. GOME

The instrument GOME (Global Ozone Monitoring Experiment) makes ozone measurements based on the principle of DOAS (see earlier). GOME flies on the second European Remote Sensing Satellite (ERS-2) of the European Space Agency (ESA); ERS-2 was launched in April 1995. GOME measurements can been used to calculate ozone profiles and they allow one to measure other substances. More sophisticated satellite instruments based on the same technique are scheduled for launch in the near future. However, clouds are causing interferences of such satellite measurements of tropospheric ozone. Global measurements of vertical profiles of tropospheric ozone (at three to four levels and a vertical resolution of 5 to 2.3 km) will be made by the Tropospheric Emission Spectrometer (TES), which will be part of the Earth Observing System's CHEM platform (launch scheduled for 2002). The use of satellite measurements for trend determination in the upper troposphere will need data sets of sufficient lengths, data quality, and spatial coverage.

2. Tropospheric Ozone Amount from Satellite Measurements

Total Ozone Mapping Spectrometers (TOMS) measure solar irradiance and radiance backscattered by the earth's atmosphere at six wavelengths extending from approximately 310 to 380 nm. TOMS measurements can be used to estimate the tropospheric ozone amount by the calculation of the difference of total ozone minus stratospheric ozone amount measured by another satellite instrument (SAGE: Stratospheric Aerosol and Gas Experiment), in particular when stratospheric ozone variations are small, usually over the tropics. However, the residual tropospheric ozone calculation has some limitation in accuracy because the tropospheric ozone amount is calculated as a small difference of two large numbers (total and stratospheric ozone). Measurements of the TOMS instrument on Nimbus 7 satellite have been used to estimate tropical tropospheric ozone trends from 1978 to 1992.

V. TREND DETERMINATION

Different methods have been developed for ozone trend determination and the most appropriate approach depends

on the scope of the study and the availability of the measurements. The first step of trend analysis must include the strict control of the data quality and the homogeneity of the measurements which is sometimes a difficult task, in particular when using historical data. Furthermore, the availability of the data must be considered. Many monitoring series contain substantial gaps because of instrument malfunctions or problems in calibration. Therefore the percentage of reliable measurements must be considered when selecting ozone series for trend analysis, and the treatment of missing values must be properly described. The most simple and widely used statistical trend model includes a linear trend over the period of available measurements. The rate of change (increase or decrease) can be defined relative to the average of the period or relative to the ozone concentrations at the start of the time series. The adopted definition needs to be clarified when the changes in concentrations over the period of measurements are large. In this article we present ozone changes relative to the average of the period (if not clarified in the text).

The main scope of many trend studies of surface ozone is to document the success of the reduction of the ozone precursors in a special area. However, the substantial variability due to different meteorological conditions complicates the documentation of air pollutant measures. Therefore methods have been developed to reduce the influence of meteorological conditions based on statistical models of different degrees of sophistication. These concepts are usually based on multiple regression models. For surface ozone measurements it has been found that the temperature is the most suitable variable to minimize the influence of meteorological conditions (see Section III). Alternative statistical techniques include the use of the Kolmogorov-Zurbenko filter, which is based on the iterations of a simple moving average, allowing one to separate components of different lengths including seasonal and short-term variation in a time series. The determination of the statistical significance of the trend results depends on the sophistication of the respective trend model. Usually the *p* values are used to describe the error probability to reject the null hypothesis (no significant trend). Simple student *t* tests have been used for the calculation of the significance of linear trends. However, ambient air concentrations of gaseous trace components often show a strong autocorrelation in the residuals of the regression models which makes the correct statistical treatment of significance tests and the calculation of the confidence intervals a more difficult task. This problem can be solved, if the error term of the regression model is described by an ARIMA (Integrated Autoregressive Moving-Average) process.

Only afternoon values during the summer are included in many studies dealing with photooxidant pollution since the anthropogenic influence on regional photooxidant air

quality maximizes during these conditions. Ozone concentrations at receptor sites show characteristic variabilities which can be described by frequency distributions. From a point of view of statistics it is more useful to analyze median than mean values and the analysis of quantiles (e.g., the highest 10% of the measurements) is more appropriate than the analysis of the extreme values which more strongly depend on the meteorological condition.

The selection of ozone data for trend analysis is often based on human health considerations or on ecological criteria. It is believed that ozone maxima are important for effects of ozone on human health including the deterioration of the respiratory functions, whereas the mean exposure over the vegetation time is more adequate to study the damage by high ozone concentration on plants, in particular the losses of agricultural yields. For these studies ozone exposure accumulated over threshold (AOT) is commonly used. The AOT40 value denotes the sum of hourly mean concentrations above 40 ppb recorded during daylight hours over the vegetation period of 3 months. Other ozone trends analyses were undertaken to document whether the air quality with respect to national air pollution legislation has improved. In these studies data are often selected according to the requirements of the national legislation, which are usually based on some limiting values and the number of days exceeding these values, etc. The air pollutant legislation and the requirements for ambient air quality vary between different countries. Unfortunately the respective data selection reduces sometimes the comparability of the results of ozone trend studies.

The ozone trend results should be compared with the changes in the primary emissions in the respective source region for the most appropriate determination of the success of measures to reduce the primary emissions. The selection of the appropriate emission source region is sometimes rather difficult and at remote sites the hemispheric emissions of an entire continent need to be considered.

The scope of trend analysis of background ozone and of ozone in the free troposphere usually differs from studies related to ambient air quality. Such studies are related to global changes because they are important in the context of the greenhouse gas warming. In these investigations, ozone not only during summer but also during the entire year have to be included and the mean values instead of extreme values are important.

Equation (19) shows the statistical model commonly used in the analysis of ozone sonde measurements

$$Y_t = \sum_{i=1}^{12} \mu_i I_{i,t} + \sum_{i=1}^{12} \beta_i I_{i,t} + \sum_{n=1}^k \xi_n \Theta_{n,t} + N_t, \quad (19)$$

where: *t* is index of the month, counted from the beginning of the measurements; *i* is index of calendar month

in the year (i.e., $i = 1$ for January, $i = 12$ for December); Y_t is measured ozone monthly mean concentrations of the month t ; μ_i is mean ozone concentration at the i th month (average of all months i of the years in the considered period); $I_{i,t} = 1$ if month t corresponds to the i th month of the year, otherwise 0; β_i is trend of the i th month of the year; $\Theta_{n,t}$ is explanatory variable (monthly mean values), coefficient ξ_n , $n = 1, \dots, k$; N_t is the noise autocorrelation term, $N_t = F_1 N_{t-1} + F_2 N_{t-2} + \varepsilon_t$, with $\varepsilon_t \sim N(0, 1)$.

The explanatory variables $\Theta_{n,t}$ are introduced to describe natural variability. Usually the Quasi Biennial Oscillation and the 11-year solar cycle are considered in ozone sonde time series analysis and a linear trend, started at the beginning of the 1970s, is assumed. By statistical models breaks in several long-term ozone series were identified. However, it is a difficult task to distinguish between unknown natural variability and data quality problems.

The scientific value of statistically significant trends of short series (less than 10 years) is somewhat doubtful, because the special meteorological conditions at the beginning and at the end of the measurements can strongly influence the results of the statistics.

VI. SURFACE OZONE TRENDS IN NORTH AMERICA (INCLUDING MEXICO)

A. Trends in Urban Plumes

Photochemical air pollution was first detected in the Los Angeles basin where to our knowledge the largest ozone concentrations were found (see Table I). The population, industrialization, and road traffic emissions rapidly increased in Los Angeles and its surroundings after World War II. Part of the problem is that the area is very badly ventilated—it is surrounded by a ring of mountains that help to contain the pollution. Figure 5 shows that ozone concentration maxima in Southern California have strongly decreased since the start of the monitoring measurements. The fluctuations of the ozone maxima from year to year are large which mainly reflects the prevailing meteorological conditions. Also the number of days in which some threshold values are exceeded in the entire monitoring network of Southern California shows a strong decreasing trend between 1976 and 1998. In another study ozone hourly mean values of each afternoon (13 to 16 local time) were extracted for summer days (June through August). Median and 90th percentiles of these concentrations were calculated for the period 1980 to 1995. The measurements of Pasadena in southern California show a significant decreasing trends of the 90th percentile values

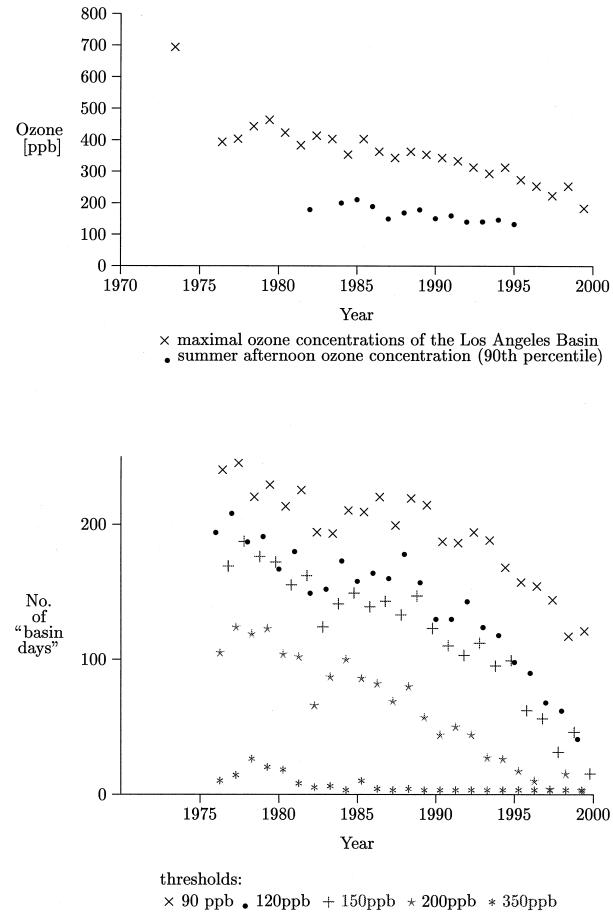


FIGURE 5 Time series of surface ozone concentrations in the Los Angeles basin. (a) Maximal concentrations and summer afternoon 90-percentiles; (b) Number of days on which a standard was exceeded (“basin days”) anywhere in the Los Angeles basin based on the measurements of the monitoring network. (The data are taken from the EPA homepage and from Fiori et al. (1998). *J. Geophys. Res.* **103**, 1471–1480.)

(see Fig. 5) ($R^2 = 0.68$, slope = -5.1 ppb per year). The state of California has implemented emission controls beyond the requirements of the rest of the United States. Total VOC emissions in the Los Angeles basin decreased by 40% between 1980 and 1997, while NO_x emissions were reduced by 10%. It appears that the large decrease in ozone concentrations is mainly due to the emission reduction of VOCs as expected from the VOC sensitivity of regional ozone production (see Section II).

In the region of New York City, prevailing winds in the summer are from the southwest and most of the monitoring sites in the area downwind of the agglomeration of New York City show significant decreasing trends in the 90th percentiles and in the medians. They spread over Connecticut up to 400 km downwind of New York city. Ozone in the Chicago metropolitan area which is the third largest agglomeration in the United States shows

significant downward trend from 1980 to 1995 in the 90th percentile but not for the medians. These trends also seem to be grossly consistent with VOC sensitivity of regional ozone production downwind of large cities because anthropogenic VOC emissions of the entire United States decreased by 12% from 1980 to 1995 while NO_x emissions remained almost constant. The analysis of meteorologically adjusted trends in daily maximum ozone values (1-hr means) of 11 urban sites of Ontario (Canada) of the period 1980 to 1990 yielded an increase of 1.2% per year. When analyzed individually, three of the sites showed a decrease and one no trend. In another analysis ozone measurements of 10 urban sites in the Lower Fraser Valley (Canada), three sites in the Atlantic region, and three sites in the region of Montreal were analyzed also taking into account meteorological effects. The composite trend of all sites showed an increase in daily maxima of 0.45% per year from 1985 to 1992 in the Lower Fraser Valley, an increase of 2.2% per year from 1985 to 1992 in the Atlantic region and a decrease of 0.87% per year from 1981 to 1993 in Montreal.

To our knowledge the second highest ozone concentrations in the world have been reported from the surroundings of Mexico City where they reached maximal concentrations of up to 477 ppb in 1992 (see Table I). Ozone values are presently higher in this region than in the Los Angeles basin, because of the decreasing trends in southern California. During the last years extreme values in the area of Mexico City show some decreasing tendency. However, present maximal values are still very large (above 300 ppb) (personal communication, S. Wakamatsu).

B. Trends in Rural and Remote Sites

Reliable long-term surface ozone series for trend determination of rural sites are rather limited. Various attempts of ozone trend analysis of different periods including available measurements of rural sites in agricultural areas did not reveal a clear picture of increasing nor decreasing trends in the United States. Also measurements of rural sites in the United States during the period 1980 to 1995 were included in the analysis of summer afternoon trends discussed in the last section. In this study also daily temperatures were used to account for the variability in O₃ concentrations associated with temperature. Trends were generally insignificant except for the sites downwind of large agglomerations both before and after segregating the O₃ measurements by temperature. This is consistent with the lack of decreasing NO_x emissions in the United States during this period because NO_x sensitivity in ozone production is expected for areas which are not under the direct influence of urban plumes (see Section II).

In another study regional ozone trends were estimated based on data of 121 sites in Canada and the United States with the majority of sites being in relatively rural areas. A principal component/time series filter analysis was employed while meteorological adjustments were not considered. Mean daily maximum ozone from May to September decreased from 1985 to 1993 by 0.05–0.27% per year for seven Canadian regions, increased by 0.07% per year in one region, and showed no trend in one region.

The longest continuous series of surface ozone measurements at a rural site in the United States started in 1974 at the Whiteface mountain which is located in the north of New York in the eastern United States. Over the period 1974 to 1995 an increase of $0.45 \pm 0.22\%$ per year was found (see Fig. 6). This relatively small increase is much reduced during the second half of the record. However, some questions about the homogeneity of the series were raised. Measurements of Barrow, Alaska, show insignificant trends ($0.18 \pm 0.24\%$ per year) in the period 1975–1995 (see Table II). At this station, marine planetary boundary air without any recent contact with anthropogenic emissions sources is sampled.

VII. SURFACE OZONE TRENDS IN EUROPE

A. Trends in Urban and Suburban Areas

Also in the surroundings of large European cities high ozone concentrations have been reported (see Table I). Maximal concentrations in the surroundings of Milan and Athens are coming close to the present measurements in southern California. In the surroundings of Milan and Athens the ventilation of ozone and its precursors seems to be poor. The 50-percentiles of the summer monthly mean values of two monitoring series in the Athens basin showed decreasing O₃ concentrations in the period from 1987 to 1998.

From 1985 to 2000 anthropogenic NO_x emissions decreased by approximately 30% in Switzerland mainly because of new legislation which requires one to install catalytic converters in new gasoline-driven cars. A rather sophisticated statistical model was used to minimize the influence of meteorological factors for analysis of the measurements of 24 Swiss monitoring sites from 1987 to 1993. Annual means of NO_x concentrations of urban sites showed a significant decrease with rates of up to 10% per year. Ozone concentrations at the same urban sites showed a significant increase with annual changes of 3 to 7%. However, O_x concentrations (see Eq. (4)) showed no significant changes in this period which indicates that the increase in O₃ concentration was caused by the decrease in

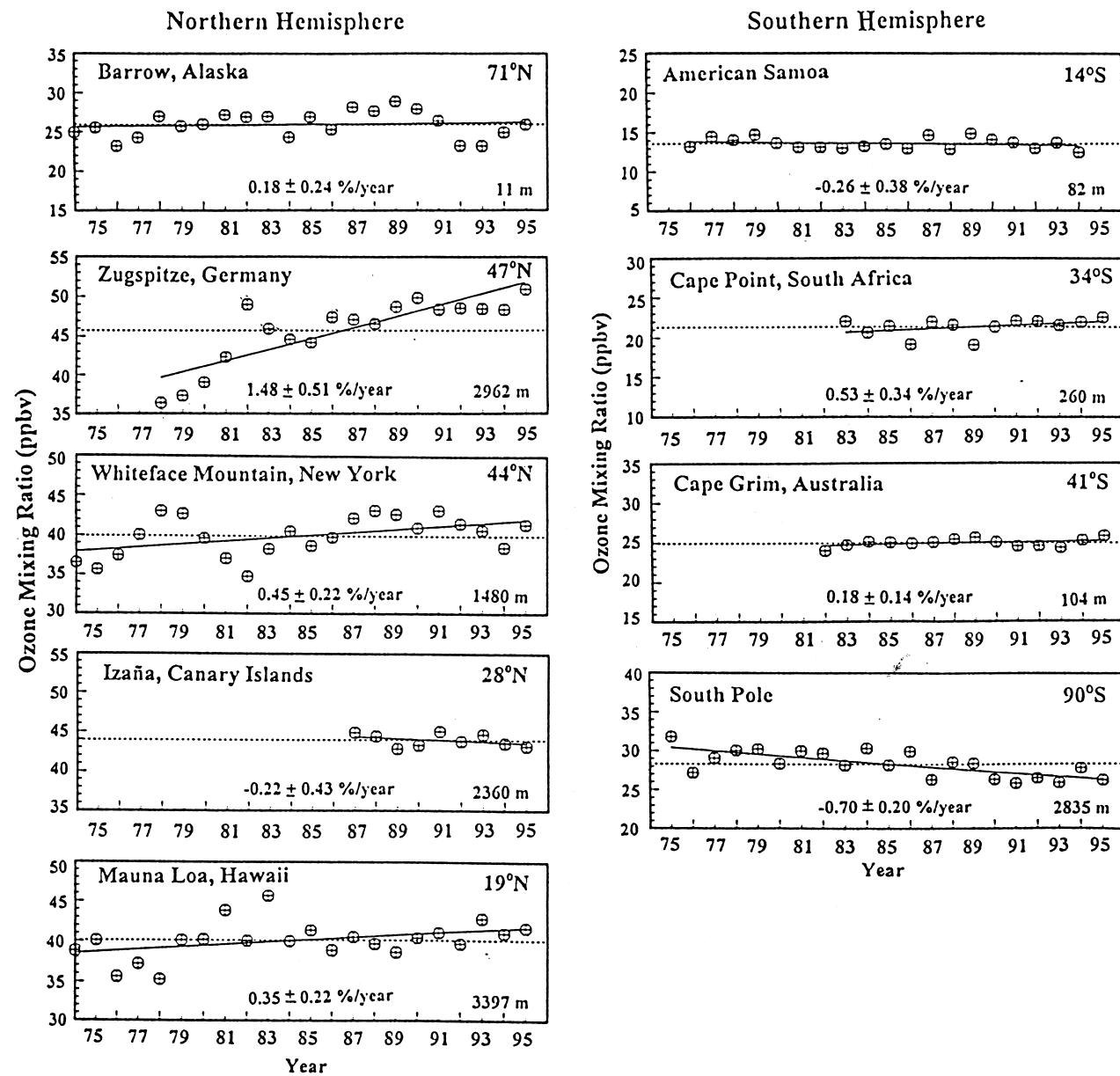


FIGURE 6 Annual average ozone concentrations (in ppb) for selected surface ozone measuring sites which are not directly influenced by local photooxidant pollution (see Table I). The dashed line is the long-term average. The solid line is the linear least squares fit of the monthly anomalies. The linear trend and the 95% confidence in percentage per year is given in the plot for each location (From Oltmans *et al.* (1998). *Geophys. Res. Lett.* **25**, 139–142.)

the yield of reaction (3) which destroys ozone close to the emission sources. This result implies that some increase of ozone concentrations in urban sites has to be accepted when NO_x emissions are strongly decreased (without large simultaneous VOC reductions). However, this increase in ozone concentration in urban areas is hardly a significant problem from the perspective of environmental protection since annual mean ozone values at these urban sites were only between 10 and 15 ppb. They are therefore much

smaller than those in the surrounding suburban and rural sites.

B. Trends in Rural and Remote European Sites

Historical ozone measurements are valuable in order to document the extent of phototchemical air pollution, and they are important in the context of greenhouse gas forcing. At Arosa ($47^\circ\text{N}/9^\circ\text{E}$, 1880 amsl.) in the Swiss Alps a

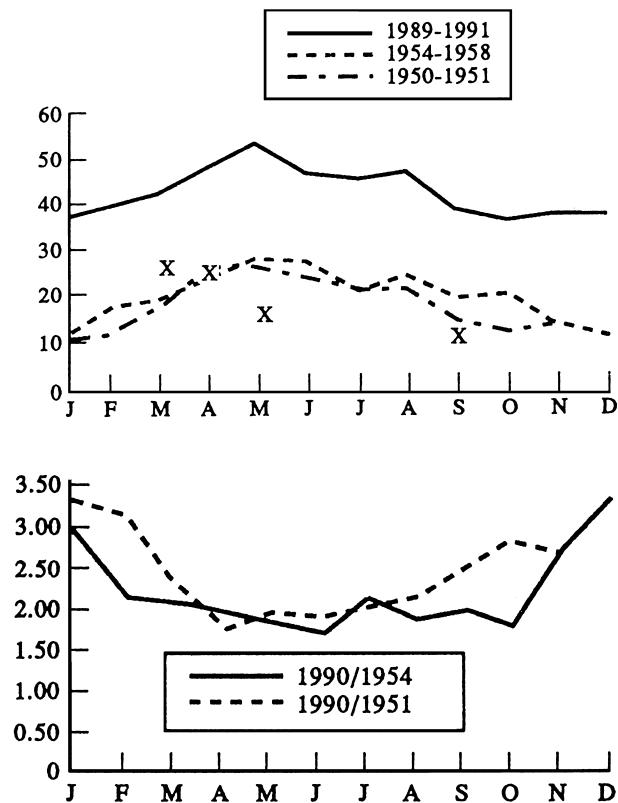


FIGURE 7 Comparison of averaged seasonal variation of surface ozone (monthly mean values at Arosa (Switzerland) during different time periods. (a) Concentrations in ppb, x: averaged concentrations calculated from the single measurements made in the 1930s during clear nights. (b) seasonal differences of the ratios from the recent measurements and the measurements of the 1950s (From Staehelin *et al.* (1994). *Atmospheric Environment* **28**, 75–87.)

few measurements by open path absorption spectrometry were made in the 1930s. In the 1950s extended surface ozone measurements were made by the KI-method. Interference of the ozone measurements by SO₂ was most probably negligible at this Alpine site. Thereafter, only a few measurements were performed at Arosa until continuous ozone measurements started again in 1989. Figure 7 shows an increase by more than a factor of 2 since the late 1950s and 1989, whereas ozone concentrations stayed constant in the following decade. The relative concentration increase was much larger in winter than in the other seasons. During summer the lifetime of ozone is much shorter than that in winter and therefore the increase in summer reflects the influence of the European emissions, whereas the winter increase is expected to be rather representative for the hemispheric ozone background.

During the 1930s and the beginning of the 1940s single ozone measurements were also made at many other European sites. For August and September a data set

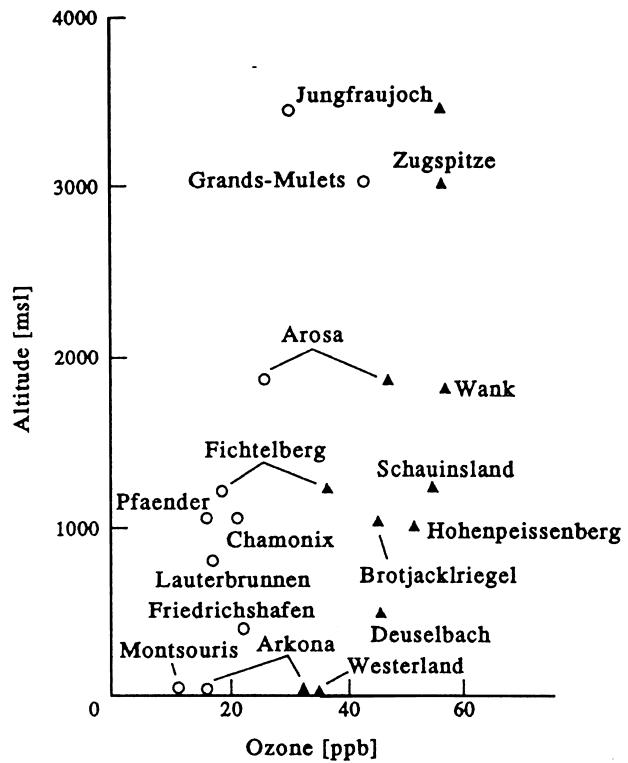


FIGURE 8 Historical (circles) and recent (triangles) surface ozone concentrations of August/September from different locations in Europe as a function of altitude. The historical measurements from the different sites also include measurements collected over short periods, whereas the recent data of 1988–1991 are based on continuous monitoring measurements. (For data sources, see Staehelin *et al.* (1994). *Atmospheric Environment* **28**, 75–87.)

was obtained which allowed one to construct a plot of ozone concentration versus altitude by including all available European measurements from rural and remote sites which were made before the end of 1950s. We conclude from Fig. 8 that background ozone concentrations in the planetary boundary layer over Europe has increased by at least a factor of 2 from the late 1950s to the end of the 1980s. In this period anthropogenic NO_x emissions strongly increased in the industrialized world (by more than a factor of 4 in Europe and by more than a factor of 2 in the United States).

The documentation of changes in background ozone in the time before the World War II is more difficult. Many surface ozone measurements were made by a chemical method in the last century at Montsouris, an observatory close to Paris. Careful tests showed that only SO₂ is a significant interference of the method, which is very similar to the widely used KI-method (arsenite was used instead of thiosulfate, see KI-method). Wind measurements of the same observatory were used to exclude measurements,

polluted by SO₂, which were expected to originate from Paris. The mean ozone concentration of the Montsouris data was 10 ppb with a significant proportion of measurements below 5 ppb, some even close to zero. The seasonal variation in the Monstosuris data shows a maximum in spring, which is different from the seasonal variation in the present polluted troposphere, where ozone peaks in summer (compare Section III). However, the large fraction of very low values is believed to be a regional effect probably caused by dry deposition in the basin of Paris or possibly by the interference of SO₂ which was not completely removed by the available wind measurements. Therefore, care should be taken especially in treating the mean value of 10 ppb of the Montsouris record as being representative for the unpolluted mid-latitude background air.

Continuous long-term surface ozone measurements were made at several sites in the former German Democratic Republic; two began in the 1950s. We consider here only the results of the rural site Arkona at the Baltic sea because the measurements were made by the KI-method which could be affected by SO₂ and its trends. The measurements show an average increasing ozone trend of 1.44% per year over the period from 1956 to 1991, which is caused by large trends in the period from 1956 to 1984. This large increasing trend in background ozone concentrations during this period is consistent with the results of the measurements at Arosa (Switzerland). The series at the observatory of Hohenpeissenberg in Southern Bavaria (Germany) shows an increase in the annual mean values of approximately 1% per year over the period of 1971 to 1992. The observatory is sometimes under the influence of the ozone plume of Munich. However, some concerns about the homogeneity of the series were raised.

Large increasing ozone trends ($1.48 \pm 0.51\%$ per year for 1978–1995) are found at the high Alpine station at Zugspitze in Germany (see Fig. 6) when only night-time data are considered. Night-time measurements at high Alpine sites during downslope conditions are believed to be representative for the free troposphere. The largest part of this increase took place in the 1970s and the early 1980s. Surface ozone monitoring started at the Jungfraujoch Observatory (Switzerland) at 3580 amsl. in 1986 but reliable measurements have been available since 1988. They show no significant trend of the annual mean values for the period 1988–1996, in agreement with the series of Zugspitze during the same period.

Ozone monitoring measurements of 20 different European sites (remote, rural, suburban, and urban) in Germany, United Kingdom, and the Netherlands, were analyzed for the period 1978–1988. Relatively few of the trends were statistically significant and no dominant regionwide trends were found. Marine stations showed negative trends in summer average concentrations, while

the stations located further in the interior exhibited positive trends. Most of the stations showed positive trends in winter averages. During the 1990s ozone background concentrations (45–55 ppb) increased at the rural sites in Switzerland, whereas the frequency of ozone maxima and minima decreased. Studies with trajectories (i.e., path of air parcels calculated from meteorological models used for weather prediction) showed that present ozone concentrations in summer at Arosa systematically increase with the residence time of the air over Europe. From such an analysis it was concluded that European ozone production adds approximately 12 ppb to the present intercontinental background of 35 ppb (see Table II). In the European project TOR-2 (Tropospheric Ozone Research) mean values or 50-percentiles of 40 rural surface ozone sites were analyzed for monthly mean trends using a statistical test (Mann–Kendall test). The measurements extended from the late 1980s to the late 1990s. Most of the northern European sites (from Sweden, Norway, and Finland) showed significant increasing trends for winter months (M. Roeamer, personal communication). This was tentatively attributed to the simultaneous decrease in NO_x-emissions, because NO reacts with ozone (see reaction (3)), although an increase in ozone background concentrations could not be ruled out. Most European sites showed decreasing concentrations in May. The higher percentiles in summer seem to decrease at most series, which are mainly located in the northwest and in central Europe.

VIII. SURFACE OZONE TRENDS IN ASIA

Very high ozone concentrations were measured in the 1970s in the very densely populated agglomeration of Tokyo and even higher concentrations occurred in the 1990s in the surroundings of Seoul (South Korea) (see Table I). An investigation of Wakamatsu (manuscript in preparation) of Kanto including 340 air pollution monitoring sites of Tokyo, the six surrounding prefectures and several remote stations found that peak ozone concentrations significantly decreased in this area over the past decade. Also at low concentrations significant changes occurred. However, annual average ozone concentrations are increasing both at urban and at rural sites in Japan at a rate of approximately 0.5–1.0 ppb per year, which probably reflects the large increase in ozone precursor emission of the Asian continent (see following). At the present time, the ambient air pollutant monitoring network in China is not very dense. The largest concentrations are up to 120 ppb reported from Linan and Hong Kong in spring and fall. They are believed to be caused by ozone precursor emissions in this area.

In North America and Europe, O₃ concentration usually peaks in summer. The Asian monsoon circulation

depresses summertime surface ozone concentrations in India and Eastern China. Extended surface ozone measurements were performed at the western edge of the city Ahmedabad in India (23°N , 72.6°E) from 1954 to 1955. By comparison with later measurements it was concluded that interferences such as SO_2 hardly influenced the measurements, which were performed by the KI-method. The comparison with measurements of 1991–1993 (performed with the UV-method) showed an averaged increase from 14.7 ppb in 1954–1955 to 23.5 ppb in 1991–1993. The increase at 14 pm from 20 ± 6 ppb to 41 ± 12 ppb at a rate of $1.9 \pm 0.04\%$ per year is believed to be caused by the local ozone production. Average ozone concentrations at 21 pm and 6 am increased by $0.49 \pm 0.37\%$ per year during the monsoon season, which was attributed to the increase in background ozone concentrations.

Surface ozone measurements from 1992 to 1997 of Cape Hebdo (26°N , 128°E) at the northern tip of Okinawa Island were analyzed in order to study ozone trends in northeast Asia. The measurements from October to March were selected, because during this time the site was often in the outflow of the continental air masses. The cases which represent polluted air masses advected from the continent were selected by trajectory analyses. From this collective an increasing trend of ozone of $2.6 \pm 2.0\%$ per year was calculated in the polluted air of the continent. An increasing trend of the annual mean values was also found in the background air of Japan (see previous discussion). A further increase in ozone concentrations in east Asia is predicted because the NO_x emissions of the Asian continent are expected to increase by a factor of 3 from 1990 to 2010 if no measures are taken for emission reduction.

IX. SURFACE OZONE TRENDS IN THE TROPICS AND SOUTHERN HEMISPHERE

Large photooxidant concentrations occur in the tropical city, Mexico City, and in Sao Paulo in the southern hemisphere (see Table I). In the tropics biomass burning is an important source of the emission of ozone precursors aside from fossil fuel combustion. It is believed that ozone concentrations as high as 100 ppb observed in South Africa are mainly caused by the primary emission of biomass burning.

Measurements of background ozone in the tropics show either slightly increasing or insignificant trends (see Fig. 6). The long-term ozone measurements of the Mauna Loa observatory in Hawaii show an increasing trend of $0.35 \pm 0.22\%$ per year in the period 1974–1995, when only downslope conditions are considered, whereas the

measurements of the site of American Samoa and the shorter record of Izania on the Canary Islands show slightly decreasing but insignificant trends (see Fig. 6). Two long-term ozone records of remote stations in the southern mid-latitudes indicate small increasing trends over the 1980s and the 1990s (Cape Point in South Africa by $0.53 \pm 0.34\%$ from 1983 to 1995, and Cape Grim in Australia by $0.18 \pm 0.14\%$ per year in the period of 1982 to 1995).

Surface ozone measurements at the South pole at 2835 m amsl. show a significant decrease of $0.70 \pm 0.2\%$ per year in the period of 1975 to 1995. This was attributed to the effect of the ozone hole, which leads to the decreasing supply of ozone from the stratosphere.

X. OZONE TRENDS IN THE FREE TROPOSPHERE

In the 1950s and 1960s it was believed that tropospheric ozone production was only a regional air pollution problem. Paul Crutzen was among the first who suggested at the beginning of the 1970s that ozone in the free troposphere could increase due to anthropogenic ozone precursor emissions. Ozone in the upper troposphere is most difficult to measure because of its low concentration (see Fig. 1). Trend analysis of continuous series are restricted to regular ozone sonde measurements which started in America (Canada), Europe, and Japan in the second half of the 1960s (see Table IV). Measurements of a few continuous LIDAR series which include the upper tropospheric ozone have been available only since the second half of the 1980s. At the present time the availability of the LIDAR data is too sparse and the series are too short to be used in trend analysis.

Table IV shows the sites where ozone sondes are launched. The suitability of the records for long-term ozone trend determination depends on the lengths of the record, the frequency of the measurements (which strongly varies from record to record, see Table IV), and data quality. Some problems are caused by the lack in the homogeneity of the measurements. The change from BM to ECC sondes of the Canadian stations around the end of the 1970s introduced breaks in these series. At Payerne (Switzerland) the launch time was changed several times leading to breaks in ozone in the planetary boundary layer because of its systematical diurnal variation.

The recent update of the ozone long-term trend determination of the sonde records in SPARC yielded the following results. Trends in tropospheric ozone are highly variable and they depend on the region.

The European stations show strong increases of 5 to 25% per decade which are significant for 1970 to 1996.

This is qualitatively in accordance with the few available surface ozone measurements of rural stations (see Section VII) and with the strong increase in European ozone precursor emissions which were particularly large in the 1970s. However, the three long-term European stations show different trends in the period after 1980 and only Payerne shows a significant increase in the 1980s.

The increase at the **Japanese** stations is largest near surface, with 10–15% per decade for 1970–1996, decreasing with altitude. They are insignificant at 9 km at the two northern stations (Sapporo (43°N) and Tskuba (Tateno) (36°N)), and at Kagoshima (32°N) at 12 km. The station at Tskuba, which includes most measurements, yields no tropospheric ozone trends for 1980–1996. Tskuba is strongly affected by the photochemical air pollution in the Tokio area (see Section VIII). After 1980 the ozone increases at Sapporo and Kagoshima are not always significant. Increases in ozone in the planetary boundary layer and the information from surface ozone measurements are in accordance with the very large increase in ozone precursor emissions that occurred in the Asian continent during this period. This was confirmed by the analysis of the ozone sonde measurements of 1989 to 1997 of Naha (Okinawa island), which is located in the Pacific rim region. Ozone concentration of the altitude of 0 to 2 km above surface increased by $2.5 \pm 0.6\%$ per year if the measurements were selected by trajectory analysis to sample the air of the continental outflow of Asia. This result is consistent with Japanese surface measurements (see Section VIII).

No significant ozone trends were found at the stations in the **United States**, at Wallops Island and at Boulder, from where data are only available since the end of the 1970s. Also no significant changes at rural surface sites in the United States were observed since 1980 which agrees with the lack of significant trends in NO_x emissions of the United States during this time (see Section VI).

The **Canadian** stations show decreases or insignificant trends for 1970–1996 and decreases of –2 to –8% per decade in the mid-troposphere in the period of 1980 to 1996. A more recent update including measurements until 1999 reveals less negative tropospheric ozone trends caused by a considerable increase in tropospheric ozone since 1993. The data analysis further indicates a strong correlation between ozone in the lower stratosphere and in the troposphere. This finding suggests that stratospheric ozone depletion causing a decrease in the amount of ozone transported from the stratosphere into the troposphere was possibly an important contribution to the tropospheric decrease of the measurements since 1980.

No significant ozone trend was found in the **tropical** station Hilo in Hawaii which provides regular ozone sondes since 1982.

Ozone is continuously measured only at one station in the **southern hemisphere** (Lauder, New Zealand) since 1986. From 1986 to 1996 increasing tropospheric ozone trends of 5% per decade were found.

We might conclude that the temporal evolution of free tropospheric ozone of the European and the Japanese stations since the late 1960s is qualitatively in accordance with the changes in continental NO_x emissions of Europe and Asia. However, the justification of this comparison might be questioned because of the considerable lifetime of ozone in the free troposphere which can be in the order of many weeks. The relation of vertical and horizontal transport with respect to free tropospheric ozone needs further investigation. Long-term changes in dynamics influencing transport of ozone and its precursors could turn out to be significant for the proper interpretation of interdecadal tropospheric ozone changes.

The database of the ozone sondes is very limited. At several stations the number of sondes launched per month is hardly sufficient for the representative characterization of the monthly mean values making the detection of significant trends a very difficult task.

The anthropogenic ozone precursor emissions of the industrialized countries started to increase rapidly after World War II. Regular ozone sonde programs started only in the late 1960s. From a few airplane measurements performed in December and August 1940 in Bavaria (Germany) and the comparison with the ozone sondes of Hohenpeissenberg, it was concluded that free tropospheric ozone already started to increase before the 1970. Single ascents of ozone sondes were performed at Arosa Switzerland in July and August 1958. The comparison with the ozone sondes of Payerne (Switzerland) suggests an increase by 50% of free tropospheric ozone (500 hPa) between the late 1950s (considering the values of the late 1950s as 100%) and the middle of the 1990s which is much less than the ozone increase in the polluted European planetary boundary layer in the same period (see Section VII).

XI. SUMMARY AND CONCLUSIONS

In the industrialized world photochemical air pollution problems started to occur after World War II. Urban plumes with high ozone concentrations were discovered in many areas in America, Europe, and Japan. National legislation led to a stabilization or a decrease in ozone precursor emissions in the industrialized world (since the late 1970s in the United States and the middle of the 1980s in Europe). Advanced technologies in industrialized countries allowed the stabilization of NO_x emissions despite large increases in traffic volume. The largest decreasing

trend of surface ozone concentrations was achieved in southern California due to very strict emission regulations. Some decreasing surface ozone trends, affected by urban plumes, were also found in other areas in the United States. These findings are in accordance with the decrease in anthropogenic VOC emissions. Decreasing peak ozone values were reported from other areas which are downwind of urban ozone precursor emissions.

Surface ozone measurements of Europe show that background ozone concentrations in the polluted planetary boundary layer of Europe increased by more than a factor of 2 since World War II. European NO_x emissions increased by more than a factor of 4 from World War II to the middle of the 1980s. No consistent picture of positive or negative ozone trends from rural stations in the United States not directly affected by the outflow of large agglomerations have been found in U.S. measurements since the 1980s. This is in agreement with the lack of any significant NO_x emission trends in the same period. Winter surface ozone values of northern European sites show increasing concentrations since the late 1908s, possibly caused by the decreasing NO emissions and/or an increase in background concentrations.

A large increase of surface ozone between the middle of the 1950s and the early 1990 was reported from the city Ahmedabad in India, which was partially attributed to the local photooxidant formation and partially to the increase in background ozone concentration. The very large photooxidant pollution in the agglomeration of Tokyo peaked in the 1970s and decreased thereafter. However, background ozone concentrations in Japan increased in the last decade, attributed to the large increase in ozone precursor emissions of the Asian continent. The emissions of ozone precursors in this region are predicted to continue to increase, because of the further expected rapid growth in economics and increase in population. An increase of NO_x emissions of a factor of 3.5 in the period of 1990 to 2020 was calculated for East Asia if no measures for air pollution reduction are taking place. This large increase is expected to lead to a further increase in background ozone concentrations in the Pacific region.

At the present time, surface ozone measurements in the industrialized world generally provide a sufficient database to monitor changes in ozone concentrations relevant for air pollution. To our knowledge the ozone monitoring program in other parts of the world is much less complete. An extension of the ozone monitoring activities is recommended for the early detection of photooxidant pollution problems downwind of large agglomerations in developing countries. Significant photooxidant air pollution problems are expected for the future in very large agglomerations if no resources for adequate reduction of primary emissions are available. This might be a particu-

lar problem in relation to the increasing demand of road traffic.

Ozone trends in the free and in particular in the upper troposphere are important in the context of the greenhouse gas forcing by ozone. The respective information is very limited in space and time and the picture is much more restricted than for stratospheric ozone depletion. Stratospheric ozone decrease caused by the anthropogenic release of ozone depleting substances started at the beginning of the 1970s, and satellite measurements, which became available at the end of the 1970s, allowed one to monitor stratospheric ozone with almost global coverage. Ozone precursor emissions of the industrialized countries started to increase after World War II, but regular ozone balloon measurements only started in the late 1960s. Nevertheless, tropospheric ozone profile trends observed in Europe and Japan are grossly consistent with the temporal evolution of ozone precursor emissions in these continents since the beginning of the 1970s. Ozone sondes of Northern America show no significant trends in the United States or significant tropospheric decreases in Canada in the period 1980–1996 when American NO_x emissions were approximately constant. Recent updates and analyses of the Canadian station revealed a strong correlation between tropospheric and lower stratospheric ozone suggesting that the stratospheric ozone decrease might have been a significant contribution to the observed tropospheric ozone trends. In the southern mid-latitudes data are only available from one station in New Zealand. It seems too early to speculate about the possible reasons of the observed free tropospheric ozone increase in the period 1986–1996.

The ozone lifetime in the free troposphere is long enough for significant intercontinental transport and therefore the possible changes in hemispheric ozone background are expected to become an important topic in future research. The turnaround of stratospheric ozone trends is expected to occur in the following decades, whereas upper tropospheric ozone in the northern hemisphere is expected to increase further because of the predicted increase in ozone precursor emissions, in particular, of East Asia. Ozone measurements from commercial aircraft in combination with the information of the satellite and LIDAR measurements are expected to contribute to the detection and quantification of these climatically relevant trends.

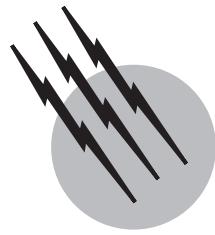
SEE ALSO THE FOLLOWING ARTICLES

ENVIRONMENTAL GEOCHEMISTRY • GREENHOUSE EFFECT AND CLIMATE DATA • NITROGEN CYCLE, ATMOSPHERIC • METEOROLOGY, DYNAMIC (TROPOSPHERE)

- POLLUTION, AIR • REMOTE SENSING FROM SATELLITES
- TROPOSPHERIC CHEMISTRY

BIBLIOGRAPHY

- Adams, R. M., Glycer, J. D., and McCarl, B. A. (1988). The NCLAN Economic Assessment: Approach, Findings and Implications. In "Assessment of Crops Loss from Air Pollutants" (W. W. Heck, O. C. Tayloer, and D. T. Tingey, eds.), pp.473–504, Elsevier Science London/New York.
- Chameides, W. L., Xingsheng, L., Xiaoyan, T., Xiuji, Z., Chao, L., Kiang, C. S., John, J. St., Saylor, R. D., Liu, S. C., Lam, K. S., Wang, T., and Giorgi, F. (1999). "Is ozone pollution affecting crop yields in China?" *Geophysical Res. Lett.* **26**, 867–870.
- Fiori, A. M., Jacob, D. J., Logan, J. A., and Yin, J. H. (1998). "Long-term trends in ground level ozone over the contiguous United States, 1980–1995," *J. Geophys. Res.* **103**, 1471–1480.
- Harris, N. R. P., Ancellet, G., Bishop, L., Hofmann, D. J., Kerr, J. B., McPeters, R. D., Prendez, M., Randel, W. J., Staehelin, J., Subbaraya, B. H., Volz-Thomas, A., Zawodny, J., and Zerefos, C. (1997). "Trends in stratospheric and free tropospheric ozone," *J. Geophys. Res.* **102**, 1571–1590.
- Lee, S. H., Akimoto, H., Nakane, H., Kurnosenko, S., and Kinjo, Y. (1998). "Lower tropospheric ozone trends observed in 1989–1997 at Okinawa, Japan," *Geophys. Res. Lett.* **25**, 1637–1640.
- Lelieveld, J., Thompson, A. M., Diab, R. D., Hov, O., Kley, D., Logan, J. A., Nielsoen, O. J., Stockwell, W. R., and Zhou, X. (1999). Tropospheric Ozone and related Processes, Chapter 8, In "WMO, Scientific Assessment of Ozone Depletion: 1998," Global Ozone Research and Monitoring Project, Rept. 44, Geneva, Switzerland.
- Lippmann, M. (1991). "Health effects of tropospheric ozone," *Environ. Sci. Technol.* **25**, 1954–1961.
- Logan, J. A. (1999). "An analysis of ozonesonde data for the troposphere: Recommendations for testing 3-D models and development of a gridded climatology for tropospheric ozone," *J. Geophys. Res.* **104**, 16.115–16.149.
- Logan, J. A., Megretskia, I. A., Miller, A. J., Tiao, G. C., Choi, D., Zhang, L., Stolarski, R. S., Labow, G. J., Hollandsworth, S. M., Bodeker, G. E., Claude, H., DeMuer, D., Kerr, J. B., Tarasick, D. W., Oltmans, S. J., Johnson, B., Schmidlin, F., Staehelin, J., Viatte, P., and Uchino, O. (1999). "Trends in the vertical distribution of ozone: A comparison of two analyses of ozonesonde data," *J. Geophys. Res.* **104**, 26.373–399.
- Low, P. S., Kelly, P. M., and Davies, T. D. (1992). "Variations in surface ozone trends over Europe," *Geophys. Res. Lett.* **19**, 1117–1120.
- Naja, M., and Lal, S. (1996). "Changes in surface ozone amount and its diurnal and seasonal patterns, from 1954–55 to 1991–93, measured at Ahmedabad (23°N), India," *Geophys. Res. Lett.* **23**, 81–84.
- Oltmans, S. J., Lefohn, A. S., Scheel, H. E., Harris, J. M., Levy, H. II, Galbally, I. E., Brunke, E.-G., Meyer, C. P., Lathrop, J. A., Johnson, B. J., Shadwick, D. S., Cuevas, E., Schmidlin, F. J., Tarasick, D. W., Claude, H., Kerr, J. B., Uchino, O., and Mohnen, V. (1998). "Trends of ozone in the troposphere," *Geophys. Res. Lett.* **25**, 139–142.
- Schere, K. L., Hidy, G. M., and Singh, H. B. (2000). (guest eds.) "The NARSTO Ozone Assessment—Critical Reviews," *Atmospheric Environment* **34**, 1853–2332.
- SPARC/IOC/GAW, Assessment of Trends in the Vertical Distribution of Ozone, edited by N. Harris, R. Hudson and C. Phillips, *SPARC Report No. 1*, WMO Ozone Research and Monitoring Project No. 43, May 1998.
- Staehelin, J., Thudium, J., Buehler, R., Volz-Thomas, A., and Graber W. (1994). "Trends in surface ozone concentrations at Arosa (Switzerland)," *Atmospheric Environment* **28**, 75–87.
- Supercities: Environmental Quality and Sustainable Development (1996). *Atmospheric Environment* **30**, 675–816.
- Thouret, V., Marenco, A., Logan, J. A., Nédélec, P., and Grouhel, C. (1998). "Comparison of ozone measurements from the MOZAIC airborne program and the ozone sounding network at eight locations," *J. Geophys. Res.* **103**, 25, 695–720.
- Volz, A., and Kley, D. (1988). "Evaluation of the Montsouris series of ozone measurements made in the nineteenth century," *Nature* **332**, 240–242.



Radiation, Atmospheric

Knut Stammes

Stevens Institute of Technology

Gary E. Thomas

University of Colorado

- I. Introduction
- II. Sources of Atmospheric Radiation
- III. Atmospheric Vertical Structure
- IV. Basic Radiative Processes and the Radiative Transfer Equation
- V. The Role of Radiation in Climate
- VI. Conclusion

GLOSSARY

Albedo Ratio of the energy reflected to the energy incident on a surface.

Blackbody An ideal hypothetical surface that emits radiation according to Planck's law. The emitted radiation depends only on frequency and temperature.

Greenhouse effect Radiation emitted by the surface is absorbed by atmospheric greenhouse gases. This leads to a warming of the planet called the greenhouse effect.

Kirchhoff's law States that for an opaque surface the absorption coefficient is equal to the emission coefficient. For an extended medium such as the atmosphere, assumed to be in local thermodynamic equilibrium, Kirchhoff's law relates the thermal volume emission coefficient to the Planck function.

Lapse rate The rate at which the temperature decreases with altitude in the atmosphere.

Planck's law By assuming that energy is quantized, Planck arrived at his famous law describing how a body

of a given temperature emits radiation. This marked the beginning of the “quantum revolution” in physics.

Radiative equilibrium If thermal emission is balanced locally by the rate of heating due to absorption at all wavelengths, local radiative equilibrium prevails. Averaging the energy gains due to absorption and the energy losses due to emission over all wavelengths, over the entire planet and over a suitably long time interval, leads to a close balance known as planetary radiative equilibrium.

Radiative forcing The difference between the net incoming solar energy and the energy emitted by the Earth at the top of the atmosphere is called radiative forcing. Note that zero radiative forcing defines planetary radiative equilibrium, which implies a stable climate.

Stefan–Boltzmann's law This law states that the energy emitted by a body at a given temperature is proportional to the fourth power of its temperature. This law can be derived from classical thermodynamics, but an integration of Planck's radiation formula over all wavelengths yields the same result.

LIFE ON EARTH began with light. Solar radiation, illuminating the primordial atmosphere, gave rise to chemical reactions that are the basis for biological evolution. This eventually led to photosynthesis, a prerequisite to life as we know it. Solar radiation continues to sustain life on Earth through interactions with the Earth's atmosphere and its land and ocean surface. While the details of these interactions may be different today than they were when life began on our planet, they are still significant factors influencing the evolution of life on Earth.

Two prominent problems currently receiving much attention in atmospheric and environmental science demonstrate the importance of understanding atmospheric radiation and its processes and effects: the possibility of widespread ozone depletion and the potential for global warming. Both of these problems have the urgency of immediate concern, and the examination of one of them—global warming—provides a useful framework for our discussion of atmospheric radiation.

I. INTRODUCTION

The Earth's climate is determined by a balance between the energy it receives from the Sun, and the energy it emits to space. It is important to realize that essentially all the energy received and emitted by our planet, and thus shaping climate, is electromagnetic or *radiative* in nature. (The geothermal energy production is only a small fraction ($\sim 0.01\%$) of the solar input and can therefore be neglected.) As explained below, one may think of both the Sun and the Earth as a blackbody whose emitted energy is proportional to the fourth power of its temperature. Radiative energy equilibrium (or balance) between these two blackbodies determines our climate. Currently, the radiative energy balance between the Sun and Earth yields an average surface temperature of about 288 K. If the Earth were to receive more (less) energy from the Sun than it can emit to space, increasing (decreasing) its temperature will change the energy emitted until a new radiative energy balance is established between the Sun and the Earth. Likewise, for a constant solar input, if atmospheric composition or the surface emission or reflectance of the Earth were to change the balance between the net energy received from the Sun and that emitted to space, then the temperature of the Earth's atmosphere and surface—the *climate system*—must adjust until a new balance is established.

The net incoming energy is the absorbed solar radiation $(1 - \bar{\rho})\bar{F}^s$, where $\bar{\rho}$ is the overall reflectance of the planet called the *spherical albedo*, and \bar{F}^s is the average solar irradiance falling on Earth. The difference, \bar{N} , between the net incoming solar energy and the energy emitted by the Earth at the top of the atmosphere

(TOA), $\bar{N} = (1 - \bar{\rho})\bar{F}^s - \bar{F}_{TOA}$, where \bar{F}_{TOA} denotes energy emitted to space, is called the *radiative forcing*. Thus, our energy balance requirement ensures that the time- and space-averaged radiative forcing, $\langle \bar{N} \rangle$, is close to zero. In fact, imposing $\langle \bar{N} \rangle = 0$ defines planetary *radiative equilibrium*, which implies a stable or unchanging climate.

The bulk of the Earth's atmosphere (99% by mass) consists of molecular nitrogen and oxygen, which are radiatively inactive (homonuclear, diatomic) molecules that have negligible impact on the radiative energy balance. Trace amounts of polyatomic molecules are responsible for atmospheric absorption and emission of radiation in several hundred thousands of individual spectral lines arising from rotational and vibrational transitions. Of these trace gases, water vapor, carbon dioxide, and ozone are the main absorbers (and emitters) contributing to warming and cooling of the atmosphere and underlying surface. These so-called greenhouse gases strongly absorb and emit infrared radiation and thereby trap radiative energy that would otherwise escape to space. The bulk of the ozone gas resides in the stratosphere. It interacts with ultraviolet/visible radiation as well as with thermal infrared (terrestrial) radiation. A thinning of the stratospheric ozone layer renders the stratosphere more transparent in the $9.6 \mu\text{m}$ region, thereby allowing more transmission and less backwarming of surface emission. Thus, ozone depletion cools the surface and tends to partially mitigate warming from increased concentration of carbon dioxide. Ozone is very important because it absorbs ultraviolet radiation that is harmful to life on our planet. Thus, a thinning of the ozone layer will have serious biological implications. A discussion of these implications is, however, beyond the scope of the present article.

In the remainder of this article we will focus on the global warming issue and discuss how radiative interactions with the atmospheric greenhouse gases impact the radiative energy balance of the Earth and hence climate. They warm our planet by absorbing radiation emitted by the surface; without them, the Earth would be some 33°C colder than at present and therefore uninhabitable. Hence, the greenhouse effect is very important for life itself. In addition to water vapor, carbon dioxide, and ozone, several other atmospheric trace gases, notably chlorofluorocarbons and methane, are infrared-active. These additional greenhouse gases make smaller contributions to warming/cooling of the atmosphere and surface. Some have natural origins, while others are partially (such as methane) or wholly (such as the chlorofluorocarbons) anthropogenic. An estimate of the radiative forcing caused by selected greenhouse gases is provided in Fig. 1.

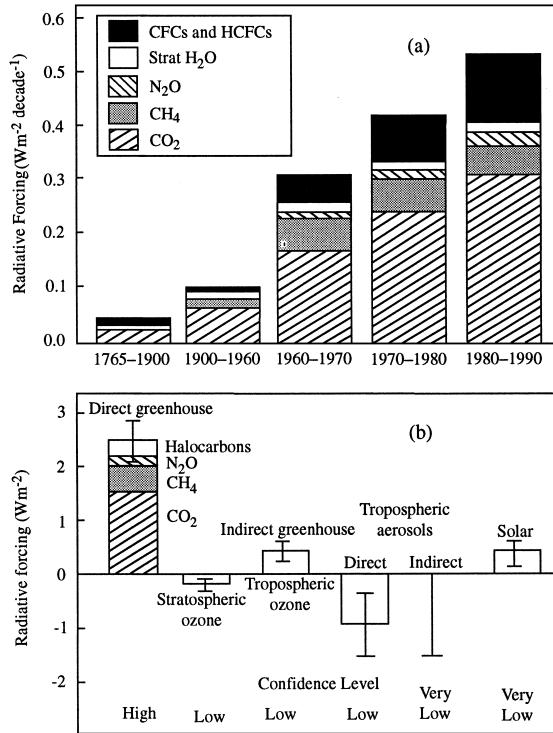


FIGURE 1 (a) The degree of radiative forcing produced by selected greenhouse gases in five different epochs. Until about 1960, nearly all the forcing was due to CO₂; today, the other greenhouse gases combined nearly equal the CO₂ forcing. (b) Estimates of the globally averaged radiative forcing due to changes in greenhouse gases and aerosols from pre-industrial times to the present day and changes in solar variability from 1850 to the present day. The height of the bar indicates a mid-range estimate of the forcing. The lines show the possible range of values. An indication of relative confidence levels in the estimates is given below each bar.

II. SOURCES OF ATMOSPHERIC RADIATION

The spectral variable is the wavelength $\lambda = c/v$, where c is the speed of light and v is the frequency [s⁻¹] or [Hz]. In the infrared (IR), λ is usually expressed in micrometers (or more commonly microns, where $1 \mu\text{m} = 10^{-6} \text{ m}$). In the ultraviolet (UV) and visible spectral range, λ is expressed in nanometers ($1 \text{ nm} = 10^{-9} \text{ m}$).

In Fig. 2, we show the spectral irradiance of the Sun's radiative energy measured onboard an Earth-orbiting satellite, beyond the influences of the atmosphere. Integration over all frequencies yields the total solar irradiance, $S \approx 1368 \text{ W} \cdot \text{m}^{-2}$, which is the basic "forcing" of the Earth's "heat engine."

Also shown in Fig. 2 are spectra of an ideal blackbody at several temperatures. Requiring that the total energy emitted be the same as a blackbody, one finds that the Sun's effective temperature is 5778 K. If the radiating

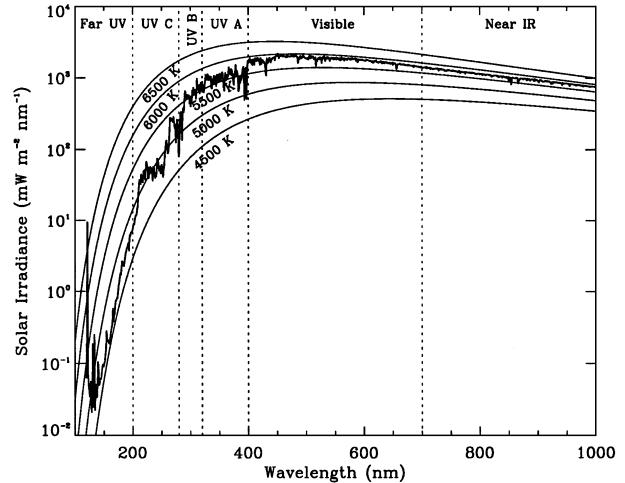


FIGURE 2 Extraterrestrial solar irradiance, measured by a spectrometer onboard an Earth-orbiting satellite. The UV spectrum ($119 < \lambda < 420 \text{ nm}$) was measured by the SOLSTICE instrument on the UARS satellite (modified from a diagram provided by G. J. Rottmann, private comm. 1995). The vertical lines divide the various spectral subranges defined in Table I. The smooth curves are calculated blackbody spectra for a number of emission temperatures.

layers of the Sun had a uniform temperature at all depths, its spectrum would indeed match one of the theoretical blackbody curves exactly. The interesting deviations seen in the solar spectrum can be said to be a result of emission from a nonisothermal atmosphere.

We can explain the visible solar spectrum qualitatively by considering two characteristics of atmospheres: (1) their absorption opacity $\tau(\nu)$ depends upon frequency, and (2) their temperature varies with atmospheric depth, and one basic rule—that a radiating body emits its energy to space most efficiently at wavelengths where the opacity is approximately unity. This rule is explained in terms of the competing effects of absorption and emission. In spectral regions where the atmosphere is transparent ($\tau(\nu) \ll 1$), it neither emits nor absorbs efficiently. In contrast, where it is opaque ($\tau(\nu) \gg 1$), its radiative energy is prevented from exiting the medium; that is, it is reabsorbed by surrounding regions. At $\tau(\nu) \approx 1$, a balance is struck between these opposing influences.

An understanding of radiative transfer is also essential for understanding the energy output of the Earth, defined to be the spectral region $\lambda > 3.5 \mu\text{m}$. Figure 3 shows the IR emission spectrum measured from a down-looking orbiting spacecraft, taken at three different geographic locations. Also shown are blackbody curves for typical terrestrial temperatures. The spectral variable in this case is wavenumber $\tilde{\nu} = 1/\lambda$, commonly expressed in units of cm⁻¹. Again, as for the solar spectrum, the deviations are attributed to the nonisothermal character of the Earth's

TABLE I Subregions of the Spectrum (adapted from Thomas and Stamnes, 1999)

Subregion	Range	Solar variability	Comments
X-rays	$\lambda < 10 \text{ nm}$	10–100%	Photoionizes all thermosphere species.
Extreme UV	$10 < \lambda < 100 \text{ nm}$	50%	Photoionizes O ₂ and N ₂ . Photodissociates O ₂ .
Far UV	$100 < \lambda < 200 \text{ nm}$	7–80%	Dissociates O ₂ . Discrete electronic excitation of atomic resonance lines.
Middle UV, or UV-C	$200 < \lambda < 280 \text{ nm}$	1–2%	Dissociates O ₃ in intense Hartley bands. Potentially lethal to biosphere.
UV-B	$280 < \lambda < 320 \text{ nm}$	<1%	Some radiation reaches surface, depending on O ₃ optical depth. Damaging to biosphere. Responsible for skin erythema.
UV-A	$320 < \lambda < 400 \text{ nm}$	<1%	Reaches surface. Benign to humans. Scattered by clouds, aerosols, and molecules.
Visible, or PAR ^a	$400 < \lambda < 700 \text{ nm}$	$\leq 0.1\%$	Absorbed by ocean, land. Scattered by clouds, aerosols, and molecules. Primary energy source for biosphere and climate system.
Near IR	$0.7 < \lambda < 3.5 \mu\text{m}$		Absorbed by O ₂ , H ₂ O, CO ₂ in discrete vibrational bands.
Thermal IR	$3.5 < \lambda < 100 \mu\text{m}$		Emitted and absorbed by surfaces and IR active gases.

^a Photosynthetically active radiation.

atmosphere. The spectral regions of minimum emission arise from the upper cold regions of the Earth's troposphere where the opacity of the overlying regions is ~ 1 . Those of highest emission originate from the warm sur-

face in transparent spectral regions ("windows"), with the exception of the Antarctic spectrum, where the surface is actually colder than the overlying atmosphere (see Fig. 3). In this somewhat anomalous situation, the lower opacity region is one of higher radiative emission because of the greater rate of emission of the warm air. Again, the deviations from blackbody behavior can be understood qualitatively in terms of the temperature structure of the Earth's atmosphere and the variation with frequency of the IR absorption opacity.

As shown in Fig. 4 there is little overlap between the radiation spectra of the Sun and the Earth. Therefore, except for applications where the region of overlap ($3\text{--}4 \mu\text{m}$) is of special interest, we may treat the two spectra separately. We note the absence of strong absorption by the major atmospheric gases throughout the visible spectrum. The major shortwave interaction is in the UV spectrum below 300 nm where sunlight never reaches the surface, being absorbed in the middle atmosphere by ozone.

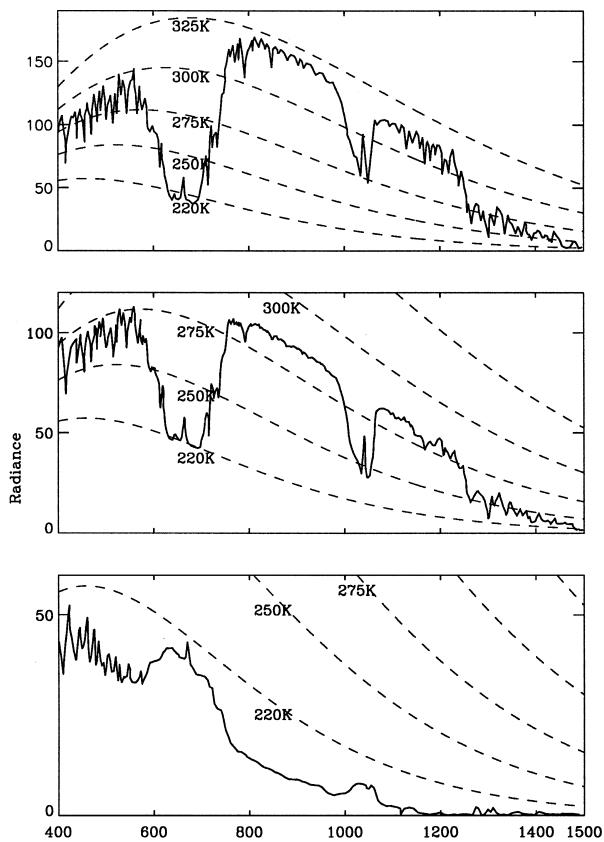


FIGURE 3 Thermal emission spectra of Earth measured by the IRIS Michelson interferometer instrument on the Nimbus 4 spacecraft. Shown also are the radiances of blackbodies at several temperatures. (top) Sahara region, (middle) Mediterranean, (bottom) Antarctic.

III. ATMOSPHERIC VERTICAL STRUCTURE

The stratified vertical structure of the bulk properties of an atmosphere is a consequence of hydrostatic balance. For an atmosphere in a state of rest, the pressure, p , must support the weight of the atmosphere above it. By equating pressure forces and gravitational forces, one finds that $dp = -g\rho dz$, where g is the acceleration due to gravity, ρ is the air density, and dp is the differential change in pressure over the small height interval dz . Combining this equation with the ideal gas law $\rho = \bar{M}p/RT = \bar{M}n$, one finds upon integration:

$$p(z) = p_0(z) \exp \left[- \int_{z_0}^z dz / H(z) \right]$$

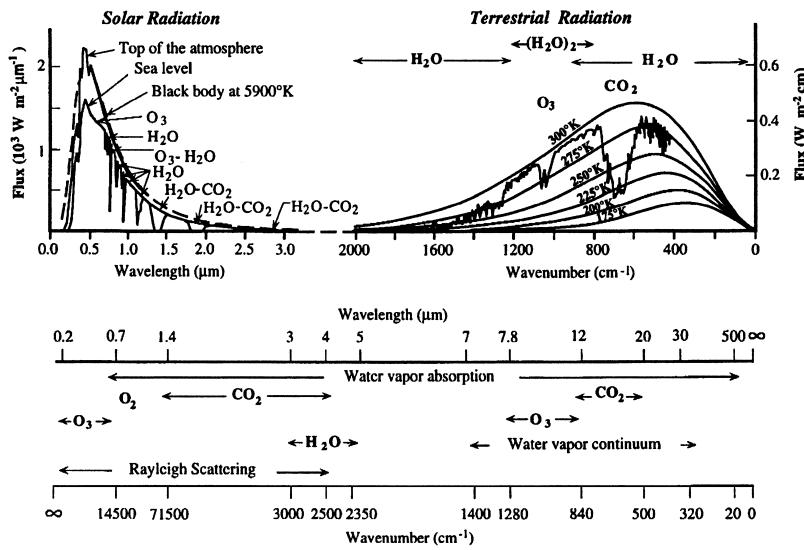


FIGURE 4 Spectral distribution of solar (shortwave) and terrestrial (longwave) radiation fields. Also shown are the approximate shapes and positions of the scattering and absorption features of the Earth's atmosphere.

where the atmospheric scale height $H = RT/\bar{M}g$. The ideal gas law allows us to write similar expressions for the density, ρ , and the concentration, n . Clearly, from a knowledge of surface pressure $p(z_0)$ and the variation of scale height $H(z)$ with height, z , the hydrostatic equation above allows us to determine the bulk gas properties at any height. This hydrostatic equation applies to gases that are well mixed, but not to short-lived species such as ozone, which is chemically destroyed or created, and water, which undergoes phase changes on short time scales.

IV. BASIC RADIATIVE PROCESSES AND THE RADIATIVE TRANSFER EQUATION

A. Definitions of Irradiance and Radiance

The net radiative energy flow, or power, per unit area within a small frequency range ν to $\nu + d\nu$ is called the spectral *net irradiance*. The spectral net irradiance F_ν is the net energy d^3E crossing a surface element dA (with unit normal \hat{n}) per unit time and per unit frequency:

$$F_\nu = \frac{d^3E}{dA dt d\nu} \quad [\text{W} \cdot \text{m}^{-2} \cdot \text{Hz}^{-1}]$$

The irradiance is positive if the energy flow is into the hemisphere centered on the direction \hat{n} and negative if the flow is into the opposite hemisphere. Thus, we may define spectral *hemispherical* irradiances $F_\nu^+ = d^3E^+/dA dt d\nu$ and $F_\nu^- = d^3E^-/dA dt d\nu$, so that the spectral net irradiance becomes $F_\nu = F_\nu^+ - F_\nu^-$. Integration over all frequencies yields the *net irradiance* $F = \int_0^\infty d\nu F_\nu$ [$\text{W} \cdot \text{m}^{-2}$].

More specific information on *directional dependence* of the energy flow is obtained by considering any small subset of the energy d^4E that flows within a solid angle $d\omega$ around direction $\hat{\Omega}$ in the time interval dt and within the frequency range $d\nu$. If this subset of radiation has passed through a surface element dA (with unit normal \hat{n}), then the energy per unit area per unit solid angle, per unit frequency, and per unit time defines the *spectral radiance* I_ν :

$$I_\nu = \frac{d^4E}{\cos \theta dA dt d\omega d\nu} \quad [\text{W} \cdot \text{m}^{-2} \cdot \text{sr}^{-1} \cdot \text{Hz}^{-1}]$$

where θ is the angle between the surface normal \hat{n} and the direction of propagation $\hat{\Omega}$. It is clear from these definitions that $F_\nu^+ = \int_+ d\omega \cos \theta I_\nu$ and $F_\nu^- = -\int_- d\omega \cos \theta I_\nu$, where the subscript (+) on the integral sign denotes integration over the hemisphere defined by $+\hat{n}$, and (-) denotes integration over the hemisphere defined by $-\hat{n}$. Thus, the spectral net irradiance can be expressed as: $F_\nu = F_\nu^+ - F_\nu^- = \int_{4\pi} d\omega \cos \theta I_\nu$.

B. Absorption, Scattering, and Extinction by Molecules and Particles

A beam of light incident on a thin atmospheric layer will interact with matter in that layer. It is found experimentally that if the layer has thickness ds , then the light is attenuated so that the differential loss in radiance is $dI_\nu = -k I_\nu ds$. Thus, the loss is proportional to the incident light and k is called the *extinction coefficient*. Integration along the beam path through the layer yields:

$$I_v(s, \hat{\Omega}) = I_v(0, \hat{\Omega}) \exp[-\tau_s(v)]. \quad (1)$$

Here, $\hat{\Omega}$ denotes the propagation direction of the beam, and the dimensionless extinction optical path or opacity along the path s is given by $\tau_s(v) \equiv \int_0^s ds' k(v)$. Attenuation of a light beam in a specific direction can be caused by either absorption or scattering. The extinction optical path of a mixture of scattering/absorbing molecules and particles is defined as the sum of the individual scattering optical path, $\tau_{sc}(v)$, and the absorption optical path, $\tau_a(v)$. Thus, $\tau_s(v) = \tau_{sc}(v) + \tau_a(v)$, where $\tau_{sc}(v) = \sum_i \int_0^s ds' \sigma^i(v, s')$ and $\tau_a(v) = \sum_i \int_0^s ds' \alpha^i(v, s')$. The sum is over all optically active species, and σ^i and α^i are the scattering and absorption coefficients. These can be defined as $\sigma^i(v, s) = \sigma_m^i(v, s) \rho_i(s) = \sigma_n^i(v, s) n_i(s)$ and $\alpha^i(v, s) = \alpha_m^i(v, s) \rho_i(s) = \alpha_n^i(v, s) n_i(s)$, where ρ_i and n_i are the mass densities and concentrations, respectively, of the i th optically active species (either molecule or particle). The quantities α_m^i (σ_m^i) and α_n^i (σ_n^i) are the mass absorption (scattering) coefficient and the absorption (scattering) cross section of the i th constituent (molecule or particle), respectively.

C. Angular Scattering by Molecules and Particles

The total scattering cross section, σ_n , is defined as the total power per unit area scattered in all directions divided by the power per unit area of the incident beam. Similarly, the scattered power per unit area per steradian (sr) in a particular direction of observation divided by the power per unit area of the incident beam is called the *angular scattering cross section*, $\sigma_n(\Theta)$. Here, the angle Θ between the directions of incidence $\hat{\Omega}'$ and observation $\hat{\Omega}$ is given by $\cos \Theta = \hat{\Omega}' \cdot \hat{\Omega} = \cos \theta \cos \phi + \sin \theta \sin \phi \cos(\phi - \theta)$. Thus, in spherical coordinates, θ' and ϕ' are the polar and azimuthal angles of the incident beam, and θ and ϕ those of the scattered beam. For an optically active species (molecule or particle) of number density n_i [m^{-3}], $\sigma_n^i(\cos \Theta)$ [$\text{m}^2 \cdot \text{sr}^{-1}$] is the angular scattering cross section per particle and $\sigma(\cos \Theta) = \sum_i n_i \sigma_n^i(\cos \Theta)$ [$\text{m}^{-1} \cdot \text{sr}^{-1}$] is the angular scattering coefficient. The sum extends over all optically-active species in the medium. We define the phase function as the normalized angular scattering cross section:

$$p(\cos \Theta) = \frac{\sigma(\cos \Theta)}{\int_{4\pi} d\omega \sigma(\cos \Theta) / 4\pi}$$

which has the normalization $\int_{4\pi} d\omega p(\cos \Theta) / 4\pi = 1$.

Treating a molecule as a classical harmonic oscillator, one finds that the phase function for molecular (Rayleigh) scattering becomes:

$$p_{Ray} = \frac{3}{4}(1 + \cos^2 \Theta)$$

For particles not small compared with the wavelength, the interaction properties must be found by solving a complicated boundary-value problem for the electric and magnetic fields. For spherical particles, this theory is well established and the scattering and extinction cross sections are calculated from the multipole expansions.

D. Longwave Absorption and Emission by the Surface and the Atmosphere

The Earth's atmosphere is in contact with land and ocean surfaces, which vary greatly in their visible-light reflectance and absorptance properties. In many applications, their strong continuous absorption in the IR allows them to be treated as thermally emitting blackbodies.

An ideal black surface emits radiation according to Planck's law:

$$I_v^{BB} = B_v(T) \equiv \frac{m_r^2}{c^2} \frac{2hv^3}{(e^{hv/k_B T} - 1)}$$

where h is Planck's constant, c is the speed of light, m_r is the real index of refraction, and k_B is Boltzmann's constant. The frequency-integrated hemispherical irradiance leaving a black surface is given by the *Stefan–Boltzmann law* $F^{BB} = \int_0^\infty dv \int_{2\pi} d\omega \cos \theta I_v^{BB} = \pi \int_0^\infty dv B_v(T) = \sigma_B T^4$ where $\sigma_B = 2\pi^5 k_B^4 / 15h^3 c^2 = 5.6703 \times 10^{-8}$ [$\text{W} \cdot \text{m}^{-2} \cdot \text{K}^{-4}$] is the Stefan–Boltzmann constant.

The spectral *directional emittance* is defined as the ratio of the energy emitted by a surface of temperature T_s to the energy emitted by a blackbody at the same frequency and temperature $\epsilon(v, \hat{\Omega}, T_s) \equiv I_{ve}^+(\hat{\Omega}) \cos \theta d\omega / B_v(T_s) \cos \theta d\omega = I_{ve}^+(\hat{\Omega}) / B_v(T_s)$. In general, ϵ depends upon the direction of emission, the surface temperature, and the frequency of the radiation, as well as other physical properties of the surface (index of refraction, chemical composition, texture, etc.). A surface for which ϵ is unity for all $\hat{\Omega}$ and v , is a *blackbody*, by definition. A hypothetical surface for which $\epsilon = \text{constant} < 1$ for all frequencies is a *graybody*. Similarly, we define the *spectral directional absorptance* as the ratio of absorbed energy to incident energy of the beam $\alpha(v, -\hat{\Omega}', T_s) \equiv I_{va}^-(\hat{\Omega}') \cos \theta' d\omega' / I_v^-(\hat{\Omega}') \cos \theta' d\omega' = I_{va}^-(\hat{\Omega}') / I_v^-(\hat{\Omega}')$. *Kirchhoff's law* states that for an opaque surface $\alpha(v, -\hat{\Omega}, T_s) = \epsilon(v, \hat{\Omega}, T_s)$. Finally, Kirchhoff's law for an extended medium such as the atmosphere relates the *thermal volume emission coefficient* j_v^{th} to the Planck function (assuming local thermodynamic equilibrium):

$$j_v^{th} = \alpha(v) B_v(T)$$

E. Shortwave Surface Reflection and Transmission

Knowledge of the visible reflectance of underlying land and ocean surfaces is necessary for calculating the diffuse radiation field. The reflectance and transmittance depend upon both the angles of incidence and reflection or transmission. For an angular beam of radiation with radiance $I_v^-(\hat{\Omega}')$ within a cone of solid angle $d\omega'$ around $\hat{\Omega}'$, the energy incident on a flat surface with normal \hat{n} is $I_v^-(\hat{\Omega}') \cos \theta' d\omega'$, where θ' is the angle between $\hat{\Omega}'$ and \hat{n} . Denoting by $dI_{vr}^+(\hat{\Omega})$ the radiance of reflected light leaving the surface within a cone of solid angle $d\omega$ around the direction $\hat{\Omega}$, we define the *bidirectional reflectance distribution function* (BRDF) as the ratio of the reflected radiance to the energy in the incident beam $\rho(v, -\hat{\Omega}', \hat{\Omega}) \equiv dI_{vr}^+(\hat{\Omega}) / I_v^-(\hat{\Omega}') \cos \theta' d\omega'$. Adding the contributions to the reflected radiance in the direction $\hat{\Omega}$ from beams incident on the surface in all downward directions, we obtain the total reflected radiance:

$$I_{vr}^+(\hat{\Omega}) = \int dI_{vr}^+(\hat{\Omega}) = \int_{2\pi} d\omega' \cos \theta' \rho(v, -\hat{\Omega}', \hat{\Omega}) I_v^-(\hat{\Omega}')$$

Thus, the reflected radiance is the integral of the energy in each incident direction times the BRDF for that particular combination of incidence and observation angles under consideration. The BRDF, or in short the *reflectance*, plays a central role in the remote sensing of planetary surfaces and is important for the correct assessment of their albedo.

Purely diffuse reflection occurs at microscopically irregular surfaces, while purely specular reflection occurs when the surface is perfectly smooth, like a mirror. If the reflected radiance from a surface is completely uniform with angle of observation, it is called a *Lambert surface*. The BRDF for a Lambert surface is independent of both the direction of incidence and the direction of observation. Then, the reflectance simplifies to $\rho(v, -\hat{\Omega}', \hat{\Omega}) = \rho_L(v)$, where ρ_L is the *Lambert reflectance*. Specular reflection from and transmission through a smooth dielectric surface can be calculated from *Snell's law* and *Fresnel's equations*, given the optical constants of air and the dielectric material.

F. The Equation of Radiative Transfer

For an atmospheric layer of thickness ds that not only attenuates but also emits radiation, we find that the differential change in radiance becomes $dI_v = -kI_v ds + j_v ds$ or $dI_v/d\tau_s = -I_v + S_v$ where the source function, S_v , is the sum of two terms: $S_v = j_v^{th}/k(v) + j_v^{sc}/k(v) = S_v^{sc} + S_v^{th}$. Here, the volume emission coefficient for scattering is $j_v^{sc} = \sigma(v) \int_{4\pi} (d\omega/4\pi) p(\hat{\Omega}', \hat{\Omega}) I_v(\hat{\Omega}')$, while (from Kirchhoff's law) that for thermal emission is just $j_v^{th} =$

$\alpha(v)B_v(T)$. Thus, the complete radiative transfer equation, which includes both multiple scattering and absorption, becomes:

$$\frac{dI_v}{d\tau_s} = -I_v + [1 - \alpha(v)]B_v(T) \\ + \frac{\alpha(v)}{4\pi} \int_{4\pi} d\omega' p(\hat{\Omega}', \hat{\Omega}) I_v(\hat{\Omega}') \quad (2)$$

where $a(v) = \sigma(v)/k(v)$, the *single-scattering albedo*.

Equation (2) has a simple physical interpretation: The term on the left side is the change in the radiance along the path ds , the first term on the right side is the loss of radiation due to extinction, the second term is the gain due emission, and the third term is the gain due to multiple scattering.

G. Plane Parallel or Slab Geometry

The Earth's atmosphere is inhomogeneous, both in the vertical and the horizontal. The horizontal variation is caused by nonuniform distributions of aerosols and cloud fields. For the subsequent discussion we shall assume that atmospheric properties varies only in the vertical direction z . Because we are interested primarily in energy transfer, rather than the directional dependence of the radiation, it is sufficient to work with the azimuthally averaged radiance $I_v(z, \mu)$, where z denotes the height in the atmosphere, and $\mu = \cos \theta$, θ being the polar angle. It is convenient to split the radiation field into two parts: (1) the *direct* solar beam, which is exponentially attenuated upon passage through the atmosphere and ocean, and (2) the *diffuse* or scattered radiation.

According to Eq. (1), the penetration of the direct solar beam through the atmosphere may be written (dropping the v -subscript) as $I^s(z) = I^s(\tau(z)) = F^s e^{-\tau_s}$. Here, F^s is the solar irradiance (normal to the solar beam direction) incident at the top of the atmosphere. We introduce the *vertical optical depth* τ defined as $\tau(z) \equiv \tau_s / Ch(\tau(z), \mu_0) \equiv \sum_i \int_z^\infty dz' k(v, z')$, or $d\tau(z) = -k(v, z) dz$, where $Ch(\tau, \mu_0)$ is a geometrical factor required for a curved atmosphere, which is unity when the Sun is overhead. Unless the Sun is close to the horizon, it is sufficient to use plane geometry for which $Ch(\tau, \mu_0) = 1/\mu_0$, where $\mu_0 = \cos \theta_0$, and θ_0 is the solar zenith angle. The transfer of diffuse radiation through a stratified atmosphere is described by ($\mu = \cos \theta$):

$$\mu \frac{dI(\tau, \mu)}{d\tau} = I(\tau, \mu) - [1 - a(\tau)]B(\tau) \\ - \frac{a(\tau)}{2} \int_{-1}^1 du' p(\tau, \mu', \mu) I(\tau, \mu') - S^*(\tau, \mu) \quad (3)$$

The last term, $S^*(\tau, \mu) = (a F^s / 4\pi) p(\tau, \mu_0, \mu) e^{-\tau/\mu_0}$, is the solar pseudo-source proportional to the attenuated solar beam, which “drives” the diffuse radiation. In the plane-parallel approximation (or *plane geometry*) we ignore the curvature of the atmosphere and assume that its optical properties vary only in the vertical.

H. Heating Rate

The rate at which radiation exchanges energy with matter is expressed in terms of the *spectral radiative heating rate* $\mathcal{H}_v = - \int_{4\pi} d\omega (dI_v/ds)$, which is (minus) the rate of change of the radiative energy per unit volume. In plane geometry, if we substitute $dI_v/d\tau_s = -\mu dI_v/d\tau$ in the radiative transfer equation, Eq. (2), use $k(v) = \sigma(v) + \alpha(v)$, and integrate over the sphere, we obtain $\mathcal{H}_v = -\partial F_v/\partial z = 4\pi\alpha(v)[\bar{I}_v - B_v(T)]$. Here, $F_v = F_v^+ - F_v^-$ is the spectral net irradiance in the z direction, and \bar{I}_v is the angular average of the radiance. Integration over frequencies therefore yields:

$$\begin{aligned}\mathcal{H} &= - \int_0^\infty dv \frac{\partial F_v}{\partial z} = 4\pi \int_0^\infty dv \alpha(v) \bar{I}_v \\ &\quad - 4\pi \int_0^\infty dv \alpha(v) B_v(T)\end{aligned}$$

Thus, the radiative heating rate is the rate at which radiative energy is absorbed less the rate at which radiative energy is emitted.

When $\mathcal{H} = 0$, the volume absorption rate exactly balances the volume emission rate, and a state of *radiative equilibrium* is said to exist. When $\mathcal{H} \neq 0$ an imbalance of heating occurs. The rate of change in the thermal energy per unit volume of a gas which is free to expand against its surroundings is therefore given by $\rho c_p (\partial T / \partial t)_p$. Thus, in the absence of other heating or cooling processes, the rate of change of temperature of a parcel at constant pressure, which we call the warming rate \mathcal{W} , is

$$\mathcal{W} = \left(\frac{\partial T}{\partial t} \right)_p = -\frac{\mathcal{H}}{\rho c_p} = 86,400 \frac{\mathcal{H}}{\rho c_p} \quad [\text{K per day}]$$

as there are 86,400 seconds per day.

V. THE ROLE OF RADIATION IN CLIMATE

The most important radiative interaction in the Earth’s system is the greenhouse effect, without which the Earth would be so cold that it would probably be in a state of permanent glaciation. Water vapor is the most important greenhouse gas, but the well-documented increase in CO₂ abundance, above what is believed to be the natural level existing in the pre-industrial era, has been a matter of considerable concern, because enhanced levels of CO₂ (and

other greenhouse gases) absorb and trap terrestrial radiation that would otherwise escape to space. This *extra* absorption causes an imbalance between the energy received and emitted by the planet. As explained in the introduction, if the planet were to receive more energy from the Sun than it is able to emit to space, then an increase in its temperature will increase the energy emitted (by the Stefan–Boltzmann law) until a new radiative equilibrium between the Sun and Earth is established. Hence, this additional trapping of terrestrial radiation by the enhanced levels of greenhouse gases is expected to lead to a warming so as to make the net energy emitted by the planet equal to that received.

A. Radiative Equilibrium

This simplest useful approach to energy balance is based on the assumption that the atmosphere has negligible absorption for visible radiation. The surface is reflective in the visible, and assumed to be black in the IR. Thus, the surface is heated by incoming solar radiation and by downward IR radiation from the atmosphere. The atmosphere is heated by IR radiation, emitted by both the surface and by surrounding atmospheric layers, and it radiates to space with a globally averaged effective temperature T_e determined by the overall energy balance. For a rotating planet, one finds by equating the incoming solar energy, $S(1 - \bar{\rho})/4$, to the energy leaving the planet, $F_{TOA} = \sigma_B T_e^4$, that the effective temperature is given by:

$$T_e = \left[\frac{S(1 - \bar{\rho})}{4\sigma_B} \right]^{1/4}$$

where S is the total solar irradiance (1368 W · m⁻²), $\bar{\rho}$ is the spherical albedo (30%), and σ_B (5.67×10^{-8} [W · m⁻² · K⁻⁴]) is the Stefan–Boltzmann constant. With no “blanketing” atmosphere to trap radiation emitted by the surface, the effective temperature is equal to the surface temperature. For the Earth, the effective temperature is $T_e = 255$ K, or -18°C .

An algebraic expression for the atmospheric temperature profile may be found by solving a simplified differential equation of infrared radiative transfer. Some of the assumptions may seem extreme but are appropriate for a conceptual model. Nevertheless, the end result is instructive and even somewhat realistic for a cloud-free planetary atmosphere.

Assumptions: (1) the IR radiation interacts only with a single absorbing gas, which absorbs equally at all IR wavelengths (i.e., it is a gray absorber with absorption coefficient α); (2) no scattering occurs ($a = 0$); (3) the atmosphere has plane geometry, in which the optical properties vary only in the vertical direction; and (4) the IR radiation

field ($\lambda > 3.5 \mu\text{m}$) is spectrally separate from the visible radiation field ($\lambda < 3.5 \mu\text{m}$).

Under assumptions (1) and (2), the source function is identical to the Planck function and may be integrated over frequency to yield the Stefan–Boltzmann expression:

$$S(\tau) = \int_0^\infty d\nu B_\nu[T(\tau)] = \sigma_B T^4(\tau)/\pi. \quad (4)$$

In the above, the IR optical depth is the vertical position variable, $\tau(z) = \int_z^\infty dz \alpha(z)$. Setting $a=0$ in Eq. (3) and integrating over IR frequencies, we find that the radiative transfer equation simplifies to:

$$\mu \frac{dI(\tau, \mu)}{d\tau} = I(\tau, \mu) - S(\tau) \quad (5)$$

where I is the azimuthally averaged radiance, integrated over all IR frequencies. Integration over all solid angles yields the IR irradiance:

$$\begin{aligned} F(\tau) &= \int_0^{2\pi} d\phi \int_0^\pi d\theta \sin \theta \cos \theta I(\tau, \theta) \\ &= 2\pi \int_{-1}^1 d\mu \mu I(\tau, \mu) = 2\pi \int_0^1 d\mu \mu [I^+(\tau, \mu) \\ &\quad - 2\pi \int_0^1 d\mu \mu I^-(\tau, \mu)] \end{aligned} \quad (6)$$

where I^\pm denotes the radiances directed into the upper (+) and lower (−) hemispheres, respectively. If we also integrate Eq. (5) over a solid angle, we obtain an expression for the irradiance derivative:

$$\int_0^{2\pi} d\phi \int_{-1}^1 d\mu \mu \frac{dI(\tau, \mu)}{d\tau} = \frac{dF(\tau)}{d\tau} = 4\pi [\bar{I}(\tau) - S(\tau)] \quad (7)$$

where $\bar{I}(\tau)$ denotes the angular average of the radiance, given by:

$$\begin{aligned} \bar{I} &= \frac{1}{4\pi} \int_0^{2\pi} d\phi \int_{-1}^{+1} d\mu \mu I(\tau, \mu) \\ &= \frac{1}{2} \int_0^1 d\mu I^+(\tau, \mu) + \frac{1}{2} \int_0^1 d\mu I^-(\tau, \mu) \end{aligned} \quad (8)$$

We now introduce a fifth assumption critical to determining an analytic solution to Eq. (5): We assume that the radiance is slowly varying with the variable μ , so that we can replace the entire angular distribution (at a given level) with two values, I^+ and I^- . This is the gist of the famous *two-stream approximation*. It allows us to make the following simplifications:

$$\begin{aligned} I(\tau, \mu) &= I^+(\tau) & \text{if } \mu > 0 & \text{and} \\ I^-(\tau) & \text{if } \mu < 0 & & (9) \\ F(\tau) &\approx 2\pi \bar{\mu} [I^+(\tau) - I^-(\tau)] & \text{and} \end{aligned}$$

$$\bar{I}(\tau) \approx \frac{1}{2} [I^+(\tau) + I^-(\tau)] \quad (10)$$

Here, $\bar{\mu}$ is the absolute value of the cosine of the mean inclination of the rays. Its value is somewhat arbitrary but usually varies between 1/2 and 1/ $\sqrt{3}$.

The final assumption, (6), is that of *radiative equilibrium*. Here, we assume that the volume absorption rate of IR radiation ($\alpha \bar{I}$) is equal to the volume rate of emission ($\alpha \sigma_B T^4/\pi$). This was discussed in Section IV in terms of the net heating rate \mathcal{H} , which is zero in radiative equilibrium (RE). This assumption requires that radiation alone heats or cools the atmosphere. It ignores the important process of convection, which we will include later.

From Eq. (7), the irradiance gradient is zero in RE, which implies a constant (net) irradiance in the vertical direction. As discussed previously, the irradiance leaving the atmosphere (at $\tau=0$) is equal to the irradiance at any arbitrary optical depth and is given in the two-stream approximation, $F(\tau) \approx 2\pi \bar{\mu} [I^+(\tau) - I^-(\tau)] = 2\pi \bar{\mu} \sigma_B T_e^4(\tau)/\pi$. Note carefully that for consistency with the two-stream approximation the exact result ($\sigma_B T_e^4$) has been replaced by the approximate result ($2\bar{\mu} \sigma_B T_e^4$). But, also note that if we set $\bar{\mu}$ equal to 1/2, the results are identical. This is the value we will adopt here, but we will retain the notation $\bar{\mu}$.

We find that the difference between the hemispherical radiances, $I^+ - I^-$, must be constant with optical depth. A second relationship between I^+ and I^- is obtained from the radiative transfer equation. Because the source function $S(\tau) = (1/2)[I^+(\tau) + I^-(\tau)]$, we write Eq. (5) as:

$$\bar{\mu} \frac{dI^+(\tau)}{d\tau} = I^+ - \frac{1}{2}(I^+ + I^-) = \frac{1}{2}(I^+ - I^-) \quad (11)$$

$$-\bar{\mu} \frac{dI^-(\tau)}{d\tau} = I^- - \frac{1}{2}(I^+ + I^-) = \frac{1}{2}(I^- - I^+) \quad (12)$$

Subtracting Eq. (12) from Eq. (11) yields:

$$\bar{\mu} \frac{(I^+ + I^-)}{d\tau} = I^+ - I^- \quad (13)$$

But, from RE, $S(\tau) = (1/2)(I^+ + I^-)$, and from Eqs. (13) and (10), we find:

$$\frac{dS(\tau)}{d\tau} = \frac{\sigma_B T_e^4}{2\pi \bar{\mu}} = C_1 = \text{constant} \quad (14)$$

It is clear that a general solution of the above equation is $S = C_1 \tau + C_2$, where C_2 is a second (so far arbitrary) constant. We may determine C_2 by invoking the principle (see Section II) that the IR radiation escapes the medium from an optical depth of unity. Averaging over all upward ray directions, this implies that the *effective emission height* occurs at an optical depth $\tau/\bar{\mu} = 1$. Since the escaping radiation has an effective temperature of T_e , we therefore assert that the source function at this level is given by

$S(\tau = \bar{\mu}) = C_1\bar{\mu} + C_2 = \sigma_B T_e^4 / \pi$. Thus, we can solve for $C_2 = \sigma_B T_e^4 / 2\pi$. Finally, substituting for C_1 and C_2 , we obtain the desired algebraic solution:

$$S(\tau) = \frac{\sigma_B T_e^4}{2\pi} (1 + \tau/\bar{\mu}) \quad (15)$$

The RE temperature profile is therefore (see Eq. (4)):

$$T_{re}(\tau) = T_e(1/2 + \tau/2\bar{\mu})^{1/4} \quad (16)$$

In order to interpret Eq. (16), it is desirable to determine the relationship of τ with the height variable z . For the Earth's troposphere, the dominant absorber is water vapor, for which the density distribution is approximately exponential with an absorber scale height H_a of 2 km. The frequency-integrated optical depth is then given by $\tau = \tau^* e^{-z/H_a}$ where z is expressed in kilometers above the surface ($z = 0$). Here, τ^* is the total optical depth, or opacity. It is impossible to define the value of a gray opacity from first principles, so we consider τ^* to be a parameter to be "tuned" to observation. In this case, we find that inserting $\tau^* = 0.63$ in Eq. (17) yields the observed mean surface temperature of 288 K. Note that even this modest optical thickness causes an appreciable warming to occur, relative to the effective temperature (which would apply to the surface of an airless planet of the same albedo and solar distance). This *greenhouse warming* thus prevents Earth from being a permanently glaciated world. Figure 5 displays a family of RE solutions, shown as dashed curves. These different curves can be thought of as applying to differing water vapor column amounts.

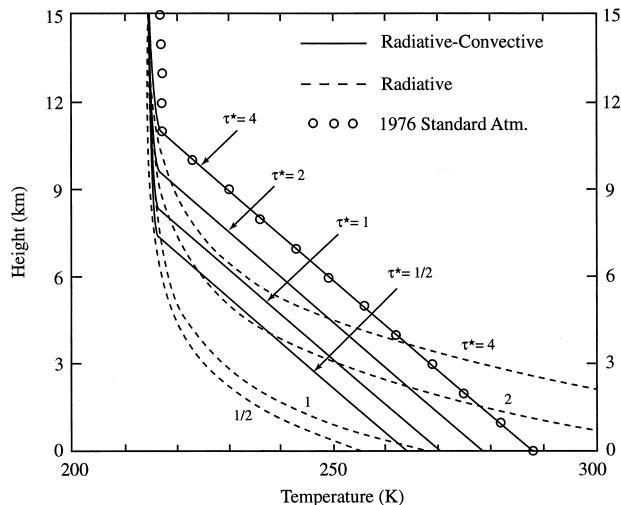


FIGURE 5 Pure-radiative (dashed lines) and radiative-convective equilibrium (solid lines) temperature profiles for four different optical depths, and for $\bar{r} = 0.30$. Open circles represent the 1976 standard atmosphere. The radiative-convective adjustment procedure is discussed in the text.

It is often stated that the greenhouse effect is a trapping of IR radiation in the opaque parts of the spectrum. This is somewhat misleading, as the radiative energy ultimately escapes from the top of the atmosphere. A more physically sound explanation is that the surface is back-warmed by the emitting atmosphere. Thus, the surface is heated not only by the Sun but also by this extra heat source. The additional downward irradiance is easily shown to be given by $F^-(\tau^*) = \sigma_B T_e^4 \tau^*$. Because the surface absorbs all this incident irradiance, under our assumption of a blackbody absorber, the extra heating is seen to be proportional to the IR optical depth. We can find the surface temperature T_s by equating the total surface heating (IR plus solar) to the total emitted energy. This may be shown to yield the result:

$$T_s = T_e(1 + \tau^*)^{1/4}. \quad (17)$$

Note the important difference between the above result and the atmospheric temperature near the surface, given by $T_{re}(\tau^*) = T_e(1/2 + \tau^*)^{1/4}$ (where we set $\bar{\mu} = 1/2$). This peculiar result occurs because of the additional heating of the surface by the Sun, as opposed to the atmosphere, which is heated only by IR radiation. To conserve total radiative energy (visible plus IR) across the interface, a discontinuity in temperature must occur. This tendency is present in the real atmosphere but is almost entirely eliminated by convective heat transport.

At the top of the atmosphere ($\tau \rightarrow 0$), T_{re} approaches the *skin temperature*, $T_e(2^{-1/4})$. The factor of $2^{-1/4}$ stems from the fact that in this optically thin region, air is heated by photons from the lower hemisphere but cools into both hemispheres. The skin temperature for Earth is about 214 K and is actually a good estimate for the minimum temperature of the region between the troposphere and the warmer stratosphere (the tropopause).

B. Radiative-Convective Equilibrium

In a real, optically thick planetary atmosphere, the radiative equilibrium solution yields an unstable temperature gradient. When this gradient is surpassed, the atmosphere responds spontaneously to maintain a *lapse rate* equal to that given by adiabatic equilibrium. This value varies with the water vapor amount present, because of phase changes that occur in rising and falling air parcels. The average adiabatic lapse rate for the Earth is $\Gamma = -6.5 \text{ K/km}$. An artifice that has been frequently used in the past is to assume that in regions where $dT_{re}/dz < \Gamma$ (i.e., has larger negative values) the atmosphere spontaneously adjusts to maintain the constant lapse rate. This correction acts to cool the lower optically thick regions and also largely eliminates the interface discontinuity. However,

this simple adjustment does not conserve energy at the intersection of the two solutions, because the upward irradiance issuing from the lower region is lower than in RE. It is necessary to move the intersection point upwards until the upward irradiances match. The height of this transition region, which we identify as the tropopause in our conceptual model, is located at heights between 7 and 11 km (see Fig. 5), depending upon the optical depth.

In our more general radiative-convective solution, it is once again necessary to tune the optical depth to match the observations. A larger opacity is needed because of the greater efficiency of upward heat transport. Figure 5 shows that $\tau^* = 4$ provides a very good fit to the empirical 1976 model at all heights in the troposphere.

An algebraic expression for the radiative-convective solution T_{rc} is also readily obtained. We invoke once again the concept of the emission height, the level of maximum radiative cooling. By requiring that the temperature at this height (denoted z_e) be equal to the effective temperature, T_e , it is trivial to find the expression for the surface temperature:

$$T_s = T_e + |\Gamma|z_e = T_e + |\Gamma|H_a \ln(\tau^*/\tau_e) \quad (18)$$

where τ_e is the optical depth at the emission height z_e . As discussed previously, $\tau_e = \bar{\mu}$. We may therefore combine the convective solution in the lower region and the RE solution in the upper region so that:

$$T_{rc} = T_e + |\Gamma|[H_a \ln(\tau^*/\bar{\mu}) - z] \quad (z \leq z_t) \quad (19)$$

$$T_{rc} = T_e(2)^{-1/4} \quad (z > z_t) \quad (20)$$

where the tropopause height z_t is given by $z_t = z_e + 0.159 \times T_e/|\Gamma|$. Note that because the temperature of the transition region is very close to the skin temperature, we have ignored this small difference in Eq. (20).

We have obtained a realistic solution for the mean state of the atmosphere, in terms of the optical depth, the adiabatic lapse rate, and the scale height of the absorber. It is instructive to consider how the solution varies as the above parameters vary. For example, in the tropics the atmosphere is more humid, so τ^* is greater than average. Our solution predicts that the surface is hotter for a more opaque atmosphere. It also predicts that the tropical tropopause is higher than average. Both of these predictions agree with observations. However, the solution predicts the same tropopause temperature everywhere on the Earth, whereas it is well known that the tropical tropopause is both higher and cooler than at other latitudes. At this point, the failure of the simple one-dimensional model is to be expected, because it ignores horizontal transport of heat. In fact, the tropics is a net source of energy, whereas the polar regions are a net sink, as a result

of horizontal transport by both the atmosphere and the ocean.

C. Effects of Clouds and Aerosols

Atmospheric particles affect both the shortwave and long-wave radiation fields. The climatic effects of greenhouse gases described above can be amplified, or damped, depending upon whether the clouds lead to a warming (due to an enhanced greenhouse effect) or to a cooling (due to an increased shortwave albedo). However, we can make some very general statements, and refer the reader to the literature for details. Particles (water droplets, ice crystals, or dust) cause an enhancement of the IR opacity and thus tend to raise the effective level of radiative cooling. The higher levels, being at lower temperature, cool less readily, thus the particles in this case cause a greater greenhouse warming. This principle applies to high clouds, such as cirrus. On the other hand, introduction of an opaque cloud at low levels increases the emissivity in spectral window regions, which puts the effective emission height below the clear-sky value, where the emission at relatively warmer temperatures causes a net cooling of the atmosphere. In addition, clouds tend to increase the shortwave albedo, thus reducing the solar heating. All of these effects are sensitive to particle size and density, as well as the effects of aerosols on cloud properties.

VI. CONCLUSION

We have provided a brief summary of the most salient features of atmospheric radiation including its interaction with scattering, absorbing, and emitting atmospheric molecules and particles, as well as the underlying surface that reflects, absorbs, and emits radiation. To provide a suitable framework for our discussion we chose to focus on the global warming issue in order to demonstrate how atmospheric greenhouse gases influence our climate. To this end, we have described a simple, conceptual model for the atmospheric temperature profile in radiative equilibrium, for the greenhouse effect, the combined effects of radiation and convection, and the tropopause. All these factors are described by an algebraic equation in which the parameters (optical depth, absorber scale height, planetary albedo, and adiabatic lapse rate) may all be varied to understand their gross effects on the planetary heat budget. The model helps to elucidate the most basic ideas of the one-dimensional radiative energy budget. Enhancements to this basic model would include the presence of clouds, with the cloud optical parameters and their height distribution as adjustable parameters. An additional

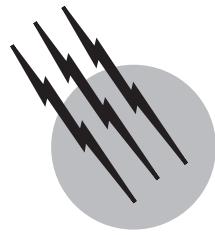
enhancement would include the effects of shortwave absorption, such as those due to ozone or aerosols of volcanic origin. If shortwave absorption is sufficiently strong, an *anti-greenhouse effect* can occur, with depressed surface temperatures and elevated temperatures aloft. The effects of diurnal variations and latitudinal and time dependence emphasize the importance of the ocean and its capacity for moderating variations of the surface temperature. Such influences are beyond the abilities of simple conceptual models and obviously require complex computer simulations. We conclude that realistic climate models that accurately capture the complex interactions and feedbacks among atmospheric radiation, chemistry, and dynamics in the climate system are not yet available. Creating such models, that can be relied upon for future climate predictions, will remain a challenge for years ahead and requires access to measurements of appropriate climate variables for testing and verification.

SEE ALSO THE FOLLOWING ARTICLES

AEROSOLS • CLOUD PHYSICS • ENVIRONMENTAL RADIATION • GREENHOUSE EFFECT AND CLIMATE DATA • GREENHOUSE WARMING RESEARCH • IMAGING THROUGH THE ATMOSPHERE • OZONE MEASUREMENT AND TRENDS (TROPOSPHERE) • RADIATION SOURCES • SOLAR THERMAL POWER STATIONS

BIBLIOGRAPHY

- Houghton J. T. et al., eds. (1995). "Climate Change 1995: The Science of Climate Change," Cambridge Univ. Press, Cambridge, U.K.
Kiehl, J. T., and Trenberth, K. E. (1997). "Earth's annual global mean energy budget," *Bull. Am. Meteorol. Soc.*, 197–208.
Smith, W. L., and Stammes, K., eds. (1997). "IRS96: Current Problems in Atmospheric Radiation," A. Deepak Publishing, Hampton, Va.
Thomas, G. E., and Stammes, K. (1999). "Radiative Transfer in the Atmosphere and Ocean," Cambridge Univ. Press, Cambridge, U.K.



Solar Terrestrial Physics

L. J. Lanzerotti

Bell Laboratories, Lucent Technologies

- I. Introduction
- II. Solar Processes: The Source for Sun–Earth Couplings
- III. Interplanetary Medium: The Mechanism for Sun–Earth Plasma Couplings
- IV. Magnetosphere of Earth: The Extension of Magnetic Fields and Plasmas into Space
- V. The Sun–Earth Connection: Implications for Human Activity

GLOSSARY

Alfvén wave Basic hydromagnetic wave in a plasma containing a magnetic field; the plasma displacement is transverse to the magnetic field, with propagation directed along the field.

Aurora Lights in the upper atmosphere (from about 90 to 300 km altitude) produced by the excitation of atmospheric gases by energetic particles; the localized areas in the two polar regions where aurora are typically observed are determined by the detailed topology of the magnetosphere.

Collisionless shock Discontinuity formed on the sunward side of the earth in the solar wind because of the interposed obstacle of the magnetosphere.

Coronal holes Regions in the solar corona in which the solar magnetic fields are not closed but are open into the interplanetary medium.

Coronal mass ejection Expulsion of hot gas from a localized region in the solar corona into the interplanetary medium.

Ecliptic plane Plane of the apparent annual path of the sun on the celestial sphere.

Geomagnetic storm Extended interval of many hours to as much as a day or more of severe auroral, geomagnetic, and magnetospheric activity, concurrent with a growth of the magnetospheric ring current.

Geomagnetic substorm Interval of approximately 1 to 3 hr of auroral, geomagnetic, and magnetospheric activity, usually followed by a several-hour interval of relative quiescence.

Heliosphere Region around the Sun in our galaxy (the Milky Way galaxy) influenced by the solar wind.

Ionosphere Ionized region of the upper atmosphere, generally taken to be from about 90 to about 1000 km in altitude; the area beyond this is defined as the magnetosphere.

Magnetopause Boundary between the flowing solar wind plasma and the magnetosphere.

Magnetosphere Region of space around the Earth in which the terrestrial magnetic and electric fields

usually dominate the transport and motions of charged particles.

Magnetosphere cusps Regions in the northern and southern polar areas separating the magnetic fields that form the dayside magnetopause from those fields that stretch into the magnetotail.

Magnetotail Region of the magnetosphere in the anti-sunward direction; if visible, the magnetotail would look somewhat like the tail of a comet.

Photosphere Visible solar surface with a temperature of about 6400 K.

Plasma Gas (atomic and/or molecular species) that is ionized. Naturally occurring plasmas usually contain magnetic fields.

Plasma sheet Sheet of plasma, several earth radii in thickness in the tail of the magnetosphere, that separates magnetic fields of opposite magnetic polarity.

Plasmapause Outermost boundary of the plasmasphere.

Plasmasphere Region of cold (about 1 eV) and dense (about 10^3 cm^{-3}) plasma of ionosphere character in the magnetosphere.

Radiation belts Localized regions in the magnetosphere that contain energetic charged particles whose motions are primarily controlled by the magnetic field of earth.

Solar active regions Regions on the sun of several or many sunspots

Solar corona Region, beginning about 2000 km above the photosphere, with a temperature of over 1 million degrees Kelvin.

Solar flare Sudden brightening, for several minutes to several hours, of a small area of the solar photosphere that contains a group of sunspots.

Solar wind Expansion of the solar corona, primarily hydrogen ions, into the interplanetary medium.

Sunspot cycle Variation with time of the appearance of the number of sunspot groups on the visible solar photosphere.

Sunspots Darkened, cooler areas of the solar photosphere usually occurring in groups and containing intense magnetic fields.

SOLAR-TERRESTRIAL PHYSICS is the study of the interactions of nonoptical solar emissions with the magnetic field and atmosphere of earth and the results and implications of these interactions for human technologies. The major nonoptical solar emission is the solar wind, a super-Alfvénic flow of charged particles boiled off the top of the sun's hot outer atmosphere, the corona. Other important solar emissions striking earth include ultraviolet light, γ rays, X rays and the energetic charged

particles produced by solar activity, including flares and coronal mass ejections.

I. INTRODUCTION

Although unrecognized as such, appearances of light emissions high in the Earth's atmosphere during nighttime, the "aurora," have announced to human observers the solar-terrestrial connection for hundreds and thousands of years. Descriptions of the aurora are common in the oral histories and legends of northern peoples. Although large auroral displays are infrequently seen at very low latitudes, descriptions of phenomena in the nighttime sky that can be interpreted as the aurora are found in ancient writings by such authors as Aristotle and Seneca, as well as perhaps in the first chapter of Ezekiel. Old Chinese texts describe auroral observations in the Orient. Natural science volumes from the middle ages often contain fanciful illustrations of auroral displays over villages and towns. The reality of the solar-terrestrial connection, however, was first identified only in the late 19th century. The extent and full implications of this connection are unknown even to this date.

One of the first "modern" manifestations of the solar-terrestrial connection, unrecognized at the time, was the puzzling report in the late 1830s by W. H. Barlow of the spontaneous appearance of anomalous currents measured on electrical telegraph lines in England. Perhaps the first scientific recognition of the connection was by the British astronomer Richard Carrington. On September 1, 1859, while sketching sunspots in the course of his studies of solar phenomena, Carrington suddenly observed an intense brightening in the region of one of the spots. This happening so excited him that he quickly called his associates to the telescope to witness the event. Within a day, violent fluctuations in the geomagnetic field and intense auroral displays were observed at various locations on earth, with reports of aurora being sighted as far south as Honolulu. Carrington was very intrigued about the possible link between his white light flare and the subsequent aurora. Nevertheless, he urged caution in connecting the two, commenting "one swallow does not a summer make."

During the several days of enhanced auroral displays that followed Carrington's solar event, telegraph systems throughout Europe and in eastern North America suffered severe impairments in operation, and even complete disruptions of service at times. The new technology of telegraphy had never experienced such widespread impacts on service, occurrences that were attributed by telegraph engineers to electrical currents flowing in the earth, somehow associated with the auroral displays.

Considerable engineering attention was devoted for many years to studies of the phenomena of electrical currents in the Earth. While the engineering literature

often discussed the relationships of these “telluric” currents to enhanced fluctuations in the terrestrial magnetic field (magnetic storms), and possibly to disturbances on the Sun, many natural scientists of the time were less inclined to see such a cosmic causal connection. While the British scientist E. Walter Maunder seriously discussed such a connection, the dominant authority of the time, Lord Kelvin, thought otherwise. In discussions of a particular magnetic storm (June 25, 1885), he definitively stated that “it . . . is absolutely conclusive . . . that terrestrial magnetic storms are [not] due to magnetic action of the sun.” He further concluded that “the supposed connection between magnetic storms and sunspots is unreal.”

The first half of the 20th century, which saw the implementation of transatlantic communications by radio and the discovery of the Earth’s ionosphere, gave rise to more considered and quantitative discussions of the sun–earth connection. Such considerations were warranted, not only out of scientific curiosity, but for very practical reasons as well. For example, transatlantic telephone traffic via low-frequency radio waves was often disrupted during magnetic storms, a situation that prompted many scientific and engineering experiments and publications in the 1930s, as well as being partially responsible for the laying of the first transatlantic telephone cable in the 1950s.

Nearly 100 years after Carrington’s discovery, at the threshold of the space age, voice traffic on the first transatlantic telecommunications cable was severely affected on February 11, 1958, by a magnetic storm that followed closely on the heels of a particularly large solar event. Toronto suffered an electrical blackout during the same geomagnetic disturbances. In August 1972, a series of large solar events resulted in such extensive and large magnetic disturbances that there was a complete disruption of a transcontinental communications cable in the midwestern United States. The entire province of Quebec suffered an electrical power outage following a solar event in March of 1989.

Studies from the Earth of the ion tails of comets led the German astronomer Ludwig Biermann to suggest in 1951 that the sun emitted invisible gases that controlled the orientations of the tails as the comets traversed the solar system. The advent of scientific spacecraft in the space age confirmed this hypothesis. Measurements made at considerable distances from earth demonstrated conclusively that solar–terrestrial connections occur not only through the optical emissions of the sun but also through the “invisible” tenuous mixture of ions, electrons, and magnetic fields that are now called the solar wind. Spacecraft also provided measurements of the connection through the much more energetic particles produced by coronal mass ejections and solar flares. Thus, the space age has provided the means to dispatch robotic instrumentation into the en-

vironment around earth and at great distances from earth in order to measure and study the near earth and interplanetary regions that can produce both the dramatic, visual spectacle of the aurora and disruptions to technology.

The advances made by the use of spacecraft in understanding the solar–terrestrial environment have been enormous. However, even as the basic morphology and some physical processes of this environment have become better known, it has also become clear that there are many physical processes occurring in it—from the sun through the interplanetary medium, and at the earth—that yet defy complete understanding. Thus, predictions regarding the state of the environment and its possible effect on the Earth remain rather rudimentary. The physics of the operative processes is complicated. However, the increasing sophistication of instruments and experiments (ground-based, rocket, and spacecraft) and of theoretical and computational capabilities has begun to provide significant insights into many of the fundamentals of the solar–terrestrial system.

II. SOLAR PROCESSES: THE SOURCE FOR SUN–EARTH COUPLINGS

Visible radiation from the sun provides the heat and light that have enabled life to originate, evolve, and thrive on our planet. The constant, unvarying nature of the source of the life-sustaining light and heat was an important element in the mythologies of many ancient civilizations, which attributed godlike qualities to the sun. This myth was shattered when Galileo’s telescope revealed that the sun was not “perfect”: Spots varying with time and with location marred the solar surface. Since Galileo’s time the sun has been found to have a number of changeable features, many of which can affect the Earth. The existence of a cyclic variation in the number of sunspots was firmly established in the mid-19th century by Heinrich Schwabe, a German druggist and amateur astronomer.

The periodic variation in sunspot numbers is shown in Fig. 1. While the number of spots varies with an approximately 11-year period as illustrated, the fundamental cycle, based upon the magnetic polarity of the sun, is approximately 22 years. That is, the north–south magnetic polarity of the sun changes with an approximate 22-year period, or an interval twice the number cycle. At the beginning of a new number cycle the spottedness begins first at latitudes of about 30° on both sides of the solar equator. As the cycle continues to develop and the number of spots increases, the spots in both hemispheres migrate slowly toward the equator until, near the end of the cycle, those nearest the equator begin to disappear. During and

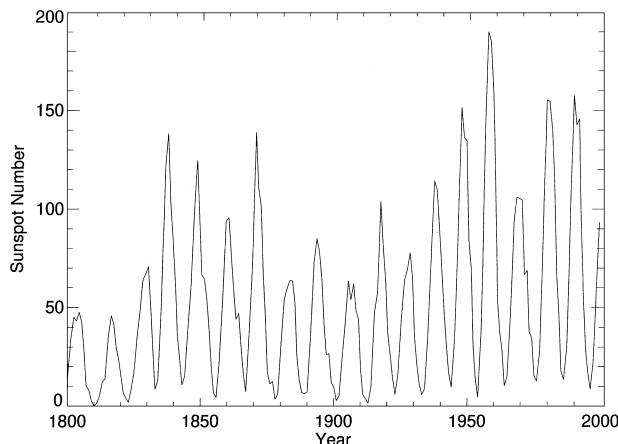


FIGURE 1 Sunspot cycle, from the early 19th century to the present.

following the disappearance, spots again begin to appear at latitudes of $\pm 25\text{--}30^\circ$. For a significant interval after their discovery, there was a long hiatus in the last half of the 17th century when the number of spots was very small, or even zero. This so-called Maunder minimum period, named after E. Walter Maunder who first drew attention to it, is a significant enigma in terms of basic understanding of solar variability.

In addition to the periodic changes—migrations—of spots on the sun during the sunspot cycle, the overall solar magnetic field configuration undergoes large changes during a cycle. During solar minimum conditions, when spots are few, the sun resembles a large bar magnet in the sky with north and south poles of the magnetic field lodged in one or the other of the solar rotational polar regions. (This resembles the magnetic configuration of earth, but in the earth's case the magnetic poles flip polarity erratically over time scales of hundreds of thousands to millions of years.) This is the time when solar coronal holes dominate the polar regions of the sun as discussed in the next section. During solar maximum conditions, in contrast, the sun's poles appear to migrate to the equatorial regions of the sun, making the solar magnetic field configuration very confused and complicated.

Sunspots are not just regions of cooler gases on the solar surface. They are also regions where the magnetic fields from the solar interior penetrate the surface, producing very large, up to several thousand gauss, intensities (compared with the magnetic field intensity at the surface of the earth of approximately one-third gauss). Sunspots occur in groups of two to several, with opposite magnetic polarities, north and south.

Magnetic fields arch from one spot to the other. Under conditions that are poorly understood, the magnetic fields of the sunspot structures can become very unstable, ultimately producing a solar flare: intense brightening of a

region on the solar surface occurs abruptly as atomic ions and electrons in the solar atmosphere are suddenly energized. The energy content of a solar storm, $10^{32}\text{--}10^{33}$ ergs, is probably derived from conversion of energy stored in the magnetic fields. Some energized particles impact the solar atmosphere, producing optical and ultraviolet light as well as X rays and γ rays. Other charged particles are emitted into interplanetary space and propagate to the orbit of earth and beyond. Intense bursts of radio frequency emissions, easily detectable at earth, accompany the particle energization as well as their impacts on the sun's atmosphere.

Solar flares are one example of violent solar activity. The other important one is the episodic injection into the interplanetary medium of large masses of matter from the solar corona. In a so-called coronal mass ejection (CME) event, approximately 10^{12} kg or more mass of ions can be expelled from the corona. Such coronal transients were first identified from data acquired during the Skylab space station era of solar research in the mid-1970s, and their studies were continued with other missions such as the Solar Maximum Mission (SMM) in the 1980s. The launch of the Solar and Heliospheric Observatory (SOHO) spacecraft in the mid-1990s revolutionized the studies of the sun, and especially the phenomena of CMEs.

Measurements by the SOHO spacecraft have shown that there is a definite solar cycle dependence of the number of CMEs that are ejected from the sun. During the solar maximum interval of the 23rd cycle (2000–2001), several CMEs were seen each day. During the quiet interval leading up to the maximum of this cycle, less than one per day was commonly observed. The relationship between the production of a CME and the occurrence of a solar flare remains uncertain. While some CMEs appear to be clearly associated with flare events, other cases are more ambiguous. The magnetic fields that thread CMEs sometimes appear to be connected back to the sun at the time that the event is measured near 1 AU, and at other times the field configuration seems to be disconnected from the sun. A launched CME can produce highly disturbed interplanetary conditions, with the outward propagating shock wave accelerating ambient interplanetary particles to much higher energies.

III. INTERPLANETARY MEDIUM: THE MECHANISM FOR SUN-EARTH PLASMA COUPLINGS

Measurements of the solar atmosphere by atomic spectroscopy techniques reveal that the outermost region—the solar corona from which CMEs are expelled from the sun—has a much lower density and a much higher temperature than the underlying surface (the photosphere). The

precise mechanisms for heating this high upper solar atmosphere are unknown and have been a matter of scientific controversy for many years. It is still not definitively determined, for example, if the corona is heated primarily near the photosphere and the heat then travels upward to the entire corona, or if the corona may be heated more uniformly in altitude. Possible heating methods could include the damping of acoustic waves generated by the convection of heat from the solar interior, heating by the damping of Alfvén waves produced in the solar photosphere that propagate outward from the sun, and heating by the occurrence of “micro” solar flares. Data from the SOHO spacecraft have shown that small “micro” flares are common on the sun and may be very important in the heating process.

The plasma structure of the solar corona, even without the occurrence of a CME, is complicated and dynamic, often exhibiting in eclipse photographs long, intense radial streamers and huge arches (Fig. 2). Large changes in the corona occur during the course of the solar cycle. During solar minimum, the corona across each of the two polar regions of the sun is considerably cooler than in the equatorial region. Cooler coronal regions, referred to as coronal holes, are regions in which the solar magnetic fields tend to be more open and extend into interplanetary space. During sunspot maximum times, coronal holes tend to be very much more limited in extent and can occur at any location on the sun.

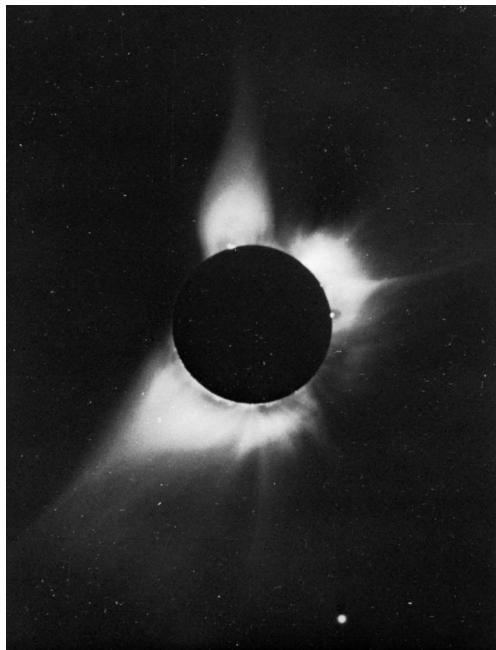


FIGURE 2 Eclipse photograph of the sun made from the surface of earth, showing coronal streamers and coronal structure.

The end result of heating the corona is that the very hot outer atmosphere continually expands away from the sun, forming the solar wind. At a distance of a few solar radii above the photosphere the velocity of expansion of the corona becomes larger than the Alfvén velocity

$$V_A = \left(\frac{B^2}{4\pi\rho} \right)^{1/2}, \text{ cm sec}^{-1},$$

where B is the magnetic field intensity in gauss and ρ the ion mass density (g cm^{-3}). That is, the velocity of expansion becomes faster than the velocity with which information can be transferred in the highly conducting (ionized) gas. Beyond this distance the solar wind expands throughout the solar system, carrying with it the magnetic field from the solar surface. The speed of this wind is typically 400–600 km/sec, although much higher velocities have been measured following some CMEs and large flares. Solar wind velocities from coronal holes are of the order of 600–700 km/sec and more. The region (whose extent is not known as yet) around the sun that is influenced by the solar wind is called the heliosphere. It is thus the solar wind that provides the primary plasma link between the sun and the earth and that is the source of geomagnetic disturbances on earth. Lord Kelvin, in his categorical statements of nearly a century ago, could not have been aware of this invisible, crucial link between the sun and the planets.

The outward expansion of the solar wind, combined with the rotation of the sun, causes an important physical phenomenon in the heliosphere: The solar magnetic field is tied firmly to the sun by the highly conducting solar atmosphere. At the same time, the magnetic field is firmly embedded (frozen) in the radially outward-flowing solar plasma (the solar wind). The field in interplanetary space thus forms a spiral pattern similar to that produced by a rotating garden water sprinkler. At the orbit of earth, the interplanetary magnetic field makes an average angle in the ecliptic plane of approximately 45° to the radially outward direction. The angle of the interplanetary field with respect to the magnetic field of earth plays a key role in determining the level of geomagnetic activity.

Solar flares, CMEs, the boundaries between coronal holes and more quiescent coronal locations, and other disturbances on the sun can significantly disrupt the tranquility of the interplanetary medium. The greatly enhanced solar wind streams that are produced by solar activity of all types have a higher velocity and a greater particle number density than the normal solar wind. Shock waves can be formed in the interplanetary medium between the boundaries of the faster and slower moving winds. These shock waves themselves can accelerate interplanetary particles to higher energies and can greatly agitate the magnetosphere.

At the orbit of earth the solar wind carries an energy density of about 0.1 ergs/cm^2 , a factor about 10^7 smaller than the solar energy incident on the earth in the visible and infrared wavelengths ($\sim 1.4 \times 10^6 \text{ erg/cm}^2$). Yet, as will be seen, this low energy density, when applied over the magnetosphere, can have profound effects on the terrestrial space environment. At great distances from the sun it is likely that the heliosphere, with the embedded sun and planets, forms a kind of magnetosphere itself in the local interstellar medium. A boundary will be established between the outward flowing solar wind and the interstellar plasmas and magnetic fields in the direction in which the sun is moving relative to the nearby stars. The boundary is expected, from present ideas, to occur between 100 and 150 earth-sun distances. The boundary is likely to be a turbulent region, with perhaps a shock wave established in the interstellar medium, similar to the situation for the earth in the solar wind (see following).

IV. MAGNETOSPHERE OF EARTH: THE EXTENSION OF MAGNETIC FIELDS AND PLASMAS INTO SPACE

A. Overall Morphology

The earth manifests its presence in the solar wind in a manner analogous to that of a supersonic airplane traversing the atmosphere. Since the solar wind velocity is faster than the characteristic speed in the ionized gas, a shock wave (the bow shock) is established. The average location of the shock is some 10–14 earth radii outward on the sunward side (Fig. 3). On the earthward side of the shock wave is a region of more turbulent solar wind flow, the magnetosheath. Then, at an average altitude above the earth of some nine earth radii, the magnetopause exists

as a thin, current-carrying boundary separating regions controlled by the terrestrial magnetic field (Fig. 3). To a reasonable approximation, the solar wind flow around the earth at the magnetopause is similar to the air flow in a wind tunnel past an aerodynamic object. Also to a good approximation, the subsolar magnetopause location can be approximated by equating the dynamic pressure $C\rho V^2$ of the flowing solar wind and the sum of the magnetic field and plasma pressure inside the magnetosphere.

$$C\rho V^2 = \frac{B^2}{8\pi} + P.$$

Here C is a constant, V and ρ are the solar wind velocity and ion number density, respectively, $B^2/8\pi$ is the pressure due to the magnetic field of the magnetosphere at the boundary, and P (usually negligible in the case of the terrestrial magnetosphere) the gas pressure inside the magnetosphere.

The actual physical processes forming the configuration of the magnetosphere, while approximated by these simple concepts, are in fact much more involved in detail. The formation of the shock wave in front of the magnetopause is more complicated than the simple airplane analogy. For one thing, the magnetic field imbedded in the solar wind is intimately associated with establishing the detailed characteristics of the shock. Furthermore, the shock wave is a collisionless shock; that is, the solar wind particles do not interact by collisions among themselves. Rather, the distribution of plasma particles becomes unstable because of the magnetic field and the plasma wave conditions produced by the presence of the obstacle. Waves generated by the formation of the shock can propagate in the solar wind back toward the sun, making the interplanetary environment near earth turbulent and quite complicated.

Similarly, the magnetopause is a complex physical region. While the magnetosphere has vast extent, the magnetopause boundary itself is a very thin region where electrical currents flow in the plasma. These currents define a thin, cellular boundary separating the two very different plasma regimes of the solar wind and the magnetosphere. This cellular-like boundary is not completely impenetrable, and its location in space can be quite variable depending upon solar wind conditions. The boundary can move inward, toward the earth, under higher solar wind velocities, and can move outward during intervals of lower solar wind speed conditions. Nevertheless, the two plasma regimes separated by the magnetopause boundary can be treated individually in analyzing many problems involving internal processes in each system. The interplanetary magnetic field also plays a key role in the formation of the boundary. A reconnection of the earth's internal magnetic field and the interplanetary field appears to occur

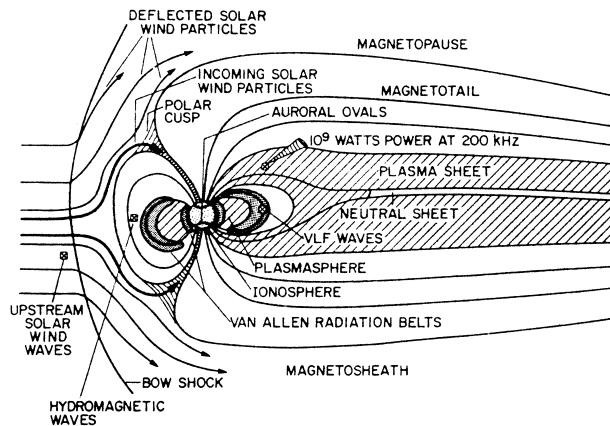


FIGURE 3 Schematic illustration of the magnetosphere of earth, the nearby plasma environment formed by the interactions of the solar wind with the geomagnetic field.

sporadically at the boundary, allowing plasma to escape from the magnetosphere and solar wind plasma to become entrapped in the terrestrial magnetosphere.

The solar wind, by a viscous interaction (which may include magnetic field reconnection processes) with the magnetosphere plasma and magnetic field, forms the magnetosphere into a long, comet-like (but invisible) object (Fig. 3). A large dawn-to-dusk electric field is created across the magnetosphere by this interaction. The energy transfer rate from the solar wind into the magnetosphere is about 10^{19} ergs/sec. The comet-like magnetotail may extend to more than a thousand earth radii (some 6×10^6 km) in the anti-sunward direction. In the magnetotail, a sheet of highly conducting plasma separates the magnetic fields that originate in the southern hemisphere of earth from those that terminate in the northern hemisphere (Fig. 3). An electrical current flows from dawn to dusk through this plasma sheet. The earthward extensions of this plasma sheet protrude along magnetic field lines that connect with the night-side auroral zone (ionosphere currents) in both hemispheres. The power dissipated in these currents is about 10^{18} ergs/sec (10^{11} W). On the front side of the magnetosphere, the separations in each hemisphere between magnetic field lines closing on the dayside and those extending into the magnetotail form a cusplike region through which solar wind plasma can reach the upper atmosphere, forming auroral emissions and electrical currents in the ionosphere. An instantaneous topology of the auroral zone can be seen in Fig. 4 in a far-ultraviolet light picture taken by the wide band imaging camera on the NASA IMAGE satellite over the northern hemisphere. This picture, made during the time of impact of a large interplanetary shock wave on the earth's magnetosphere on July 15, 2000, provides a striking, visible measure of the solar-terrestrial connection.

During intervals of geomagnetic substorms, the auroral zone can extend to lower latitudes, and the electrical currents flowing in the ionosphere can be significantly enhanced in intensity and in latitudinal extent. The energy dissipated during a geomagnetic substorm, which may last 2–3 hr, is about 10^{19} ergs/sec. These magnetic disturbance conditions seem to occur from the energization of the plasma sheet in the magnetotail, often triggered by changed solar wind conditions. The energization is believed to occur from the conversion of magnetic field energy into plasma particle kinetic energy through a process of reconnection of magnetic field lines across the plasma sheet. This enhanced plasma is then transported into the auroral zones. Additional energization of charged particles occurs along magnetic field lines above the aurora. Occasionally the interplanetary medium is disturbed for many hours, even days. At such times the geomagnetic

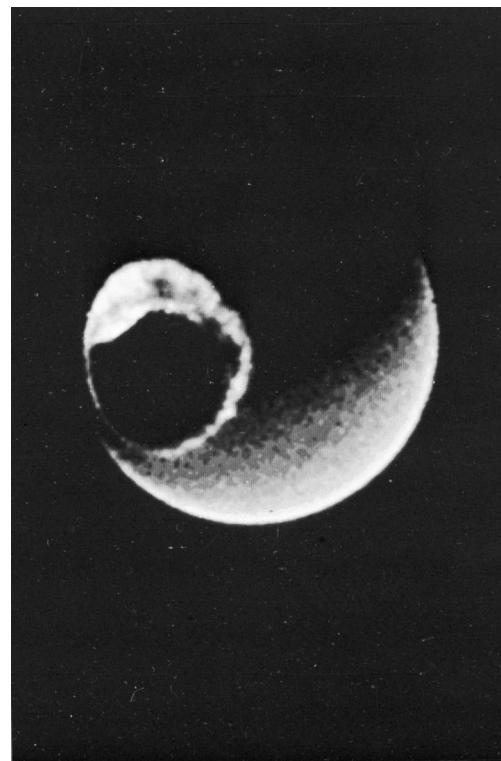


FIGURE 4 Auroral zone of the earth measured in ultraviolet light on the NASA IMAGE satellite at the time of an interplanetary shock wave striking the magnetosphere (S. Mende and H. Frey, University of California, Berkeley).

disturbances in the magnetosphere can also last for a day or more, a condition called a magnetic storm.

During intervals of sustained high geomagnetic activity, the aurora and other disturbances can continue for a day or more. During these periods of geomagnetic storms, the electrons and protons in the plasma sheet are convected deep into the magnetosphere from the night side. Eventually an earth-encircling ring of particles (ring current) with a peak energy of a few hundred kilo-electron volts forms. The altitude of this current ring above the earth is commonly at 2 to 3 earth radii. This current ring serves to depress the geomagnetic field in the equatorial and mid-latitude regions. The geomagnetic field depression slowly decreases over a few days as the particles in the ring current disappear. The loss of the ring particles is caused by their interactions with the residual neutral atmosphere and by their interactions with plasma waves in the magnetosphere. It is during large geomagnetic storms that the solar-terrestrial impacts on technology most often occur (Section V).

This overall concept of the terrestrial magnetosphere was rapid in developing once spacecraft were flown above the atmosphere. Soon after the discovery of the radiation belts (Fig. 3) by Van Allen and his students using

data from their instrumentation on the Explorer 1 satellite in 1958, the magnetosphere boundary and shock were detected. The following decades have seen the concept of a magnetosphere greatly extended. The discovery and detailed measurements of the characteristics of the magnetospheres of the giant planets Jupiter, Saturn, Uranus, and Neptune have occurred. The mini-magnetosphere of Mercury and that formed by the solar wind interaction with the ionized upper atmosphere of Venus have been delineated. Magnetosphere-like interactions of several of the moons of Jupiter with Jupiter's own magnetosphere have been identified. Hence, the concept of a magnetosphere has become quite general in cosmic physics. Indeed, many exotic astrophysical objects, which can only be studied remotely by detection of emitted electromagnetic radiation (e.g., pulsars and some radio galaxies), are now commonly discussed in magnetosphere terms.

B. Plasma Processes

There are many localized and small-scale plasma processes throughout the global magnetosphere system which provide the underlying mechanisms that determine the state of the system. A large variety of plasma waves exist in the magnetosphere. These waves can be electromagnetic, that is, having both electric and magnetic properties; or they can be electrostatic, resulting from the separation of oppositely charged plasmas. Lightning storms in the atmosphere create radio frequency waves that not only produce static in radios but also propagate, with a whistling tone, along geomagnetic field lines extending into the outer magnetosphere, even to the opposite hemisphere. Other waves, with wavelengths on the order of the length of a magnetic field line (several earth radii or more) are known as Alfvén waves. These waves are generated by the flow of the solar wind past the magnetic field of the earth and by instabilities that can occur in the magnetosphere plasma. Still other radio waves are generated above the auroral regions (within perhaps one earth radius altitude), where electrons and protons are accelerated into and out of the ionosphere.

A most interesting aspect of the magnetosphere, as well as of solar flares, is the possible existence of regions in which magnetic field energy is converted directly to plasma particle energy. Such regions are believed to be those in which magnetic field lines of opposite polarities can suddenly reconnect and, in the process of reconfiguring, release energy to the embedded plasma. As noted earlier, these regions can exist at the front side of the magnetosphere, possibly in the cusp regions, and in the magnetotail. In the magnetotail, the magnetic field lines in the plasma sheet are always oppositely directed. If the plasma sheet conditions should change in such a manner that the

plasma became less conducting, reconnection of the magnetic field lines might occur spontaneously. Alternatively, the magnetic field lines could be pushed together by an external force (such as the solar-wind interaction) that would slowly compress the plasma sheet. As the plasma in the center of the region became more dense, collisions among plasma particles or the onset of certain plasma instabilities might alter the conductivity, providing an environment for reconnection to occur. Many theoretical considerations and computer simulations of the reconnection process have been carried out, and it is one of the most active areas of basic theoretical magnetospheric plasma research at present.

The ionosphere of the earth is an intriguing plasma environment, in which both neutral and ionized gases are threaded by the geomagnetic field. At times, the ionization layers become unstable, producing patchy conditions with different ionization densities. This can be particularly prevalent in the electrical currents that flow in the auroral zone. In the equatorial regions of the ionosphere, where the magnetic field lines are parallel to the ionization layers (as well as to the surface of the earth), the ionosphere layers can become unstable; plasma bubbles can form and rise through the ionosphere to the upper levels. Studies of the basic plasma physics of such bubbles and ionization patches have led to significant new insights into cosmic plasma processes.

The background plasma density in the magnetosphere varies from several thousand particles per cubic centimeter within the first few earth radii altitude to only a few per cubic centimeter at higher altitudes. There is ordinarily a rather sharp discontinuity between the two plasma regimes, and the boundary is called the plasmapause ([Fig. 3](#)). The boundary is formed approximately at the location where there is a balance between the electric fields produced by the rotation of the geomagnetic field and the large-scale electric field imposed across the magnetosphere by the solar wind flow. This discontinuity in the plasma distribution can be a source of magnetosphere plasma waves and can significantly affect the propagation of Alfvén waves. The background plasma density inside the plasmapause results from ionosphere plasma diffusing up into the magnetosphere during local daytime conditions. Outside the plasmapause, the cold plasma from the ionosphere is swept (convected) out of the magnetosphere by the cross-magnetosphere electric field.

New spacecraft instrumentation techniques have now allowed imaging of the plasmasphere, as illustrated in [Fig. 5](#). In this figure, which is a view downward onto the north pole, the circumference of the earth is indicated by the white circle. The sun is to the upper right. The northern auroral zone, imaged by the extreme ultraviolet imager on

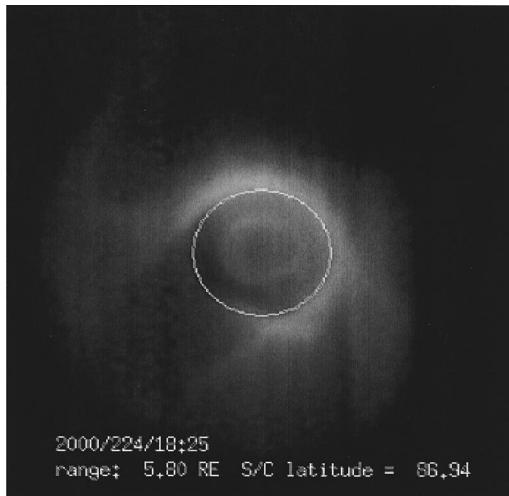


FIGURE 5 The plasmasphere of earth as viewed in ultraviolet light from the NASA IMAGE satellite above the North Pole (B. R. Sandel and T. Forrester, University of Arizona).

the NASA IMAGE satellite, can be seen faintly in the center of the circle. In addition to the direct detection of the aurora, the imager has detected the solar extreme ultraviolet photons that are resonantly scattered by singly ionized helium (which constitutes about 10 to 15% of the plasmasphere's ion population; protons essentially constitute the remainder) that is contained within the plasmasphere. The plasmasphere appears in this image as a cloud surrounding the earth; the earth's shadow appears as a dark biteout in the plasmasphere emission to the lower left. A faint tail of plasma is seen extending sunward in the upper left of the image.

The radiation belts are populated largely by particles accelerated out of the upper atmosphere and ionosphere and by solar wind ions and electrons. The relative importance attributed to these two sources of radiation belt particles has varied during the history of magnetosphere research, with much emphasis being placed at present on the importance of the ionospheric source. The ions and electrons can be accelerated to their radiation belt energies by internal plasma instabilities and by large-scale compressions and expansions of the magnetosphere under action of the variable solar wind. In the innermost part of the magnetosphere the decay of neutrons, produced by high-energy cosmic rays that strike the upper atmosphere, yields electrons and protons to the trapped radiation belts. The motions of radiation belt ions and electrons are controlled by the terrestrial magnetic field. The inner Van Allen belt of electrons is located earthward of the plasmapause, while the outer belt is outside.

A sketch of major current systems in the magnetosphere-ionosphere system is shown in Fig. 6. The auroral current system is linked to the magnetosphere and the

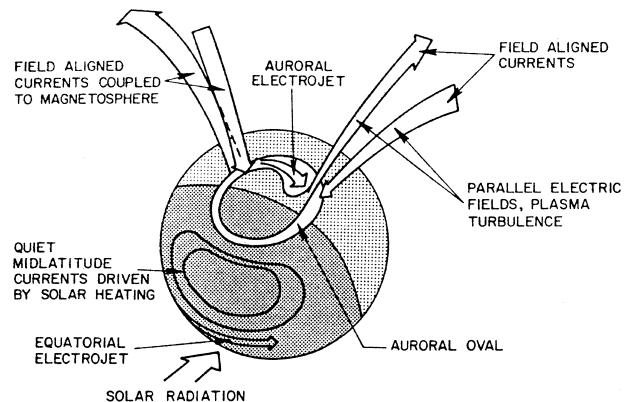


FIGURE 6 Simplified picture of ionospheric current systems and the connection of the auroral ionosphere current systems to the magnetosphere via electrical currents flowing along geomagnetic field lines.

plasma sheet via currents that flow along geomagnetic field lines. Solar radiation in the visible and ultraviolet wavelengths produces ionospheric electrical currents on the dayside of the earth, including an intense current in the equatorial regions.

The auroral electrical current systems can heat the upper, neutral atmosphere of earth, altering the overall circulation patterns in these regions. Without auroral heating, upper atmospheric circulation would tend to flow from the hotter, equatorial regions toward the colder, polar regions. Increased levels of geomagnetic activity (caused by disturbances in the solar wind) produce increased auroral heating. As illustrated in the three panels of Fig. 7, which show the results of theoretical calculations of the effect for the December solstice, increasing the intensity of the auroral electrical currents causes the upper atmosphere circulation patterns to reverse at some latitudes. The latitude of reversal depends upon the intensity of auroral current heating.

V. THE SUN-EARTH CONNECTION: IMPLICATIONS FOR HUMAN ACTIVITY

The magnetic disturbances that followed Carrington's white light flare had a dramatic impact on the modern technology of the time, the telegraph. Indeed, during some particularly intense auroras the telegraph lines running from Boston to points both north and south could be operated for several hours without benefit of the battery supplies. Even in daylight when the aurora could not be seen directly, some telegraph lines were disrupted, indicating to the engineers that the aurora must still be present. During a night-time episode, the extraneous voltage measured on a Boston telegraph line was reported to vary with an approximately 30-sec periodicity, coincident with variations in the auroral display.

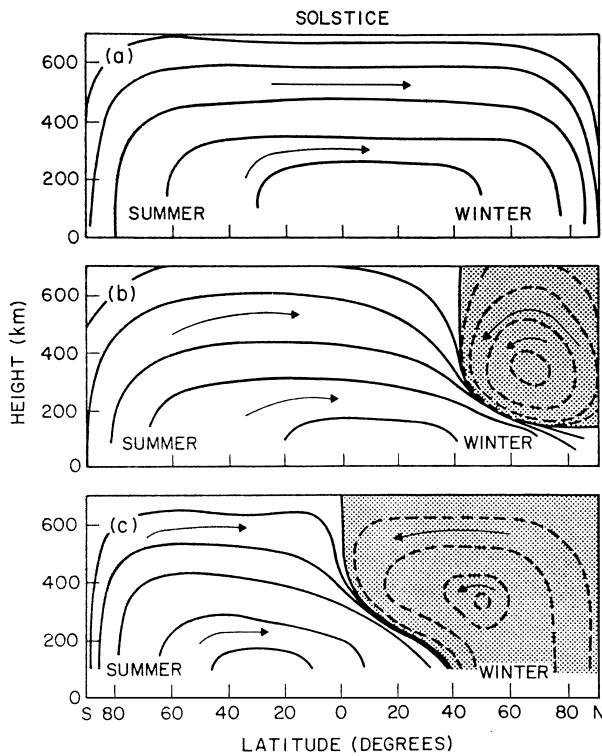


FIGURE 7 Diagrams of the modeled mean meridional atmospheric circulation in the thermosphere at the time of the winter solstice in the northern hemisphere, for three levels of auroral (geomagnetic) activity. The intensity of auroral heating of the upper reaches of the atmosphere increases from the top to the lowest panel. As auroral heating of the upper thermosphere increases, the influences of the heating are felt at lower latitudes, inhibiting circulation from the summer to the winter hemispheres.

Many of the effects of the sun–earth connection on technological activities involve those that use long conductors, such as telephone and telegraph lines, and electric power grids. These phenomena occur because the time-varying geomagnetic field, produced by the variable solar wind, causes electrical currents to flow in the earth. These earth currents enter the long conductors at the points where they are grounded to the earth and can produce deleterious effects on the connected electronics. For example, the building of the Alaskan oil pipeline directly across the auroral zone meant that considerable effort was expended in studying the problem of the induction, by auroral electrical currents, of earth currents in the pipe. Procedures had to be adopted to avoid potential difficulties. While earth currents have been much studied for over a century, their description remains largely empirical for two major reasons. First, it is difficult, if not impossible as yet, to predict the exact spatial scales of the disturbed, time-varying magnetic fields. Second, the inhomogeneities in the conductivity structure of the earth are poorly known, so that

regions of expected higher earth current flow for a given geomagnetic field variation cannot easily be predicted.

The magnetic storm of March 13, 1989, during the 22nd solar cycle caused the most widespread disruptions and damages to electrical power systems in North America yet observed. Montreal and most of the Province of Quebec, Canada, were without electrical power for as much as 9 hr due to the effects of auroral-produced currents on the long power lines stretching from Hudson's Bay to the south. High-voltage transformers at power generation stations in New Jersey were burned out. The northeastern United States was spared an electrical “blackout” condition only by a small margin, primarily because the largest effects occurred during the middle of the night and because the weather was somewhat mild. During the magnetic storm, anomalous voltage fluctuations as large as 700 to 800 V were measured on the repeater power line on the TAT-8 fiber optics cable across the Atlantic Ocean. Large repeater line voltage fluctuations continued for more than a day.

Prospecting techniques for mineral resource deposits using air and satellite surveys often employ magnetic sensing. Natural geomagnetic field fluctuations can seriously impair the interpretation of such surveys; the spurious fluctuations must be removed by various analytic techniques to resolve geologically important features. It is particularly difficult to perform subtractions of magnetic disturbances in the auroral zones, and ground and aerial survey parties are scheduled, in so far as feasible, to operate during geomagnetically quiet intervals. The ability to predict such intervals is still somewhat rudimentary but is, of course, of considerable economic importance to prospecting companies.

The development of more sophisticated technologies on the ground and the extension of entirely new technologies into space has brought a new and significant set of practical concerns. Illustrated schematically in Fig. 8 are many of the contemporary technologies that are impacted by the solar–terrestrial environment. As illustrated, impacted technologies range from radio transmissions to global positioning to communications to human flight in aircraft and in spacecraft.

When communications satellites were originally proposed, it certainly was not recognized that the environment in which they would operate would be anything but benign. However, it was quickly realized that the magnetospheric plasmas (including the energetic particles) can significantly alter the properties of solar cells and spacecraft thermal control blankets. They can cause anomalous charging on and inside spacecraft materials. The energetic particles also can damage semiconductor components and produce anomalies in memory devices. Spacecraft have been known to go suddenly into a noncontrolled state because of radiation effects on electronic components.

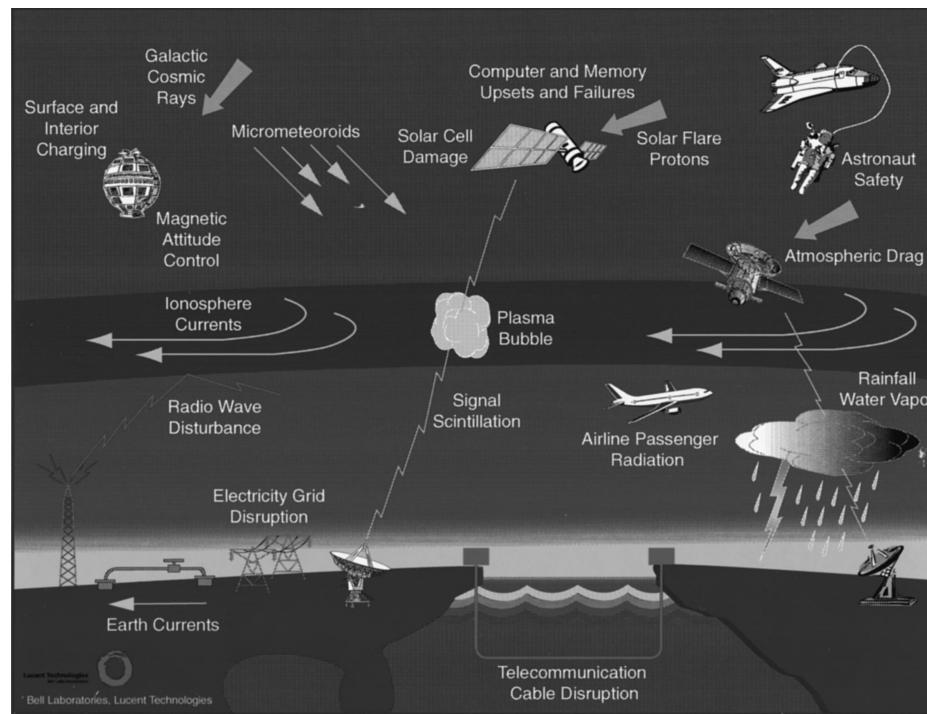


FIGURE 8 Schematic illustration of contemporary technologies that can be affected by solar-terrestrial processes.

The space shuttle and the international space station, flying some 200–300 km above the surface of earth, and other spacecraft in similar low-altitude orbits, have been found to have a visible glow around them. This glow is produced by the interaction of the vehicle with the space environment at these altitudes, mostly with oxygen ions. However, the precise physical and chemical processes involved are not well understood. Such a glow, if present around earth-sensing and astronomical telescopes, can affect the sensitivity and operational conditions of these instruments.

The possibility that humanity will eventually establish a permanently inhabited base on the moon and undertake flight to the planet Mars continues to stimulate the imagination of many people, including scientists and engineers. At the same time, discussions of the tradeoffs in benefits between robotic exploration versus sending humans continues to excite much debate. The matter of human safety in such exploration endeavors is of foremost importance. Exposure to particle radiation is high on the list of safety concerns. During solar maximum conditions the probability for large solar flares and CMEs, and hence large fluences of radiation, is highest. However, solar maximum conditions suppress the intensities of galactic cosmic radiation reaching the earth's orbit. The particle radiation is highly variable within a solar cycle, and from cycle to cycle. For example, in the 22nd solar cycle, the total fluence of solar radiation from solar activity during 1989 exceeded

the fluence measured during the entire previous solar cycle. Such fluence levels could be fatal to astronauts. The as yet largely unpredictable occurrences of solar flares and CMEs mean that there is a large human safety uncertainty at present in undertaking a moon–Mars program, an uncertainty that requires research into biological responses to radiation, solar prediction schemes, and engineering shielding possibilities.

The possible influence of the variable features of the sun—solar optical radiation (the solar constant), solar activity, and solar wind—on the lower atmosphere of earth, where weather patterns are established, remains a great enigma. Statistically, certain elements of climate (and weather) in some regions appear to be related to solar variability, particularly the longer-interval solar cycles. The several elements of solar variability may all be interrelated in complex ways, which thus far have inhibited understanding of the largely statistical results. Variability in the solar constant can be incorporated into modern atmosphere circulation models to test for cause and effect relations. On the other hand, relating variabilities in solar activity and the solar wind to weather and climate (if these quantities are uncoupled from the solar constant) is fraught with difficulties (to say nothing of controversies) and presents the significant problem of identifying the physical driving mechanism(s). If such a mechanism (or mechanisms) exists to couple the solar particle and magnetic field flows to the lower atmosphere, it remains to be

discovered in the complex and interrelated atmosphere-ionosphere-magnetosphere system. Some evidence for this coupling has been revealed in the last decade or so. Optical measurements above thunderclouds, to altitudes as high as the ionosphere, have demonstrated that there is some type of a coupling, at least on occasion. But the importance of this for the weather/climate problem remains uncertain.

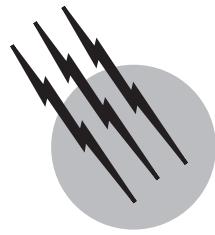
SEE ALSO THE FOLLOWING ARTICLES

AURORA • GEOMAGNETISM • IONOSPHERE • MAGNETIC FIELDS IN ASTROPHYSICS • SPACE PLASMA PHYSICS • SOLAR PHYSICS • SOLAR SYSTEM, MAGNETIC AND ELECTRICAL FIELDS

BIBLIOGRAPHY

Burch, J. L. (April 2001). "The fury of space storms," *Scientific Am.* **284**.
 Campbell, W. H. (1997). "Introduction to Geomagnetic Fields," Cambridge Univ. Press, New York.
 Cravens, T. E. (1997). "Physics of Solar System Plasmas," Cambridge Univ. Press, New York.
 Foukal, P. (1990). "Solar Astrophysics," Wiley, New York.
 Friedman, H. (1986). "Sun and Earth," Scientific American Books, New York.

- Gombosi, T. I. (1998). "Physics of the Space Environment," Cambridge Univ. Press, New York.
 Hargreaves, J. K. (1992). "The Solar-Terrestrial Environment," Van Nostrand, Reinhold, New York.
 Hastings, D., and Garrett, H. (1996). "Spacecraft-Environment Interactions," Cambridge Univ. Press, New York.
 Herman, J. R., and Goldberg, R. A. (1978). "Sun, Weather, and Climate," NASA SP-426, National Aeronautics and Space Administration, Washington, DC.
 Kallenrode, M.-B. (1998). "Space Physics," Springer Verlag, Heidelberg.
 Kappenman, J. G., Albertson, V. D., and Lanzerotti, L. J. (1990). "Bracing for geomagnetic storms." *IEEE Spectr.* **(3)**27.
 Kennel, C. F., Lanzerotti, L. J., and Parker, E. N. (eds.) (1979). "Solar System Plasma Physics," 3 vols., North Holland, Amsterdam.
 Kivelson, M. G., and Russell, C. T. (eds.) (1995). "Introduction to Space Physics," Cambridge Univ. Press, New York.
 Lanzerotti, L. J., and Krimigis, S. M. (1985). "Comparative Magnetospheres," *Physics Today* **38**(11), 24.
 Lanzerotti, L. J., Thomson, D. J., and MacLennan, C. G. (1999). Engineering Issues in Space Weather, In "Modern Radio Science 1999," Oxford Univ. Press, p. 25.
 National Research Council (2000). "Radiation and the International Space Station," National Academy Press, Washington, DC.
 Odenwald, S. (2001). "The 23rd Cycle," Columbia Univ. Press, New York.
 Parker, E. N. (2000). "The physics of the sun and the gateway to the stars," *Physics Today* **43**(6).
 Tribble, A. C. (1995). "The Space Environment," Princeton Univ. Press, Princeton, NJ.
 Volland, H. (1984). "Atmospheric Electrodynamics," Springer Verlag, Heidelberg.



Terrestrial Atmospheric Electricity

Leslie C. Hale

Pennsylvania State University

- I. Historical Note
- II. Classical Global Electrical Circuit
- III. Atmospheric Electrical Conductivity
- IV. Thunderstorm Generator
- V. Fair-Weather Field
- VI. Coupling with Ionospheric, Magnetospheric, and Solar Effects
- VII. Recent Developments

GLOSSARY

Air–earth current Current to the surface of the earth from the atmosphere, usually given as a current density. The fair-weather air–earth current is several picoamperes/(meter)² (pico = 10^{-12}).

Columnar resistance Total resistance of a 1-square-meter column of air between the earth and the “ionosphere.” This is largely determined by the first few kilometers above the surface and is typically $1-2 \times 10^{17} \Omega$.

Field changes Discrete temporal jumps in the electric field of thunderstorms coincident with lightning strokes or flashes, caused by the neutralization of charges due to lightning currents.

Galactic cosmic rays Very high-energy particles from outside the solar system, many of which penetrate to

the surface of the earth and are the principal source of atmospheric ionization in the troposphere and stratosphere.

Ionospheric potential Electric potential with respect to earth of the highly conducting regions of the upper atmosphere, assumed to be equipotential (which is not always true).

Lightning flash Most spectacular visible manifestation of thunderstorms, consisting of one or more discrete lightning strokes and frequently resulting in the neutralization of tens of coulombs of previously separated charge, with the release of the order of 10^8 to 10^9 joules (J) of energy.

Magnetic storm Variations in the earth’s magnetic field associated with enhanced auroral activity, usually following solar flares and solar wind enhancement (solar activity).

Magnetosphere Region of the earth's magnetic field above the ionosphere, which is also the site of the high-energy trapped electrons constituting the Van Allen radiation belts.

Maxwell current density Mathematical curl of the magnetic field vector \mathbf{H} , equal to the vector sum of all current densities, which in the atmosphere is usually limited to conduction, convection, diffusion, lightning, and displacement current terms.

Relaxation time Characteristic time (electrical) for the atmosphere below the ionosphere, equal to the free space permittivity ϵ_0 divided by the electrical conductivity σ . Also known as the screening time.

Schumann resonances Characteristic frequencies of the earth-ionosphere cavity, which are approximately multiples of 7 Hz.

Solar wind Outflow of charged particles from the sun's corona. The solar wind tends to exclude galactic cosmic rays from the earth's environment.

Solenoidal field Field whose flux lines close and whose flux is divergenceless. The divergence of the curl of any vector field is zero. Any field that is the curl of a vector field is solenoidal.

Transient luminous event (TLE) Visible optical events in the upper atmosphere associated with terrestrial lightning; includes "red sprites," "blue jets," "elves," "halos," and others.

TERRESTRIAL ATMOSPHERIC ELECTRICITY is concerned with the sources of atmospheric electrification and the resulting electrical charges, fields, and currents in the vicinity of the earth. The earth and ionosphere provide highly conducting boundaries to the global electrical circuit, which consists of thunderstorm generator and the fair-weather load. Although originally conceived as a dc circuit confined principally to the lower atmosphere, it is now known to have ac components over a broad frequency range and to penetrate deeply into space, coupling with extraterrestrial phenomena.

I. HISTORICAL NOTE

In 1752 T. D'Alibard and B. Franklin independently confirmed the electrical nature of thunderclouds; in the same year L. Lemonnier observed that electrical effects occurred in fine weather. Not until the late nineteenth century was the electrical conductivity of air established by several workers, and the discovery of cosmic ray ionization by V. Hess and the postulation of the global electrical circuit by C. T. R. Wilson came in the early twentieth century. Now known as the classical global cir-

cuit, this concept has stood the test of time and is still generally agreed to be correct. However, some observations cannot be fully explained by classical theory. In any case the classical theory provides a convenient framework in which to discuss topics in atmospheric electricity and solar-terrestrial relationships of an electrical nature.

II. CLASSICAL GLOBAL ELECTRICAL CIRCUIT

The classical concept of the global electrical circuit of the earth is of an electrical system powered by upward currents from thunderstorms all over the earth (totaling ~ 1500 A) reaching and flowing through the "ionosphere" (an equipotential conducting layer in the upper atmosphere) and returning to earth in "fairweather" regions. This viewpoint is generally attributed to C. T. R. Wilson and was used by the late R. P. Feynman as a premier example of the triumph of scientific methodology. The key evidence is that the fair-weather electric field, measured in polar, oceanic, and high-mountain regions not subject to certain local effects, generally shows a diurnal variation (frequently called the Carnegie curve, from an early expedition on which it was observed) that is relatively stationary in universal (UT) rather than local time (LT) and that shows a peak at ~ 1900 UT. Since the atmospheric convective activity that produces thunderstorms is greater over land than over water and since thunderstorm activity peaks in the afternoon, the maximum current to the global circuit would be expected to occur when it is afternoon at the longitude of the centroid of the largest concentration of continental thunderstorm activity, which is observed. More detailed analysis shows bumps corresponding to particular continents (Fig. 1).

Considerable time averaging is implicit in these curves. They are much less stationary on a day-to-day basis and are clearly affected by various geophysical events, as described later.

The theory outlined above, frequently called the classical theory of atmospheric electricity, has stood for many decades and is still held by consensus to be generally true.

We shall examine the following topics: atmospheric electrical conductivity; the thunderstorm as a generator of both dc and ac currents to the ionosphere; the fair-weather field, including its variability; and possible coupling between global circuit elements and solar-magnetospheric effects. The author pleads guilty to over-simplification in order to discuss topics that have run to book-length references, which are recommended to the reader. Some recent observations and theory are discussed in VII.

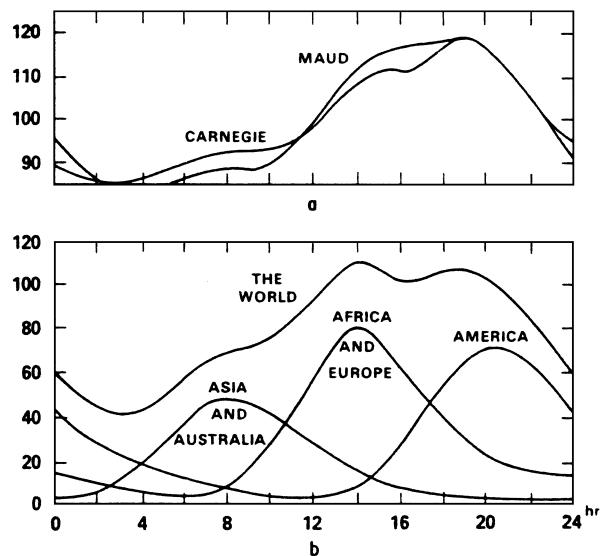


FIGURE 1 (a) Relative diurnal variation in fair-weather field in the Arctic Ocean in northern winter (Maud) and in all oceans (Carnegie). [From Parkinson, W. C., and Torrenson, O. W. (1931). "The diurnal variation of the electrical potential of the atmosphere over the oceans," *Compt. Rend. de l'Assemblée de Stockholm, 1930*, IUGG (Sect. Terrest. Magn. Electr. Bull. **8**, 340–345).] (b) Diurnal variation in expected thunderstorm areas in units of 10^4 km^2 . [After Whipple, F. J. W., and Scrase, S. J. (1936). "Point discharge in the electric field of the earth," *Geophys. Memoirs (London)* **8**, 20.]

III. ATMOSPHERIC ELECTRICAL CONDUCTIVITY

An early theory postulated that the earth is electrically charged because it was created that way. We now know that the air has sufficient electrical conductivity that such a charge would leak off in less than an hour without a recharging mechanism.

This conductivity is due to electrically charged particles initially created by ionizing radiation, primarily galactic cosmic rays, which create ion-electron pairs. In the lower atmosphere the electrons attach rapidly to form negative ions. Subsequent photochemical reactions, attachment to and charge exchange with aerosol particles, and clustering with water molecules tend to create ions whose size varies over several orders of magnitude. Their eventual loss is due primarily to ion-ion recombination. The electrical mobility of even multiply charged aerosol particles is generally much smaller than that of small molecular ions, hence charge immobilized on aerosol particles contributes less to conductivity, causing aerosol-laden air to possess much lower conductivity. Other sources of atmospheric ionization include natural and anthropogenically induced radioactive material (e.g., soil radioactivity and radioactive gases from the earth and ^{85}Kr routinely

released from nuclear reactors) and "point discharge" or corona ions emitted from surface objects in regions of high electric fields near thunderstorms (e.g., "St. Elmo's fire").

Galactic cosmic rays in the gigaelectronvolt energy range are believed to be the principal source of atmospheric ionization in the undisturbed atmosphere below ~ 60 km altitude. They are subject to some variability in flux with the 11-yr sunspot cycle, and from high latitudes to equatorial regions their ionization rate in the lower atmosphere decreases by a factor of ~ 10 , due to screening effects of the earth's magnetic field. Solar activity (an enhanced "solar wind") tends to decrease the galactic cosmic ray flux, and when this occurs over a few days the phenomenon is known as a Forbush decrease. Although Forbush decreases at the surface can exceed 10%, a much more common response to the frequently occurring solar activity that causes "magnetic storms" is a variation in ionization rate in the lower atmosphere of the order of 1%. Major solar proton events send large fluxes of tens of megaelectronvolt-range protons toward the earth several times per year. These events produce enhanced ionization as low as 20 km at high latitudes above $\sim 60^\circ$ magnetic (referring to a coordinate system based on the earth's magnetic field).

At altitudes above ~ 50 km the variability in electrical conductivity is much greater. Ionizing radiation (principally hydrogen Lyman- α) from the sun penetrates to about this altitude, as do high-energy electrons associated with aurora. X-rays from solar flares also penetrate the "mesosphere" (about 50 to 85 km), and *bremssstrahlung* from high-energy electrons sometimes even deeper. Conversely, aerosol clouds that form in the mesospheric polar night, in "noctilucent clouds," and to a lesser extent in the undisturbed nighttime mesosphere are expected to decrease the electrical conductivity greatly, at least at night. (In the daytime, aerosol particles exposed to the strong ionizing radiation of the upper mesosphere can be sources as well as sinks for free electrons, hence their effect on conductivity is complex.)

Figure 2 shows typical profiles of electrical conductivity with altitude under a number of conditions. The enhancement due to solar proton events is principally at high latitudes.

IV. THUNDERSTORM GENERATOR

Thunderstorms (and other electrified clouds) have generally been regarded as generators of slowly varying or "dc" current to the global circuit; lightning is of more interest as a radiator of electromagnetic energy in the kilohertz to megahertz range. It also appears that there is a substantial variability in the thunderstorm source currents on time

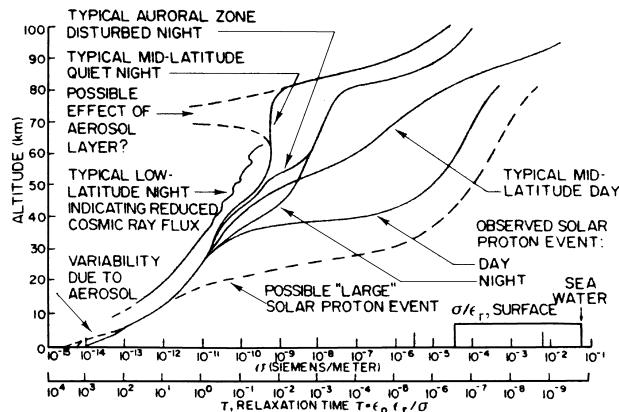


FIGURE 2 Approximate atmospheric electrical conductivity profiles under various conditions.

scales of the order of 10^{-3} to 10^3 sec, constituting a large “ac” component of current to the global circuit.

Meteorological processes in thunderstorms are believed to separate positive from negative charge. Current theories involve the separation of charge by collisional interactions between various forms of ice and water particles, with the aggregation of negative charge on larger hydrometeors and positive charge on smaller particles. The larger negative particles are affected to a greater extent by gravity, tending to fall faster, and the smaller positive particles may move more rapidly upward in regions of convective up-draft. The net effect is to produce more positive charge near the top of the cloud and more negative near the bottom (Fig. 3a). Frequently a net negative charge is produced on the cloud, as positive charge may leak off more rapidly to the ionosphere due to the conductivity gradient of the atmosphere. A smaller positive charge is sometimes found near the bottom of the cloud, but the overall effect can be approximated by a “dipole” with the positive charge several kilometers above the negative. The schema of Fig. 3b has been used by numerous workers to predict dc fields and currents. The meteorological charge separation is represented by a current generator, maintaining the up-

per ($+Q_u$) and lower ($-Q_l$) charges, which continuously dissipate in the conducting atmosphere. An early application of this model used a symmetric dipole $Q_u = Q_l$ to explain electric field data obtained in an aircraft overflight (Fig. 4a). It should be pointed out that this overpass was one of the relatively few not showing evidence of at least one lightning flash (Fig. 4b) and that the passes indicating the largest currents showed evidence of many flashes (Fig. 4c). By the use of a “lightning-free” model, equations (known as the Holzer-Saxon equations) were derived that describe the fields of the model of Fig. 3b for an assumed exponential conductivity profile. In this model, using the equilibrium relation $Q = I\tau$, where $\tau = \varepsilon_0/\sigma$ is the relaxation time or lifetime of the charge, yields $Q_u < Q_l$, which is more commonly observed. It has been pointed out by J. S. Nisbet that since, according to the Holzer-Saxon equations, usually nearly all of the current to the upper charge center reaches the ionosphere, the net fraction of I_m that reaches the ionosphere in the lightning-free case is highly dependent on the height of the lower charge center. The Holzer-Saxon equations usually show that over half of I_m reaches the ionosphere and global circuit (I + GC). It should be pointed out that cloud-to-earth lightning, corona or point-discharge, and precipitation currents simply provide alternative paths for I_m to reach the I + GC; they are not independent sources. Intracloud lightning, on the other hand, reduces the current that reaches the I + GC and thus has an important effect on the dc global circuit.

It will be noted that the single lightning flash of Fig. 4b produced a major perturbation of the electric field observed above a thunderstorm, with a pulse of ~ 10 -sec duration, a result confirmed later by other aircraft overflights. Data obtained from a rocket-launched parachute-borne probe over a thunderstorm in 1981 (Fig. 4d) revealed that pulses of similar width occurred as high as 47 km altitude. This led to the conclusion that, rather than the expected (by some) evanescent transient, these waveforms represented pulses that flow through the entire global circuit. Their source can be explained with the crude equivalent circuit of Fig. 3c. The meteorological generator I_m charges the cloud capacitance C_c directly and, in parallel, charges the cloud-to-earth capacitance C_e through the entire ionosphere and global circuit. It is then a race to see which capacitor discharges first, but since both cloud-to-earth and intracloud flashes do occur they must be charged to comparable potentials. Thus, when one capacitor discharges (creating a lightning flash), the other one dumps a comparable amount of energy into the entire ionosphere and global circuit, including the earth. It will be noted that charging and discharging C_c contributes nothing to the net dc global circuit current I_{gc} , but both the charging current and lightning discharge of C_e (causing C_c to discharge partially through the global circuit) contribute to I_{gc} .

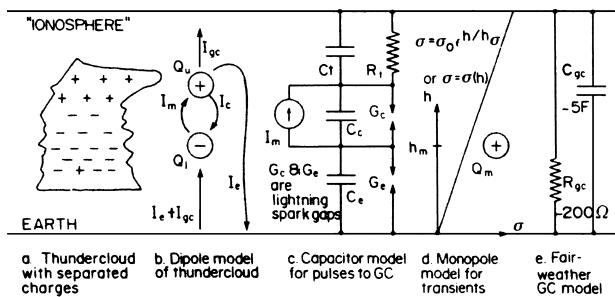


FIGURE 3 Approximate models for describing thundercloud and global circuit (GC) as electrical entities.

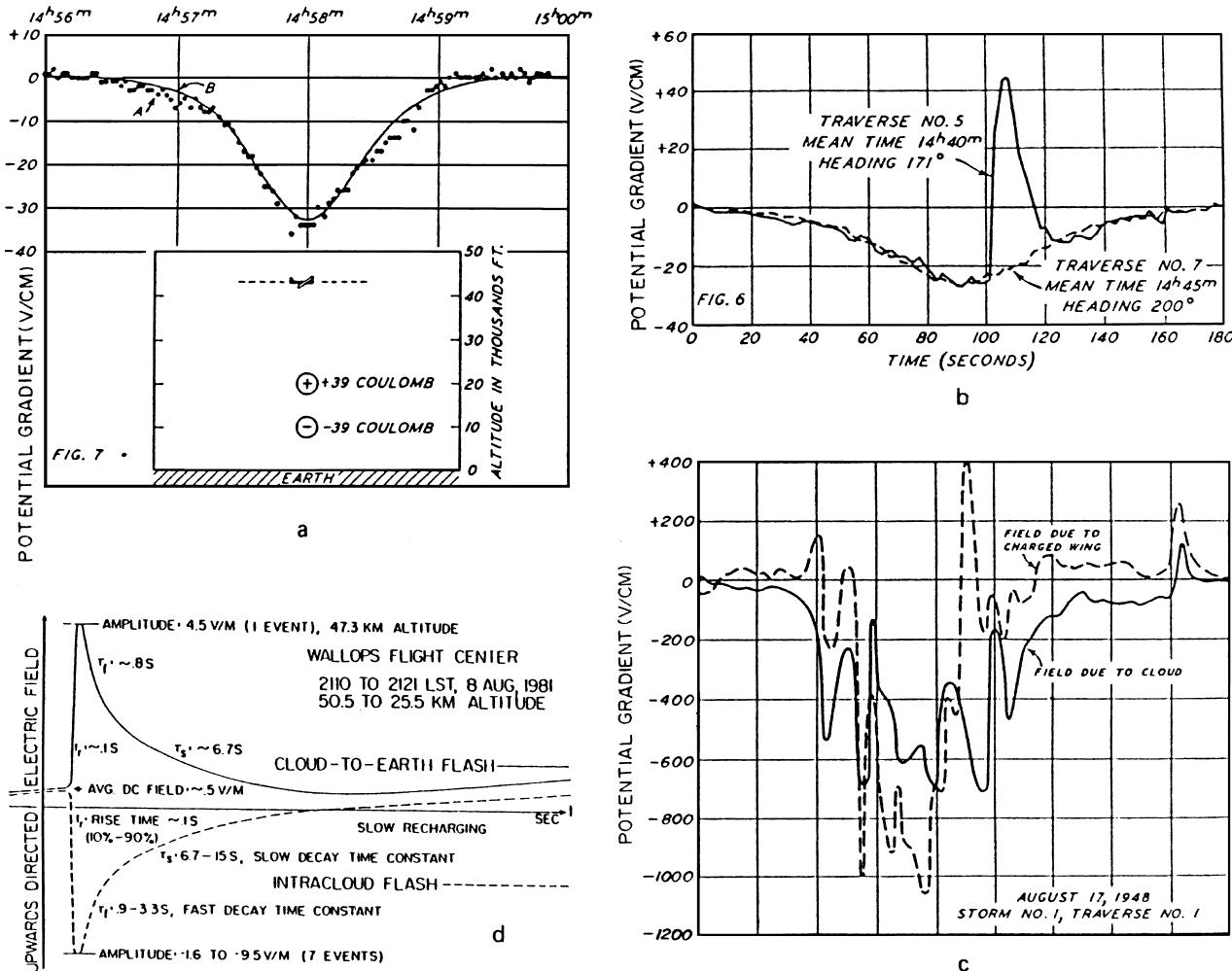


FIGURE 4 Measurements of electric fields above thunderstorms. (a) Potential gradient (negative of electric field) for aircraft traverse above thunderstorm. A, Observed; B, calculated for model shown. (b) Potential gradients over thunderstorms showing effect of single lightning flash. (c) Potential gradient over storm of large current output to global circuit. Note enhanced magnitudes and evidence of multiple lightning flashes. (d) Schema of electric fields observed at very high altitudes showing persistence of wide pulses. [Graphs (a), (b), and (c) are aircraft measurements of Gish, O. H., and Wait, G. R. (1950). *J. Geophys. Res.* **55**, 473. (d) Rocket-parachute measurements reported by Hale, L. C. (1983). In "Weather and Climate Responses to Solar Variations" (B. M. McCormac, ed.). Colorado Associated University Press, Boulder.]

It is noted that storms with frequent lightning supply more dc current to the I+GC (see Fig. 4). This is probably because such storms have more intense "meteorological" generators (I_m), and not because the lightning acts as a separate source.

Although the crude model of Fig. 3c predicts the general characteristics of the large pulses to the ionosphere, it does not accurately predict their shape, which is generally nonexponential, with rise times slower than what might be expected from lightning. It should be possible to predict accurately the shape of these pulses with numerical and possibly analytical modeling. A model for calculating transients is shown in Fig. 3d. It is related to the fact that

the "field changes" observed near thunderstorms can be explained by adding charge to the preflash charge distribution and calculating the fields due to the added charge, a procedure first suggested by C. T. R. Wilson. For a cloud-to-earth flash one "monopole" is added to the lower charge center, equal and opposite to the charge lowered to earth by the flash. A number of published papers have used a similar monopole model to study the transient problem, and an update on this work is provided in Section VII. Figure 3e is an approximate model for the global circuit for direct and alternating currents at frequencies below a few hertz.

It is generally agreed that several thousand simultaneously active thunderclouds contribute an average of a

fraction of an ampere each to the global circuit. It has also become apparent that rapid variability, due mainly to lightning, also contributes a comparable amount of alternating current. Most of the energy due to these sources dissipates locally near the thunderstorm, with the dc power dissipating in the high-resistance lower atmosphere and the ac power capacitively coupled to higher altitudes dissipating at the altitudes where the atmospheric electrical relaxation time τ corresponds to the reciprocal of the source angular frequency $1/\omega$. Lightning phenomena also radiate electromagnetic energy at frequencies principally in the kilo- to megahertz range, forming the source of much “radio noise.”

The vast majority of lightning flashes consist of several short duration “strokes,” and lower negative charge (up to several coulombs) to earth. Lightning can be “normal” or “negative;” however, a small fraction of lightning is “positive.” These events are generally larger and consist of a single stroke. They tend to possess longer duration currents in the lightning channel after the stroke and to be primarily responsible for TLE such as “red sprites.” Because of the longer “continuing currents” (milliseconds or greater), positive lightning is regarded as more dangerous for starting fires.

The upward dc currents flow with little spreading in a rapidly increasing conductivity profile to the ionosphere, where they combine to establish the ionospheric potential and return to earth in fair-weather regions. In the ionosphere, the very lowest frequency ac currents will follow similar trajectories. The ionosphere and earth form a cavity resonator with characteristic frequencies that are multiples of ~ 7 Hz, so that above a few hertz these “Schumann” resonances must be considered. At higher frequencies, complex modes dominate the ac propagation. However, much of the ac energy is at lower frequencies, where a simpler viewpoint prevails.

A useful viewpoint is that a relatively invariant quantity associated with thunderstorms is the Maxwell current density defined as $\text{curl } H$, where H is the magnetic field due to the thunderstorm currents. This quantity is divergenceless ($\text{div curl } H = 0$) and solenoidal, and hence can be followed through any circuit. In the atmosphere outside thunderclouds this generally consists of the sum of conduction current σE and displacement current $\varepsilon_0 \partial E / \partial t$ (E , electric field; σ , electrical conductivity; ε_0 , permittivity). The divergenceless nature of the Maxwell current means that, following it around a flux tube, the magnitude of the individual components of flux may vary with their sum remaining constant. For example, it has been shown that the Maxwell current near a thunderstorm is invariant through a layer of high-conductivity air caused by point discharge (corona) ions near the earth’s surface, with increased conduction current balanced by decreased dis-

placement current. It is not true that the Maxwell current viewpoint always leads to great simplification, because during transients the shape of the solenoidal field may vary with time. On the other hand, it may be dangerous to use the clearly valid approximate condition $\text{curl } E \approx 0$ to obtain unique solutions, because $\text{curl } E \equiv 0$ mathematically yields a transient solution permitting zero time-varying Maxwell current density, which is contrary to experimental data, although this solution has frequently been used incorrectly. Model studies have shown that the global circuit currents are not crucially dependent on cloud conductivity and are not, to first order, affected by the “screening” charge that forms at sharp conductivity gradients, such as cloud boundaries.

From typical measurements of the vertical electric field (Figs. 4c,d), it can be seen that, for appropriate conductivity values ($\sim 10^{-12}$ S/m for Fig. 4c and $\sim 10^{-10}$ S/m for Fig. 4d) the maximum displacement current density ($8.85 \times 10^{-12} \partial E / \partial t$) is greater than the maximum conduction current density. For the large storm of Fig. 4c the root-mean-square ac Maxwell current is comparable to the dc current. Thus, large thunderstorms may be as effective in generating ac currents as dc currents, and much more ac energy is coupled to the upper atmosphere. This ac generator mechanism may be the principal source of atmospheric noise in the ELF range between the “micropulsation” region below ~ 1 Hz and electromagnetic radiation from lightning currents in the kilohertz range and above.

V. FAIR-WEATHER FIELD

Thunderstorm currents totaling ~ 1500 A flow to the ionosphere, which is an approximately equipotential layer in the upper atmosphere, and return to earth in fair-weather regions. The resistance between the ionosphere and earth is ~ 200 Ω , giving rise to an ionospheric potential of ~ 300 kV. The vertical electric field near the earth’s surface is of the order of 100 V/m, and the associated current density is typically several picoamperes/m². The ac component of the fair-weather field is relatively much smaller than the ac component of individual thunderstorm currents.

The upward currents from individual thunderstorms that do not return to earth or the bottom of the cloud locally flow to the ionosphere, where they combine to establish the ionospheric potential. The height of the ionosphere used to be taken, for global circuit purposes, to be ~ 60 km, but it is now generally agreed that the horizontal currents that close the global circuit generally flow in the much higher ionosphere of the radiophysicist, above 100 km. The principle is that they tend to spread very little, flowing upward in a monotonically increasing conductivity profile, which frequently persists to the ionosphere. Above ~ 70 km the

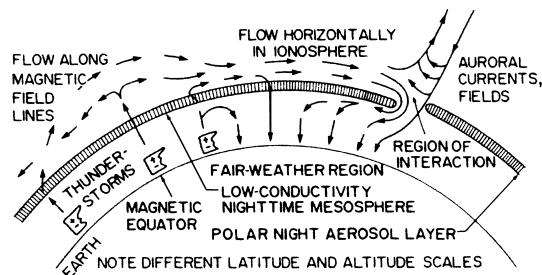


FIGURE 5 Longitudinal cross section showing global circuit current paths at night with one possibility for coupling with auroral (Birkeland) currents (morning sector).

conductivity ceases to be a simple scalar quantity but possesses directional properties, with the highest conductivity along the direction of the magnetic field, whereby the currents can flow directly into the magnetosphere along magnetic flux lines (Fig. 5). Returning along the conjugate flux line in the opposite hemisphere in a decreasing conductivity profile, the currents tend to spread very rapidly in the ionosphere. Ultimately, the largest horizontal currents are thought to flow in the region of highest conductivity transverse to magnetic field lines, which occurs between about 100 and 120 km. Here, they must merge with the very much higher magnitude ionospheric current systems (typically 10^5 A) which also flow at these altitudes. The thunderstorm currents return to earth in fair-weather regions, which include the entire surface of the earth not directly involved in thunderstorms. The ionospheric potential is approximately given by the current (~ 1500 A) multiplied by the total resistance between ionosphere and ground, which has been determined to be $\sim 200 \Omega$, and hence is ~ 300 kV.

There are numerous factors that modify this general picture of the fair-weather field. Except in the polar regions, local convective activity in the earth's boundary layer produces a current that frequently obscures the Carnegie curve in electric fields and currents measured in the lower atmosphere over land. This is generally attributed to the effects of local atmospheric convective activity on space charge (due to an excess of ions of one sign of charge, usually positive, which is a necessary consequence of the global circuit currents in a conductivity gradient). Frequently, there is a strong correlation with water vapor, which is controlled by temperature and convective "exchange" processes. Water vapor reduces conductivity by forming ionic clusters of lower mobility. Aerosol particles from many sources (volcanoes, industry, fires, clouds, blowing dust and snow, etc.) tend to reduce the conductivity and increase the magnitude of the electric field, often with relatively constant current density. There is a clearly identified "sunrise" increase in electric field, which has also been considered to be caused by convective exchange processes.

In addition to locally induced variations in the fair-weather field, there are variations of much larger extent. There is a well-established seasonal dependence in the ionosphere potential, with a maximum in Northern Hemisphere winter. The reasons for this are not clear but are probably related to thunderstorm activity in tropical regions, which are believed to be the source of most of the global circuit current. Long-term observations of the ionospheric potential have shown variations in the range of 200 to 400 kV, probably due to variations in source current and total resistance. An expected 11-yr variation due to the effects of the sunspot cycle on galactic cosmic rays has been obscured by other factors, including aerosol particles from volcanoes. It has been predicted by W. L. Boeck that, even without nuclear disasters, the routine release of radioactive ^{85}Kr from nuclear power plants may eventually substantially affect the global circuit by increasing atmospheric conductivity, if the nuclear power industry expands. If, as some have suggested, the fair-weather conductivity (atmospheric ionization) and/or electric field play a role in initial thunderstorm development, this could potentially alter weather and climate. Some evidence exists for the effects of the release of radioactivity from nuclear reactors. In the weeks following the Chernobyl incident in May 1986, an order of magnitude enhancement in lightning frequency was observed in Central Sweden, along the path of radioactive "fallout," as compared with stationary patterns observed over many years.

The dc currents from thunderstorms add in an arithmetical manner to produce the total fair-weather current, with the majority of storms producing upward currents. The ac currents, however, tend to have random phase and add vectorially to produce a much smaller resultant, inasmuch as they are independent. This, combined with filtering due to the global circuit capacitance, means that the ac currents are not a major factor in the fair-weather electric field, although they may contribute to the noise background in the hertz to kilohertz range. (Other factors here are turbulent local space charge, fields of nearby thunderstorms, and radiated electromagnetic fields from distant lightning.) Large, independent individual events, such as "superbolts" or correlated bursts of lightning flashes, may appear as distinct perturbations of the fair-weather field.

VI. COUPLING WITH IONOSPHERIC, MAGNETOSPHERIC, AND SOLAR EFFECTS

It has been known for decades that there is coupling between elements of the global circuit and magnetospheric and solar phenomena. The modulation of galactic cosmic

rays by solar activity affecting atmospheric electrical conductivity is possibly the best understood relationship, but there are many others that are accepted to various degrees. Some that cause “downward” coupling have come to the fore in the search for sun-weather relationships, but there are also cases in which thunderstorms affect the ionosphere and magnetosphere.

For many decades it has been known that lightning produces audio-frequency “whistlers.” These are very low-frequency radio waves whose frequency decreases over a period of the order of 1 sec, produced by the interaction of very low-frequency radio waves from lightning with magnetospheric plasma (ionized particles) to produce amplified waves. It has been established by several groups that lightning-related events can produce precipitation or “dumping” of high-energy trapped electrons from the Van Allen radiation belts (e.g., Fig. 6). These electrons may create enough ionization enhancement in the ionosphere to affect ionospheric radio propagation measurably for periods of the order of 1 min. Such events are called Trimpie events after their discoverer, M. L. Trimpie. The data show that similar “dumping” events occur in the daytime when very low-frequency waves cannot easily propagate to the magnetosphere. It is an open question whether ac thunderstorm currents, which easily reach the magnetosphere, can trigger such events.

In the magnetosphere, the tensor (directional) electrical conductivity perpendicular to magnetic field lines is very small. Thus, “transverse” fields can “map” downward into the ionosphere. Horizontal fields, of magnetospheric or ionospheric origin, penetrate farther down into the atmosphere with an efficiency dependent on the “scale size” of the field. Fields extending over several hundred kilometers can map nearly to the surface with little attenuation. Such fields can clearly perturb the local ionospheric potential, because it can no longer be equipotential. Effects of large magnitude, however, were usually thought to be confined to auroral zone and polar cap regions. However, there is some evidence for relatively large effects penetrating to lower latitudes.

A number of both earth-based and balloon or aircraft measurements at mid-latitudes have indicated 10–30% variations in local ionospheric potential, electric field, and air–earth current, persisting for up to several days following moderate solar or magnetic events that occur more than once per month. These typically involve an approximate doubling of solar wind speed, moderate magnetic and enhanced auroral activity, and a small Forbush decrease in galactic cosmic ray flux. (The electrical events have also been correlated with movements of the earth across solar magnetic field sector boundaries. Since these also tend to be correlated with solar magnetic events, there has been some difficulty in establish-

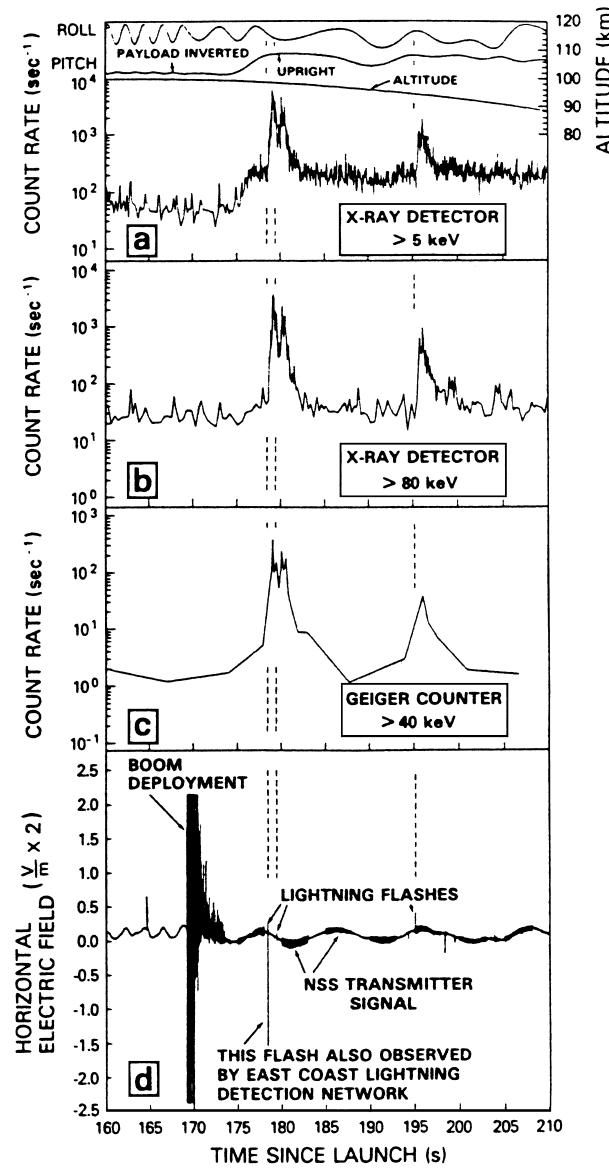


FIGURE 6 Rocket measurements showing effects of lightning on high-energy detectors, indicating lightning-induced precipitation of particles from Van Allen belts. [After Goldberg et al. (1986). *J. Atmos. Terr. Phys.* **48**, 293.]

ing cause and effect.) It has not been conclusively established whether the electrical effects are due to an overall change in ionospheric potential or to “external” electric fields in the atmosphere creating horizontal gradients in the local ionospheric potential. Possibly both sorts of effects are involved. Stratospheric balloon measurements have indicated horizontal fields of tens of millivolts per meter, which are much greater than readily permitted by classical global circuit theory. Carried over thousands of kilometers, these fields could produce the tens of kilovolts necessary to explain the data. An alternative explanation

for the local variations in the fair-weather field is modulation of the overall ionospheric potential due to variations in the thunderstorm source current. If this could be proved true, it would represent a very important sun-weather relationship. Perhaps transient events in the magnetosphere couple to individual thunderclouds and modulate their lightning rates, thus affecting the total global generator current.

VII. RECENT DEVELOPMENTS

Much of the recent work in this field has involved the consolidation of existing knowledge, which has led in some cases to the reaffirmation of earlier theories. For example, the basic dc global circuit has undergone some rehabilitation, along with clarification.

In recent years, partisans of cloud-to-earth lightning, corona or “point discharge,” and currents carried by precipitation have all argued for the relative importance of these various “sources.” A consensus is developing that these are not independent sources, but different processes by which the currents originating from the basic “meteorological” generator are completed through the ionosphere and global circuit (I+GC). An exception to this “consolidation” viewpoint is the role of intra-cloud (IC) lightning, which definitely tends to weaken the global meteorological generator by discharging thunderclouds locally. Thus, IC lightning plays a critical role in determining the currents to the I+GC. (This viewpoint was first expressed to this author by Lothar Ruhnke.)

To recapitulate the basic dc global circuit, air is partially electrified by a number of different processes, and charge separation by meteorological processes produces a number of local electrical generators, primarily in thunderstorms, which act in concert to establish an “ionospheric potential” (IP). This IP then drives a “fair-weather” current back to earth in storm-free regions, with the current density determined by the IP divided by the “columnar resistance.” This latter factor depends largely on aerosol loading of the atmosphere and on orography, with much larger relative currents to mountains and elevated regions such as the Antarctic Plateau. These factors have been embodied in numerous computer models (see B. K. Sapkota and N. C. Varshneya, *J. Atmos. Terr. Phys.* **52**, 1, 1990).

In the last decade or so the most (literally) spectacular observations have been of a variety of visible effects in the upper atmosphere above thunderstorms, the most colorful of which are “red sprites” observed in the nighttime mesosphere. Such phenomena were originally suggested by C. T. R. Wilson (*ca.* 1925), were first observed by J. Winckler and colleagues at the University of Minnesota

in 1989, and were later confirmed by triangulated measurements from two aircraft over the U.S. Great Plains by D. Sentman and a group from the University of Alaska. Both optical and electrical measurements have confirmed the “electrical breakdown” nature of these emissions, which generally initiate at about 75 km and spread both upward and downward, probably by “streamer” mechanisms suggested by V. Pasko and a Stanford group. Sprites are nearly always associated with “positive” lightning, possibly because such events are generally larger and tend to occur in single strokes.

The electrodynamics of such sprites depend on a number of things. First, the extremely low conductivity of the nighttime mesosphere (see Fig. 2) allows the penetration of electromagnetic energy to about 80 km. A quasi-static field due to the large Wilson monopole injected by positive lightning establishes a field between the earth and ionosphere (at about 80 km) sufficient to produce electrical breakdown at about 70 km or above. In conjunction with establishing this field, a “millisecond slow tail” is launched in the earth-ionosphere waveguide (Fig. 7). This waveform can be observed in the electric or magnetic field at thousands of kilometers horizontal distance, and

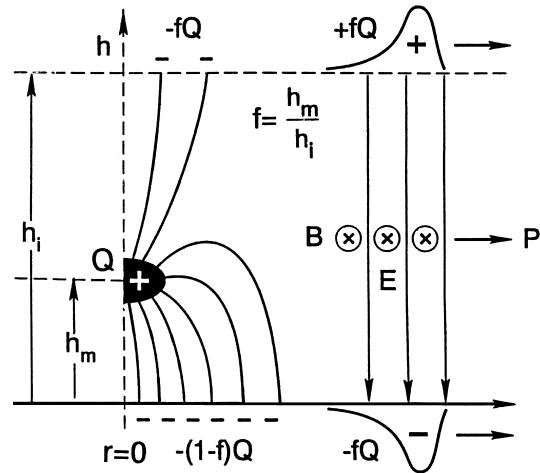


FIGURE 7 (Left) Establishment of a quasi-static field between the earth and the ionosphere at effective height h_i (typically about 80 km at night) by a Wilson monopole Q injected at height h_m by a lightning stroke. The fraction of electric field lines reaching the ionosphere is fQ , where $f = h_m/h_i$. For typical low nighttime mesospheric conductivities, such fields can persist for up to tens of milliseconds, allowing time for most sprites to develop. (Right) In order to satisfy the post-stroke boundary conditions of the field, a roughly one-millisecond “slow tail” is launched in the earth-ionosphere waveguide in the radial TEM mode (flat earth approximation). The initial charge associated with these unipolar wavelets is fQ but actually increases with distance from the source. (The polarities shown are for normal “negative” lightning; for positive lightning, the polarities of all charges are reversed.)

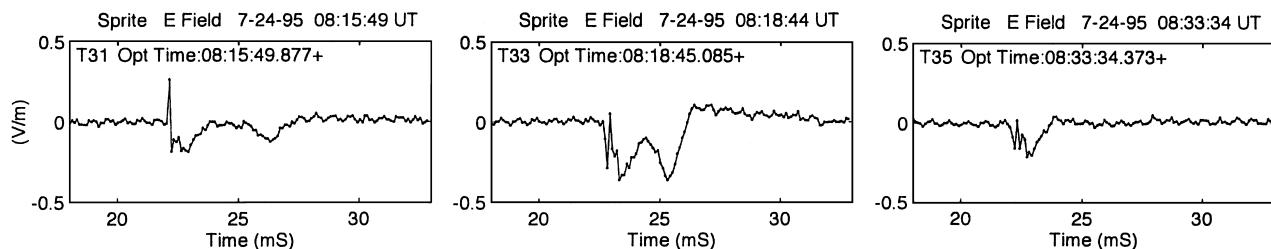


FIGURE 8 Vertical electric field waveforms observed at several thousand kilometers from related optically observed red sprites. The sprites occurred in West Texas and the fields were observed from central Pennsylvania. Similar waveforms were observed in the magnetic field. The initial slow tails in each case were due to the parent lightning, as indicated by the higher frequency components on the leading edge. The subsequent smooth tails are due to the charge separated in the sprite. Note that the middle event has less time delay and is larger in amplitude, and that the one on the right shows no detectable waveform due to the sprite itself, which is frequently the case. [Adapted from Marshall, L. H. et al. (1998). *J. Atmos. Solar-Terr. Phys.* **60**, 771.]

is in a mode (TEM) that can be observed at any altitude, thus not requiring balloons or other spacecraft. (This is not true for mesospheric electrical conductivity profiles; their observation during potential sprite-producing conditions would be extremely useful, but does require rocket launches.)

The development of the sprite usually takes several milliseconds after the establishment of the quasi-static field. This is probably because the breakdown process must start from a few “seed” electrons and requires many “e-folding” times to reach observable amplitude. The sprite discharge separates charge and produces a Wilson monopole at sprite altitudes, thus launching a millisecond slow tail similar to the one launched by the parent lightning. These are frequently seen on the same record (Fig. 8). Sprites are predominantly red and take a variety of shapes. They can appear over tens of kilometers in the mesosphere (~ 50 to 80 km altitude) and although appearing to be primarily vertical in orientation can occur over similar horizontal distances. Their shapes are so varied as to defy description, but it has been suggested that some appear to be generated by “fractal” processes.

Other such phenomena include “blue jets,” which are discharges proceeding upward from a thundercloud into the stratosphere, first observed by E. Westcott and the Alaska group from aircraft and explained as propagating upward discharges by U.S. Inan and Stanford colleagues.

“Elves” are much shorter discharges at higher altitudes, generally 100 km or greater, that were first observed by Fukunishi and colleagues from Tohoku University of Japan in Colorado. These events, which last only a few hundred microseconds, are generally attributed to direct breakdown caused by electromagnetic radiation from lightning.

Observations of red sprites, blue jets, elves, and a number of other phenomena that are currently under study

have largely been made from the Yucca Ridge facility of W. A. Lyons near Fort Collins, CO. This superb facility has been used over several summers by a number of groups to do coordinated studies of these interesting new phenomena. The work has been largely reported at American Geophysical Union meetings (particularly Fall) and published in numerous articles, largely in *Geophysical Research Letters*, the *Journal of Geophysical Research*, and the *Journal of Atmospheric and Solar-Terrestrial Physics*. Such phenomena are becoming known collectively as transient luminous events (TLEs).

The millisecond slow tails described in Fig. 7 may enter into other phenomena. Similar electric field waveforms have been observed to penetrate into the highly conducting ionosphere, parallel to the earth’s magnetic field, and much stronger than had initially been expected according to “shielding” considerations. L. Hale has suggested that this is due to polarization of the ionospheric and magnetospheric plasma as the slow tails pass below. Furthermore, it was suggested that this polarization, which occurs on a global basis as the “slow tails” propagate with relatively little attenuation, deposits opposite polarity charge in the “conjugate” mesosphere (at the other end of a magnetic field line), where it returns toward earth. It was further suggested that such polarization is responsible for the large (up to several volts/meter) mesospheric electric fields observed frequently by Russian and U.S. groups. However, little is known about the interaction of such polarization waves with magnetospheric plasma, so the theory cannot yet be recognized as established. The situation does suggest a definitive experiment, however, observing the electric field transients coupled to the surface at a location magnetically conjugate from located lightning. Suitable venues for doing this include between Southern Africa and Central Europe and also between Australia and various locations in Asia.

SEE ALSO THE FOLLOWING ARTICLES

CLOUD PHYSICS • IONOSPHERE • RADIATION, ATMOSPHERE • THUNDERSTORMS, SEVERE

BIBLIOGRAPHY

The history of this subject and much of the earlier work is covered in: Israël, H. (1973). "Atmospheric Electricity," rev. ed., Keter Press, Jerusalem.

Much of the more recent work has been covered in work presented at symposia of the International Commission on Atmospheric Electricity in 1948, 1952, 1956, 1963, 1968, 1972, 1978, 1982, 1986, 1992, 1996, and 1999. Proceedings of all but the last three symposia were published, and subsequent papers have been pub-

lished in the open literature, largely in the *Journal of Geophysical Research*. Current work, particularly in the field of transient luminous events, also tends to appear in *Geophysical Research Letters*.

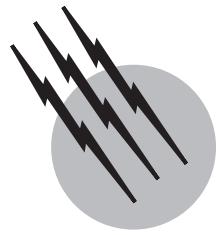
The most comprehensive book on the subject of lightning is Uman, M. A. (1987). "The Lightning Discharge," Academic Press, San Diego, CA.

Papers mainly on the electromagnetics involved have been presented at meetings of the International Union of Radio Science (URSI) and are frequently published in *Radio Science*.

Two papers that give an extended bibliography for many of the subjects discussed herein are

Hale, L. C. (1994). "The coupling of ELF/ULF energy from lightning and MeV particles to the middle atmosphere, ionosphere, and global circuit," *J. Geophys. Res.* **99**, 21089.

Marshall, L. H., et al. (1998). "Electromagnetics of sprite and elve Associated Sferics, *J. Atmos. Solar-Terr. Phys.* **60**, 771.



Thunderstorms, Severe

R. Jeffrey Trapp

National Oceanic and Atmospheric Administration

- I. Introductory Remarks
- II. Severe-Thunderstorm Climatology
- III. Meteorological Setting for Severe Thunderstorms
- IV. Characteristics of Severe Thunderstorms
- V. Severe-Thunderstorm Forecasts and Warnings

GLOSSARY

Adiabatic process Thermodynamic process in which there is no exchange of heat or mass between a metaphorical parcel of air and its surroundings; thus, responding to the decrease in atmospheric density with height, rising air cools adiabatically due to expansion and sinking air warms due to compression.

Dew-point temperature Temperature at which air becomes saturated (100% relative humidity) when cooled at a constant pressure and constant water-vapor content.

Gust front Leading edge of relatively cool, horizontal airflow (or outflow) that originates from the downdraft region(s) of a thunderstorm.

Hodograph Curve connecting the tips of horizontal wind vectors (in order of increasing height aboveground) that are plotted from a common origin.

Hydrometeor Liquid (rain, drizzle) or solid (hail, graupel/snow pellet, ice pellet/sleet, snow) precipitation particle formed from water vapor in the atmosphere.

Mesocyclone Region of cyclonic (counterclockwise) rotation, about a quasi-vertical axis in a supercell thunder-

storm, that provides the background rotation required for tornado formation.

Multicell storm Single thunderstorm comprised of several updrafts of moderate intensity (vertical wind speeds of 20–30 m sec⁻¹), at sequential stages of evolutionary development.

Propagation (thunderstorm) Component of storm movement that deviates from individual updraft movement owing to the discrete formation of subsequent updrafts; each new updraft is on the right or left side (or “flank”) of the previous updraft.

Radar, conventional Electronic ranging device that transmits radio signals and then detects returned (scattered and reflected) signals; the return signal strength (normalized for radar range) is proportional primarily to the size, type, and concentration of hydrometeors within the radar sampling volume.

Radar, Doppler Coherent radar that not only measures the strength of the received signal from hydrometeors (as does a conventional radar) but also measures the Doppler shift in transmitted frequency, due to the component of hydrometeor motion toward or away from the radar.

Radar reflectivity factor (Z_e) Product of the number of hydrometeors per cubic meter and the average sixth power of their diameters expressed as millimeters; conventionally presented in units of dBZ, defined as $10 \log_{10} Z_e$. For long-wavelength weather radar, Z_e is a measure of the reflectivity of hydrometeors.

Supercell storm Long-lived thunderstorm with an intense (vertical windspeeds of $30\text{--}50 \text{ m sec}^{-1}$), rotating updraft. Most prolific producer of large hail, damaging winds, and tornadoes.

Vorticity Vector measure of local rotation within fluid flow; the vertical component of vorticity (in this chapter, vertical vorticity), which represents rotation about a vertical axis, is an important quantity in the discussion of supercell storms.

Wind shear Local variations of the wind vector over some distance; in this chapter, it generally refers to variation of the horizontal wind with increasing height above the ground.

A THUNDERSTORM—a lightning-producing cumulonimbus or, in some instances, a cumulus congestus cloud—is classified as severe when it has the capability of causing significant damage to life and property on the earth’s surface. The U.S. National Weather Service defines a severe thunderstorm as one that produces winds of 25.8 m sec^{-1} (50 knots) or more, hail of 1.9-cm (0.75-in.) diameter or larger, and/or a tornado. Severe thunderstorm types include the multicell storm and the supercell storm. The prominent feature of a supercell is its strong, rotating updraft, which provides the vorticity-rich environment within which tornadoes may form.

I. INTRODUCTORY REMARKS

The basic building blocks of a severe thunderstorm are vertically extensive regions of upward and downward airflow, known as updrafts and downdrafts, respectively. Typically treated as quasi-cylindrical, updrafts and downdrafts in cumulonimbus clouds have diameters that range from a few kilometers to more than 10 km, within which speeds are of the order of 10 m sec^{-1} but may exceed 50 m sec^{-1} in more intense updrafts. In the extreme example of a tornado-bearing storm, an updraft may span the $\sim 10\text{-km}$ -deep layer from cloud base ($\sim 1 \text{ km}$ above the ground) up to and exceeding the level of the tropopause (~ 10 to 15 km above the ground); downdrafts generally are shallower.

Airflow within updrafts and downdrafts is forced to a large extent by buoyancy. Strictly, the buoyant motions (or *convection*, in the lexicon of meteorologists) arise from displacements, from a state of equilibrium, of

small amounts of air (often termed “air parcels”) within a stratified atmosphere. The stratified atmosphere in the neighborhood of the displacements constitutes the *environment* and is assumed to be undisturbed by the displacements. As illustrated in Section III, growth of the displacements into updrafts and downdrafts depends on the characteristics of the environment.

When viewed by weather radar, an individual thunderstorm can be likened to an elementary microscopic organism—a cell—that grows, replicates by division, etc. Thus, adapting terminology from the biological sciences, we have “multicell” storms with multiple updrafts and downdrafts, “splitting” cells in which an initial single updraft metamorphoses into two updraft/downdraft pairs, and even “supercells” (see Section IV). These thunderstorms fall into the class of “severe” when they have the potential to inflict significant damage to life and property. Although severe storms also can be associated with frequent lightning, heavy rain, and flash floods, long-standing concerns—historically by the aviation industry—of hail and high winds motivate the following emphasis on these phenomena, in addition to that on tornadoes.

II. SEVERE-THUNDERSTORM CLIMATOLOGY

A. Severe-Thunderstorm Reports

Large-scale weather systems (spanning hundreds of kilometers and lasting days) are well documented because they affect and are viewed by a large number of official observing stations at the ground. However, severe thunderstorm events are so localized (spanning tens of kilometers) and short-lived (lasting an hour or so) that they frequently do not pass over official observing stations. They may even occur in an area or at a time of day that escapes human detection. To be counted, a severe thunderstorm must be observed, perceived as a severe event, and, most importantly, reported to local weather officials. Nonmeteorological factors such as population density, public interest, and ease of reporting affect the reporting process. There also is a tendency to report the more unusual event when more than one is observed. For example, an observer may not bother to mention that a tornado was accompanied by damaging hail and winds or that strong winds accompanied large hail.

With these caveats in mind, we present the temporal and spatial distributions of (reported) severe thunderstorm events in the contiguous United States (summarized in Table I). One expects a tornado-producing supercell storm also to have damaging hail and winds. However, only 4% of the tornadoes from 1955 through 1983 were accompanied by damaging wind reports, and another

TABLE I Severe Thunderstorm Events Reported in the Contiguous United States^a

Type	Data period	Approximate annual average	Approximate percentage
Damaging wind ^b	1955–1983	1600	50
Damaging hail	1955–1983	1000	30
Tornado	1950–1978	650	20
Severe thunderstorm	29 years	3250	100

^a Data sources: Schaefer, J. T., National Severe Storms Forecast Center; Kelly, D. L., Schaefer, J. T., and Doswell, C. A., III (1985). *Monthly Weather Rev.* **113**, 1997–2014.

^b Includes wind speeds (25.8 m sec^{-1} or higher) deduced from the extent of structural damage.

4% were accompanied by damaging hail reports. By assuming that hail and damaging wind occurred but are not parts of a tornado report and that damaging wind occurred but is not a part of a hail report, we can treat the reports in Table I as representing independent events. Thus, an average of about 3250 severe thunderstorm events per year was reported from the early 1950s through the early 1980s in the contiguous United States. Half of these events are due to damaging winds only. Damaging hail accounts for 30% and tornadoes account for 20% of the reports. Owing to improved reporting procedures (and other nonmeteorological factors identified above) since the early 1970s, the numbers of wind, hail, and tornado reports have been steadily increasing.

B. Seasonal Variations

Severe thunderstorms are most frequent during the spring and summer months in the contiguous United States but may occur any time during the year (Fig. 1). Nationwide, most tornadoes and hailstorms are reported in May and June; most damaging winds reports are received in June and July. Severe thunderstorm occurrence is at a minimum during the winter months.

C. Diurnal Variations

Solar heating of the earth's surface and subsequent heating of the air near the ground play a major role in the formation of convective storms. Although the maximum solar radiation is received at the ground by approximately noon (all times "local"), the air near the ground continues to warm throughout the afternoon. Consequently, the frequency of severe thunderstorm development increases markedly by early afternoon, reaching a peak in the late afternoon and early evening (2–3 hr before sunset; Fig. 2). Each of the three severe-weather phenomena of damaging wind, large hail, and tornado exhibits such a diurnal variation in frequency of occurrence. Some storms that form during the afternoon continue to be severe a few hours

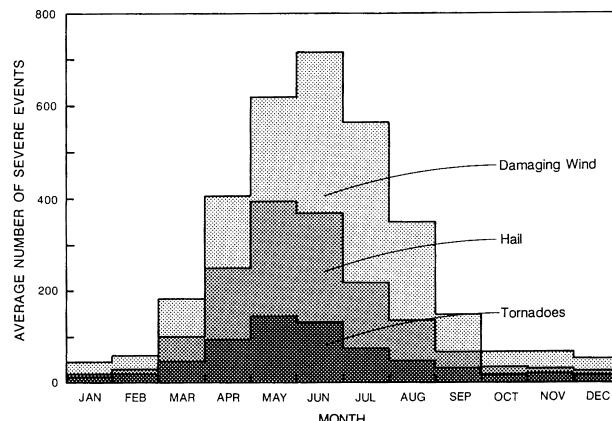


FIGURE 1 Average monthly severe thunderstorm events reported in the contiguous United States during a recent 29-year period. Damaging wind events include assumed 100% occurrences coincident with hail and tornadoes. Likewise, hail events include assumed 100% occurrences coincident with tornadoes. [Data courtesy of J. T. Schaefer, Storm Prediction Center; and from Kelly, D. L., Schaefer, J. T., and Doswell, C. A., III (1985). *Monthly Weather Rev.* **113**, 1997–2014.]

after sunset. Still fewer storms exist throughout the night; obviously, solar heating may not be implicated directly in their overnight sustenance.

D. Geographical Distribution

On average, most U.S. severe thunderstorms occur annually in a north–south-elongated region in the central portion of the country, between the Rocky Mountains and the Mississippi River; a secondary maximum extends east–west across the upper Mississippi Valley (Fig. 3).

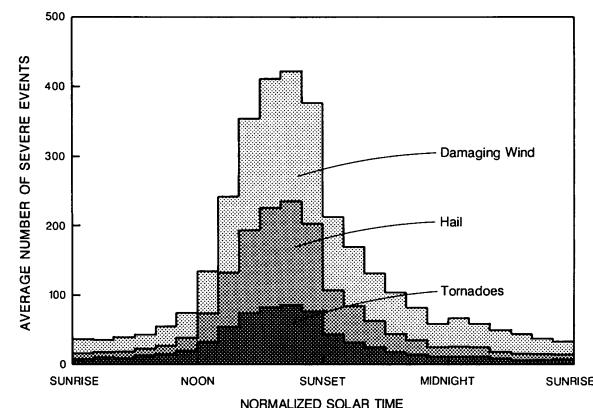


FIGURE 2 Average annual diurnal variation of severe thunderstorm events reported in the contiguous United States during a recent 29-year period. Same occurrence assumptions used as in Fig. 1. Time periods from sunrise to sunset and sunset to sunrise were separately divided into 12 time intervals. [Data courtesy of J. T. Schaefer, Storm Prediction Center; and from Kelly, D. L., Schaefer, J. T., and Doswell, C. A., III (1985). *Monthly Weather Rev.* **113**, 1997–2014.]

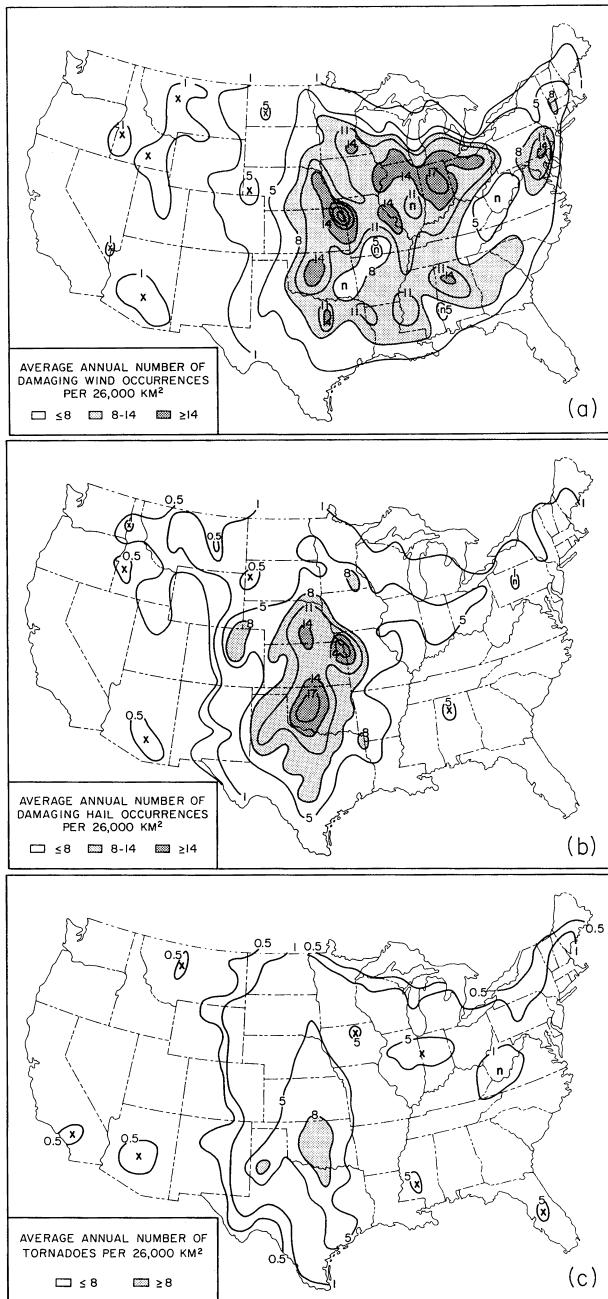


FIGURE 3 Average annual frequency of reported occurrence per 26,000 km² during a recent 29-year period for (a) damaging wind, (b) hail, and (c) tornadoes. Local maxima are indicated by 'x's; local minima, by 'n's. [Data courtesy of J. T. Schaefer, Storm Prediction Center; and from Kelly, D. L., Schaefer, J. T., and Doswell, C. A., III (1985). *Monthly Weather Rev.* **113**, 1997–2014.]

A seasonal variation in this distribution is linked essentially to the poleward (equatorward) migration in the mean position of the upper-tropospheric jet stream during Northern Hemisphere summer (winter). Accordingly, the relatively large number of damaging wind reports

(Fig. 3a) in the southeastern United States, for example, are more likely from storms during autumn–early spring, while those in the northern Great Plains and upper Mississippi Valley are more likely from storms during late spring–summer.

Two-thirds to three-quarters of all severe storms move from the southwest toward the northeast. Generally, these are the spring and early–summer severe thunderstorm events in the southern and central Great Plains, steered by the predominately southwesterly flow in the midtroposphere. A significant percentage of the summertime events in the northern Great Plains and upper Mississippi Valley, however, is associated with northwesterly flow and, hence, moves from the northwest toward the southeast.

III. METEOROLOGICAL SETTING FOR SEVERE THUNDERSTORMS

The formation of severe thunderstorms is not random but rather is regulated by the coexistence of the key ingredients of atmospheric water vapor, positive buoyancy of metaphorical “parcels” of air with respect to an ambient state (known as *buoyant instability*), and a mechanism(s) by which such vertical air motions are initiated in the lower troposphere; a “lifting mechanism” describes atmospheric phenomena such as synoptic-scale cold fronts in which horizontal convergence is concentrated at low altitudes. Given these ingredients in appropriate quantities, altitude-dependent increases and directional changes in the environmental horizontal wind (which define the *vertical wind shear*) govern, to a large degree, the physical characteristics and, hence, the “type” and severity of the thunderstorm (see Section IV).

A. Buoyant Instability of the Atmosphere

Consider a parcel of air that by definition rises in the atmosphere without mixing with the surrounding air. Decreasing air density with altitude above the ground dictates that the parcel will expand and cool adiabatically (with no loss/gain of heat to/from surroundings). If the air parcel is not saturated (relative humidity, <100%), it will cool at a rate of 9.8°C per kilometer of ascent; this process is called dry or adiabatic ascent, and this rate, the *dry adiabatic lapse rate*. If the air is saturated (100% relative humidity), it cools during ascent at a slower rate known as the *moist adiabatic lapse rate*: once the relative humidity reaches 100%, any excess water vapor condenses into cloud droplets and the concomitant release of latent heat decreases the rate of adiabatic cooling. Comparing these rates of parcel temperature change with altitude with those of the parcel’s immediate environment defines the following concept of buoyant instability.

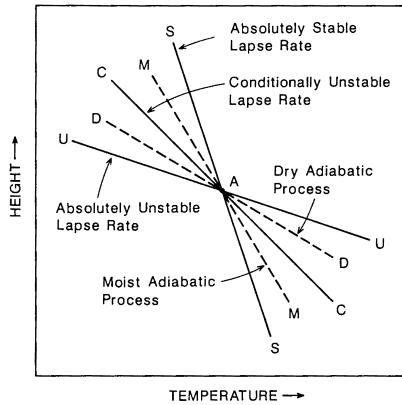


FIGURE 4 Examples of absolutely stable, conditionally unstable, and absolutely unstable environmental lapse rates (solid lines) relative to the dry and moist adiabatic process lapse rates (dashed lines) experienced by parcels of air displaced vertically from point A. Parcels follow the dry (DD) and moist (MM) adiabatic process curves.

When the parcel of air originally located at point A in Fig. 4 is displaced upward along either curves DD or MM (representing dry and moist adiabatic processes, respectively), the air becomes increasingly colder than the surrounding temperatures indicated by the SS curve. By virtue of its negative buoyancy, the parcel descends back to equilibrium point A. When the parcel is displaced downward, it becomes increasingly warmer than profile SS and positive buoyancy returns it to point A. Environmental lapse rate SS is said to be *absolutely stable* because a parcel of air displaced either dry or moist adiabatically returns to its equilibrium level.

Now consider temperature profile UU. When an air parcel at point A is displaced adiabatically upward (downward) along DD or MM, it becomes increasingly warmer (colder) than the environment and continues to rise (sink). Thus environmental lapse rate UU is said to be *absolutely unstable* because a parcel of air continues to ascend or descend once it has been displaced either moist or dry adiabatically from its equilibrium level.

It is possible for an environmental lapse rate to be both stable and unstable, depending on whether or not the air parcel is saturated. Profile CC is said to be *conditionally unstable* because it is stable for dry adiabatic processes but unstable for moist adiabatic processes.

Indices have been developed to express the instability of the environment in quantitative terms. An example is the *lifted index*, the computation of which is illustrated in Fig. 5 for an environment typical of that in which tornadic thunderstorms form. Surface air is lifted dry adiabatically (dashed curve) until it becomes saturated (at the lifted condensation level). Then it is lifted moist adiabatically to the 500-millibar (mbar) pressure level, passing, in this example, the level of free convection, above which the

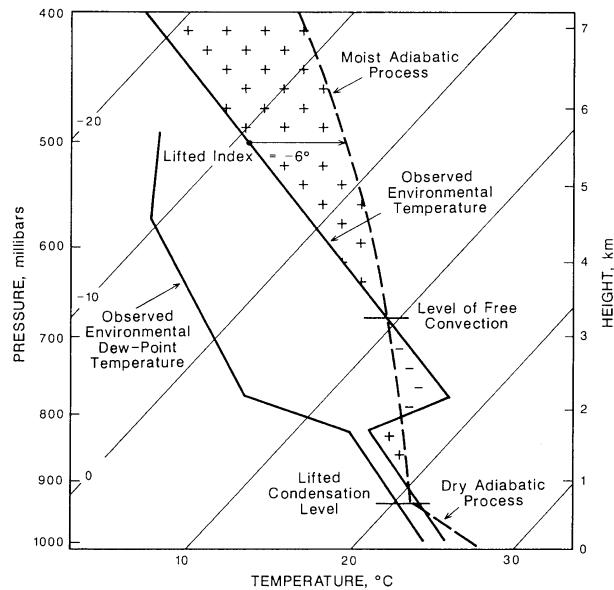


FIGURE 5 A typical tornadic thunderstorm sounding of the atmosphere plotted on a log pressure-skew temperature adiabatic chart. The thick solid lines represent environmental temperature and dew-point temperature measurements. The dashed line is the temperature of a parcel lifted from the earth's surface. Plus and minus signs indicated regions of positive and negative buoyancy experienced by the parcel. Since the parcel is -6°C warmer than the environment at 500 mbar, the lifted index (stability indicator) is defined to be -6°C . [Adapted from Fawbush, E. J., and Miller, R. C. (1953). *Bull. Am. Meteorol. Soc.* **34**, 235–244.]

parcel remains warmer than its environment. The relative warmth at the 500-mbar level is expressed in terms of the difference between the observed 500-mbar temperature and the lifted parcel's temperature: this difference defines the lifted index. In Fig. 5, the lifted index is -6°C , which indicates that the formation of severe thunderstorms is likely.

Another commonly used quantitative measure of environmental buoyant instability is the total positive buoyancy experienced by the parcel from the level of free convection (z_0) upward to the altitude where the parcel is no longer positively buoyant ($z_1 > z_0$). Such *convective available potential energy* (CAPE) may be expressed in a rudimentary form as

$$\text{CAPE} = g \int_{z_0}^{z_1} \frac{T - T'}{T} dz, \quad (1)$$

where g is the acceleration due to gravity (9.8 m sec^{-2}), T is the temperature ($^{\circ}\text{C}$) of the parcel at some altitude z , and T' is the temperature ($^{\circ}\text{C}$) of the environment at altitude z . CAPE is proportional to the gain in kinetic energy of a buoyant parcel between altitude z_0 and altitude z_1 . A nominal value associated with a severe-storm environment is $\sim 1500 \text{ m}^2 \text{ sec}^{-2}$, but this may be larger or smaller and still

be associated with a severe storm, depending on certain characteristics of the environmental wind profile that are discussed below.

The temperature profile in Fig. 5 indicates that the additional presence of a lifting mechanism is required to force an air parcel to rise through regions of negative buoyancy (between ~ 2 and 3 km in Fig. 5) near the temperature inversion (the layer where the temperature *increases* rather than decreases with height) up to the level of free convection. Indicated in the dewpoint temperature profile in Fig. 5 is the presence of dry air at altitudes above 2 km. This air originates in the elevated arid regions of the southwestern United States and northern Mexico. Evaporation of hydrometeors into the dry air favors the formation of vigorous, cool downdrafts that descend from middle altitudes to the ground. At the leading edge of the cool outflowing air, a gust front becomes the primary lifting mechanism for moist low-altitude air feeding the storm's updraft.

B. Environmental Wind Profile

The salient characteristics of the environmental wind profile are revealed by a hodograph, a curve connecting the tips of the horizontal wind vectors¹, plotted from a common origin, in order of ascending height. Qualitatively, a straight-line hodograph (as in Figs. 6a and b) is characteristic of the environment in which some severe thunderstorms including supercells form; a hodograph with curvature between the ground and the middle heights (as in Fig. 6c) characterizes an environment supportive primarily of supercell storms.

In the idealized, straight-line hodograph LMH in Fig. 6a, southeasterly low-altitude (L) winds turn clockwise and increase in speed with altitude, becoming southwesterly² at middle altitudes (M) and west-southwesterly at high altitudes (H) near the tropopause. The arrow **U** in Fig. 6a represents the motion vector for a storm's primary updraft; **U** generally is in the direction of, but at a slower speed than, the mean tropospheric wind. The arrow **S** represents the storm-motion vector, which is the vector sum of **U** and a propagation component **P** (the dashed vector in Figs. 6b and c); as illustrated in Section IV, propagation is due to the formation of a series of severe-thunderstorm updrafts, with each new one forming on the right flank of the previous one. Hence, **S** is the vector essentially of the envelope of the motions of successive updrafts. Plotted relative to the tip of **S**, storm-relative winds (Fig. 6b) are often invoked in discussions of the dynamics of thunderstorms, as are, at times, updraft-

¹The vertical component of the environmental wind vector is negligible compared to the horizontal components.

²Note that meteorologists define wind direction as the direction from which the wind blows.

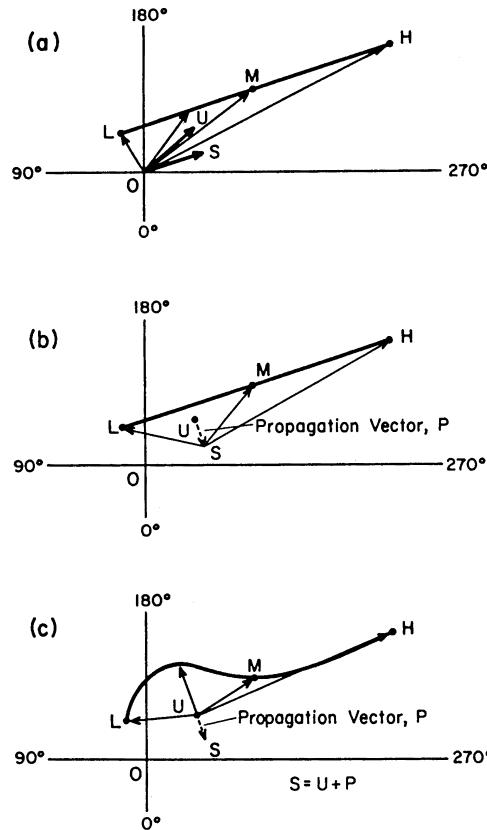


FIGURE 6 Hodographs (thick solid curves) of vertical profiles of the environmental wind constructed by connecting the tips of wind vectors (thin arrows) radiating from a common origin, O. Points L, M, and H on the hodograph indicate low-, middle-, and high-altitude wind vector locations. Thick arrows in (a) indicate updraft (**U**) and storm (**S**) motion vectors. The dashed vector in (b) and (c) is the propagation vector (**P**), representing the relative position of successive updrafts.

relative winds (plotted relative to the tip of **U**; Fig. 6c); ground-relative winds (O) are plotted in Fig. 6a.

Of importance to studies of severe thunderstorms is the vector difference between the wind at different altitudes. *Shear vector* **LM** is the difference between the middle-altitude (**OM**) and the low-altitude (**OL**) wind vectors (Fig. 6a); by definition, the shear vector at a particular height is tangent to the hodograph at that point. The magnitude of **LM** tends to correlate with the storm severity, in an environment with a sufficiently high CAPE. It is preferable, though, to consider the combined contributions of shear and CAPE, the correlation with storm severity of which is quantified through a "bulk" Richardson number,

$$Ri = \frac{\text{CAPE}}{\frac{1}{2}\bar{V}^2}, \quad (2)$$

where \bar{V} is the difference between the mean, density-weighted, horizontal wind speed over the low- to

middle-altitude layer (the lowest ~6 km above the ground) and the mean, horizontal wind speed within the low altitudes (the lowest ~500 m above the ground). As demonstrated in Section IV, supercell thunderstorms tend, for example, to be associated with only a narrow range of values of environmental Ri .

C. Large-Scale Environmental Features

The appropriate environmental wind, temperature, and humidity profiles required for severe-thunderstorm development is frequently found in a favored zone within large-scale (synoptic-scale) extratropical cyclones. These cyclones form in the band of strong, middle-latitude westerly winds and typically move toward the northeast during their mature stage.

An idealization of a middle-latitude cyclone is shown in Fig. 7. At the ground or surface of the earth, the cyclone is characterized by a warm front and cold front emanating from the central low-pressure area (L). The fronts

represent the leading edges of warm surface air moving toward the north and northeast and of cold surface air moving toward the south and southeast. The air is warm and moist in the sector between the two fronts, a result of the northward transport of subtropical air by a low-altitude jet (LJ). In the upper troposphere, a polar jet (PJ) marks the southward extent of cold air aloft; a subtropical jet (SJ) sometimes is located south of the polar jet and at a higher altitude.

Severe thunderstorms most likely form initially near region I, where the polar jet overlies the low-altitude jet. Here, vertical shear of the horizontal wind—from south-southeasterly at the surface to very strong southwesterly in the upper troposphere—is extreme, as is the buoyant instability. As the cyclone, with its attendant jets, moves toward the northeast during the afternoon and evening, severe thunderstorms accordingly become favored in the area shaded in Fig. 7.

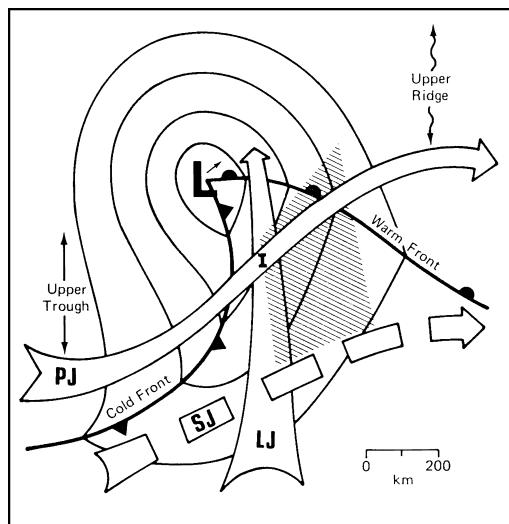


FIGURE 7 A large-scale environmental situation that favors severe-thunderstorm development. Thin curves indicate sea-level isobars (contour lines of constant pressure) surrounding the center of a low-pressure (L) of a midlatitude cyclone that is moving to the northeast (thin arrow). Broad arrows represent the low-altitude jet stream (LJ), upper-altitude polar jet stream (PJ), and still higher subtropical jet stream (SJ). The intersection (I) of LJ and PJ indicates a favored area for severe thunderstorm development owing to vertical shear of the horizontal wind and owing to the abundance of low-altitude moisture provided by the low-altitude jet stream. The hatched region represents the area of anticipated severe thunderstorms during the ensuing 6–12 hr as the large-scale low-pressure system moves to the northeast. [From Barnes, S. L., and Newton, C. W. (1986). In "Thunderstorm Morphology and Dynamics, Vol. 2. Thunderstorms: A Social, Scientific, and Technological Documentary" (E. Kessler, ed.), University of Oklahoma Press, Norman.]

IV. CHARACTERISTICS OF SEVERE THUNDERSTORMS

A. Single-Cell Thunderstorms

Long-lived severe thunderstorms do not form in an environment in which the wind is essentially uniform with height (hence no vertical wind shear), even though the CAPE may be large. Such an environment, characterized by $Ri \rightarrow \infty$ does, however, allow for the development of single-cell storms. A discussion of the simpler processes associated with this storm type provides a worthwhile introduction to the more complex processes found in most severe thunderstorms.

In the initial stages in the life cycle of a single-cell storm, a cumulus cloud develops vertically, and cloud droplets in the upper regions of its updraft grow into radar-detectable hydrometeors; radar signal or *echo* returned from such a storm is contoured in Fig. 8. After ~15 min of cloud growth, the development of an indentation (termed a weak radar-echo region; WER) in the bottom of the radar reflectivity pattern indicates that the updraft has become fairly strong (at least $15\text{--}20 \text{ m sec}^{-1}$).

Within ~20 min, the quantity and size of hydrometeors have increased to the point where the radar reflectivity factor near the top of the updraft exceeds 50 dBZ. The rising cloud top spreads laterally and the updraft weakens upon encountering the buoyantly stable region above the tropopause, which acts much like a rigid lid to the troposphere. Lack of vertical wind shear has a noteworthy effect during this stage of the storm's evolution: larger hydrometeors aloft are not evacuated or transported away from the updraft by environmental winds stronger than

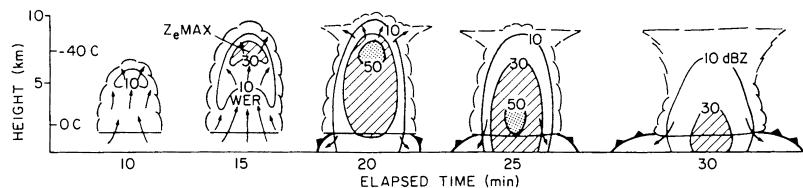


FIGURE 8 Schematic representation of the life cycle of a single-cell thunderstorm. Contours are of the equivalent radar reflectivity factor (Z_e), which is a function of the hydrometeor size, type, and concentration (proportional to rainfall rate) and which is expressed here and in subsequent figures in units of $10 \log_{10} Z_e$ (dBZ). The presence of an updraft is indicated by the weak radar-echo region (WER) indentation in the bottom of the radar reflectivity pattern. The leading edge of the surface cold-air outflow (gust front) is indicated by the bold, barbed line. [Adapted from Chisholm, A. J., and Renick, J. H. (1972). In "Hail Studies Report 72-2," pp. 24–31, Research Council of Alberta, Edmonton.]

those below, allowing the hydrometeors (in the form of rain and perhaps small hail) to fall through the updraft, slowing the updraft, and then effectively replacing it with a downrush of air. Arriving at the surface with the precipitation, such air is colder than its environment because of the evaporation of some raindrops. The associated sudden outrush of cool air produces gusty winds (which may reach severe limits), the leading edge of which forms a gust front beneath the weakening updraft and rain area. As it expands radially outward, the gust front prohibits the inflow, into the base of the updraft, of updraft-sustaining warm, unstable air. The gust front does act as the lifting mechanism for subsequent—but distinctly separate—thunderstorm formations, however. This process is evident in the photograph in Fig. 9, taken by Gemini XII astronauts over the Gulf of Mexico.

The single-cell thunderstorm is no longer active 30–45 min after its initial formation as a small cumulus cloud. Cloud droplets in the lower and middle portions of the storm begin to evaporate in response to unsaturated air that is drawn into the storm behind (above) the mass of descending precipitation. However, as may be noted in Fig. 9, portions of the upper region of the cloud that are comprised of ice crystals may continue to exist for an hour or more until all of the ice crystals sublime (or undergo a phase change from the solid state to the gaseous state).

B. Multicell Thunderstorms

Experiments with computer-simulated thunderstorms and also observations demonstrate that multicell storms most likely occur when $Ri > 30$ (Fig. 10), which may be given in an environment, for example, with a 0- to 6-km shear vector magnitude of $\sim 10 \text{ m sec}^{-1}$ and a CAPE of $\sim 2000 \text{ m}^2 \text{ sec}^{-2}$.

At any particular time, the multicell storm is comprised of three or four distinct cells at various stages of development. The newest cell is consistently found on the right side of the “complex” and the oldest cell is on the left.

Resulting storm motion is significantly to the right of individual updraft–downdraft motion (Figs. 6 and 11). While individual cells within the multicell storm are short-lived, the storm as a whole can be long-lived.

The evolution and general characteristics of multicell storms are revealed by horizontal sections of an idealized multicell, plotted in Fig. 11 as a function of time and height. The following discussion focuses on “cell 3” in this multicell storm. Radar reflectivity values greater than 40 dBZ in cell 3 are shaded for emphasis; vertical cross sections in the direction of cell 3’s motion are shown at the bottom of the figure.



FIGURE 9 Thunderstorm activity over the Gulf of Mexico in November 1966 as photographed from the Gemini XII spacecraft (partially visible in the bottom of the picture) at an altitude of about 225 km. Each cloud-free region surrounded by a (partial) ring of clouds represents the surface pool of colder air left behind by a former thunderstorm that has totally dissipated. Careful inspection of the photograph indicates that, in a few instances, the ice-crystal cirrus clouds from the top of the former thunderstorm remain above the cloud-free region. Note that new thunderstorms are forming where two or more outflow boundaries intersect. [NASA photograph courtesy of NOAA, National Environmental Satellite, Data, and Information Service.]

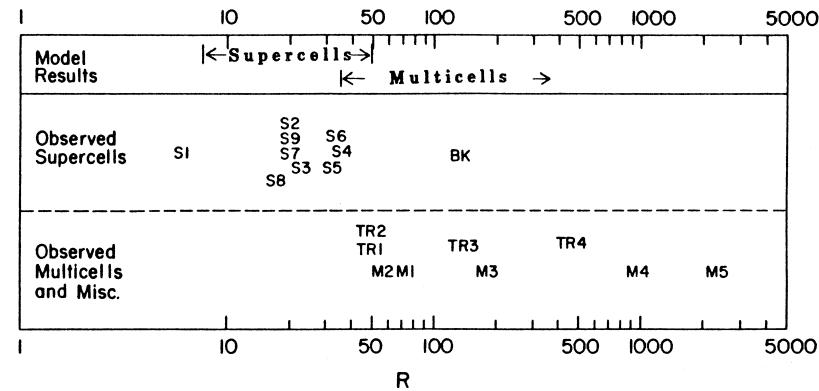


FIGURE 10 Richardson number (defined as Ri in text) of the environment of observed and numerically simulated supercell and multicell thunderstorms. S1–S9 denote observed supercell storms; M1–M9 denote observed multicell storms; TR1–TR9 denote convective storms in tropical environments. [From Weisman, M. L., and Klemp, J. B. (1986). In “Mesoscale Meteorology and Forecasting” (P. Ray, ed.), American Meteorological Society, Boston.]

At time 0 (Fig. 11), radar signals returning from hydrometeors growing in the upper portion of a new updraft (cell 3) start to appear at a height of about 9 km. This new cell is born out of the following processes: Hydrometeors forming in the middle to upper portions of the currently mature updraft (in cell 2) are carried downwind by updraft-relative environmental winds (Fig. 6c) and initially descend ahead of the updraft. As the precipitation falls, it moves toward the updraft’s left (relative to its motion vector) side or “flank” in response to the updraft-relative flow. A dome of low-altitude cold air develops in the precipitation area and spreads out behind a gust front that expands beneath cell 2’s updraft. As storm-relative environmental air approaches this storm’s right flank, it is

lifted by the gust front to form the updraft in cell 3, as illustrated in Fig. 12.

During the next 5–7 min, the increase in the radar return with height and the presence of a WER reflect the existence of a moderately strong updraft ($20\text{--}30 \text{ m sec}^{-1}$) in the developing cell 3 (see also Fig. 12). As the cell attains its maximum height of about 12 km, light precipitation reaches the ground and a gust front begins to form as increasing amounts of downdraft air reach the surface. The gust front expands with increasing rainfall rate, causing (i) the weakening of cell 3’s updraft as its lower regions are replaced by cold air and (ii) the formation of a new updraft (cell 4) to the right of cell 3. During this time period, earlier cell 2 is in its final stage of decline.

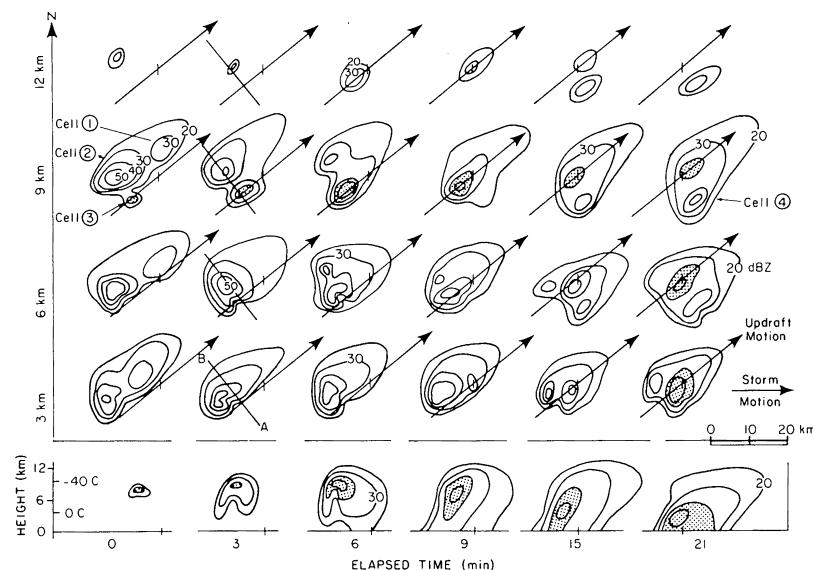


FIGURE 11 Cross sections through an idealized, multicell severe thunderstorm. Arrows through contoured Z_e indicate the direction of cell 3 (shaded) updraft motion and orientation of the vertical cross section at the bottom. Vertical cross section AB is shown in Fig. 12. [Adapted from Chisholm, A. J., and Renick, J. H. (1972). In “Hail Studies Report 72-2,” pp. 24–31, Research Council of Alberta, Edmonton.]

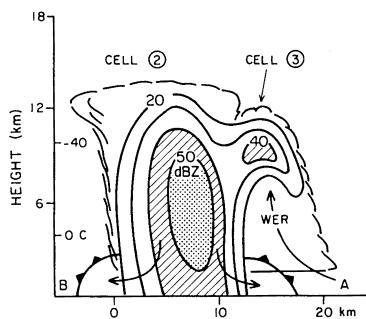


FIGURE 12 Vertical cross section AB (see Fig. 11) through an idealized, multicell severe thunderstorm. Contours are of Z_r . Bold, barbed lines show the gust front from the precipitation area of cell 2. Low-altitude environmental air approaching from the right is forced upward by the gust front to form the new cell 3; the weak echo region (WER) indicates the presence of an updraft. [Adapted from Chisholm, A. J., and Renick, J. H. (1972). In "Hail Studies Report 72-2," pp. 24–31, Research Council of Alberta, Edmonton.]

In terms of storm severity, the existence of weak to moderate shear in this case leads generally to a longer-lived storm, which may enhance the duration of damaging storm-generated surface winds behind the gust front. Moreover, large hailstones may grow from ice-crystal or frozen-raindrop embryos if their trajectories keep them in regions of the storm with large amounts of supercooled (unfrozen at temperatures less than 0°C) water drops. This occurs, for example, as embryos that develop in a new cell are swept into the updraft of a mature cell. Hail then grows in the mature updraft through accretion of the supercooled water drops until the updraft no longer can suspend the hailstone, the supply of supercooled water is depleted, or the hailstone is ejected from the updraft. Tornadoes in multicell storms are rare.

C. Supercell Thunderstorms

Supercell thunderstorms typically develop in a strongly sheared (e.g., 0- to 6-km shear vector magnitude of $\sim 30 \text{ m sec}^{-1}$) environment like those illustrated in Figs. 6 and 7. The thermodynamic environment of supercells is typified by the temperature and humidity profiles in Fig. 5. CAPE values of $\sim 2500 \text{ m}^2 \text{ sec}^{-2}$ or larger are common in such an environment and suggest the potential for strong updrafts ($30\text{--}50 \text{ m sec}^{-1}$ or more). The range of combined vertical shear and buoyancy values that support most supercells is relatively small: $10 < \text{Ri} < 40$ (see Fig. 10).

The interaction between a strong updraft and strongly sheared environmental winds yields a type of storm that outwardly does not resemble a multicell storm, though the basic physical processes are the same. Unlike the multicell storm, a supercell is distinguished by a strong, long-lived, rotating updraft. Significant rotation (about a vertical axis) at middle altitudes is responsible for the storm's long life (typically several hours) and deviant motion to the right

of the mean tropospheric environmental wind. The rotating updraft also is, in varying degrees, responsible for the conditions required for the growth of large hail, the buildup of electric charge that leads to lightning, and the genesis of long-lived, strong to violent tornadoes. Additionally, strong updrafts lead to strong precipitation-driven downdrafts, which can in turn lead to damaging surface winds.

Visual characteristics of a supercell storm are depicted in Fig. 13, a schematic representation of what may be viewed by an observer a considerable distance southeast of the storm. The presence of a strong updraft is indicated both by the "overshooting" cloud top, which may extend several kilometers above the tropopause and which is visible above the expanding anvil (see Fig. 14), and by the lowering of the cloud base in the form of a "wall cloud" that marks where moist air is converging into the base of the rotating updraft. If a tornado occurs, it forms within the updraft and first becomes visible to the observer when it descends from the rotating wall cloud or perhaps as a near-ground cloud of debris prior to the development of the characteristic funnel of condensed water vapor. The main precipitation area consisting of rain and hail is found ahead (typically northeast) and to the left (typically northwest) of the updraft. Most of the "cloud-to-ground" lightning occurs within the precipitation region, but occasionally a lightning channel will exit the cloud at middle altitudes and descend to the ground in the clear air (producing an especially dangerous situation for unsuspecting humans).

Clouds that form along the storm's gust front grow in height and size as they move toward and merge with the body of the storm; this feature is called a flanking line (Fig. 14). At times, the nonrotating updraft within one of the developing clouds concentrates the vertical component of vorticity (hereafter referred to as vertical vorticity) produced by horizontal wind shear across the gust front, and a localized, short-lived vortex (known as a "gustnado") is generated. The presence of a vortex is indicated by a dust whirl on the ground beneath the flanking line. Although capable of inflicting some light damage, this vortex is not to be confused with the destructive tornado that forms within the storm's main rotating updraft.

The mature supercell's updraft (near the intersection of lines AB and CD in Fig. 15) has a dominant influence on the radar reflectivity structure depicted in Fig. 15 and 16. The updraft is so strong that cloud droplets do not have time to grow to radar-detectable hydrometeors until they are halfway to storm top. The resulting bounded weak echo region (BWER) extends to middle altitudes in the storm. The BWER is capped by an area of high reflectivity because the largest hydrometeors form within the upper regions of the updraft. Small hydrometeors are carried downwind by strong environmental flow, forming an anvil-shaped plume. Larger hydrometeors descend

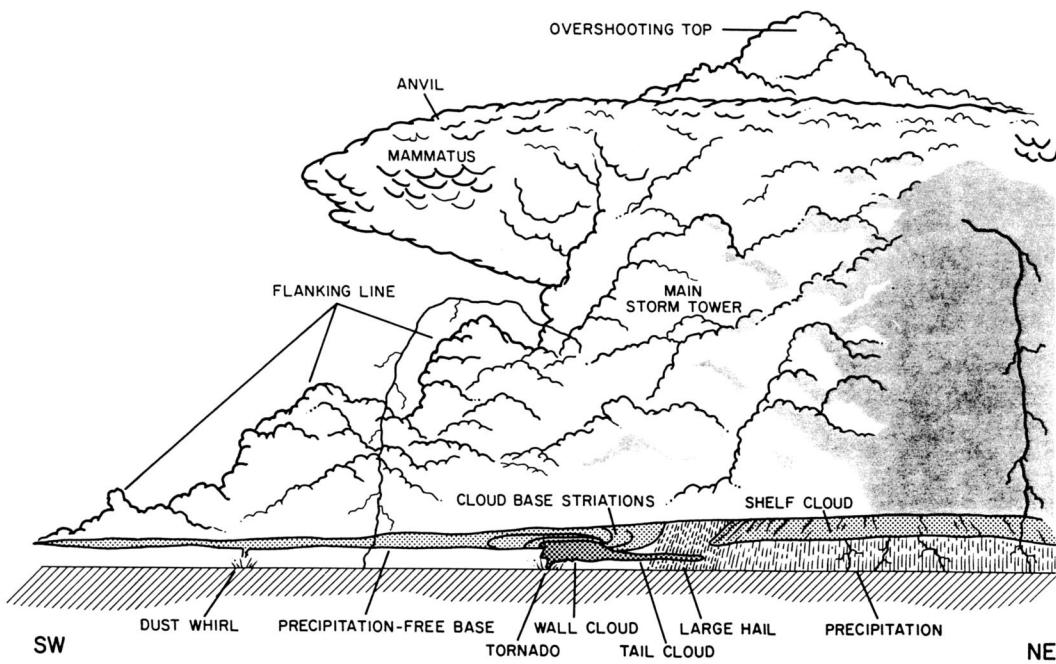


FIGURE 13 Schematic of a supercell thunderstorm as viewed by an observer on the ground, looking toward the northwest. Characteristic features are identified. [Adapted from the original; courtesy of C. A. Doswell III, National Severe Storms Laboratory.]

immediately downwind of the updraft. Responding to the environmental winds as they fall, their trajectories curve cyclonically (counterclockwise) around the updraft, forming a downdraft area on the left and left-forward flanks of the storm. Large hail, whose growth by accretion of super-cooled water drops is confined to the supercell's primary updraft, is found adjacent to the updraft along the left and rear storm perimeters. The resultant curvature of the reflectivity pattern—which often takes the shape of a “hook”—

around the rear of the updraft at low altitudes suggests the presence of cyclonic rotation, readily confirmed with time-lapse photography and Doppler-radar observations.

The region of cyclonically rotating air within the supercell's updraft is called a mesocyclone. As illustrated in Fig. 17, mesocyclones tend to form first at middle altitudes through what may be viewed as a two-step process in the environment characterized by the idealized straight-line hodograph. The first involves the environmental vorticity vector: the largely horizontal vorticity vector (which points to the left of the shear vector in Fig. 6a), due to the vertical shear of the environmental wind, is vertically tilted in horizontal gradients of the vertical air speed comprising the updraft. The result is counterrotating vortices on the left and right flanks of the updraft. The second step involves certain dynamics of the storm and this environment that helps induce a split of the initial updraft (Fig. 17). The new updraft on the right (left) flank of the initial updraft ultimately becomes spatially well correlated with the cyclonically (anticyclonically) rotating air as it tends to move to the right (left) of the mean tropospheric wind vector. The “right-moving” storm with its mesocyclone is the long-lived supercell that may then bear large hail and perhaps a tornado; the “left-moving” storm with its mesoanticyclone may, in some instances, continue to exist for 30 min or more but seldom produces a tornado. “Storm splitting” is unlikely in an environment with a strongly curved hodograph, in which typically only a right-moving storm forms and subsequently acquires net cyclonic rotation

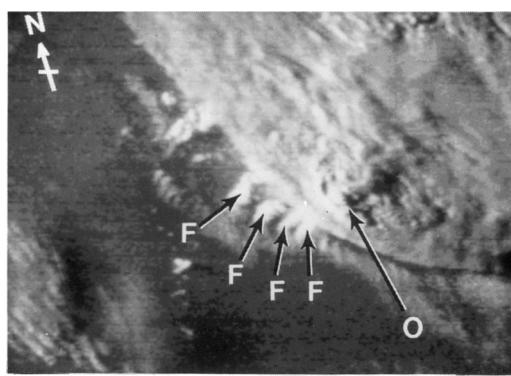


FIGURE 14 GOES-West satellite visual image of the supercell thunderstorm that produced the Wichita Falls, Texas, tornado during the afternoon of April 10, 1979. The satellite viewed the storm from the southwest from a position over the equator at 135°W longitude. The overshooting top (O) and multiple flanking lines (F's) at the rear of the storm are indicated. [Courtesy of NOAA, National Environmental Satellite, Data, and Information Service.]

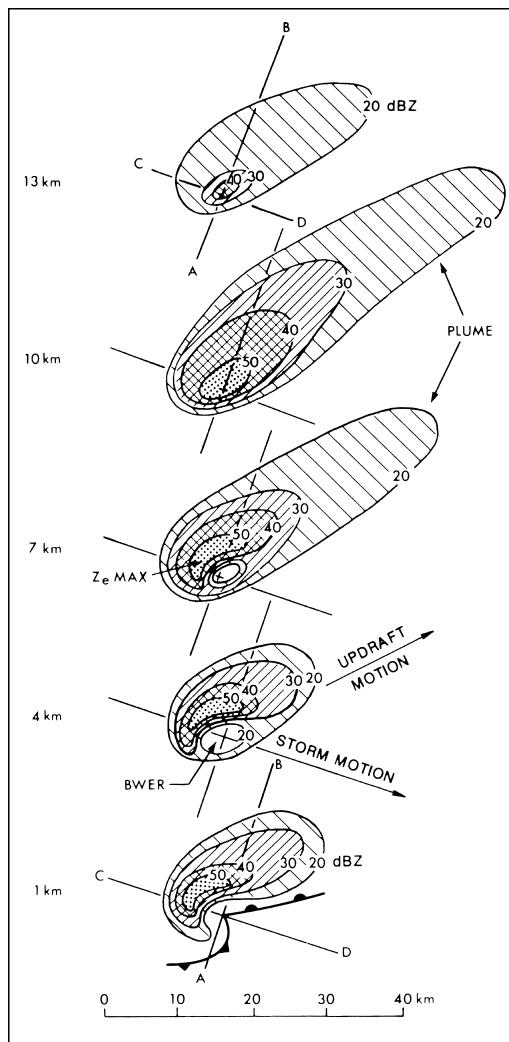


FIGURE 15 Cross sections through an idealized, supercell thunderstorm. Contours are of Z_e . A very strong updraft is indicated by the bounded weak echo region (BWER) in the reflectivity pattern. Surface gust fronts associated with the forward-flank downdraft and the rear-flank downdraft are indicated by the bold line with filled semicircular symbols and triangular symbols, respectively. Vertical cross section CD is shown in Fig. 16. [Adapted from Chisholm, A. J., and Renick, J. H. (1972). In “Hail Studies Report 72-2,” pp. 24–31, Research Council of Alberta, Edmonton.]

in its updraft; the mesocyclone-development process with a curved hodograph also involves vertical tilting of the environmental vorticity vector.

While the overall circulation of the mesocyclone extends outward 5–10 km from its axis of rotation, peak rotational wind speeds of 20–25 m sec⁻¹ typically are found at a radius of 2.5–3 km. During a mesocyclone’s organizing stage, such rotational wind speeds, or alternatively vertical vorticity ($\geq 1 \times 10^{-2}$ sec⁻¹), are found primarily at middle altitudes. Some mesocyclones do not develop beyond this stage. Those that do reach “maturity”

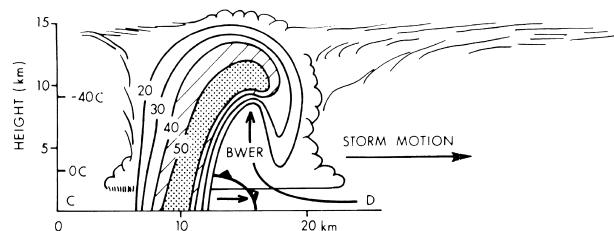


FIGURE 16 Vertical cross section CD (see Fig. 15) through an idealized, supercell thunderstorm. Contours are of Z_e . The presence of a very strong updraft is indicated by the bounded weak echo region (BWER) indentation in the reflectivity profile. The bold, barbed line shows the gust front. Low-altitude environmental air approaching from the right is forced upward by the gust front. [Adapted from Chisholm, A. J., and Renick, J. H. (1972). In “Hail Studies Report 72-2,” pp. 24–31, Research Council of Alberta, Edmonton.]

are associated with significant vertical vorticity from the ground upward through middle altitudes. The actual process by which such mesocyclone-scale vertical vorticity is generated at lower altitudes is not clear, although observations, computer model simulations, and theory are in general agreement that the process is different than that described above. One explanation requires an elongated region of low-altitude buoyancy contrast between the air cooled by rainfall associated with the “forward-flank” (with respect to the updraft) downdraft and warm, moist, inflow air (Fig. 18); a similar region of buoyancy contrast is found to the rear and left of the updraft. Parcels of air that flow within and along the narrow zone(s) of buoyancy contrast acquire horizontal vorticity generated by virtue of a solenoidal effect (that is, through a circulation that results from the tendency of warm air to rise and cool air to sink). This vorticity is tipped into the vertical, at low altitudes, as the air parcels exit a downdraft and then enter the primary updraft.

Low-altitude, converging airstreams associated with the (i) storm-relative inflow, (ii) outflow driven by the forward-flank downdraft, and (iii) outflow driven by a “rear-flank” downdraft intensify or “stretch” the vertical vorticity of the mature mesocyclone into that of a tornado, at the location marked “T” in Fig. 18. Not all mature mesocyclones bear tornadoes, however, as alluded to in the discussion in Section V. Indeed, a delicate balance between the processes that govern the converging airstreams is one of the necessary conditions for “tornadogenesis.” Consider, for example, a rear-flank downdraft and associated outflow that are too strong relative to the environmental inflow. The rear-flank gust front can then advance well into the inflow air and, ultimately, choke off the supply of moist, buoyant air to the updraft. As a consequence, the mesocyclone and the updraft to which it is coupled decay.

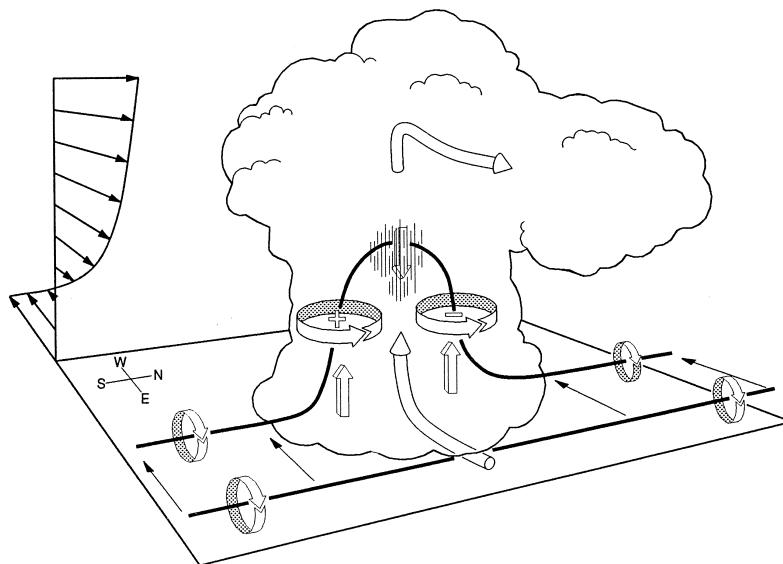


FIGURE 17 Schematic of a thunderstorm growing in an environment characterized by a straight-line hodograph. Cylindrical arrows reveal the flow within the storm. Circular-ribbon arrows depict the sense of rotation of the vortex lines (drawn as thick lines). Note the counterrotating vortices on the north and south sides of the updraft. The remaining arrows show the developing downdraft and updrafts that lead to the eventual storm split. [From Klemp, J. B. (1987). *Annu. Rev. Fluid Mech.* **19**, 369–402. Reprinted, with permission, from *Annual Review of Fluid Mechanics*, Volume 19 ©1987 by Annual Reviews.]

Note that a storm initially may allow for tornado formation, then subsequently evolve in some critical way (or move into a slightly different environment) such that the rear-flank gust front can advance and cause updraft, mesocyclone, and tornado demise. In some instances, lifting of the moist unstable air by the gust front at a location nominally southeast of the initial updraft can lead to a new updraft and storm regeneration (see Fig. 18). Since the new updraft develops generally within the region of “residual” cyclonic vertical vorticity, it rapidly concentrates the vorticity (through convergence) to produce a new mesocyclone and perhaps a new tornado. The sequence of gust-front advance, updraft/mesocyclone decay, and storm regeneration/new mesocyclone development may repeat itself several times such that a single supercell storm may in effect spawn a series of tornadoes.

D. Other Configurations

Severe thunderstorms occur most frequently in the form of multicell and supercell storms and hybrids incorporating features of both. Thus, at times, storm classification is quite arbitrary. The archetypal examples presented above provide a basis for interpreting the more complicated real-life configurations.

Consider the thunderstorms that become organized into long “squall lines.” These lines can propagate as an apparent entity for a number of hours or for a few days in extreme cases. Most of the individual elements of a squall line are

of the multicell type. However, some cells in the line, particularly at the south end or at breaks in the line (where there is not competition from nearby cells for the moist inflow air required for storm sustenance), may be supercellular. A broad area of light to moderate precipitation frequently is found behind the strong updrafts and down-drafts at the leading edge of the squall line.

The most common type of severe weather associated with storms in this configuration is hail and very strong “straight-line” winds that occur just to the rear of the updrafts. The latter is especially prevalent when a portion of the line accelerates forward, resulting in a convex structure that, when viewed by weather radar, is associated with a bow-shaped echo (or “bow echo”). The occasional genesis of a tornado is favored just north of the apex of such a bow echo (Fig. 19); prior to the bow-echo stage, a squall line may engender tornado formation in other locations, particularly at the southern end of the line.

Another noteworthy configuration is that of a cumulus congestus growing over a narrow zone of low-level vertical vorticity. Likened to a “vortex sheet,” the vertical-vorticity zone is generated on a larger scale, for example, by terrain effects and the influence of background or planetary rotation. Stretching of the vertical vorticity in the growing cumulus congestus updraft leads to tornado development in the absence of a bona fide mesocyclone. Such a “non-mesocyclone” tornado tends to be less intense than its supercellular or mesocyclone counterpart. Nevertheless, its parent cloud—which in many regards is a single-cell

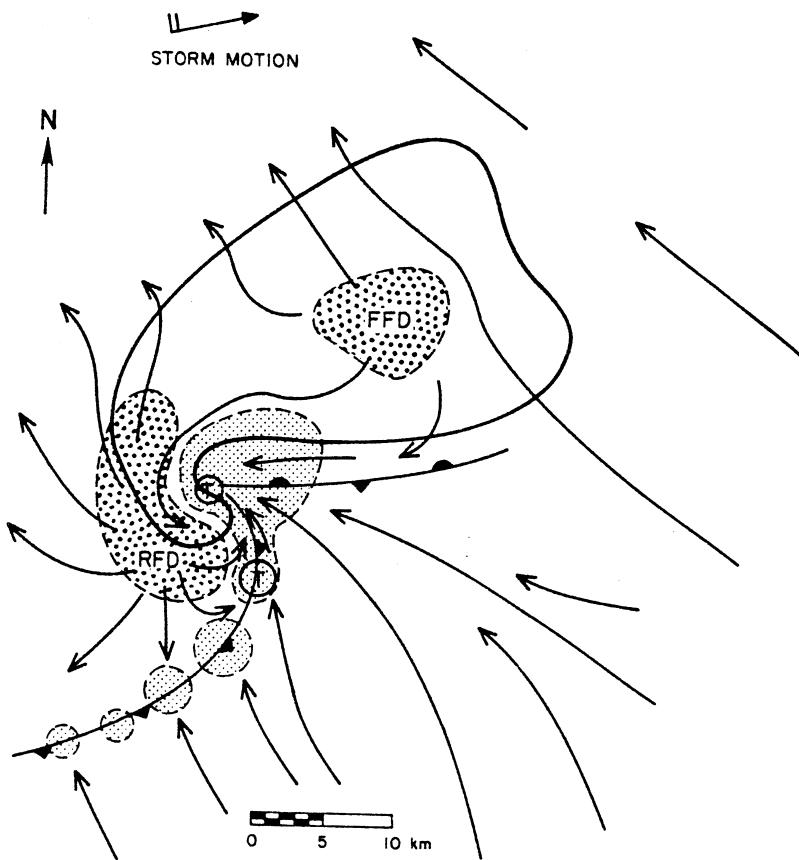


FIGURE 18 Structure of an idealized tornadic supercell thunderstorm at the ground. The radar echo, including the characteristic “hook” appendage, is drawn as a thick line. Arrows are streamlines of horizontal storm-relative airflow. FFD, forward-flank downdraft; RFD, rear-flank downdraft. Positions of surface gust fronts associated with the FFD and the RFD are given by the bold line with filled semicircular symbols and triangular symbols, respectively. Lightly stippled regions represent updrafts. The T within the hook echo indicates the current location of the tornado; the other T indicates the possible location of a subsequent tornado. [From Davies-Jones, R. (1986). In “Thunderstorm Morphology and Dynamics, Vol. 2. Thunderstorms: A Social, Scientific, and Technological Documentary” (E. Kessler, ed.), University of Oklahoma Press, Norman.]

thunderstorm—is certainly a member of the severe-storm family.

V. SEVERE-THUNDERSTORM FORECASTS AND WARNINGS

The temperature, humidity, and wind profiles of the atmosphere, in addition to the large-scale environmental features discussed in Section IV.C, provide clues concerning the likelihood of the occurrence of severe thunderstorms, with attendant lightning, hail, damaging winds, and/or tornadoes. In the United States, highly specialized severe-thunderstorm forecasters at the Storm Prediction Center of the National Weather Service use this and other information to alert the public to the *potential* for damaging and life-threatening storm development. The initial alert comes in the form of an increasingly accurate “watch”: the percentage of verifiable severe-weather watches has increased from 63% in 1972 to 90% in 1996.

More localized “warnings” that existing storms are severe or will soon become severe are issued in the United States by individual field offices operated by the National Weather Service and depend on the visual observations of trained (and typically volunteer) storm “spotters” and on the interpretation of weather radar data. The existence of a hook-shaped appendage on the right-rear side of the radar echo at low altitudes (Fig. 15) has been used since the mid-1950s to help identify the tornado’s parent mesocyclone. Unfortunately, trajectories of radar-detectable precipitation particles do not always result in hook-shaped echoes, and, many times there are similarly shaped appendages that do not represent rotation. Doppler weather radars, on the other hand, provide a positive and unambiguous identification of a mesocyclone. When a mesocyclone signature is detected from the ground to the middle or upper portions of a storm, there is virtually a 100% probability that the storm is producing (or will soon produce) damaging wind and/or hail. There is, at best, a 50%

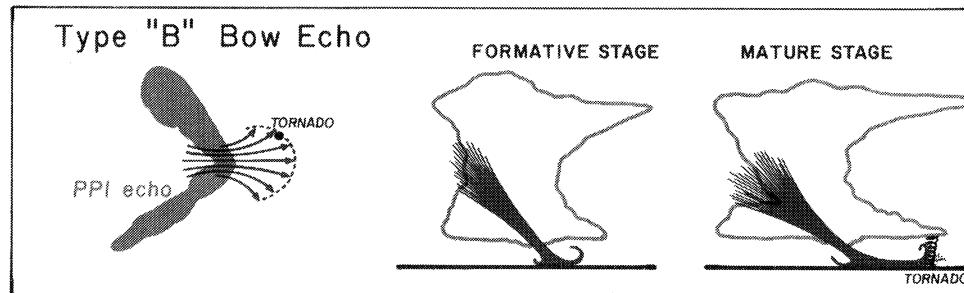


FIGURE 19 Conceptual model of tornado formation within a bow echo. Left panel: Radar echo. Note the tornado position to the north of the apex of the bow. Right panels: A vertical slice through the bow echo, in a location north of the apex. The tornado develops at the leading edge of the outflow. [From Fujita, T. T. (1985). "The Downburst," Satellite and Mesometeorology Research Project (SMRP), Department of Geophysical Sciences, University of Chicago, Chicago.]

probability that the storm will produce a tornado, but by using Doppler-radar measurements one is able to eliminate a number of other storms from the category of possible tornado producers.

As mentioned at the beginning of this article, the severe thunderstorm is such a localized phenomenon that it continues to be a challenge to forecast the expected time and locale of storm formation and, also, the anticipated storm type and severity. The key to improved forecasting is an improved knowledge of basic severe-thunderstorm processes and of the state of the atmosphere. New technological advances hold promise for such improved knowledge: During the coming decade(s), remote sensors on the ground and in space likely will provide the human forecaster as well as the numerical forecast models with more frequent and densely spaced observations of the atmospheric variables associated with storm initiation and subsequent character and severity. Once a storm forms, data from a recently implemented (and eventually upgraded) nationwide network of Doppler radars, which feed and subsequently are augmented by automated storm- and storm-attribute-detection algorithms, will continue to improve the timeliness and accuracy of severe-thunderstorm and tornado warnings.

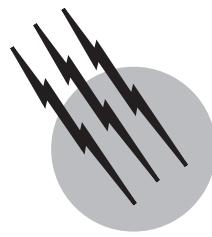
SEE ALSO THE FOLLOWING ARTICLES

ATMOSPHERIC DIFFUSION MODELING • ATMOSPHERIC TURBULENCE • CLIMATOLOGY • CLOUD PHYSICS •

METEOROLOGY, DYNAMIC (STRATOSPHERE) • METEOROLOGY, DYNAMIC (TROPOSPHERE) • RADAR

BIBLIOGRAPHY

- Atlas, D. (ed.) (1990). "Radar in Meteorology," American Meteorological Society, Boston.
- Bluestein, H. B. (1993). "Synoptic-Dynamic Meteorology in Midlatitudes. Volume II: Observations and Theory of Weather Systems," Oxford University Press, New York.
- Cotton, W. R., and Anthes, R. A. (1989). "Storm and Cloud Dynamics," Academic Press, San Diego, CA.
- Doviak, R. J., and Zrnic', D. S. (1993). "Doppler Radar and Weather Observations," Academic Press, Orlando, FL.
- Emanuel, K. A. (1994). "Atmospheric Convection," Oxford University Press, New York.
- Foote, G. B., and Knight, C. A. (eds.) (1977). "Hail: A Review of Hail Science and Hail Suppression," Meteorological Monograph, Vol. 16, No. 38, American Meteorological Society, Boston.
- Houze, R. A., Jr. (1993). "Cloud Dynamics," Academic Press, San Diego, CA.
- Kessler, E. (ed.) (1983a). "The Thunderstorm in Human Affairs, Vol. 1. Thunderstorms: A Social, Scientific, and Technological Documentary," 2nd ed., University of Oklahoma Press, Norman.
- Kessler, E. (ed.) (1983b). "Thunderstorm Morphology and Dynamics, Vol. 2. Thunderstorms: A Social, Scientific, and Technological Documentary," 2nd ed., University of Oklahoma Press, Norman.
- Kessler, E. (ed.) (1983c). "Instruments and Techniques for Thunderstorm Observation and Analysis, Vol. 3. Thunderstorms: A Social, Scientific, and Technological Documentary," 2nd ed., University of Oklahoma Press, Norman.
- Ray, P. S. (ed.) (1986). "Mesoscale Meteorology and Forecasting," American Meteorological Society, Boston.



Tropospheric Chemistry

Peter Warnecke

Max-Planck-Institut für Chemie

- I. The Troposphere as a Chemical Reactor
- II. Trace Gases
- III. Photochemical Processes
- IV. Basic Gas-Phase Free Radical Chemistry
- V. The Oxidation of Methane
- VI. Hydrocarbon Oxidation Mechanisms
- VII. Ozone
- VIII. Nitrogen Oxide Chemistry
- IX. The Oxidation of Sulfur Compounds
- X. Aerosols and Clouds
- XI. Chemical Reactions in Clouds
- XII. Tropospheric Chemistry, Integration by Models

GLOSSARY

- Aldehyde** Organic compound with a terminal carbonyl ($-CHO$) group.
- Alkane** Fully saturated hydrocarbon.
- Alkene** Hydrocarbon containing one $>C=C<$ double bond.
- Aromatic compound** Benzene derivative.
- Biomass burning** The combustion of living and dead plant organic matter.
- Boundary layer** Region closest to the earth surface, often capped by a temperature inversion layer.
- Catalyst** Substance promoting the rate of a chemical reaction or set of reactions.
- Concentration** Amount, mass, or number of molecules of a substance per unit volume.

Dry deposition Removal of a trace substance from the troposphere by downward transport and absorption by materials at the earth surface.

First-order reaction The reaction rate is proportional to the concentration of the substance being removed.

Free troposphere The tropospheric domain above the boundary layer.

Ketone Organic compound with a carbonyl group ($>CO$) positioned between two carbon atoms.

Mixing ratio Amount of a substance per unit volume divided by the total amount of all substances in the same volume. Also known as amount fraction or mole fraction.

pH Negative decadic logarithm of the hydrogen ion activity in an aqueous solution.

Photodissociation The process in which a molecule

absorbs a photon endowed with sufficient energy to cause the molecule to split into fragments.

pK The pH at which the dissociation of a weak acid in aqueous solution is half-way completed.

Radical (free radical) A reactive atom or a fragment of a stable molecule characterized by an odd number of bonding electrons. Free radicals in the atmosphere are generated by photochemical and subsequent processes.

Reaction rate Change in concentration of a substance per unit time.

Second-order reaction The reaction rate is proportional to the product of two concentrations, that of the substance being removed and that of a second reactant.

Wet deposition Removal of a trace substance from the troposphere by incorporation into cloud, fog, or raindrops followed by their precipitation.

THE MAIN CONSTITUENTS of air are nitrogen, oxygen, and argon, with (dry air) volume fractions of N₂ 78.0%, O₂ 20.95%, and Ar 0.95%, approximately. This distribution is preserved at altitudes up to ~100 km. All other components are minor ones with many of them occurring only in traces. In the troposphere, the atmospheric region closest to the earth surface, the main constituents are chemically inert at ambient temperatures and photochemically inactive. Oxygen participates in oxidation processes, but it initiates few chemical reactions on its own. Tropospheric chemistry, therefore, deals primarily with trace substances. They may occur as gases or in the form of particles. Many trace gases are emitted from the biosphere so that they are natural constituents of the troposphere. Human activities have caused additional emissions that in some cases have led to marked changes in atmospheric composition. Atmospheric chemistry explores the origins of trace compounds and processes governing their abundances, spatial distributions, chemical transformations, and their removal from the atmosphere. This article will focus on sources and chemical reaction pathways.

I. THE TROPOSPHERE AS A CHEMICAL REACTOR

The troposphere extends from the surface up to the tropopause, which varies in altitude from about 8 km over the Poles to 17 km over the equator. The next higher region, the stratosphere, extends to 50 km altitude. The barometric law causes the pressure to decrease nearly exponentially with height from about 1000 hPa at the surface to 120 hPa at the tropical tropopause. The temperature de-

creases as well, from an average 287 K at the surface to ~200 K. The tropopause is defined by a reversal of the temperature gradient, in the stratosphere the temperature rises again. The principal mode of heat transfer in the lower troposphere is convection, in the stratosphere it is infrared radiation. Because convection ensures rapid turbulent mixing of air, the troposphere is vertically well mixed. As vertical mixing in the stratosphere is much slower, the tropopause represents a natural boundary of transport between both regions. This makes the troposphere a fairly well-defined geochemical reservoir. It is not a well-mixed reservoir, however.

The transport of trace components generally follows the mean air currents established by the general circulation patterns in the troposphere. These are directed westward in the tropics and eastward in mid-latitudes. Transport in other directions occurs primarily via turbulent mixing processes. Turbulent transport is directed from higher to lower mixing ratios of a trace compound, that is, along the gradient of the mixing ratio. Such gradients arise because of the competition between transport and chemical removal rates. The large difference in vertical and horizontal scales, 15 versus 10,000 km in each hemisphere, makes the troposphere highly anisotropic with regard to the degree of turbulence and the rate of transport. Typical time scales within each hemisphere are 2–3 weeks for vertical mixing, 4 weeks for zonal transport, and 3 months for transport between Pole and equator. The exchange between the northern and southern hemispheres takes about 1 year. The troposphere is a reasonably well-mixed reservoir only for long-lived trace compounds, whereas others with lifetimes less than 1 yr will develop spatially nonuniform distributions.

The lifetime τ of a compound is inversely proportional to its total removal rate: $\tau = 1/k = 1/\sum k_i [s]$, where the k_i are first-order rate coefficients associated with the individual processes contributing to the removal. Classical first-order processes are radioactive decay and photochemical destruction. The uptake of a trace gas by materials at the earth surface (soils and vegetation) also represents a first-order process. Most chemical reactions follow second-order kinetics. In this case, the reaction rate depends on the concentrations of two species, that of the compound being removed and that of a second reactant. First-order kinetic conditions can nevertheless be approximated by setting the concentration of the second reactant constant. This approach requires that spatial and temporal variations of its concentration are smoothed out to derive an average value. This procedure is frequently applied to estimate the lifetime of a compound in the troposphere. For longer-lived trace gases that permeate the entire troposphere, the reservoir theory holds. With the assumption that the global input rate and the total removal rate are approximately in

balance, the residence time is $\tau = G/Q$, where G is the total trace gas content in the reservoir and $Q = \sum Q_i$ is the global input rate resulting from the individual sources. As input and removal rates usually fluctuate considerably in space and time, it is necessary also in this model to work with suitable averages.

II. TRACE GASES

Water vapor, H_2O , and carbon dioxide, CO_2 , are minor components of air. Both are chemically rather inert in the troposphere, but H_2O participates in the generation of OH radicals, which initiate many important oxidation reactions. Water vapor enters the troposphere by evaporation from the ocean and transpiration of land plants, and it is removed by cloud formation and precipitation. The local abundance of H_2O is limited by the saturation vapor pressure, which depends on temperature, so that the H_2O mixing ratio decreases with altitude from about 7% over the tropical oceans to $3\text{ }\mu\text{mol mol}^{-1}$ in the lower stratosphere. The abundance of CO_2 has been rising in the last century owing to emissions from the combustion of fossil fuels. The average mixing ratio currently is about $365\text{ }\mu\text{mol mol}^{-1}$. Carbon dioxide is fairly evenly distributed in the atmosphere. Interaction with the biosphere leads to local and seasonal variations amounting to a few percent of the total abundance.

Table I presents an overview on important trace gases in the troposphere, summarizing data on approximate mixing ratios, spatial distributions, and dominant sources and removal processes. Most of the gases are emitted at the earth surface from multiple sources. Ozone, which is produced entirely within the atmosphere, is an outstanding exception. The flux of a trace gas through the troposphere depends on the strength of the surface-bound sources, whereas its abundance is determined by the total rate of the removal processes. A steady state is achieved when total emission rates and removal rates balance, and this condition determines the global budget. An important reagent for many trace gases is the hydroxyl radical. A global average OH concentration of $(8 \pm 2) \times 10^5$ molecule cm^{-3} has been derived from the observation of trace gases that have only anthropogenic sources with a well-known global emission rate, such as methyl chloroform (1,1,1-trichloroethane), and the known rate coefficient for reaction with OH radicals. Because all sources and removal processes are subject to diurnal and seasonal variations, the budget usually refers to the time period of 1 year. In the following, the budgets of several important trace gases are briefly described.

A. Methane

Methane is generated by anaerobic bacteria in swamps, rice fields, and in other anaerobic media. The global emission rate currently is about 500 Tg yr^{-1} . Natural emissions contribute 27%, agriculture 44%, landfills and sewage 15%, and coal mining and natural gas production 14%, approximately. Ice-core records show that the CH_4 mixing ratio has risen from a preindustrial value of about $0.7\text{ }\mu\text{mol mol}^{-1}$ 200 years ago to $1.7\text{ }\mu\text{mol mol}^{-1}$ today, which demonstrates the impact of anthropogenic sources. With a tropospheric residence time of 8.5 yr, the CH_4 mixing ratio is globally nearly uniform. The northern hemisphere contains a slight excess of $\sim 7\%$. The principal process removing methane from the troposphere is reaction with OH radicals. A small fraction escapes to the stratosphere and is oxidized in that atmospheric domain.

B. Carbon Monoxide

The CO mixing ratio of 50 nmol mol^{-1} in the southern hemisphere corresponds closely to that produced from the oxidation of methane. The northern hemisphere contains nearly three times more CO than the southern. The uneven distribution results from the fairly short lifetime of CO combined with an excess of sources in the north. The global source strength is 2500 Tg yr^{-1} . The oxidation of CH_4 contributes 600 Tg, another 800 Tg is produced from the oxidation of other natural hydrocarbons where formaldehyde occurs as an intermediate. Fossil fuel-related sources generate 450 Tg and the burning of biomass 700 Tg. Fossil fuel consumption occurs mainly in the northern industrialized regions; biomass burning is an agricultural practice in the tropics. The removal of CO from the troposphere occurs to 85% by reaction with OH radicals. The remainder is bacterial consumption in soils and loss to the stratosphere.

C. Hydrocarbons

In addition to methane, the troposphere harbors a great variety of volatile hydrocarbons, with carbon numbers ranging at least up to C_{30} . Low molecular weight species occur with the highest abundances. The removal of hydrocarbons takes place predominantly by reaction with OH radicals. The reaction rate increases with the carbon number of the compound, and the lifetime decreases accordingly. On the continents, where most of the sources are located, the mixing ratios of most hydrocarbons decrease nearly exponentially with increasing altitude. Ethane, acetylene, and propane have lifetimes on the order of 56, 18, and 12 days, respectively, which is long enough for transport around the globe. Large gradients are observed between

TABLE I Approximate Molar Mixing Ratios, Global Distribution, Sources and Sinks, and Lifetimes of Several Important Trace Gases in the Troposphere^a

Trace gas	Mixing ratio ^b	Distribution	Major sources (Tg y ⁻¹)	Major sinks (Tg y ⁻¹)	Residence time
Methane, CH ₄	1.7 ppm Rising	Uniform	Rice paddy fields, 75 Domestic animals, 100 Swamps/marshes, 120 Biomass burning, 50 Fossil sources, 90	Reaction with OH, 430 To the stratosphere, 40 Uptake by soils 30	8.5 yr
Hydrogen, H ₂	0.5 ppm	Uniform	CH ₄ oxidation, 20 Oxidation natural VOC, 18 Biomass burning, 20 Fossil fuel use, 17	Reaction with OH, 16 Uptake by soils 70	2 yr
Carbon monoxide, CO	150 ppb 50 ppb	NH SH	Anthropogenic, 450 Biomass burning, 700 CH ₄ oxidation, 600 Oxidation natural HC, 800	Reaction with OH, 2000 To the stratosphere, 100 Uptake by soils, 300	2 m
Ozone, O ₃	15–50 ppb	Low near equator Rising to the poles	Influx from stratosphere, 600 Photochemical production, 4200	Dry deposition, 1350 Photochemical loss, 3450	1 m
Nitrous oxide, N ₂ O	0.3 ppb Rising	Uniform	Emission from soils, 10 Emissions from oceans, 6 Anthropogenic, 9	Loss to stratosphere, 19	110 yr
Nitrogen oxides, NO, NO ₂	30 ppt 300 ppt 5 ppb	Marine air Continental remote Continental rural	Fossil fuel-derived, 21 Biomass burning, 8 Emissions from soils, 7 Lightning, 5	Oxidation to HNO ₃ by OH and O ₃ , removal by wet and dry deposition	2 d
Nitric acid, HNO ₃	0.1–2 ppb 70 ppt 50–130 ppt	Continental air Marine air Free troposphere	Oxidation of NO ₂ , 180	Formation of NO ₃ ⁻ aerosol Wet and dry position	6 d
Peroxyacetyl nitrate, CH ₃ C(O)OONO ₂	120–180 ppt 10–90 ppt 7–10 ppt	Free troposphere, cont. Marine, NH Marine, SH	Chemical production, represents Major part of NO _x reservoir	Thermal dissociation	2–100 d
Ammonia, NH ₃	50–90 ppt 5 ppb	Marine air Continental air	Domestic animals, 25 Emissions from vegetation, 6 Emissions from oceans, 7 Fertilizer use, 6	Dry deposition, 15 Conversion to NH ₄ ⁺ aerosol and wet deposition, 30	3 d
Carbonyl sulfide, OCS	500 ppt	Uniform	Soils and marshes, 0.3 Emission from ocean, 0.3 Oxidation of CS ₂ and DMS, 0.5	Uptake by vegetation, 0.5 Reaction with OH, 0.13 Loss to the stratosphere, 0.1	7 yr
Hydrogen sulfide, H ₂ S	5–30 ppt 50–100 ppt	Marine air Continental rural, Higher in urban air	Emission from soils, 0.5 Emission from vegetation, 1 Volcanoes, 1	Reaction with OH	3 d
Carbon disulfide, CS ₂	<30 ppt 35–190 ppt	Marine air Continental air	Emission from ocean, 0.4 Anthropogenic, 0.6	Reaction with OH	7 d
Dimethyl sulfide, CH ₃ SCH ₃	20–150 ppt	Marine air Continental rural, higher in urban air	Emission from oceans, 50 Soils and vegetation, 4	Reaction with OH Reaction with NO ₃	2 d
Sulfur dioxide, SO ₂	20–90 ppt 0.2–10 ppb	Marine air Continental rural	Fossil fuel-derived, 160 Volcanoes, 16 Oxidation of sulfides, 45	Dry deposition, 100 Oxidation to SO ₄ ²⁻ aerosol and wet deposition, 140	4 d
Ethane, C ₂ H ₆	1.3 ppb 0.4 ppb	NH SH	Natural gas losses, 6 Biomass burning, 7	Reaction with OH	60 d
Propane, C ₃ H ₈	1 ppb 0.2 ppb	NH SH	Mainly anthropogenic sources, 23	Reaction with OH	12 d
Isoprene, C ₅ H ₈	0.2–5 ppb	Continental surface air	Emissions from deciduous trees, 570	Reaction with OH	0.2 d

(continued)

TABLE I (continued)

Trace gas	Mixing ratio ^b	Distribution	Major sources (Tg y ⁻¹)	Major sinks (Tg y ⁻¹)	Residence time
Terpenes, C ₁₀ H ₁₆	0.03–2 ppb	Continental surface air	Emissions from coniferous and deciduous trees, 140	Reaction with OH	0.4 d
Methyl chloride, CH ₃ Cl	600 ppt	Uniform	Emissions from ocean, 3 Biomass burning, 0.7	Reaction with OH	1.3 yr

^a yr = year; m = month; d = day; NH = northern hemisphere; SH = southern hemisphere.

^b Approximate amount fractions: ppm = $\mu\text{mol mol}^{-1} = 10^{-6}$; ppb = nmol mol⁻¹ = 10^{-9} ; ppt = pmol mol⁻¹ = 10^{-12} .

the northern and southern hemisphere. Ethane occurs in marine air of the northern hemisphere with an average mixing ratio of 1.4 nmol mol⁻¹ compared to 0.4 nmol mol⁻¹ in the southern. Similarly uneven distributions are observed for acetylene and propane. Global emission rates are reasonably well defined only for ethane. Natural gas losses and biomass burning are the major sources. Light alkenes, such as ethene, C₂H₄, or propene, C₃H₆, have lifetimes of ~ 1 d. In the continental boundary layer, their mixing ratios are about 1 nmol mol⁻¹. Mixing ratios in marine air are about 100 pmol mol⁻¹ in both hemispheres. These hydrocarbons would not survive transport from continental to marine regions. They occur in marine air because they are released from the ocean.

Estimates for emissions from anthropogenic sources of hydrocarbons fall in the range 70–180 Tg yr⁻¹. Three types of sources are most important: The production of liquid fuels from petroleum and road traffic contributes about 54% in developed countries and 36% worldwide. Solvent use adds another 15%. The third major source is biomass burning, which contributes 34% on a global scale (16% in developed countries). About 45% of the emissions are alkanes, 35% are alkenes, and 17% are aromatic compounds.

The foliage of trees and other vegetation provides a natural source of many volatile organic compounds. The dominant hydrocarbons are isoprene (2-methyl-1,3-butadiene), which is emitted only from deciduous trees, and monoterpenes that are emitted from conifers and many deciduous trees as well. Monoterpenes are C₁₀H₁₆ compounds, mostly cyclic alkenes, with α -pinene, β -pinene, and d-limonene being prominent representatives. Both isoprene and the monoterpenes derive from the same metabolic pathway, but the production of isoprene is associated with the photosynthetic process, so that the emission rate depends on the light intensity, in contrast to that of monoterpenes. The emission rates of isoprene and the monoterpenes also increase with rising temperature, so that they undergo strong diurnal variations. Isoprene has been estimated to contribute 570 Tg yr⁻¹ to the source strength of hydrocarbons on a global scale; the contribution of monoterpenes is 140 Tg yr⁻¹. Both compounds

react rapidly with OH radicals, and lifetimes are very short. This limits their occurrence to the continental boundary layer. Mixing ratios typically range from 0.1 to 5 nmol mol⁻¹, with strong diurnal variations.

D. Nitrous Oxide

The uniform and rather stable mixing ratio of N₂O in the troposphere, *ca.* 300 nmol mol⁻¹, suggests a long residence time for this compound. N₂O is an example for a chemically nonreactive compound, because no mechanisms are known that destroy N₂O in the troposphere. The removal occurs by transport to the stratosphere where N₂O undergoes photolysis. Long-term measurements have shown that the tropospheric N₂O concentrations increase by about 0.25% annually. Comparison with air trapped in the great polar ice sheets indicated that the level before 1800 was lower, about 285 nmol mol⁻¹. Nitrous oxide is released naturally from soils and ocean waters, where it is produced by nitrifying and denitrifying bacteria that are active within the microbial nitrogen cycle. These sources contribute 10 and 6 Tg yr⁻¹, respectively. Human activities have resulted in a variety of additional sources summing to 9 Tg yr⁻¹. The stratosphere is estimated to destroy 16–19 Tg yr⁻¹; the remainder accumulates in the troposphere and is responsible for the observed rise in mixing ratio.

E. Nitric Oxide and Nitrogen Dioxide

Both compounds must be considered together (NO_x = NO + NO₂) because photolysis by sunlight converts NO₂ partly to NO, and the concentrations of both species are related by a photostationary state. With a lifetime of NO_x on the order of a few days and with continental sources being dominant, the NO_x mixing ratio in the troposphere is highest in the continental surface boundary region and lowest in remote marine air. The global budget of NO_x of some 40 Tg yr⁻¹ results in 75% from human activities, that is, the combustion of fossil fuels and biomass burning. The remainder is due to natural sources from soils and lightning. The emission occurs mainly in the form of

NO in all cases. In the troposphere, NO is first oxidized to NO_2 by ozone, and NO_2 is then further oxidized to nitric acid, HNO_3 , which partly forms nitrates that become associated with aerosol particles, and both HNO_3 and particulate nitrate return to the earth surface by wet and dry deposition.

F. Ammonia

The overriding source of ammonia in the troposphere is bacterial decomposition of urea excreted by domestic animals, which is estimated to release about 25 Tg yr^{-1} . The use of fertilizer and the escape from agricultural soils also is an appreciable source. Thus, almost 70% of total emissions derive from anthropogenic sources. Emissions from vegetation provides a natural source. A compensation point exists: a concentration such that plants assimilate ammonia when the ambient concentration is higher, and release it when the concentration is lower, than the compensation point. The compensation point for most trees is about 0.4 nmol mol^{-1} , whereas continental NH_3 mixing ratios usually are higher, so that forests tend to remove ammonia. On the other hand, grassland and crops generally act as sources. The ocean also is a natural source of ammonia. In the marine atmosphere, ammonia combines with sulfuric acid to form sulfates, and returns to the ocean mainly by wet deposition. Over the continents the removal occurs either by dry deposition to vegetation or by the formation of ammonium sulfates and wet deposition. The reaction of OH radicals with ammonia is slow and comparatively insignificant.

G. Reduced Sulfur Compounds

The principal reduced sulfur compounds occurring naturally in the troposphere are carbonyl sulfide, OCS, hydrogen sulfide, H_2S , carbonyl disulfide, CS_2 , and dimethyl sulfide, CH_3SCH_3 (DMS). Ocean waters, coastal wetlands, soils and vegetation contribute to emissions of reduced sulfur compounds. Carbonyl sulfide, which is the least reactive, occurs with the highest mixing ratio and with a uniform distribution throughout the troposphere. It is produced in sea water from organic sulfur compounds, most copiously in coastal and shelf regions, and by bacteria in marshes and continental soils. The oxidation of DMS and CS_2 induced by hydroxyl radicals provides another significant source. The main removal process appears to be assimilation by plants. This process converts OCS partly to H_2S , which is released. A small fraction of OCS is removed by reaction with OH and by loss to the stratosphere. Hydrogen sulfide is emitted from soils and marshes, from vegetation, and from volcanoes. Emissions from the oceans are negligible. Source strengths are diffi-

cult to quantify, however. Estimates for the global emission rate range from 1 to 5 Tg yr^{-1} ; the data in Table I must be considered approximate. The main loss process for H_2S in the troposphere is reaction with OH radicals. Carbon disulfide is emitted from the ocean, which is the dominant natural source, from soils and marshes, and from anthropogenic processes such as the manufacture of regenerated cellulose and cellophane. Also in this case, the emission rates are difficult to quantify.

Dimethyl sulfide is emitted mainly from the ocean where it is released from phytoplankton. Estimates of emission rates range from 30 to 68 Tg yr^{-1} . Soils and vegetation contribute comparatively little to the global emission rate. The rate of DMS emissions evidently exceeds that of all other reduced sulfur compounds. Although this makes DMS the most important reduced sulfur compound globally, its impact is essentially confined to the marine atmosphere. The removal of DMS occurs primarily by reaction with OH radicals.

H. Sulfur Dioxide

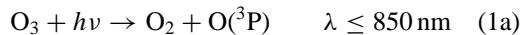
Anthropogenic activities leading to emissions of sulfur dioxide are the combustion of fossil fuels in electric power plants and the smelting of nonferrous metals. Emission rates have been rising in the period 1860–1980. In Europe and North America they have decreased in recent years as a result of pollution control measures. In China, however, the emissions currently are rising. The global emission rate of 160 Tg yr^{-1} shown in Table I refers to 1985. Volcanoes represent a significant natural source of SO_2 . Continuous emissions contribute about 16 Tg yr^{-1} , but a similar amount may be released in a single violent eruption. The oxidation of sulfides also leads to the production of SO_2 . The oxidation of DMS makes the most significant contribution, whereas that of other sulfides is negligible in comparison. At least 85% of DMS is oxidized to sulfur dioxide, leading to a production rate of ~ 45 Tg yr^{-1} . The main fate of SO_2 in the troposphere is oxidation to sulfuric acid, occurring predominantly by aqueous reactions in clouds, followed by association with ammonia to produce ammonium sulfate aerosol when the cloud evaporates. Eventually, when the aerosol particles are incorporated in a raining cloud, they undergo wet precipitation. On the continents, where SO_2 concentrations are high, the direct dry deposition of SO_2 is an important removal process.

III. PHOTOCHEMICAL PROCESSES

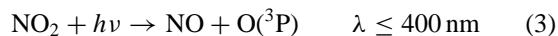
Photochemical processes play a dominant role in triggering chemical reactions in the troposphere. Photodissociation requires that the compound to be photolysed absorbs

solar radiation in the accessible wavelength region, and that the energy transferred to the molecule is sufficient to break a chemical bond. The latter requirement is usually met only by radiation in the ultraviolet region of the solar spectrum. The most energetic extreme ultraviolet part of the solar spectrum is absorbed by oxygen in the upper atmosphere (<170 nm) and by ozone in the stratosphere (<300 nm), so that only radiation at wavelengths >300 nm penetrates downward into the troposphere. This subdivides photolytic processes into those that occur only in the stratosphere or above and those that can occur at all altitudes. N₂O, for example, is one of the few compounds that are chemically inert in the troposphere and that have the onset of optical absorption at wavelengths below 300 nm, so that they must rise to the stratosphere to become susceptible to photodissociation. The number of chemical compounds that absorb solar radiation and undergo photolysis at wavelengths reaching the troposphere is fairly small. Most important are ozone, nitrogen dioxide, and formaldehyde. Each of these substances plays a critical role in the overall chemical mechanisms.

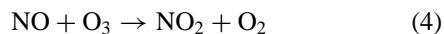
The photodissociation of ozone generates mainly ground-state oxygen atoms, O(³P). Their major fate is the third body-assisted recombination with molecular oxygen



which keeps the steady-state concentrations of oxygen atoms at a level too low for any of their reactions to gain significance in the oxidation of trace gases. The photolysis of nitrogen dioxide leads also to the formation of oxygen atoms



which recombine with O₂ to form ozone. Outside the continental boundary layer, sufficiently far away from surface emissions of NO_x, photolysis of NO₂ is the only source of NO in the troposphere. NO reacts readily with ozone regenerating NO₂



but it also enters into competing reactions with peroxy radicals (see further below), which cause photolysis of NO₂ to be a source of ozone.

At wavelengths below ~320 nm, the character of O₃ photolysis changes. Solar photons in this wavelength region carry enough energy for the formation of oxygen atoms excited to the first electronic state, O(¹D), which lies about 190 kJ mol⁻¹ above the ³P ground state. The ¹D state has a fairly long radiative lifetime (~140 sec), so that it is largely deactivated by collisions with the major

constituents of air. However, a fraction of O(¹D) reacts with water molecules to produce hydroxyl radicals



The last reaction is an important source of OH radicals in the troposphere. Hydroxyl radicals do not undergo reactions with any of the major constituents of air, but they react rapidly with many trace compounds and represent the most significant oxidant in the troposphere. Up to 25% of the O(¹D) atoms are converted to OH under favorable atmospheric conditions. The spectral window available for the production of O(¹D) and OH radicals is quite narrow. Toward short wavelengths, it is limited by the cut-off of the solar spectrum near 300 nm; at wavelengths near 320 nm, it becomes ineffective by the diminishing quantum yield for O(¹D) formation in the photolysis of ozone. Figure 1 illustrates the overlap between solar radiation intensity, the O₃ absorption spectrum, and the O(¹D) quantum yield. The area underneath the action curve shown at the bottom of the figure represents the photodissociation coefficient

$$J(\text{O}(\text{D}^1)) = \int q(\lambda)\sigma(\lambda)I_0(\lambda) d\lambda \quad [\text{s}^{-1}]$$

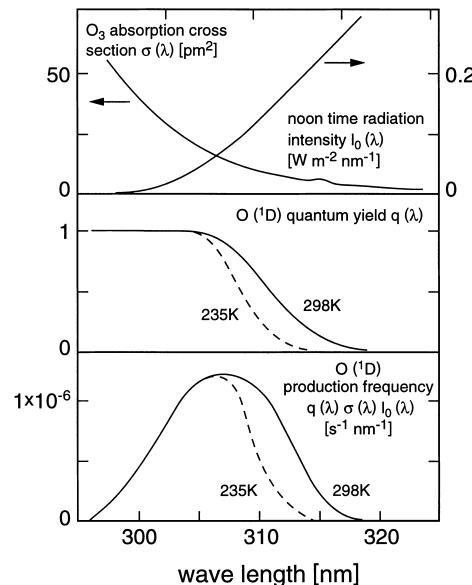


FIGURE 1 Photodissociation of ozone in the near ultraviolet spectral region. Overlap of solar radiation actinic intensity, ozone absorption cross section, and O(¹D) quantum yield to derive the O(¹D) production frequency as a function of wavelength. [From Zellner, R., ed. (1999). "Global Aspects of Atmospheric Chemistry," Steinkopff/Springer, Darmstadt, Germany.]

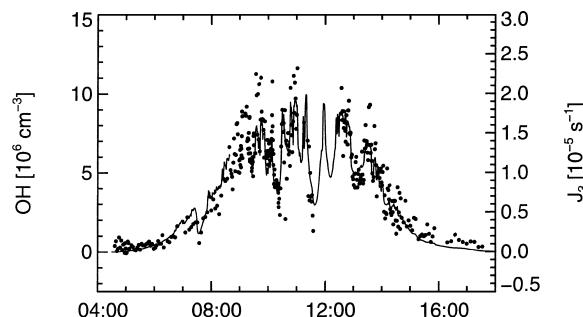


FIGURE 2 Diurnal variation of OH concentration (against universal time) measured by laser-induced fluorescence in August 1994 in northern Germany (points, left-hand scale), compared with the ozone photolysis frequency to form O(¹D) (thin line, right-hand scale). [From Zellner, R., ed. (1999). “Global Aspects of Atmospheric Chemistry,” Steinkopff/Springer, Darmstadt, Germany.]

where the absorption cross section $\sigma(\lambda)$ and the quantum yield $q(\lambda)$, both as a function of wavelength λ , are quantities determined by appropriate laboratory experiments, and $I_0(\lambda)$ is the so-called actinic intensity, which includes not only the direct solar radiation but also scattered light reaching the measurement point from all sides. Production rates and concentrations of OH radicals maximize in the lower troposphere, because the highest H₂O concentrations are found there, and in the equatorial region where the ultraviolet solar flux is least attenuated by the stratospheric ozone layer. In addition, the OH production rate varies diurnally, maximizing at local noon and vanishing at night. Figure 2 shows recent measurements that demonstrate the close correspondence of the diurnal variation of OH concentration and O(¹D) formation rate in the northern hemispheric rural continental atmosphere in summer.

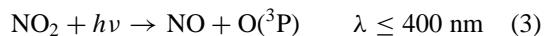
IV. BASIC GAS-PHASE FREE RADICAL CHEMISTRY

The major loss reactions of OH radicals involve carbon monoxide (50–70%) and methane (10–20%). These reactions lead to the formation of hydroperoxy radicals, HO₂, that can regenerate OH radicals, either by reaction with nitric oxide, NO, or by reacting with O₃, for example

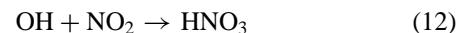
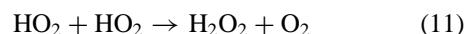


The rate of the reaction with NO is three orders of magnitude faster than that with O₃. Over the continents, where

the mixing ratio of NO under daytime conditions usually exceeds 0.1 nmol mol⁻¹, the HO₂ radical reacts predominantly with NO. In the marine background troposphere both reactions are equally important. Nitrogen dioxide formed in reaction (9) is photolysed



whereby both NO and O₃ are regenerated. Reactions (7–9) represent a system of chain reactions. In this system the HO_x = OH + HO₂ radicals are recycled several times, until the cycle is interrupted by a termination reaction. In the troposphere, the most important HO_x termination reactions are

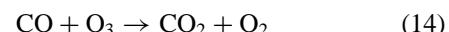


The products H₂O₂ and HNO₃ are well soluble in water, and they are removed from the troposphere fairly rapidly by absorption in cloud water and precipitation. Both H₂O₂ and HNO₃ can also be photolysed, whereby OH radicals are regenerated, but the photodissociation rates are fairly small, and wet removal predominates.

Because the HO_x chain cycle proceeds several times before the radicals enter a termination reaction, the reaction sequence (7–9) followed by (3) and (2) can be combined to result in the net reaction



demonstrating the formation of excess ozone. On the other hand, the combination of reactions (7–8) followed by (10) results in the net reaction



leading to ozone destruction. In the marine background atmosphere, NO_x concentrations are low, and HO_x radicals react with O₃ most of the time, causing its destruction. Over the continents, at higher NO_x concentrations, ozone is produced. At very high concentrations of NO_x, such as occurring under urban conditions, the termination reaction (12) causes appreciable losses of HO_x, so that the chain reaction is quenched.

V. THE OXIDATION OF METHANE

Figure 3 shows the reaction sequence for the OH radical-induced oxidation of methane. The first product following hydrogen atom abstraction is a methyl radical. Rapid addition of oxygen converts the methyl radical to methyl peroxy

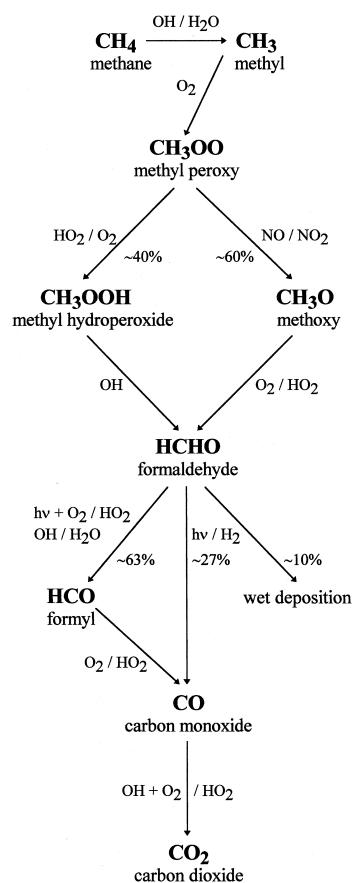
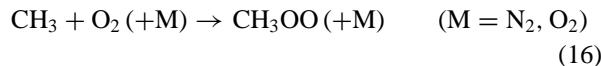
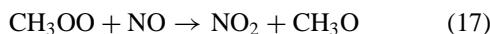


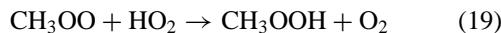
FIGURE 3 Mechanism for the oxidation of methane.



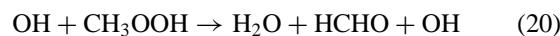
It can react either with NO or with HO₂. The first reaction causes the abstraction of an oxygen atom and produces a methoxy radical, which reacts with oxygen to produce formaldehyde and HO₂



The second reaction leads to methyl hydroperoxide as product.

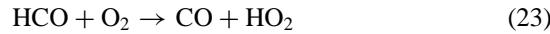
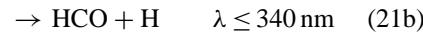
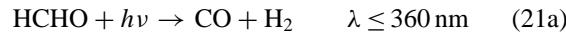


Methyl hydroperoxide is much less soluble in cloud water than H₂O₂, so that CH₃OOH reacts largely with OH radicals

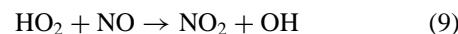


Alternatively, CH₃OOH may undergo photodissociation leading to the same products.

Formaldehyde, as noted earlier, is subject to photolysis, and it reacts with OH as well. Two channels exist for photodissociation. One leads to CO and H₂ as stable products, the other channel produces a formyl radical, which reacts further with oxygen to produce CO and an HO₂ radical



Under conditions of sufficiently high NO_x concentration, the reactions are followed by



so that the OH radical lost in reaction (22) is regenerated. Moreover, the production of H and HCO associated with HCHO photolysis results in a net gain of HO_x by 40–70%. The steady-state mixing ratio observed in clean marine air, where the oxidation of methane provides the dominant HCHO source, is ~200 pmol mol⁻¹, in agreement with that calculated in model simulations. Over the continents, formaldehyde is produced not only from methane but also in the oxidation of other hydrocarbons, and it is emitted directly from automobiles. Under these conditions, the photolysis of formaldehyde provides a significant source of HO_x. All pathways of HCHO oxidation in the gas phase lead to CO. The end products of methane oxidation are H₂O and CO₂. The former is deposited at the earth surface, the latter returns to the pool of atmospheric CO₂, which the biosphere had used to produce CH₄ in the first place. In this manner, the elements carbon and hydrogen are cycled through the atmosphere back to the inorganic reservoir from where they originated.

VI. HYDROCARBON OXIDATION MECHANISMS

Although a great number of different hydrocarbons occur in the troposphere and individual oxidation schemes can be quite complex, the mechanisms follow the same overall principle. The process is initiated by reaction with an OH radical, proceeds via organic peroxy and alkoxy radicals as intermediates, and leads to carbonyl compounds as oxidation products. In this respect the individual reaction steps are similar to those discussed above for methane.

A. Alkanes

Figure 4 shows a simplified mechanism for the oxidation of alkanes in the troposphere. Hydrogen abstraction

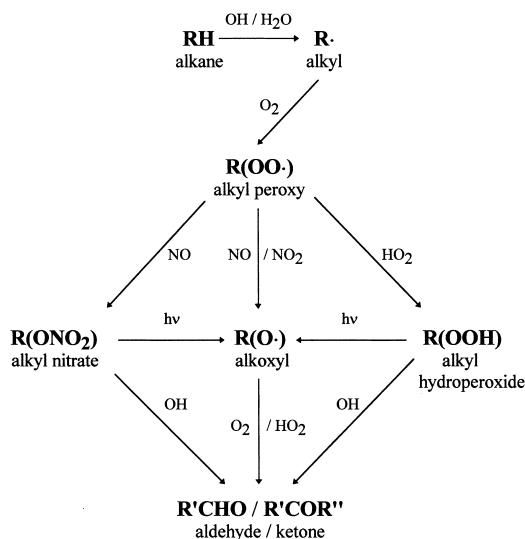
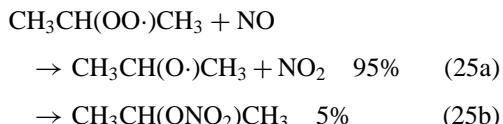
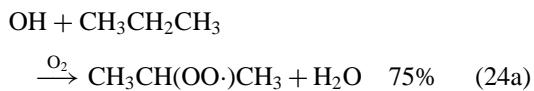


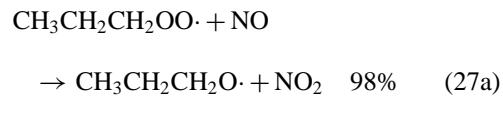
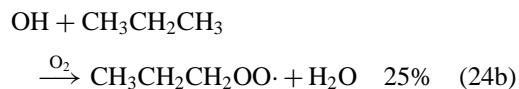
FIGURE 4 Schematic oxidation mechanism for hydrocarbons.

is followed by the addition of oxygen to the reaction site. The resulting alkylperoxy radical, RO_2 , reacts predominantly with NO to form an alkoxy radical, RO, which then reacts further with O_2 to produce either an aldehyde or a ketone. A parallel reaction of RO_2 with NO forms alkyl nitrates. For the lower alkanes, this is a minor reaction channel, but it increases in importance for the higher alkanes. The interaction of RO_2 with HO_2 radicals, which leads to alkylhydroperoxides, becomes important when NO_x concentrations are low. Subsequent reactions of alkyl nitrates and hydroperoxides also produce aldehydes or ketones. Alkylperoxy radicals can also associate with NO_2 to form alkylperoxy nitrates. These are thermally unstable, however, and provide only a temporary reservoir of RO_2 radicals.

In contrast to methane, the hydrogen atoms of alkanes are not equivalent. Three different types exist: primary H-atoms associated with CH_3 groups, secondary H-atoms attached to $-\text{CH}_2-$ chain links, and tertiary H-atoms located at chain branching points. The last two are removed more easily than primary H-atoms. The oxidation of propane may serve a simple example to illustrate this principle. The reaction sequence following abstraction of a secondary H-atom is



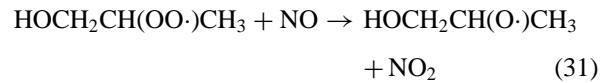
whereas the abstraction of a primary H-atom leads to



In the case of propane, the frequency of removal of a secondary H-atom is 75% and that of a primary H-atom 25%. In the first case the final product is acetone, in the second case it is propionaldehyde.

B. Alkenes

Hydroxyl radicals react with alkenes mainly by addition to the double bond. This generates a new reaction site at the neighboring carbon atom where oxygen is added. The resulting hydroxy-alkylperoxy radical reacts with NO in a manner similar to that of alkylperoxy radicals to produce hydroxy-alkoxy radicals, for example, in the case of propene

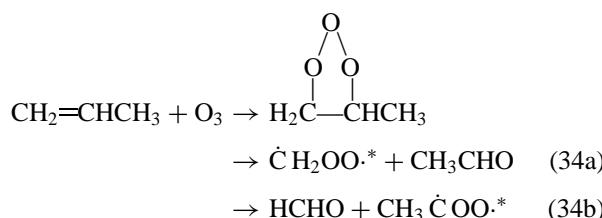


Unlike alkoxy radicals, however, the hydroxy-alkoxy radicals undergo mainly decomposition rather than reaction with oxygen

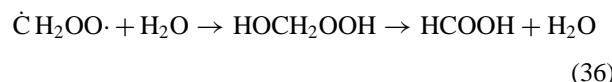
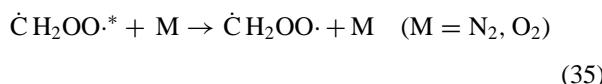


which leads to the formation of formaldehyde and a second aldehyde for all alkenes featuring a terminal double bond. The OH radical may attach either at the terminal position or at the adjacent carbon atom. The first pathway is favored and occurs approximately 75% of the time.

Alkenes can also be oxidized by ozone. The addition of O_3 to the double bond forms an unstable ozonide, which decomposes to yield an aldehyde and a Criegee radical, for example



The Criegee radicals contain excess energy and decompose to yield a variety of fragments unless the radicals are stabilized by collisions. At atmospheric pressure, about 50% of the radicals are stabilized and react predominantly with water to yield acids, for example



with formic acid as the final product. This is one of a few pathways leading to the formation of acids in the gas phase. Among the fragments resulting from the decomposition of a Criegee radical is the OH radical. As tropospheric reactions between alkenes and ozone occur at night, such reactions provide a nighttime source of OH radicals.

C. Aromatic Compounds

As in the case of other hydrocarbons, the oxidation of aromatic compounds is induced by OH radicals. Figure 5 illustrates prominent oxidation pathways for toluene. Two routes may be distinguished. One is abstraction of a hydrogen atom from the methyl group, which leads to benzaldehyde as the main product. The other pathway is the addition of OH to the aromatic ring to form

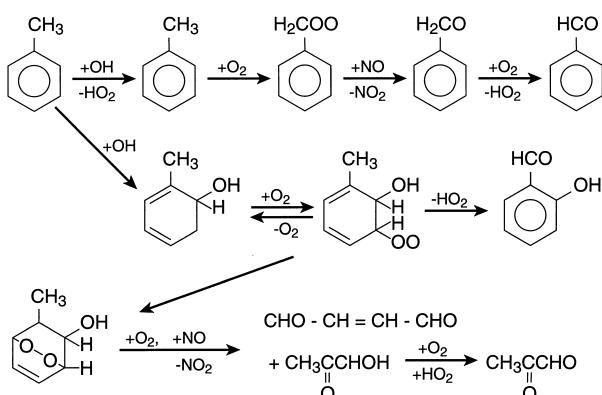
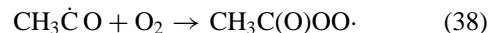


FIGURE 5 Oxidation mechanism for toluene. The pathway leading to ring opening is not fully understood, only one of several possibilities is indicated. [From Warneck, P. (1999). "Chemistry of the Natural Atmosphere," Academic Press, San Diego.]

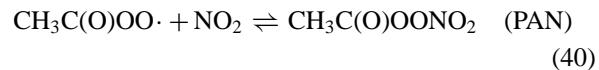
the methyl-hydroxy-cyclo-hexadienyl radical (several isomers), which reacts further with oxygen to produce cresols with about 35% yield. Other reaction products, such as formaldehyde, acetaldehyde, methyl glyoxal, glyoxal, and 1,4-dicarbonyl compounds, result from ring cleavage in a mechanism that remains to be fully elucidated.

D. Aldehydes and Ketones

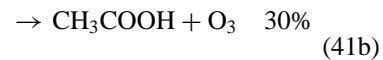
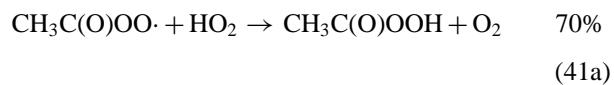
Acetaldehyde and acetone are prominent representatives of carbonyl compounds produced by the oxidation of hydrocarbons. In contrast to formaldehyde, the higher aldehydes are photochemically less active, but all of them react rapidly with OH radicals. The reaction with acetaldehyde may illustrate the reaction sequence



Acetyl peroxy radicals resulting from the addition of oxygen to $\text{CH}_3\cdot\text{CO}$ react further with NO_x

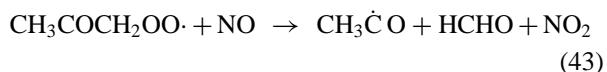
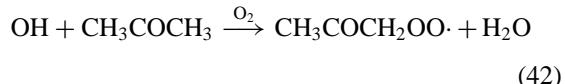


The reaction with NO leads to the formation of CO_2 and a methyl radical that is oxidized to formaldehyde by reactions (16)–(20). In addition, the oxidation of CH_3 regenerates HO_x so that the oxidation cycle continues. Association with NO_2 produces peroxyacetyl nitrate (PAN). Its lifetime is longer than that of alkylperoxy nitrates, but strongly temperature dependent, ranging from 1 hr at 298 K to 140 d at 250 K. Thus, PAN can be transported over a great distance before undergoing thermal decomposition. Under conditions of low NO_x concentrations acetyl peroxy radicals interact also with HO_2 radicals

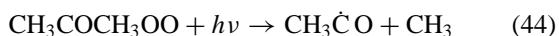


The reaction produces acetic acid and peroxyacetic acid. Both are eventually absorbed by cloud water and leave the troposphere by wet precipitation.

Acetone in the troposphere reacts with OH radicals and undergoes photodissociation. In the first case, the reaction



produces formaldehyde and an acetyl radical, which then enters into the reaction sequence (38)–(41) and forms CO₂, PAN, and acetic acids. The photolysis of acetone, in turn, produces acetyl and methyl radicals



The former enters into the oxidation sequence (38)–(41); the latter undergoes oxidation to formaldehyde. Photolysis generates two new radicals and therefore augments the oxidation chain, whereas oxidation induced by OH will do so only partly by the subsequent photolysis of formaldehyde.

E. Isoprene

Isoprene and monoterpenes react very rapidly with OH radicals. The oxidation pathways are complex and have not yet been fully delineated. The oxidation of isoprene is similar to that of the alkenes discussed above, but as an alkadiene, it permits the addition of OH radicals at four different sites of the two double bonds. Figure 6 sketches the two reaction sequences following terminal OH addition.

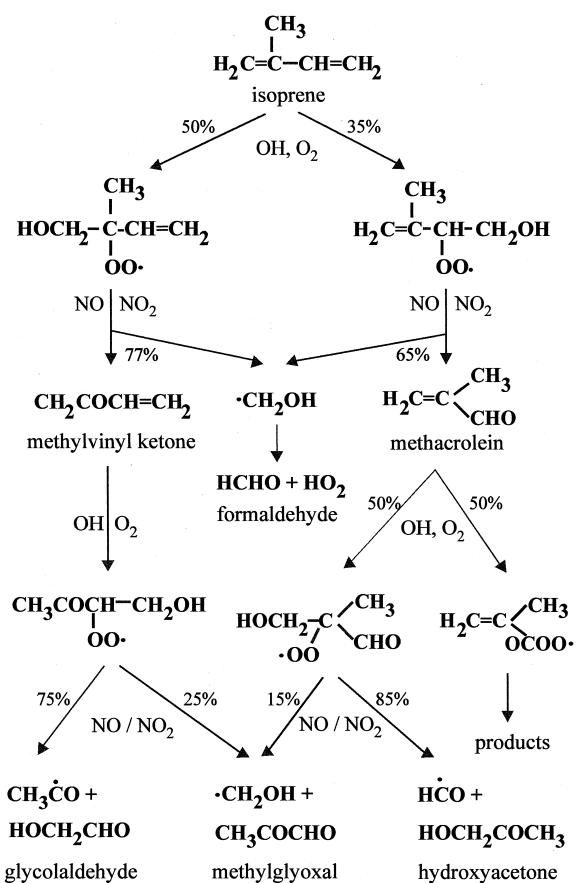
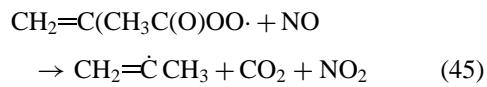


FIGURE 6 Major pathways for the oxidation of isoprene.

tion. The probability for terminal addition at the methylated double bond is 90% and that at the nonmethylated one, at least 75%. As in the case of the alkenes, the reaction proceeds by the addition of oxygen at the neighboring carbon atom followed by reaction with nitric oxide, which removes an oxygen atom from the peroxy group. The resulting hydroxy-alkoxy radicals undergo mainly decomposition to form methyl vinyl ketone and methacrolein. The other fragment, the CH₂OH radical, reacts with oxygen to yield formaldehyde and an HO₂ radical. Methyl vinyl ketone and methacrolein, incidentally, also are major products resulting from the reaction of isoprene with ozone.

Methyl vinyl ketone and methacrolein are further oxidized. The addition of OH radicals to the double bond of methyl vinyl ketone and subsequent reactions produce glycolaldehyde and methylglyoxal. The reaction of OH with methacrolein follows two routes: the OH radical either adds to the double bond or it abstracts a hydrogen atom from the carbonyl group. In the first case, the reaction sequence leads to methylglyoxal and hydroxyacetone as products. In the second case, the methacrylperoxy radical is formed. Reaction with NO₂ converts it to peroxymethacryl nitrate. This species is thermally unstable, however. Ultimately, the methacrylperoxy radical is expected to react with NO, split off carbon dioxide, and form a methylvinyl radical



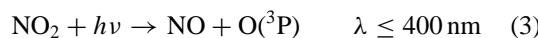
which is further oxidized to yield formaldehyde and an acetyl peroxy radical



The production of formaldehyde in appreciable yield is an important feature of isoprene oxidation. Large amounts of isoprene are naturally emitted from trees, making formaldehyde a prominent by-product. In addition, HCHO photolysis provides a significant source of HO_x radicals that are necessary for hydrocarbon oxidation. The other products, glycolaldehyde, methylglyoxal, and hydroxyacetone, react with OH radicals and can partly undergo photolysis, but the reaction products and their relative yields are not fully known.

VII. OZONE

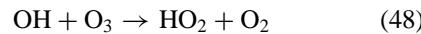
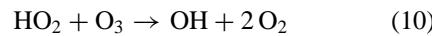
Ozone in the troposphere derives from two sources. One is the influx of ozone from the stratosphere, the other is the interaction of nitrogen oxides with HO₂ and RO₂ radicals that occur as intermediates in hydrocarbon oxidation



In the northern hemisphere, tropospheric ozone exhibits a widespread increase since preindustrial times due to the rise of anthropogenic emissions of nitrogen oxides and hydrocarbons. The increase affects both the boundary region and the free troposphere. In densely populated regions, boundary layer ozone mixing ratios in summer can rise to more than 100 nmol mol⁻¹ under adverse meteorological conditions, appreciably above background levels of 20–35 nmol mol⁻¹.

On a global scale, the influx of ozone from the stratosphere is estimated to contribute about 600 Tg yr⁻¹ to the budget of tropospheric ozone. This is about 1% of that produced in the stratosphere. Total photochemical production of ozone within the troposphere amounts to about 4200 Tg yr⁻¹. The reaction of HO₂ with NO makes the largest contribution, about 3100 Tg yr⁻¹, the reaction of CH₃OO· with NO adds 800 Tg yr⁻¹, and the reaction of other RO₂ radicals, derived largely from the oxidation of isoprene, adds another 300 Tg yr⁻¹. The total production of ozone is clearly dominated by processes in the troposphere, but the import from the stratosphere cannot be neglected. The downward flux features a maximum in later winter, and it is partly responsible for the spring maximum of tropospheric ozone.

Ozone is removed from the troposphere by chemical reactions and by dry deposition. On the continents, the destruction of ozone at the earth surface occurs mainly by vegetation via the leaf stomata, with the uptake rate undergoing diurnal variations (the stomata open during the day and close at night). The solubility of ozone in water is low, so that dry deposition to the ocean surface is less effective than dry deposition on land. Dry deposition is estimated to remove about 1350 Tg yr⁻¹ of ozone from the troposphere, with two-thirds of it being destroyed in the northern hemisphere due to the greater land mass there. The principal chemical reactions responsible for the removal of ozone in the tropospheric air space are



Photodissociation of ozone to form O(¹D) atoms and their reaction with water vapor consumes about 1850 Tg yr⁻¹; the reaction of ozone with HO_x radicals removes another

1400 Tg yr⁻¹. Miscellaneous other reactions, such as the reaction with nitric oxide or that with alkenes, take up another 200 Tg yr⁻¹, approximately. The total amount of ozone removed by chemical reactions is 3450 Tg yr⁻¹. The excess of photochemical production over losses by chemical reactions is 750 Tg yr⁻¹, which is of the same magnitude as the influx from the stratosphere. The residence time of ozone in the troposphere, which is determined by the average concentration and the sum of all sources, is about 20 d. As the turnover is dominated by photochemical processes, the local lifetime depends on latitude and season. Winter lifetimes can be considerably longer, and tropical lifetimes are shorter.

The diurnal variation of the ozone mixing ratio at continental surface sites is determined by boundary layer meteorology. Downward transport of ozone by turbulent mixing, which compensates surface destruction, is more rapid during the day than at night. Flatland stations, therefore, feature a maximum of the mixing ratio in the early afternoon and a minimum at night. Low latitude island sites (e.g., Samoa) display a different diurnal variation. Here, the mixing ratio reaches a maximum in the early morning and a minimum in the late afternoon. Smaller losses by dry deposition to the sea surface, low concentrations of NO_x, and high solar ultraviolet fluxes combine to destroy ozone photochemically during the day. Surface concentrations of ozone over the tropical ocean generally are lower than those at higher latitudes.

VIII. NITROGEN OXIDE CHEMISTRY

[Figure 7](#) shows the relation between different compounds involved in the tropospheric nitrogen oxides cycle. The importance of NO as a catalyst in the oxidation of methane and other hydrocarbons has already been discussed. The principal reactions determining the concentration of NO are photolysis of NO₂ which generates NO and O₃, and reoxidation of NO by O₃ as well as by a suite of RO₂ radicals with HO₂ being most important. Nitric oxide can also associate with OH radicals to form nitrous acid, HNO₂. Under daylight conditions, HNO₂ is rapidly photolysed to regenerate NO and OH. Photostationary conditions favor low concentrations of HNO₂. Nitrous acid is also formed when NO₂ reacts with water adsorbed on surfaces. The continental ground surface as well as the integrated surface of aerosol particles are involved in this reaction. The HNO₂ mixing ratio is observed to increase during the night until sunrise, when photodissociation sets in and the HNO₂ mixing ratio declines. In the boundary layer, the continuous production of HNO₂ at the ground surface followed by photolysis during the day may greatly augment the formation of OH radicals from other sources.

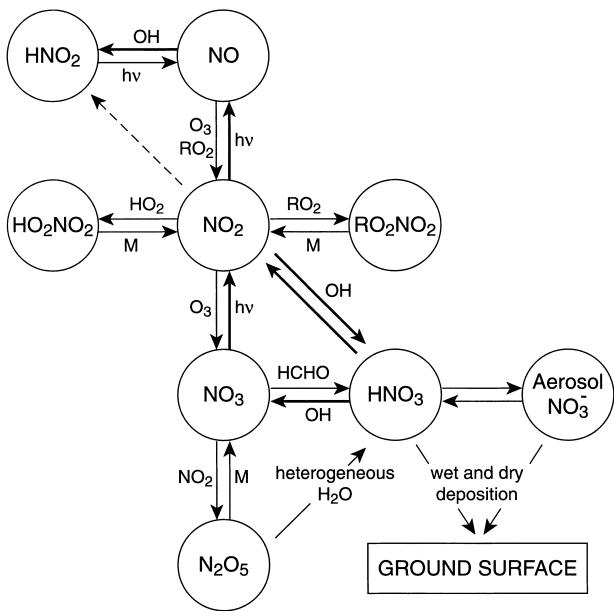
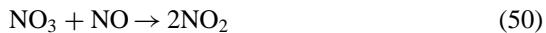
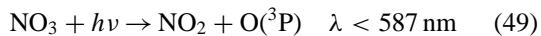


FIGURE 7 Compounds participating in the tropospheric nitrogen oxides cycle and their interrelations. [From Warneck, P. (1999). "Chemistry of the Natural Atmosphere," Academic Press, San Diego.]

Nitrogen dioxide can be oxidized to nitric acid in two ways. One is reaction with OH radicals during the day



which forms HNO_3 directly; the other occurs by the reaction with ozone, which forms the nitrate radical, NO_3 , as an intermediate. During the day, the main fate of NO_3 is either photodissociation or reaction with NO, whereby NO_3 is destroyed

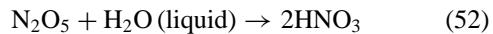


At night, the concentration of NO decreases rapidly to low values, and photolysis also ceases. Under these conditions the main fate of NO_3 is the association with NO_2 to form nitrogen pentoxide

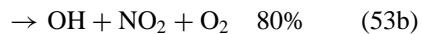
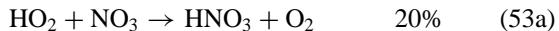


in a reversible reaction. A temperature-dependent equilibrium is established favoring N_2O_5 whenever NO_2 concentrations are high and/or temperatures are low. Under nighttime conditions, temperatures in the continental boundary layer usually decrease sufficiently to convert an appreciable fraction of NO_3 to N_2O_5 . In the upper troposphere, where temperatures are much lower, NO_3 exists mainly in the form of N_2O_5 . Nitrogen pentoxide is the anhydride of

nitric acid, and N_2O_5 undergoes rapid hydrolysis in liquid water. The reaction with water vapor, in contrast, is immeasurably slow. In the lower troposphere, 75% of N_2O_5 reacts heterogeneously with liquid water that is associated with aerosol particles



to form nitric acid. The remainder undergoes thermal decay to NO_3 , which is photolysed after sunrise. The nitrate radical, NO_3 , reacts slowly with alkanes, at moderate rates with alkenes, aldehydes, and aromatic compounds, and quite rapidly with several terpenes. In such reactions NO_3 shows a behavior similar to that of the OH radical. The subsequent hydrocarbon oxidation reactions lead to the formation of RO_2 and HO_2 radicals, but because of the low nocturnal NO concentration, the reaction mechanism differs from that occurring in daylight. The reactions between peroxy radicals and NO_3 are fairly rapid, and they may convert RO_2 to RO radicals at night. The reaction between HO_2 and NO_3 has two channels



The second channel produces OH radicals, which continue the reaction chain, whereas only 20% of the reaction enters the termination channel. However, this nighttime chemistry gains significance only for high NO_2 concentrations. Then the radicals approach concentrations typical of daylight conditions.

A fraction of nitric acid in the troposphere becomes associated with aerosol particles and forms particulate nitrates. The interaction of HNO_3 with virgin sea salt particles stabilizes nitric acid in the form of sodium nitrate, NaNO_3 . Over the continents, the interaction with ammonia leads to the formation of solid ammonium nitrate, NH_4NO_3 . This compound exists in a temperature-dependent equilibrium



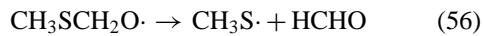
which leads to the formation of ammonium nitrate at sufficiently low temperatures. The interaction with sulfuric acid also liberates HNO_3 from NH_4NO_3 and produces ammonium bisulfate, NH_4HSO_4 , which is a solid in the absence of water, yet deliquesces at relative humidities near 40%, so that in the lower troposphere most of it occurs as an aqueous liquid. Observations show that in the boundary layer gaseous nitric acid and particulate nitrate are about equally abundant, whereas in the free troposphere nitric acid contributes 70–80% to the total concentration of nitrate.

Figure 7 shows that NO_2 forms adducts with HO_2 and with RO_2 radicals leading to temperature-dependent

equilibria with peroxy nitric acid, HO_2NO_2 , and organic peroxy nitrates, RO_2NO_2 , respectively. HO_2NO_2 has not yet been observed in the troposphere, although it should be present in the upper troposphere where the temperature is low. Alkylperoxy nitrates are thermally rather unstable and their concentrations are insignificant. The most important organic peroxy nitrate is PAN, which derives from the association of NO_2 with acetylperoxy radicals, $\text{CH}_3\text{C}(\text{O})\text{OO}$. Acetaldehyde and acetone, the oxidation products of ethane and propane, are important precursors of acetylperoxy radicals. The concentrations of ethane and propane in marine air are higher in the northern hemisphere compared to the southern, and the same is true for PAN. However, still higher concentrations of PAN occur over the continents of the northern hemisphere, where additional sources of acetylperoxy radicals are active. In the low-temperature regime of the middle and upper troposphere PAN forms a reservoir containing a significant fraction of NO_x and both species occur at similar concentrations. PAN is not very reactive, so that it represents a chemically inert reservoir of NO_x . The thermal decomposition of PAN liberates NO_x , for example, when PAN-containing air is brought to warmer regions near the ground level.

IX. THE OXIDATION OF SULFUR COMPOUNDS

The oxidation of reduced sulfur compounds in the troposphere is initiated by OH radicals. Figure 8 indicates the reaction sequences in the case of dimethyl sulfide. The reaction occurs partly by addition of the OH radical to the sulfur atom, yet mainly by abstraction of a hydrogen atom from one of the methyl groups. Addition is followed by reaction with oxygen, which leads to the formation of dimethyl sulfoxide. This compound can undergo further addition of OH and reaction with oxygen to form dimethyl sulfone. Both substances have been detected in the marine atmosphere. Hydrogen abstraction is followed by the addition of oxygen. The resulting peroxy radical undergoes reaction with NO to produce formaldehyde and $\text{CH}_3\text{S}\cdot$. The methyl sulfide radical reacts readily with ozone and with NO_2 .



Both reactions produce $\text{CH}_3\text{SO}\cdot$. The addition of oxygen converts this radical to $\text{CH}_3\text{S}(\text{O})\text{OO}\cdot$. The subsequent reac-

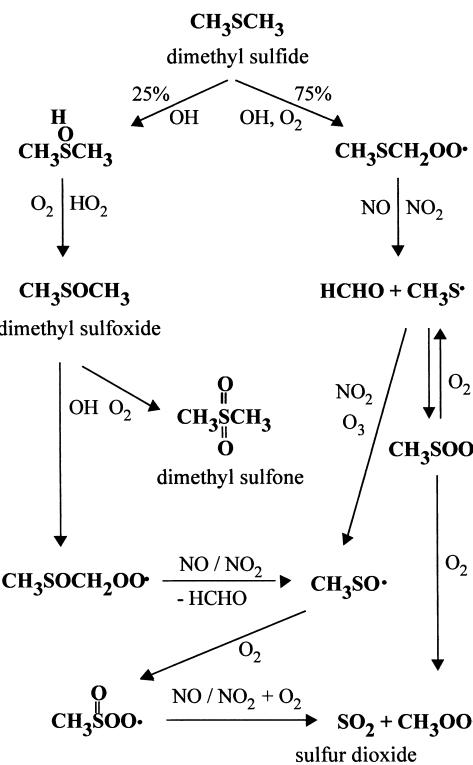
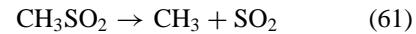
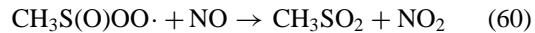
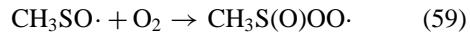


FIGURE 8 Oxidation mechanisms for dimethyl sulfide.

tion with NO yields CH_3SO_2 , which decomposes to form CH_3 and SO_2 .



The addition of oxygen to the $\text{CH}_3\text{S}\cdot$ radical leads to a reversible addition product, which reacts either with additional O_2 to yield $\text{CH}_3\text{OO}\cdot$ and SO_2 directly or with NO

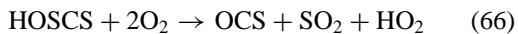


This $\text{CH}_3\text{SO}\cdot$ radical enters into the reaction sequence (59)–(61) to form CH_3 and SO_2 .

Another product resulting from DMS oxidation is methane sulfonic acid, $\text{CH}_3\text{SO}_3\text{H}$. Laboratory studies under near-atmospheric conditions have shown that the yield of SO_2 is 90% and that of $\text{CH}_3\text{SO}_3\text{H}$ about 10%. The pathway leading to $\text{CH}_3\text{SO}_3\text{H}$ is not yet established, however. The $\text{CH}_3\text{SO}\cdot$ radical appears to be an intermediate. Methane sulfonic acid is a component of the marine aerosol, and it also is present in marine cloud and rain waters. Carbonyl sulfide is another product resulting from

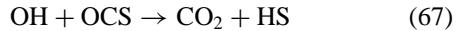
DMS oxidation in small yield, about 0.7%. The CH₃S radical appears to be an intermediate in OCS formation, but the real pathway is not established. Despite the small yield, the oxidation of DMS provides an important source of OCS in the troposphere, because emissions of DMS are the largest among all reduced sulfur compounds.

The oxidation of carbon disulfide by OH radicals proceeds via the formation of an addition complex



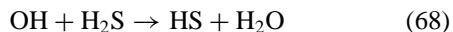
where M is a suitable third body molecule. The precise mechanism for the second step is still open, but it is well established that equal amounts of OCS and SO₂ are formed. The oxidation of CS₂ thus provides a source of OCS in the troposphere.

The reaction of OH radicals with carbonyl sulfide also proceeds via an addition complex, which rearranges to form carbon monoxide and the hydrogen sulfide radical

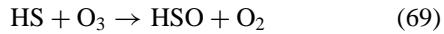


the subsequent oxidation of the HS radical will be outlined further below. The OH + OCS reaction is fairly slow, and it is a minor removal pathway in the tropospheric budget of OCS.

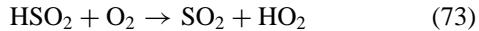
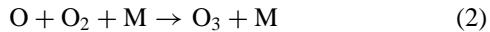
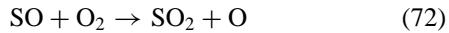
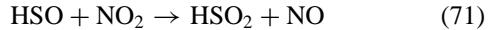
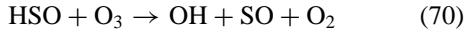
The oxidation of hydrogen sulfide by OH radicals occurs mainly by abstraction of an H-atom



The reaction of HS with oxygen is slow, whereas reactions with ozone and nitrogen dioxide are rapid. Both produce the HSO radical, for example



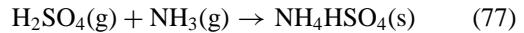
The HSO radical also does not react well with oxygen, but again, the reactions with ozone and nitrogen dioxide are rapid, leading to the following reaction sequence



with sulfur dioxide as the final oxidation product.

The above reaction mechanisms for reduced sulfur compounds make evident that sulfur dioxide is the principal oxidation product in all cases. The further oxidation of sulfur dioxide to sulfuric acid occurs in the gas phase as well as in the aqueous phase of clouds. Aqueous reactions are discussed in Section XI. The gas-phase oxidation

is induced by OH radicals and proceeds via an addition complex



where the last reaction involves gaseous reactants to produce solid ammonium bisulfate. The neutralization of sulfuric acid by ammonia may occur either in the gas phase or in the aqueous phase of aerosol particles, following condensation of sulfuric acid on the particles. In the presence of a sufficient supply of ammonia, NH₄HSO₄ reacts again with NH₃ to form ammonium sulfate, (NH₄)₂SO₄. Under conditions of low particle concentration, the newly formed solid material associates with water vapor to form molecular clusters that eventually grow to form new particles in a process called homogeneous nucleation. If the particle concentration is adequate, condensation onto existing particles takes preference.

X. AEROSOLS AND CLOUDS

In addition to trace gases the troposphere contains finely dispersed particles ranging in size from about 0.002 to 50 μm, the atmospheric aerosol. The highest number concentration of aerosol particles occurs in urban air (>10¹¹ particle m⁻³), low values are found in marine background air (3 × 10⁸ particle m⁻³), and still lower ones over the Poles. The mass concentration decreases also, from ~100 μg m⁻³ in urban air to 10 μg m⁻³ in marine air. The mass of a particle increases with its volume, so that the mass concentration of the aerosol is determined by the number of larger particles in the assembly, in the size range 0.1–50 μm. Both concentrations decrease as one goes from urban to remote regions, but mass concentration and number density are not linearly correlated because the median particle size changes and becomes larger in the remote troposphere. The aerosol concentration also decreases with altitude. In the free troposphere the decrease is roughly proportional to air density, in the boundary layer the decline is more pronounced.

Aerosol particles attract a certain amount of moisture, which rises with relative humidity. Clouds and fogs are formed when adiabatic or radiation cooling of the air raises the relative humidity (r.h.) beyond the moisture saturation point. The formation of water drops requires aerosol particles that can act as condensation nuclei. In the region of supersaturation, at r.h. >100%, the condensation of water on the particles encounters a size-dependent barrier, which

for accessible degrees of supersaturation (<0.4%) can be surmounted only by large particles (>0.1 μm). Once such particles are activated they can grow further to form cloud drops, whereas smaller particles remain present as an interstitial aerosol. When the cloud drops evaporate again, the particles are reconstituted. Most clouds dissipate, so that aerosol particles may undergo several cloud condensation cycles before they are finally incorporated in a raining cloud that removes them from the troposphere. In the middle latitudes, wet precipitation provides an efficient removal mechanism leading to an aerosol life time of 5–7 d for particles in the 0.1–10 μm size range.

A. Sources of Aerosol Particles

A large fraction of the natural aerosol originates from the action of wind forces on materials at the earth surface. On the continents, mineral dust particles are produced by the erosion of soils. The great deserts are major source regions for mineral dust on a global scale. Over the oceans, sea salt particles arise from sea spray, which is produced most copiously in breaking waves (whitecaps). Dried-up sea salt particles occur not only in marine air, but they can be carried far over the continents. Finally, wind forces are responsible for the occurrence of biogenic particles in the air. These include pollen and spores as well as fragments of plants such as cellulose, leaf waxes, and resins. Global production rates for mineral dust and sea salt are on the order of 2000 Tg yr^{-1} each; biogenic materials contribute about 100 Tg yr^{-1} .

Another fraction of the atmospheric aerosol derives from high temperature processes. Volcanoes produce mineral particles. Their impact is essentially local except when violent eruptions inject large amounts of material high into the atmosphere. High temperature anthropogenic processes are the combustion of fossil fuels for the purpose of electric power generation, which produces fly ash particles consisting of oxides of silicon, aluminum, and iron. The combustion of liquid fuels in automotive engines leads to emissions of soot and other organic particles. Biomass burning produces ash and partly burned char particles in addition to products resulting from the condensation of nonvolatile organic vapors liberated during the fire. The amount of particulate materials entering the troposphere from volcanoes is 20–90 Tg yr^{-1} ; direct emissions from anthropogenic processes add about 130 Tg yr^{-1} , which includes 5–25 Tg yr^{-1} of soot. Biomass burning contributes 50–140 Tg yr^{-1} .

Finally, a number of gas-phase chemical reactions, such as the oxidation of selected organic compounds, generate nonvolatile condensable products that associate with aerosol particles. A large fraction of the aerosol consists of ammonium sulfate, which derives from the oxidation

of sulfur dioxide to sulfuric acid and its reaction with ammonia. The global rate of aerosol sulfate production from biogenic sulfur gases is 60–110 Tg yr^{-1} , sulfate from volcanic sulfur dioxide adds about 25 Tg yr^{-1} , whereas sulfate from the oxidation of anthropogenic sulfur dioxide contributes 60–180 Tg yr^{-1} . The production of nitrates from the oxidation of nitrogen dioxide is 40–100 Tg yr^{-1} , with approximately equal contributions from anthropogenic and natural sources. The production rate of non-volatile organic matter that associates with the aerosol is highly uncertain. Estimates range from 40 to 200 Tg yr^{-1} .

B. Size Distributions

The wide range of particle sizes is an outstanding feature of the atmospheric aerosol. Although dry particles usually are nonspherical in shape, it is convenient to represent their size by an aerodynamically equivalent diameter or radius. The size distribution itself varies with time and location, as it depends on the rates of particle production and removal processes, yet the general features are always preserved. Figure 9 shows a set of typical size distributions for the rural aerosol as an example. The distribution shown on the left demonstrates that the number concentration is determined by very small particles in the size range below 0.1 μm (Aitken particles), formed either directly in combustion processes or arising from gas-to-particle conversion processes in the troposphere. Aitken particles have a high mobility and undergo sticky collisions among themselves, which causes coagulation and growth, and they

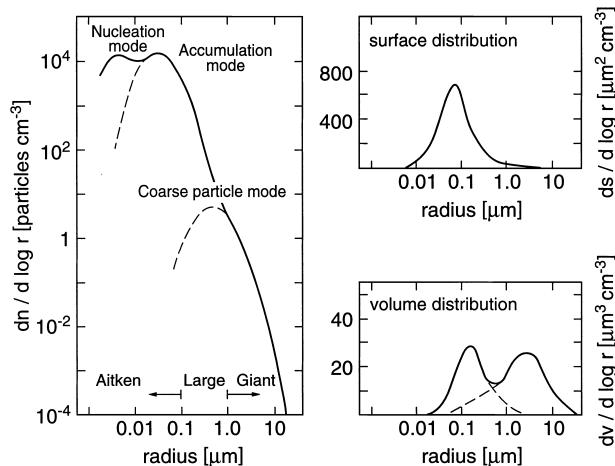


FIGURE 9 Left: distribution of particle number concentration versus aerodynamically equivalent radius for the rural continental aerosol. Right: The corresponding size distributions for the concentrations of particle surface and volume. The size distribution of mass concentration is nearly equivalent to that of volume. [From Zellner, R., ed. (1999). "Global Aspects of Atmospheric Chemistry," Steinkopff/Springer, Darmstadt, Germany.]

collide and combine with larger particles. In this manner, material is accumulated in the size range 0.01–1.0 μm . Because of their small size, Aitken particles do not contribute much to the mass concentration of the aerosol, which is determined by particles in the 0.1–20 μm size range. Particles larger than 20 μm in radius are rare. They are so heavy that they are rapidly removed by sedimentation.

As Fig. 9 shows, the size distribution of volume (and mass) is bimodal in shape, indicating that two different sources of materials contribute to total aerosol mass. From chemical analysis it is known that the coarse particle mode with radii above $\sim 0.5 \mu\text{m}$ arises largely from mineral dust over the continents, and from sea salt over the oceans. The other mode has its maximum near 0.1 μm . Here, ammonium sulfate is a prominent chemical component. The size range and shape of this mode is similar to that of the distribution of the surface of particles, suggesting that its origin is the condensation of compounds of low volatility and/or reactions of gas-phase species occurring at the surface of the aerosol particles. This size range, therefore, is called the accumulation mode. However, particles in the 0.1–0.5 μm size range also provide most of the cloud condensation nuclei, and nonvolatile products arising from chemical processes in clouds remain attached to the particles after evaporation of the cloud drops. A third maximum at radii near 0.01 μm often appears in the distribution of number concentration (see Fig. 9). Particles in this size range, called the nucleation mode, have sufficient inertia to escape collisions with larger particles for some time, and they remain present after the production of new particles ceases.

C. Chemical Composition of the Aerosol

Because a great variety of sources contribute materials to the tropospheric aerosol, it represents a complex mixture of many substances that additionally depends on the size of the particles. Source characteristics are preserved only in the vicinity of sources. The mixture may be divided into three fractions: water-soluble inorganic salts (electrolytes), water-insoluble minerals, and organic compounds, both soluble in water and insoluble. Table II shows the chemical composition of two boundary layer aerosols that are typical of marine and of rural continental air.

The marine aerosol lacks the crustal component except in regions affected by the advection of desert dust plumes. The electrolyte fraction is a mixture derived from sea salt, largely sodium chloride, and from gas-to-particle conversion processes, largely sulfates and nitrate. Sea-salt components reside in the coarse particle mode; products from gas-phase reactions occur primarily in the accumulation mode. Sea salt contains a certain amount of sulfate. Additional (excess) sulfate arises from the oxidation of SO_2

TABLE II Typical Mass Distributions of Chemical Constituents of Tropospheric Aerosols

Constituent	Marine	Continental ^a
Electrolytes [$\mu\text{g m}^{-3}$]	11.1	8.9
Anions [$\mu\text{g m}^{-3}$]	7.3	6.4
Components (excess) ^b [mass %]		
Cl^-	64 (−8)	2
SO_4^{2-}	35 (25)	77
NO_3^-	1 (1)	21
Cations [$\mu\text{g m}^{-3}$]	3.8	2.5
Components [mass %]		
NH_4^+	4 (4)	79
Na^+	78	3
K^+	3	6
Mg^{2+}	11	3
Ca^{2+}	4	9
Insoluble minerals [$\mu\text{g m}^{-3}$]	<0.1	2.9
Principal elements as oxides [mass %]		
SiO_2	—	43
Al_2O_3	—	29
Fe_2O_3	—	28
Organic compounds [$\mu\text{g m}^{-3}$]	0.8	4.1

^a Rural continental regions.

^b Excess compared with sea water is indicated in parentheses.

produced from dimethyl sulfide. Sulfur dioxide partly dissolves in sea spray and is oxidized by ozone. Another part undergoes gas-phase oxidation to sulfuric acid, which reacts with ammonia emanating from the sea surface and forms NH_4HSO_4 . Nitrate is formed by interaction of nitric acid with sea salt particles. Ammonium and nitrate are not components of sea salt. The acidification of sea salt particles following the addition of H_2SO_4 and HNO_3 displaces chlorine in the form of HCl and causes a chloride deficit.

Continental aerosol particles contain a significant fraction of minerals. The insoluble fraction consists mainly of the major crustal elements silicon, aluminum and trivalent iron, which occur as aluminosilicates, quartz, and iron oxides. Elements that are eluted from minerals by water are sodium, potassium, calcium (in part), and magnesium. The water-soluble inorganic salt fraction is dominated by ammonium sulfate. Again, sulfate arises from the oxidation of sulfur dioxide, both by gas-phase and by aqueous phase reactions. Whereas the mineral components are mainly found in the coarse particle size range, ammonium sulfate resides mainly in the accumulation mode. Nitrate occurs partly in association with ammonium in the accumulation mode, and partly together with sodium and other cations in the coarse particle mode. Thus, nitrate often shows a bimodal size distribution.

Organic compounds contribute appreciably to the total mass of tropospheric aerosols. The contribution to the continental aerosol is about 30%, and that to the marine aerosol is about $0.8 \mu\text{g m}^{-3}$ almost uniformly over the entire ocean. The number of compounds occurring in the organic fraction is large. A complete differentiation is made difficult by the fact that solvent extraction elutes less than 50% of the total organic material and that the larger part of the eluted fraction cannot be resolved by chromatographic techniques. Thus, more than 80% of organic material associated with the continental aerosol has so far remained unidentified. Morphological studies have shown that about 15% of continental aerosol particles have a biogenic origin independent of size, and this portion can account for much of the nonextractable fraction. Among the compounds identified by chromatographic analysis are *n*-alkanes ($\text{C}_{23}\text{--C}_{34}$), *n*-alkanoic acids ($\text{C}_9\text{--C}_{30}$), aliphatic dicarboxylic acids ($\text{C}_2\text{--C}_9$), aromatic polycarboxylic acids, diterpenoid acids, and polyaromatic hydrocarbons. Alkanes and *n*-alkanoic acids derive largely from the shedding of epicuticular leaf waxes by plants, with additional contributions from traffic and wood burning. An appreciable fraction of organic matter, 20–30%, is soluble in water. Among the compounds considered above, only the dicarboxylic acids are well soluble, but they account for less than 20% of water-soluble organic matter. Oxalic acid contributes about 50%. Dicarboxylic acids have been identified in gasoline and diesel engine exhaust. The larger part is thought to derive from the gas-phase oxidation of organic compounds. Dicarboxylic acids are ubiquitous in the troposphere and occur also in cloud and rain waters.

XI. CHEMICAL REACTIONS IN CLOUDS

The constituents of cloud water derive from two sources: one is material incorporated with the condensation nuclei, the other is the dissolution of gases from the surrounding air. As the numbers of particles serving as cloud condensation nuclei are most numerous in the size range of the accumulation mode, cloud water generally represents a dilute solution of this fraction of the aerosol. But the components of the aerosol fraction are already fully oxidized and, therefore, not very reactive. On the other hand, many of the gases that dissolve in cloud water have a potential for further oxidation. The aqueous concentration of such substances depends on their abundance in the gas phase before cloud condensation sets in and on the individual gas-liquid (Henry's law) partition coefficients, which causes a redistribution of the substances between the two phases. The amount of liquid water in clouds is in the range $0.1\text{--}0.5 \text{ g/m}^{-3}$, so that the volume of liquid

compared to that of the gas phase is less than 10^{-6} . Only highly soluble substances (e.g., ammonia, acids, hydrogen peroxide, or formaldehyde), enter the aqueous phase to such an extent that they are greatly depleted in the gas phase. Gas-phase concentrations of slightly soluble gases, such as CH_4 or CO , remain unaffected, and their concentrations in the aqueous phase are not significant. Oxygen, although only slightly soluble, is a major solute in the aqueous phase because of its high concentration in the air. A number of gas phase species interact with water to form ions via acid-base dissociation equilibria. They include CO_2 , SO_2 , NH_3 , HNO_3 , HCl , and HCOOH . Nitric acid and ammonia, which are almost completely transferred to the aqueous phase, form NO_3^- and NH_4^+ , respectively, thus adding to the store of these ions already supplied by aerosol particles. One consequence of acid-base ion equilibria is that the amount of material entering the aqueous phase is greater than that determined by simple Henry's law partitioning of the undissociated substance. Another consequence is that the interaction of the strong (fully dissociated) acids and bases determines the pH of the solution. The pH, in turn, influences the dissociation equilibria of weak (not fully dissociated) acids. Typically, the pH adjusts to values near pH 4. The exact value depends on the chemical composition of condensation nuclei and varies between individual cloud drops. While the complete transfer of material from the gas phase to the liquid phase may require up to 100 sec in individual cases, the process is generally completed and a new gas-liquid equilibrium is established before the cloud is fully developed. Ion equilibria are established very rapidly.

Clouds are multiphase chemical systems coupling the chemistry in the gas phase to that in the aqueous phase. Small clouds do not greatly attenuate the intensity of solar radiation and the gas-phase photochemistry continues, but reaction rates change because of the redistribution of gases in the presence of liquid water.

A. Basic Chemistry in Clouds

Cloud chemistry is simplest in the marine atmosphere, where the concentrations of many trace gases are small and can be neglected. The principal reactions in the gas phase were discussed in Sections IV and V. Hydrogen peroxide and formaldehyde dissolve largely in the aqueous phase and their gas-phase concentrations are reduced. Figure 10 shows the principal reagents in a cloud drop exchanging material with the surrounding gas phase. HO_x radicals, which enter mainly from the gas phase, assume an important role also in the aqueous phase. Photolysis of H_2O_2 , which is not a good OH source in the gas phase, becomes more important in the aqueous phase because it is concentrated there. In aqueous solution, HO_2 acts as an

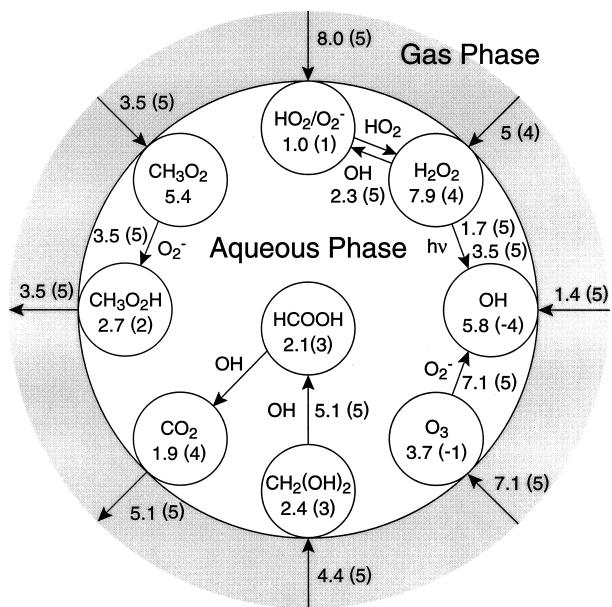


FIGURE 10 Aqueous phase concentrations (nmol dm^{-3}) of several species in a marine cloud (liquid water content 0.3 g/m^{-3}). Fluxes to and from the gas phase are given in molecule $\text{cm}^{-3} \text{ s}^{-1}$, fluxes in the aqueous phase are gas-phase equivalents in the same units. Orders of magnitude are shown in parentheses. [From Zellner, R., ed. (1999). "Global Aspects of Atmospheric Chemistry," Steinkopff/Springer, Darmstadt, Germany.]

acid with O_2^- as the conjugate anion. It reacts rapidly with dissolved ozone to form an ozonide anion, which reacts further with hydrogen ion to generate OH radicals

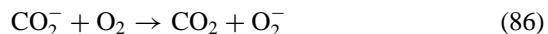


This mechanism is responsible for about 60% of OH radicals in the aqueous phase of marine clouds, photolysis of H_2O_2 contributes about 30%, and the intrusion of OH radicals from the gas phase contributes the remaining part. Formaldehyde dissolved in water occurs mostly in the form of a hydrate, $\text{CH}_2(\text{OH})_2$, which reacts with OH to produce formic acid

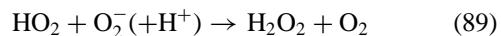


In contrast to HCHO, $\text{CH}_2(\text{OH})_2$ does not absorb ultraviolet light. Therefore, it is not photolysed and CO is not a product in the aqueous phase. Formic acid reacts fur-

ther with OH radicals and becomes oxidized to carbon dioxide



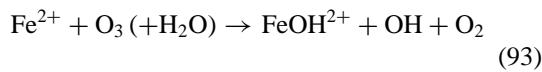
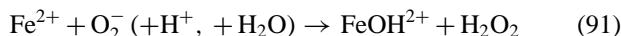
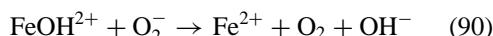
Most of the CO_2 is released to the gas phase. A portion of it forms the bicarbonate anion, HCO_3^- , which is already present in the aqueous phase at appreciable concentrations because of the dissolution of atmospheric CO_2 in cloud water. The HCO_3^- anion reacts with OH, but the reaction is slow and without significant consequence. The above reactions of OH radicals with formaldehyde and formic acid regenerate HO_2/O_2^- radicals, which continue the reaction until the chain is terminated when HO_2 radicals recombine to form H_2O_2



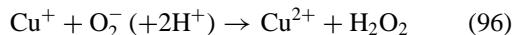
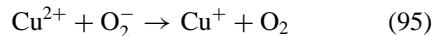
In sunlit clouds the concentrations and fluxes in the multiphase system come to a steady state within 30 sec. Figure 10 indicates aqueous concentrations and fluxes to and from the gas phase. The reaction mechanism leads to the destruction of ozone by reaction with O_2^- . Most of the formic acid produced in reactions (82, 83) is destroyed again by reaction (85) with OH radicals. Thus, the production of formic acid from the oxidation of formaldehyde is not significant compared to other, earth-bound sources of formic acid.

Continental clouds differ from marine clouds in several ways: The high mineral content of continental aerosol particles leads to the release of ions of trace metals to the aqueous phase that can interact with other solutes in the system; and the higher abundance of sulfur dioxide in continental air must be taken into account.

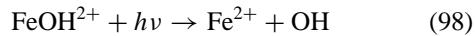
The metals iron, manganese, and copper are most frequently considered. As transition metals, they feature two oxidation states and represent one-electron redox species. Trivalent iron, Fe(III), is the most abundant metal. It occurs largely as an insoluble oxide, Fe_2O_3 , but it can be brought partly into solution either by the formation of complexes or by reduction to Fe(II). Copper and manganese, which appear as Cu(II) and Mn(II), are well soluble in water. Although their concentrations are smaller than that of iron, the dissolved fractions often are equivalent. In the region $\text{pH} > 2$, the Fe^{3+} ion forms a complex with the hydroxyl ion and occurs largely as FeOH^{2+} . The interaction of iron ions with HO_x radicals leads to the following reactions



The last reaction is not very effective, because OH radicals react preferentially with formaldehyde and formate anions. HO₂ reacts with FeOH²⁺ and Fe²⁺ in the same way as O₂⁻, although more slowly. Other transition metal ions undergo similar reactions. The reactions of HO₂/O₂⁻ with copper ions are the most rapid ones, so that they become more important than those of iron despite its higher concentration. Moreover, the Cu⁺ ion reduces Fe(III) to Fe(II)



The consumption of HO₂/O₂⁻ radicals in these reactions reduces their concentrations by two orders of magnitude. The reaction of O₂⁻ with ozone is so much reduced that it becomes fairly unimportant as a source of OH radicals, whereas reactions (92) and (93) of Fe²⁺ with H₂O₂ and O₃ are more significant. These reactions contribute about 50%, photolysis of H₂O₂ about 10%, and the intrusion of OH radicals from the gas phase about 30% to the total rate of OH formation. Photolysis of the iron hydroxo-complex also represents a source of OH radicals



but it is made ineffective as a source of OH because other reactions keep the concentration of FeOH²⁺ at a low value. At night, in the absence of sunlight, Cu(I) is quickly reoxidized to Cu(II), whereas Fe(II) persists for a much longer time. In sunlight an appreciable fraction of Fe(III) is converted to Fe(II), so that both oxidation states generally are observed to coexist. Finally, it should be noted that owing to the lower concentration of O₂⁻ the rate of ozone destruction is also greatly reduced.

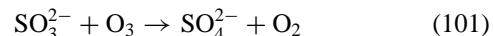
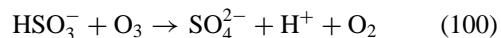
B. Oxidation of Sulfur Dioxide in Clouds

The principal anions formed by the dissolution of sulfur dioxide in cloud water are HSO₃⁻ and SO₃²⁻. Their relative abundance is determined by the pH-dependent ion equilibrium



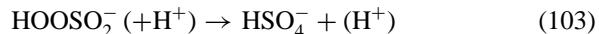
A great number of reaction pathways are known that oxidize these ions to sulfate, SO₄²⁻, but only three of them appear to be most important in clouds, namely, the reactions with ozone, with hydrogen peroxide, and with peroxy nitric acid.

The reactions with ozone are



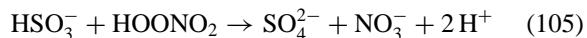
The second reaction is several orders of magnitude faster than the first, and it is preferred even at pH ~ 4 , where the concentration of SO₃²⁻ is only about 0.1% that of HSO₃⁻. Because of the pH dependence of the equilibrium (99), the concentration of SO₃²⁻ rises with increasing pH so that the overall S(IV) oxidation rate due to the reaction with ozone rises as well.

The reaction with H₂O₂ shows the opposite behavior in that the reaction rate decreases with increasing pH. The reaction of H₂O₂ with SO₃²⁻, which is important only in alkaline solution, is fairly slow. The reaction with HSO₃⁻ occurs via an acid catalyzed mechanism with peroxy sulfurous acid as intermediate



Acid catalysis occurs in the second step of the mechanism. This causes the reaction rate to increase proportionately with rising hydrogen ion concentration. The release of hydrogen ions in the third step acidifies the solution and promotes the reaction, but this effect is almost counterbalanced by a decrease in HSO₃⁻ concentration caused by the shift in the equilibrium $\text{HSO}_3^- + \text{H}^+ \rightleftharpoons \text{SO}_2 + \text{H}_2\text{O}$ forcing SO₂ to return to the gas phase.

Methyl hydroperoxide and H₂O₂ react with HSO₃⁻ at similar rates but the aqueous concentration of CH₃OOH is smaller due to the smaller Henry's law coefficient, and the reaction of CH₃OOH is much less important. More significant is the oxidation by peroxy nitric acid, which enters cloud drops from the gas phase where it is formed by the recombination of HO₂ with NO₂. The rate of the reaction



shows no pH dependence in the pH region 2–5. Above pH 5, HOONO₂ dissociates to form the NO₄⁻ anion, which is unstable and decomposes. Hydroxyl radicals react rapidly with HSO₃⁻ and SO₃²⁻ initiating a complex chemistry involving sulfuroxy radicals. These reactions, however, cannot compete with those of H₂O₂ and O₃,

because only a small fraction of OH radicals enters this pathway. Most OH radicals are scavenged by reactions with formaldehyde and formate anion. Metal ions that represent one-electron redox species can also initiate the oxidation of HSO_3^- and SO_3^{2-} . The reactions proceed via metal ion-sulfite complexes that decompose to form sulfuroxy radicals. Although these reactions are not known in every detail, it appears that they are not very effective under the conditions existing in clouds.

In continental clouds, hydrogen peroxide is the most important oxidant of sulfur dioxide dissolved in the aqueous phase, contributing about 80% to the total oxidation rate. Ozone and peroxy nitric acid oxidize up to 10% each, and the gas-phase reaction of SO_2 with OH radicals adds about 3%. Clouds are estimated to occupy about 15% of the airspace in the lower troposphere. In-cloud reactions thus oxidize 70–80% of SO_2 in the troposphere, the remaining 20–30% of SO_2 is oxidized in the gas-phase by reaction with OH radicals in cloud-free air.

XII. TROPOSPHERIC CHEMISTRY, INTEGRATION BY MODELS

Laboratory studies of reaction mechanisms, field measurements of atmospheric trace constituents, the assessment of surface source strengths, studies in general meteorology, all of these activities have provided essential elements in assembling our current knowledge and understanding of atmospheric chemistry. An overall description of tropospheric chemistry requires a synthesis of the various pieces of information. The foregoing discussion has emphasized chemical mechanisms of individual substances. An integral treatment must take into account not only the interaction of individual substances through mutual reactions, but also the transport of each substance in the troposphere. This integration is achieved by means of mathematical models.

The basic terms embodied in the differential equations of a model describe the transport properties of the troposphere, the rates of chemical reactions, and physical removal processes. Many models utilize an Eulerian description of the troposphere by subdividing the airspace into an assembly of boxes that exchange air and trace constituents with adjacent boxes in accordance with the prevailing tropospheric flow field. Surface sources of trace constituents are prescribed by appropriate boundary conditions. The equations are solved numerically on fast computers.

Depending on the purpose, models of varying degrees of complexity exist, ranging from simple box models that can deal with complex chemistry, but treat trans-

port lightly, to global transport models representing the entire tropospheric air space. Most desirable are three-dimensional transport and chemistry models that can simulate the spatial distribution and temporal behavior of chemical compounds in the troposphere. These comprehensive models are quite demanding in terms of computer power, but they are extremely useful in showing the outcome of the interaction between chemical transformations and transport on a regional scale as well as on a global scale.

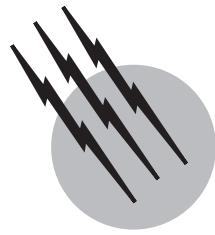
Models provide important diagnostic tools. The disagreement between results from model simulations and field observations generally indicates insufficient understanding of the system. Long-term measurements require much effort, and the simultaneous measurement of all trace components necessary to describe the system comprehensively in three-dimensional space is currently out of reach. Models, however, can promote understanding of the integrated system, even if supported by measurements at only a few strategically located observing stations. Much of our knowledge in tropospheric chemistry has been acquired in this manner.

SEE ALSO THE FOLLOWING ARTICLES

AEROSOLS • CARBON CYCLE • CLOUD PHYSICS • ENVIRONMENTAL GEOCHEMISTRY • GREENHOUSE WARMING RESEARCH • METEOROLOGY, DYNAMIC (TROPOSPHERE) • NITROGEN CYCLE, ATMOSPHERIC • OZONE MEASUREMENTS AND TRENDS (TROPOSPHERE) • RADIATION, ATMOSPHERIC

BIBLIOGRAPHY

- Brasseur, G. P., Orlando, J. J., and Tyndall, G. S., eds. (1999). *Atmospheric Chemistry and Global Change*, Oxford Univ. Press, New York.
- Calvert, J. G. (1990). "Glossary of Atmospheric Chemistry Terms," *Pure Appl. Chem.* **62**, 2167–2219.
- Finlayson-Pitts, B. J., and Pitts Jr., J. N. (2000). *Chemistry of the Upper and Lower Atmosphere*, Academic Press, San Diego.
- Seinfeld, J. H., and Pandis, S. N. (1998). *Atmospheric Chemistry and Physics, From Air Pollution to Climate Change*, Wiley & Sons, New York.
- Warneck, P. (1999). *Chemistry of the Natural Atmosphere*, 2nd ed., Academic Press, San Diego.
- Wayne, R. P. (1991). *Chemistry of Atmospheres*, 2nd ed., Clarendon Press, Oxford.
- Zellner, R., ed. (1999). *Global Aspects of Atmospheric Chemistry*, Topics in Physical Chemistry, Vol. 6, Deutsche Bunsen-Gesellschaft für Physikalische Chemie, Steinkopff, Darmstadt, Germany; Springer, New York.



Weather Prediction, Numerical

Akira Kasahara

*National Center for Atmospheric Research **

Masao Kanamitsu

National Centers for Environmental Prediction

- I. History
- II. Basic Equations of the Atmosphere
- III. Principle of Numerical Weather Prediction
- IV. Modeling of Physical Processes
in the Atmosphere
- V. Numerical Methods for Prediction Models
- VI. Observing Systems and Data Collections
- VII. Atmospheric Input Data for Prediction Models
- VIII. Applications of Numerical Prediction
- IX. Computational Issues
- X. Future Outlook

GLOSSARY

Coriolis force Apparent force acting on a moving parcel and resulting from the earth's rotation. The magnitude of the force is expressed by twice the product of the horizontal speed of parcel, the earth's angular velocity of rotation, and $\sin(\text{latitude})$. The force is directed at right angles to the movement, toward the right facing downstream in the Northern Hemisphere and upstream in the Southern Hemisphere.

Ensemble prediction To evaluate the reliability and probability of a prediction by making multiple runs with a prediction model starting from slightly different initial states.

* National Center for Atmospheric Research is sponsored by the National Science Foundation.

Finite-difference and spectral methods Two different numerical approaches to solve a prediction model by representing the model variables at discrete grid points and by collection of continuous functions, respectively.

Four-dimensional data assimilation Method of analyzing the state of the atmosphere or ocean by using a prediction model in such a way to best fit observed data during a period of 6–12 hours.

Parameterization Formulation of physical processes in terms of the variables of a prediction model as model constants or functional relations.

Predictability Ability of a perfect dynamical model to produce meaningful predictions; it is said to be lost if the separation of pairs of the solutions starting from slightly different initial states reaches the level of climatological variance.

Primitive equations Eulerian equations of motion under the effect of Coriolis force and modified by the

assumption of hydrostatic balance in the vertical equation of motion.

Satellite data Generic name for observed data of the atmosphere or the ocean surface conditions by indirect measurements with instruments aboard polar-orbiting and geostationary satellites.

Semi-Lagrangian method Another numerical time integration approach based on the Lagrangian method, which calculates the properties of fluid following the fluid parcels. To avoid the distortion of parcel trajectories, new fluid parcels are selected at regularly distributed grid points at each time step.

Singular vector and “breeding” methods Two different approaches to create a set of multiple initial states for ensemble prediction, which contain the most rapidly growing perturbation components.

THE MOTIONS of the atmosphere are governed by physical laws expressed as the equations of hydrodynamics and thermodynamics. These equations determine the time rate of changes of meteorological variables, such as wind, pressure, temperature, and water vapor and liquid water contents, which are basic elements of the weather. Various physical processes that control the weather are added to the atmospheric equations. Future evolution of the weather is predicted by integrating these atmospheric equations numerically with respect to time, starting from an initial state of the atmosphere. The present atmospheric state is analyzed objectively from the observations of meteorological elements collected worldwide. The time integration of the prediction equations and the objective analysis of initial conditions are performed numerically on high-speed electronic computers without human intervention. Forecast products are disseminated throughout the world via high-speed telecommunication lines.

I. HISTORY

The idea of weather prediction based on numerical time integration of the atmospheric equations was first suggested by the Norwegian scientist Vilhelm Bjerknes (1862–1951) at the beginning of the 20th century. He pointed out that the following conditions are necessary for a rational approach to objective weather prediction: (1) a sufficiently accurate knowledge of the state of the atmosphere at an initial time, and (2) a sufficiently accurate knowledge of the laws according to which one state of the atmosphere develops from another. Bjerknes was fully aware that these conditions were not met in those days. Meteorological observations are needed not only at the earth’s surface, but also throughout the atmosphere, the knowledge of which was meager in his time. Although the atmosphere is a physical

TABLE I Scales of Motion Associated with Typical Meteorological Phenomena

Classification of motions	Spatial scale (km)	Temporal scale (s)	Examples of weather phenomena
Large-scale	$\sim 10^4\text{--}10^3$	10^6	Planetary waves, cyclones, and anticyclones
Medium-scale	$\sim 10^3\text{--}10^2$	10^5	Hurricanes, fronts, and tropical disturbances
Mesoscale	$\sim 10^2\text{--}10$	10^4	Mesoscale convective systems, squall lines, and severe thunderstorms
Small-scale	$\sim 10\text{--}1$	10^3	Thundershowers, cloud bands, tornados, waterspouts, and downdrafts
Microscale	$\sim 1\text{--}10^{-2}$	$\sim 10^2\text{--}1$	Thermals, wind gusts, and dust devils

system, weather phenomena (see Table I) are complex, and it is difficult to express their processes quantitatively.

From 1913 to 1920, the British scientist Lewis Fry Richardson (1881–1953) formulated a mathematical method of weather prediction and even attempted to perform a time integration by hand using the meteorological data then available. Richardson, a Quaker and humanitarian, carried out this calculation while serving as an ambulance driver during World War I. It was no surprise that his “forecast” was unsuccessful. In retrospect, explaining why Richardson’s rational approach failed seems equivalent to describing the history of numerical weather prediction. Several major breakthroughs were necessary to produce today’s successful weather forecasts. The remarkable achievement of Richardson, however, was that he formulated once and for all the essential problems to be faced by future workers in this field and laid a thorough groundwork for their procedure.

For a long time after what Richardson called his “glaring error,” no one attempted to follow his rational approach of weather forecasting. The technology to perform the vast number of calculations needed to solve hydrodynamical equations was not in existence. Richardson wrote in 1922: “Perhaps some day in the dim future it will be possible to advance the computations faster than the weather advances and at a cost less than the savings to mankind due to the information gained. But that is a dream.”

In the mid-1940s, John von Neumann (1903–1957) of the Institute for Advanced Study in Princeton embarked upon research on the problem of numerical weather prediction as the most complex physical problem yet conceived and whose solution would require the fastest computing devices for many years to come. At the Institute, a meteorological research group was set up under Jule Gregory Charney (1917–1981) to attack the problem of numerical weather prediction through a step-by-step

investigation of models designed to approximate more closely the real properties of the atmosphere. In the spring of 1950, the first successful numerical prediction experiments were conducted on the electronic numerical integrator and calculator (ENIAC) at Aberdeen Proving Ground, Maryland. This was the dawn of the numerical weather prediction era.

One important factor in the Princeton Group's success in weather prediction seems to lie in distinguishing different spatial and temporal scales of motion governing weather systems and in appropriate simplification of the basic hydrodynamic equations by rational approximations to describe motions of a particular phenomenon.

Table I shows the spatial and temporal scales of motion associated with various meteorological phenomena. Large-scale motions, such as planetary waves, were discovered from upper-air observations during World War II. The work of Jacob Bjerknes (1897–1975), Carl-Gustav Rossby (1898–1957), and others led to the important discoveries that (1) planetary-scale waves are clearly related to weather systems, and (2) planetary-scale waves are predominantly horizontal and in quasi-geostrophic balance, as explained later.

The large-scale horizontal wind blows approximately parallel to the isobars on level surfaces, clockwise (counterclockwise) around a high (low)-pressure area in the Northern Hemisphere, and in the opposite direction in the Southern Hemisphere. The pressure force, directed at right angles to the isobars from high to low pressure, is approximately balanced by the Coriolis force, which is an apparent force resulting from the earth's rotation. This state of large-scale motions is referred to as *quasi-geostrophic balance*.

In 1955, then the National Meteorological Center, presently the National Centers for Environmental Prediction, Washington, D.C., began to issue numerical forecasts on operational basis by means of an electronic computer, the IBM 701. Activities on research and practice of operational numerical weather prediction soon spread like wildfire to many countries in Scandinavia, Europe, and Asia. However, because of the limited capability of electronic computers in those days, earlier prediction models had to adopt the quasi-geostrophic approximation and only limited physical processes of the atmosphere, which caused noticeable errors in their forecasts.

Today, a half-century later from the birth of numerical forecast, owing to the greatly increased speed and storage capacity of electronic computers and significant progress in numerical techniques for solving fluid dynamic equations and in understanding physical processes in the atmosphere, operational forecasting models no longer adopt the quasi-geostrophic approximations and, moreover, they include many more physical processes. Every day numerical forecasts are produced by operational centers in many nations in the world. One notable development in

the practice of operational weather prediction during the past half-century is the establishment of European Centre for Medium Range Weather Forecasts (ECMWF) in 1975 at Reading, UK, whose operations are supported by 19 member states of the European community. The collections of worldwide meteorological observations and the dissemination of numerical forecast products are coordinated by the World Meteorological Organization (WMO), Geneva, Switzerland, comprising 179 member states and six member territories, all of which maintain their own meteorological and hydrological services.

II. BASIC EQUATIONS OF THE ATMOSPHERE

We present the basic atmospheric equations in spherical coordinates of longitude, λ , latitude, ϕ , and radial distance r from the center of the earth. We assume that the coordinate system is rotating with the earth about an axis through the poles with a constant angular velocity Ω .

A. Equations of Motion and Mass Continuity

The governing laws of the state of a moving air parcel are expressed by the following equations of motion:

$$\frac{du}{dt} + \frac{uw}{r} - \frac{uv}{r} \tan \phi - fv + \hat{f}w = -\frac{1}{\rho r \cos \phi} \frac{\partial p}{\partial \lambda} - \frac{1}{r \cos \phi} \frac{\partial \Phi}{\partial \lambda} + F_\lambda \quad (1)$$

$$\frac{dv}{dt} + \frac{uw}{r} + \frac{u^2}{r} \tan \phi + fu = -\frac{1}{\rho r} \frac{\partial p}{\partial \phi} - \frac{1}{r} \frac{\partial \Phi}{\partial \phi} + F_\phi \quad (2)$$

$$\frac{dw}{dt} - \frac{u^2 + v^2}{r} - \hat{f}u = -\frac{1}{\rho} \frac{\partial p}{\partial r} - \frac{\partial \Phi}{\partial r} + F_r \quad (3)$$

where

$$f = 2\Omega \sin \phi \quad \text{and} \quad \hat{f} = 2\Omega \cos \phi$$

and the total derivative d/dt with respect to time t is

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \frac{u}{r \cos \phi} \frac{\partial}{\partial \lambda} + \frac{v}{r} \frac{\partial}{\partial \phi} + w \frac{\partial}{\partial r} \quad (4)$$

in which u , v , and w are the velocity components given by

$$\begin{aligned} u &= r \cos \phi \frac{d\lambda}{dt} \\ v &= r \frac{d\phi}{dt} \\ w &= \frac{dr}{dt} \end{aligned} \quad (5)$$

In Eqs. (1), (2), and (3), p is the pressure and ρ the density of air, and F_λ , F_ϕ , and F_z represent the frictional components (to be discussed in Section IV). The quantity Φ , called geopotential, is the sum of the earth's gravitational potential Φ^* and its centrifugal potential $-\frac{1}{2}(\Omega r \cos \phi)^2$. The terms $uw/r - (uv/r) \tan \phi$ in Eq. (1), $uw/r + (u^2/r) \tan \phi$ in Eq. (2), and $(u^2 + v^2)/r$ in Eq. (3) are apparent acceleration terms due to the curvature of the coordinate surfaces. Terms fv and $\hat{f}w$ in Eq. (1), fu in Eq. (2), and $\hat{f}u$ in Eq. (3) are also apparent acceleration (Coriolis) terms, but they result from the rotation of the coordinate system.

The equations of motion deal with the relationship between the acceleration of an air parcel and the pressure gradient force. The time rate of change of air density is determined by the conservation equation of air mass, written in the form

$$\frac{\partial \rho}{\partial t} + \frac{1}{r \cos \phi} \left[\frac{\partial(\rho u)}{\partial \lambda} + \frac{\partial}{\partial \phi}(\rho v \cos \phi) \right] + \frac{\partial(\rho r^2 w)}{r^2 \partial r} = 0 \quad (6)$$

B. Thermodynamic and State Equations

The equations of motion relate the dynamics of flow to the pressure and density fields. The equation of mass continuity determines the time rate of change in the density field in terms of kinematics of flow. We now need a relationship between the time rate of change of pressure and density; this is given by the first law of thermodynamics. The atmosphere is considered as an ideal gas, so that internal energy is a function of temperature only, and not density. We can also assume for an ideal gas that the specific heat at constant volume C_v is constant.

For an ideal gas or a mixture of ideal gases, the equation of state defines temperature T in relation to pressure p and density ρ ,

$$p = R\rho T \quad (7)$$

where R is the specific gas constant for the particular ideal gas under consideration and relates to C_v through $C_p = R + C_v$, which defines the specific heat at constant pressure C_p . Now, the first law of thermodynamics may be expressed by

$$C_p(dT/dt) - (1/\rho)(dp/dt) = Q \quad (8)$$

where Q denotes the rate of heating per unit mass per unit time.

C. Simplification in the Equations of Motion and Mass Continuity

The vertical extent of the major part of the atmosphere is approximately 100 km above the earth's surface. This

means that the thickness of the atmosphere is very small compared to the earth's radius of about 6370 km. Thus, the magnitude of the vertical velocity is expected to be much smaller than that of the horizontal velocity for large-scale motions. We can also assume that the surface of constant apparent gravitational potential Φ is approximated by a sphere, so that Φ depends only on the altitude z relative to the earth's radius a , which is treated as a constant. Incorporating these simplifications and other minor assumptions, the following equations of motion and mass continuity are used for atmospheric dynamics instead of Eqs. (1), (2), (3), and (6):

$$\frac{du}{dt} - \left(f + \frac{u \tan \phi}{a} \right) v = -\frac{1}{\rho a \cos \phi} \frac{\partial p}{\partial \lambda} + F_\lambda \quad (9)$$

$$\frac{dv}{dt} + \left(f + \frac{u \tan \phi}{a} \right) u = -\frac{1}{\rho a} \frac{\partial p}{\partial \phi} + F_\phi \quad (10)$$

$$\frac{dw}{dt} = -\frac{1}{\rho} \frac{\partial p}{\partial z} - g + F_z \quad (11)$$

$$\frac{\partial \rho}{\partial t} + \frac{1}{a \cos \phi} \left[\frac{\partial(\rho u)}{\partial \lambda} + \frac{\partial}{\partial \phi}(\rho v \cos \phi) \right] + \frac{\partial(\rho w)}{\partial z} = 0 \quad (12)$$

where

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \frac{u}{a \cos \phi} \frac{\partial}{\partial \lambda} + \frac{v}{a} \frac{\partial}{\partial \phi} + w \frac{\partial}{\partial z} \quad (13)$$

and

$$\begin{aligned} u &= a \cos \phi (d\lambda/dt) \\ v &= ad\phi/dt \\ w &= dz/dt \end{aligned} \quad (14)$$

In Eq. (11) g denotes the earth's gravitational acceleration, and we use $g = 9.8 \text{ m/s}^2$. The error committed by assuming the value of g as constant is about 3% at $z = 100 \text{ km}$.

III. PRINCIPLE OF NUMERICAL WEATHER PREDICTION

A. Time Extrapolation

Equations (7) through (12) constitute a mathematical model of the atmosphere. If frictional terms F_λ , F_ϕ , F_z in Eqs. (9), (10), and (11) and heating rate Q in Eq. (8) are expressed by functions of the velocity components, u , v , w , pressure p , density ρ , and temperature T , the atmospheric model becomes complete together with proper boundary conditions, which must be specified. Because Eq. (7) does not contain a time-dependent term, one of the

three variables, T , ρ , and p , can be determined diagnostically from the remaining two variables. For example, if we choose u , v , w , ρ , and p to be the prognostic variables, T can be determined as the diagnostic variable from p and ρ through Eq. (7). Prognostic and diagnostic variables are defined at the points of a four-dimensional lattice in space and time.

Let $\mathbf{W}(t) = (u_1, \dots, u_N, v_1, \dots, v_N, \text{etc.})$ represent a column vector consisting of N values each of prognostic variables, u , v , w , ρ , and p at time t . The number N denotes the number of grid points and is typically on the order of millions (see Section IX). Let $\mathbf{W}(t_0)$ represent the values of \mathbf{W} at $t = t_0$. Then the values of \mathbf{W} at $t_0 + \Delta t$, where Δt is a small time increment, may be extrapolated by

$$\mathbf{W}(t_0 + \Delta t) = \mathbf{W}(t_0) + (\partial \mathbf{W} / \partial t)_{t=t_0} \Delta t \quad (15)$$

where

$$t_0 \leq \tau \leq t_0 + \Delta t$$

The time derivative $\partial \mathbf{W} / \partial t$ can be obtained from discretized forms of the partial derivatives of the atmospheric model as described later. However, because we do not know the value of τ , we approximate $(\partial \mathbf{W} / \partial t)_{t=\tau}$ at $t = t_0$. In this case, the time extrapolation is *explicit*. If we approximate $(\partial \mathbf{W} / \partial t)_{t=\tau}$ at $t = t_0 + \Delta t$, the procedure of Eq. (15) is called *implicit* because the evaluation of $(\partial \mathbf{W} / \partial t)_{t=\tau}$ at $t = t_0 + \Delta t$ requires the value of \mathbf{W} at $t_0 + \Delta t$, which is still unknown at t_0 . Hence, an implicit scheme requires some kind of iteration or inversion of a matrix for the calculation of $\partial \mathbf{W} / \partial t$ and generally is time consuming to solve. In any case, repetition of Eq. (15) will yield a prediction of \mathbf{W} for any later time.

When the system of equations is solved by an explicit scheme, there is a constraint on the value of Δt . This constraint (Section V.A) states that Δt must satisfy the condition $C_m \Delta t / \Delta s < 1$, where Δs represents one of the space increments in the three-dimensional grid, and C_m denotes the fast characteristic speed of motions. In the atmospheric system, a representative fast speed is that of sound, approximately 300 m/s. If we choose a horizontal grid increment of 300 km for large-scale motions, Δt should be less than 1000 s. Sound waves in this system, however, propagate in the vertical direction as well as in the horizontal. Since we must choose a much smaller space increment in the vertical, say 3 km, then Δt must be less than 10 s. However, we know that large-scale weather systems change rather slowly and it appears unreasonable to use such a small time step for the time integration of large-scale weather patterns. In order to resolve this predicament, we need to understand the characteristics of global atmospheric motions.

B. Atmospheric Modeling

We can learn a great deal about the motions of the global atmosphere by examining the normal modes of the atmosphere. The atmosphere is a vibrating system and has natural modes of oscillations, like a musical instrument. Although the atmospheric equations are nonlinear, they can be linearized if we are interested in small-amplitude motions such as the perturbations around the atmosphere at rest with no external forcing and heating. Solutions of such a system with appropriate boundary conditions are referred to as *normal modes*.

A sound wave is manifested as one kind of the atmospheric normal modes, known as the acoustic mode, and is caused by the compressibility of air. There are two more kinds: One is called the gravity-inertia mode, which is caused by a combination of the restitutive force of gravity against thermally stable atmospheric stratification and the Coriolis force due to the earth's rotation. The other kind is called the rotational or planetary mode, which is caused by the meridional variation of the Coriolis force. The importance of the latter kind of normal mode as a prototype of upper tropospheric large-scale disturbances was clarified by C.-G. Rossby and his collaborators a little over one decade prior to the dawn of the numerical prediction era (see Section I). In retrospect, the very nature of this discovery was hidden in complicated calculations for the normal modes of the global atmospheric model. The mathematical analysis was initiated by the French mathematician Marquis de Laplace (1749–1827), and the complete solutions became clear only with the aid of electronic computers. It is remarkable that Rossby was able to capture the essence of this important type of wave motion, now referred to as the Rossby wave, from a simple hydrodynamic principle of the conservation of the absolute vorticity that is expressed by the sum of the vertical component of the relative vorticity and the planetary vorticity f .

The large-scale weather patterns are mostly governed by the behavior of Rossby waves whose phase speed is on the order of 10 m/s. In contrast, the phase speed of gravity-inertia waves can be much greater—as large as 300 m/s. These two kinds of motions propagate predominantly in the horizontal direction, while acoustic waves propagate in the vertical direction as well. That is why the integration time step Δt is constrained by the vertical propagation of acoustic waves. Therefore, we will explore two alternatives: One is to use the implicit time integration scheme (Section V) in the vertical to relax the computational constraint in the vertical. The other is to modify the atmospheric equations so that the vertical propagation of sound waves can be eliminated altogether.

For large-scale motions of the atmosphere, the horizontal extent of motion is much larger than the vertical, and

we can show by scale analysis that the vertical acceleration dw/dt and the frictional term F_z may be neglected in Eq. (11). This approximation results in the following important relationship, called the *hydrostatic equilibrium*:

$$\frac{\partial p}{\partial z} = -\rho g \quad (16)$$

The system of equations of horizontal motion [Eqs. (9) and (10)], hydrostatic equilibrium [Eq. (16)], mass continuity [Eq. (12)], thermodynamics [Eq. (8)], and the ideal gas law [Eq. (7)] is called the hydrostatic prediction model, or *primitive equations*. The hydrostatic assumption modifies the basic atmospheric prediction system in such a way as to eliminate the vertical propagation of sound waves.

Another significant aspect of the primitive equation model is that vertical velocity w is no longer a prognostic variable. We must calculate w diagnostically from the condition that the calculations of $\partial\rho/\partial t$ from Eq. (12) and $\partial p/\partial t$ from Eq. (8) must always satisfy the hydrostatic condition of Eq. (16). This was first pointed out by L. F. Richardson in 1922.

The hydrostatic equation (16) permits a unique relationship between the height of an isobaric surface z and the pressure p . This relationship can be used to transform the primitive equation system adopting p as the vertical coordinate and the temperature T as one of the prognostic variables instead of p . Thus, in the transformed system, the coordinate variables are (λ, ϕ, p, t) , the prognostic variables are (u, v, T) , and the diagnostic variables are the isobaric height z , the density ρ , and the vertical pressure velocity $\omega (=dp/dt)$ replacing the vertical velocity w . Various forms of the primitive equations are extensively adopted at many operational forecasting centers in the world for medium-range weather predictions (Section VIII).

IV. MODELING OF PHYSICAL PROCESSES IN THE ATMOSPHERE

A. Role of Atmospheric Physics

The prediction model is not complete without specific forms of heating Q in the thermodynamic equation [Eq. (8)] and frictional terms F_λ and F_ϕ in the horizontal equations of motion [Eqs. (9) and (10)] in the primitive equation system. Weather changes occur as a result of heating and cooling in the earth's environment, i.e., atmosphere, biosphere, hydrosphere, and cryosphere, and dynamical effects of the earth's orography. The accurate formulation of atmospheric physical processes become increasingly important as the forecast period extends.

The time rate of heating Q may be broken up into

$$Q = Q_a + Q_b + Q_c \quad (17)$$

where Q_a is the heating due to solar and atmospheric radiation, Q_b the vertical and horizontal diffusion of sensible heat, and Q_c the release of latent heat of condensation of water vapor.

Similarly, the frictional term \mathbf{F} given by the vector form of (F_λ, F_ϕ) may be expressed

$$\mathbf{F} = \mathbf{F}_L + \mathbf{F}_s \quad (18)$$

as the sum of the time rate of change of momentum by large-scale friction, \mathbf{F}_L , and by momentum transport due to the scales of motion smaller than the computational grid increment, \mathbf{F}_s .

Specification of heating and frictional terms in terms of dependent variables of prediction model requires detailed knowledge of the physical processes involved. These details of physical processes are studied by specialized branches of atmospheric science, such as atmospheric radiation, atmospheric chemistry, micrometeorology, hydrology, ecology, and cloud physics. The formulation of physical processes in terms of the model variables as parameters (constants or functional relations) is called *parameterization*.

B. Solar and Terrestrial Radiation

All energy of atmospheric and oceanic motions is ultimately derived from incoming solar radiation, called *insolation*. The insolation at the top of the atmosphere is approximately 1360 W m^{-2} at normal incidence, known as the *solar constant*. A mean flux of solar energy perpendicular to the earth's surface is about 340 W m^{-2} , because the surface area is four times the cross-sectional area of the sphere. Of this, approximately 30% is reflected from the atmosphere, including clouds and portions of the earth's surface. The percentage of radiative energy reflected back to space is referred to as the *planetary albedo* of the earth. Approximately 25% is absorbed in the atmosphere and 45% at the earth's surface. Because the mean temperature of the earth's system does not change appreciably from one year to the next, the radiative energy received by the earth must be sent back to space. This ultimate return of approximately 70% of radiative energy is in the form of low-temperature radiation from the earth-atmosphere system, that is, *terrestrial infrared radiation*.

At low latitudes, the earth-atmosphere system gains more heat energy per unit area by the absorption of insolation than it loses in space by the emission of infrared radiation; the reverse occurs at higher latitudes. Surplus heat energy in the tropics must somehow be carried to the poles. Otherwise, tropical regions would become steadily hotter and polar regions steadily colder. To equilibrate this

differential heating, the motions of the atmosphere and oceans occur.

The heating/cooling term Q_a in Eq. (17) can be divided into two parts: $Q_a = Q_{as} + Q_{ae}$, where Q_{as} is the heating due to absorption of insolation in the atmosphere and Q_{ae} the heating and cooling due to infrared radiation emitted from the atmosphere.

The calculations of Q_{as} are based on radiative transfer theory and depend on the known vertical distribution of absorbing or emitting gases, such as water vapor (H_2O), carbon dioxide (CO_2), oxygen (O_2), and ozone (O_3). We consider not only absorption by major absorbing gases, but also the effects of scattering by atmospheric molecules and aerosols and reflection by clouds. The insolation depends on the solar constant and the sun's zenith angle. The reflection by clouds is geometrically calculated by taking into account cloudiness as a function of height.

Similarly, the calculations of Q_{ae} are carried out based on radiative transfer theory and the known absorptivity of atmospheric constituents and the vertical distribution of temperature. The radiatively active gases are H_2O , CO_2 , O_3 , methane (CH_4), nitrous oxide (N_2O), carbon monoxide (CO), and chlorofluorocarbons (CFCs). Since both upward and downward fluxes are involved in the infrared radiation calculation, the handling of cloud effects in determining Q_{ae} becomes more complicated than the calculation of Q_{as} .

Although CO_2 is well mixed in the atmosphere, most of the radiatively active atmospheric constituents are highly variable in their spatial distributions. Moreover, they are affected by chemical processes, although the production of CO_2 is mainly influenced by biological processes and the distribution of H_2O is controlled by thermodynamic processes rather than atmospheric chemistry. Well-known tracer compounds that are radiatively active and controlled by chemical processes include O_3 , CH_4 , N_2O , and CFCs. Knowledge of atmospheric chemistry and modeling of chemical processes are necessary to complete the parameterization package for radiative calculations.

The role of atmospheric chemistry also comes in determining the distribution of aerosols affecting atmospheric radiation. Aerosols are composed of either solid or liquid particles suspended in the air. Some of them, such as sea salt, soil dust, and smoke or soot, are injected into the atmosphere by natural and anthropogenic processes. Others are composed primarily of sulfate, nitrate, and organic compounds. Sulfate aerosol is generated by the sequence of chemical processes involving sulfur dioxide (SO_2), emitted into the atmosphere naturally or anthropogenically, and water vapor through photochemistry. Nitrates follow a chemical sequence similar to that for sulfates. However, the physics and chemistry of aerosol formation

are complex. And the precise distribution of aerosols is often very difficult to obtain, because they are subjected to transport and removal by the atmospheric motions, as well as to atmospheric chemistry.

C. Role of Water Vapor and Clouds

Although the percentage of water vapor in the atmosphere is small, only 3 g per 1000 g of the air on average in the troposphere and much less in the stratosphere, the presence of water in three different forms—vapor, liquid, and solid—has a profound effect on the motion of the atmosphere. Prediction of clouds and precipitation is an important aspect of weather forecasting. Unfortunately, the problem is also difficult because the time and space scales involved are much smaller than those of large-scale motions.

The mixing ratio of water vapor q is defined by $q = \rho_w / \rho$, where ρ_w is the density of water vapor and ρ is the density of dry air. When the mixing ratio exceeds the saturation mixing ratio q_s , the excess water vapor over q_s ordinarily gives rise to condensation in the form of rain or to sublimation in the form of snow. Latent heat of condensation or sublimation is liberated in the phase transition and becomes available to heat the air. The opposite of condensation is the evaporation that acts to cool the air. The rate of heating or cooling from this process is designated by Q_c in Eq. (17).

When considering the motion of moist air that contains water vapor, the dynamical system described by Eqs. (7) to (12), which are suitable only for the motion of dry air, must be modified accordingly. For example, the density of water vapor ρ_w must be added to the density of dry air ρ in the calculation of pressure p by Eq. (16).

The first step in predicting cloud formation and precipitation is to predict the large-scale moisture field. This is done via the continuity equation of water vapor for ρ_w :

$$\partial(\rho q)/\partial t + \nabla \cdot (\rho q \mathbf{V}) + \partial(\rho q w)/\partial z = M + \rho E \quad (19)$$

where ∇ and \mathbf{V} denote the horizontal gradient operator and the horizontal velocity, respectively. This is similar to the mass continuity equation [Eq. (12)], except that Eq. (19) includes the source and sink terms of water vapor due to the process of condensation M and the vertical and horizontal diffusion of water vapor ρE .

The prediction of clouds in terms of type, height, and amount (cloudiness) based on the large-scale flow variables, such as temperature, vertical velocity, and water vapor, is the objective of cloud parameterization. As a by-product of cloud parameterization, the precipitation rate in the form of rain or snow is determined. The prediction of layered clouds, such as stratus, is important in connection

with solar and infrared radiative calculations. The latent heat released in the ensembles of deep cumulus convection in the tropics provides the major heat source for the atmospheric general circulation. Yet, the parameterization of cumulus convection is more difficult than that of layer clouds. Cumulus convection also transports momentum as well as heat and moisture, playing the role of vertical mixing. Understanding the roles of deep cumulus convection as a heat source and mixing of momentum and moisture is a key factor in numerical weather prediction and is a topic of research in its own right.

D. Atmospheric Boundary Layer

The region from the earth's surface to an altitude approximately 1000 m above is the atmospheric boundary layer. Not only does kinetic energy dissipate by friction here, but the layer normally acts as an energy source transporting sensible heat and water vapor (latent heat) from the earth's surface to the interior of the atmosphere.

The region above the atmospheric boundary layer is the free atmosphere. Here, frictional effects are generally negligible except for clear air turbulence caused by shearing instability near atmospheric fronts, cumulus convection, and upward-propagating gravity waves. These are subgrid-scale processes (see Section IV.G).

The atmospheric boundary layer may be divided into two horizontal layers. The lowest layer, extending not more than 100 m above the surface, is the *surface boundary layer* (or the Prandtl layer). Here the vertical fluxes of momentum, heat, and water vapor may be assumed independent of height, and the atmospheric structure is primarily determined by characteristics of the earth's surface, thermal stratification, and the variation of wind with height.

The domain between the surface boundary layer and the free atmosphere is the *planetary boundary layer*, or the Ekman layer. Lately, it has been referred to more often as the *mixed layer*. The top of the mixed layer is often well delineated by a solid or broken cloud cover having a stable free atmosphere above, and below which turbulent motions are confined.

The large-scale frictional force \mathbf{F}_L in Eq. (18) may be expressed by

$$\mathbf{F}_L = \rho^{-1}(\partial\tau/\partial z) + \mathbf{F}_H \quad (20)$$

where \mathbf{F}_H is the horizontal component of \mathbf{F}_L and τ , the Reynolds stress, which is the vertical flux of horizontal momentum transported by small-scale eddy motions.

The heating rate due to the vertical and horizontal diffusion of sensible heat Q_b in Eq. (17) may be expressed by

$$Q_b = \rho^{-1}(\partial h/\partial z) + Q_H \quad (21)$$

where Q_H is the horizontal component of Q_b and h the vertical flux of sensible heat transported by small-scale eddy motions.

The time rate of change in water vapor content due to eddy diffusion E in Eq. (19) may be given by

$$E = -\rho^{-1}(\partial r/\partial z) + E_H \quad (22)$$

where E_H represents the horizontal component of E and r denotes the vertical flux of water vapor transported by small-scale eddy motions.

To complete the formulation of the atmospheric boundary layer, we must specify the vertical fluxes of momentum τ , sensible heat h , and water vapor r in terms of large-scale flow variables. At the earth's surface, including sea surface, the vertical fluxes τ_s , h_s , and r_s are expressed by some formulas based on surface boundary layer theory. Here, r_s is known by a familiar name of evaporation rate. Likewise, we must assume the forms of horizontal components \mathbf{F}_H , Q_H , and E_H (Section IV.G).

E. Soil Moisture, Ground Hydrology, and Vegetation

The exchanges of sensible and latent heat between the atmosphere and the earth's surface by small-scale turbulence and convection are important heat sources for driving atmospheric motions. Since the specific heat of soil is much less than that of water, the temperature of the ground rises or falls more rapidly than that of an equivalent amount of water. Thus, diurnal and seasonal variations of the surface temperatures of land are, in general, larger than those of lakes and oceans. The surface temperature of land is calculated using a surface energy balance relation.

The characteristics of land are complex because of the content of water in the soil and the presence of rivers and groundwater aquifers. The topography and various types of vegetation add further complexity. The amount of latent and sensible heat transported from the land depends mainly on the wetness of the soil. In deserts, characterized by lack of soil moisture, heat is transported mostly in a sensible form, whereas over wetlands, heat is transported chiefly in a latent form through evaporation. The latent heat transported from land surfaces is affected by plant foliage through transpiration, the process of water changing from liquid to vapor states by plant metabolism. Thus, vegetation processes allow a direct interaction of deep layer soil water with the atmosphere.

Precipitation and air temperature are major factors controlling the content of soil moisture. If precipitation is in the form of rain, water percolates into the soil until the soil is saturated and the excess drains as runoff. If in the form of snow, it may accumulate over the land, with the snow cover changing the surface albedo. Precipitation and the

surface air temperature also influence the development of vegetation which, in turn, changes the surface albedo and the amount of evaporation and sensible heat transport at the earth's surface.

The physical processes of atmospheric and ground hydrology and vegetation are intimately related to those of the atmospheric boundary layer (Section IV.D). Accurate modeling of these land surface processes is crucial to produce reliable forecasts not only for short-term, but also for long-term weather changes. For example, there is a positive feedback between snow cover and prolonging the period of cold weather, since snow on the land contributes to increase the surface albedo, which in turn enhances atmospheric cooling.

F. Oceans and Sea Ice

The oceans influence the atmosphere more than does the land because they cover approximately 70% of the earth's surface and have a large heat capacity. Just as with land, the most important variable characterizing the air-sea interaction is sea surface temperature (SST). To determine SST, the same principle of surface energy balance holds for the oceans except that we now deal with the hydrosphere instead of the lithosphere. Knowledge of ocean circulation is needed to evaluate heat transport to the deep ocean and the net gain or loss of heat due to transport by ocean currents. For medium-range prediction up to about 10 days, it is sufficient to specify a fixed initial SST during the forecast period because the temporal variation of SST is ordinarily much slower than the time variations of atmospheric motions. However, for extended-range prediction of 1 month or more, interactions between the atmosphere and the oceans must be taken into account. This means that the prediction model includes both atmospheric and ocean components as a coupled system (see Section VIII.C).

For weather prediction in the polar areas, it is important to take into account the effects of sea ice. Approximately 2% of the total water on the earth is stored in the form of ice in polar areas and glaciers. Sea ice accounts for nearly two-thirds of the earth's ice cover in areal extent. Sea ice plays the major role of controlling the exchange of heat, water vapor, and momentum between sea and air in the polar regions. Ice cuts off heat and water vapor transport from the ocean to the atmosphere and increases the albedo. Thus, similar to snow cover over land, sea ice contributes to cooling over the ice surface, which, in turn, tends to thicken the ice—a positive climatic feedback.

In advanced treatment of predicting sea ice, it may be necessary to calculate the compactness of sea ice as well as the thickness of ice, since the albedo of sea ice and the heat and moisture transports from the sea are depen-

dent on the population of floes. The problem is analogous to predicting the population of convective-type clouds in the atmosphere. Difficulty in developing adequate sea ice parameterization is often compounded by the lack of observational data on sea ice as well as a limited knowledge of ice dynamics. However, accelerated progress has been made in sea ice modeling research owing to the advent of high-resolution observations of sea ice from polar-orbiting satellites.

G. Subgrid-Scale Processes

Any numerical model of the atmosphere must use a finite resolution in representing the continuum. Certain physical and dynamical phenomena, such as thunderstorms, cumulus convection, and gravity waves whose scales are smaller than the computational grid, are not explicitly represented in large-scale global prediction models. Such unresolvable phenomena are referred to as *subgrid-scale processes*.

Contribution of subgrid-scale motions to larger scales may be measured by the amount of energy contained in the subgrid-scale motions. Longitudinal spectra of kinetic energy in the atmosphere show that kinetic energy of a scale of motion smaller than about 3000 km varies as L^3 , where L denotes a horizontal scale of motion. Hence, the kinetic energy contained in subgrid-scale motions is fortunately small. However, this does not imply that the influence of subgrid-scale motions is negligible. Because of nonlinearity of fluid motions, kinetic energy of the small scales will cascade up toward larger scales. Small-scale errors will eventually grow within a matter of about 10 days to the magnitude of large-scale energy variance (see Section VIII.A).

If we allow an extremely fine resolution in a numerical model, the ultimate energy dissipation takes place by molecular viscosity. In practice, the smallest grid size is far larger than desired in an ideal situation. The objective of subgrid-scale parameterization is to design the physical formulation of energy sink, withdrawing the equivalent amount of energy comparable to cascading energy down at the grid scale in an ideal situation. The so-called eddy viscosity designed to simulate this energy-cascading process is based on two- and three-dimensional turbulence theory. Eddy viscosity formulation has been used to simulate the subgrid-scale dissipation terms \mathbf{F}_s in Eq. (18) and \mathbf{F}_H in Eq. (20), the horizontal diffusion of sensible heat Q_H in Eq. (21), and the horizontal diffusion of water vapor E_H in Eq. (22).

The task of subgrid-scale parameterization differs greatly depending on how realistically one wishes to forecast the atmospheric motions of various scales (see Table I). For a low horizontal resolution model (grid increment of about 300 to 500 km), the use of a simple

eddy viscosity formulation may be a practical solution. However, for a high horizontal resolution model (grid increment of 100 km or less), more sophisticated approaches are required to simulate the relevant physics and dynamics involved in unresolvable but significant small-scale motions. A notable example is the necessity to include the momentum dissipation due to gravity wave drag in high-resolution models. Vertically propagating gravity waves can be excited in the atmosphere where stably stratified air flows over an irregular lower boundary and by internal heating and wind shear. These waves are capable of transporting significant amounts of momentum from their source regions to high-altitude regions where they are absorbed or dissipated. The parameterization of this mechanism of momentum sinks is referred to as gravity wave drag in atmospheric modeling.

In summary, reliable physical parameterizations are just as important as accurate approximations of the dynamical equations in building a successful numerical prediction model. Since atmospheric phenomena are complex, it is important for modelers to decide which atmospheric processes should be included in designing a prediction model. The importance of a particular physical process can be judged by performing sensitivity experiments with and without its impact, once a quantitative formulation in terms of model variables is made.

V. NUMERICAL METHODS FOR PREDICTION MODELS

A. Time Differencing Schemes

The evolution of flow patterns is determined by integrating prediction equations with respect to time, starting from initial conditions. Because our differential equations are nonlinear, we must resort to solving them numerically. If the space derivatives of prediction equations are approximated by finite differences (Section V.B) or spectral representation (Section V.C), the result becomes a system of ordinary differential equations, involving time t as the only independent variable, which may be expressed as

$$du/dt = F(u, t) \quad (23)$$

where the dependent variable $u = u(t)$ is a vector. An initial condition

$$u(0) = u_0 \quad (24)$$

is required to complete the system.

Let us consider an oscillatory motion given by

$$du/dt = i\omega u \quad (25)$$

where ω denotes the frequency (constant). With the initial condition of Eq. (24), the solution of Eq. (25) is given by

$$u(t) = u_0 \exp(i\omega t) \quad (26)$$

We now replace the time derivative of Eq. (25) with a finite difference quotient to derive a difference equation. We define grid points $t_n = n\Delta t$ with time step Δt and $n = 0, 1, 2, \dots$. We write U_n to approximate $u(t_n)$.

By Taylor expansion, we obtain

$$\begin{aligned} U_{n+1} &= U_n + u_t \Delta t + \frac{1}{2} u_{tt} (\Delta t)^2 + \dots \\ &= \left[1 + i\omega \Delta t - \frac{1}{2} (\omega \Delta t)^2 + \dots \right] U_n \\ &= U_n \exp(i\omega \Delta t) \end{aligned} \quad (27)$$

where u_t and u_{tt} are evaluated at $t = n\Delta t$. Equation (27) can be written compactly as

$$U_{n+1} = CU_n \quad (28)$$

where C is called the amplification factor in the case of a scalar variable. Repeating Eq. (28) from the initial condition of Eq. (24) yields

$$U_n = C^n U_0 \quad (29)$$

To have a meaningful solution, we must have $|C| \leq 1$. This implies that time step Δt must satisfy a certain constraint, called the *computational stability condition*.

Let us approximate u_{n+1} by the first two terms in Eq. (27). This is Euler's scheme and is unconditionally unstable because $C = 1 + i\omega \Delta t$ and therefore $|C| > 1$ for $\omega \neq 0$. Note that the scheme is conditionally stable if it is applied to an exponentially damped problem (i.e., $\omega = iv$), where v is real and positive. In this case, the stability condition is $v\Delta t \leq 2$.

Table II shows some typical time difference schemes, used to solve Eq. (23) and corresponding amplification factors C applied to Eq. (25). We also indicate whether the scheme is explicit or implicit (see Section III.A), whether it is first- or second-order in accuracy, and whether it is computationally stable. The Crank–Nicholson scheme is accurate and stable, but because of its implicit nature, its application can be time consuming. The leapfrog scheme is most common and stable provided that $|\omega|\Delta t < 1$, but because three time levels are involved, it produces two independent solutions. In the limit $\Delta t \rightarrow 0$, $C_1^n \rightarrow e^{i\omega t}$, and $C_2^n \rightarrow (-1)^n e^{-i\omega t}$, so that the second solution is a computational mode, while the first is the physical mode. The computational mode is a false solution and must be controlled to prevent it from overwhelming the physical mode.

An implicit time integration scheme such as the Crank–Nicholson scheme, which is free from the computational constraint, has an important application to atmospheric

TABLE II Various Time Difference Schemes Applied to Eq. (23) and Their Amplification Factors Applied to Eq. (25)

Euler (explicit, first-order, unstable):
$U_{n+1} = U_n + F(U_n) \Delta t$
$C = 1 + i\omega \Delta t$
Backward (implicit, first-order, stable):
$U_{n+1} = U_n + F(U_{n+1}) \Delta t$
$C = (1 - i\omega \Delta t)^{-1}$
Crank–Nicholson (implicit, second-order, stable):
$U_{n+1} = U_n + \frac{1}{2}[F(U_n) + F(U_{n+1})]\Delta t$
$C = \left(1 + \frac{1}{2}i\omega \Delta t\right)\left(1 - \frac{1}{2}i\omega \Delta t\right)^{-1}$
Leapfrog (explicit, second-order, conditionally stable):
$U_{n+1} = U_{n-1} + 2F(U_n)\Delta t$
$C_{1,2} = i\omega \Delta t \pm [1 - (\omega \Delta t)^2]^{1/2}$

modeling. As pointed out in Section III.B, the frequency of acoustic waves in the atmosphere is considerably larger than that of Rossby waves, which are meteorologically dominant motions, but acoustic waves are essentially noise in the atmosphere, containing little energy. Yet, when an explicit time differencing scheme is used, an extremely small time step must be adopted to account for the computational stability condition. Thus, it is attractive to employ implicit time differencing schemes to stably handle the propagation of acoustic waves, while using explicit time differencing schemes to handle the propagation of Rossby waves. Such a targeted use of implicit time differencing scheme in the otherwise explicit integration formulation of a prediction model is referred to as the *semiimplicit method*. Since gravity-inertia waves also carry relatively small energy compared with Rossby waves, the semiimplicit approach is also applicable to handle the propagation of gravity-inertia waves. However, the use of a larger time step than that allowed by the computational stability condition reduces the accuracy of calculations. Therefore, it is not appropriate to use the implicit integration scheme in the modeling of weather phenomena for which accurate forecasts are desired.

B. Finite-Difference Method

We now consider a system of partial differential equations:

$$w_t + Aw_x = 0 \quad (30)$$

where $w(x, t)$ is a vector whose components are functions of x and t , and A is a matrix whose components depend on w only. We assume that the eigenvalues of A are all real. We prescribe the initial condition $w(x, 0) = w_0(x)$ in the domain $-\infty < x < \infty$.

We divide the (x, t) plane by a set of points given by $x_j = j\Delta x$ and $t_n = n\Delta t$, where Δx and Δt are the in-

crements of x and t and $j = \dots, -2, -1, 0, 1, 2, \dots$ and $n = 0, 1, 2, \dots$. An approximation to $W(x_j, t_n)$ is denoted by W_j^n .

A typical explicit second-order difference scheme, called leapfrog, is

$$W_j^{n+1} = W_j^{n-1} - (\Delta t/\Delta x)A(W_{j+1}^n - W_{j-1}^n) \quad (31)$$

To analyze its stability property, we assume A to be constant. We express the solution of Eq. (31) as

$$W_j^n = \hat{W}^n(t) \exp(ikj\Delta x) \quad (32)$$

where k is the wavenumber and \hat{W}^n is the amplitude. The solution for W_j^{n+1} can also be expressed in the same form, but the amplitude is given by

$$\hat{W}^{n+1} = G\hat{W}^n \quad (33)$$

where G is the amplification matrix. Substitutions of Eqs. (32) and (33) into Eq. (31) yield the following equation for G :

$$G^2 + 2i[A(\Delta t/\Delta x) \sin(k\Delta x)]G - 1 = 0 \quad (34)$$

To obtain bounded solutions of Eq. (31), we must require that the eigenvalues of G not exceed unity in absolute value. This is essentially the von Neumann stability condition for a finite difference scheme. To satisfy the *computational stability condition* for the leapfrog scheme, we see from Eq. (34) that

$$|a|\Delta t/\Delta x \leq 1 \quad (35)$$

where a is an eigenvalue of matrix A . In the case of atmospheric equations, a typical large eigenvalue of A corresponds to the sound wave speed. Condition (35) was first pointed out by R. Courant, K. Friedrichs, and H. Lewy in 1928.

There is a large body of literature describing a variety of difference schemes to solve equations of the same type as Eq. (30). Any difference scheme for solving atmospheric equations has some deficiencies as well as merits. For example, the leapfrog scheme is perhaps the most important among explicit schemes, but there is a problem as pointed out in Section V.A. Because the computational mesh can be divided into odd and even points and the difference calculations at the odd points proceed independently of those at the even points, careless application of this scheme can cause out-of-phase values at two consecutive grid points.

The von Neumann stability analysis applies only to difference equations with constant coefficients and periodic boundary conditions. Instability may arise from specification of boundary conditions or nonlinear terms in differential equations. Instability caused by nonlinear terms is called *nonlinear instability* and was first noticed in a numerical solution of the nonlinear barotropic vorticity equation in the early days of numerical weather prediction.

This type of instability is caused by the accumulation of so-called *aliasing errors* due to misrepresentation of nonlinear interactions involving the smallest resolvable $2\Delta s$ waves, where Δs denotes the grid increment. Aliasing errors can be eliminated by filtering out all waves with wavelengths between $2\Delta s$ and $3\Delta s$. However, it is not correct to infer that aliasing errors always induce nonlinear instability, because it is possible to construct special difference equations that are computationally stable despite the presence of aliasing interactions. The class of such difference equations is called the *quadratic conserving scheme*. The essential requirement is that the total variance of solution within the integration domain does not increase faster than by $1 + 0(\Delta t)$. Among quadratic conserving schemes, the most well-known are Arakawa Jacobians. Akio Arakawa in 1966 invented unique second- and fourth-order difference schemes that conserve the sums of vorticity, square vorticity, and kinetic energy over the domain. Arakawa's conserving schemes have been applied extensively to solve many fluid dynamics problems. In the design of numerical methods, it is important to preserve the conservation relationships of a dynamical system for mass, momentum, energy, entropy, etc., as much as possible in the corresponding discretized equations.

C. Spectral and Finite-Element Methods

So far, we discussed an approximation of a dependent variable in terms of its values at discrete points in physical space and time. An alternative is to represent the field of any dependent variable in physical space as a finite series of smooth and, preferably, orthogonal functions. In this case, the prediction equations are transformed to a set of ordinary differential equations for the expansion coefficients, which depend only on time. Since expansion coefficients are referred to as spectra, this approach is called the *spectral method*.

To present the basic idea of the spectral method, we consider the differential equation

$$\frac{\partial w}{\partial t} = F(w) \quad (36)$$

over a closed domain S with a suitable boundary condition on w along the domain boundary. The dependent variable w is a function of space variable x and time t . We approximate w by W in terms of a finite series of linearly independent basis functions $\phi_j(x)$ from $j = 1$ to N , which is the number of spectral components:

$$W(x, t) = \sum_{j=1}^N C_j(t) \phi_j(x) \quad (37)$$

The coefficients $C_j(t)$ are functions of time t .

In deriving the prediction equations for $C_j(t)$, we make use of the *Galerkin approximation*. Namely, the residue

$\partial W / \partial t - F(W)$, obtained by substitution of Eq. (37) into (36), is forced to be zero in an averaged sense over domain S with basis functions ϕ_j as weights. From this integral, we obtain a system of N equations for N unknowns dC_j / dt . The integral can be performed exactly by numerical quadratures within the accuracy of representation in Eq. (37). The use of numerical quadratures in the physical space along with the spectral representation of Eq. (37) is called the *transform method*, which is now widely adopted to solve global atmospheric equations, using the spherical harmonics as the basis functions.

One modified version of the spectral method is called the *finite-element method*. The basis functions ϕ_j of Eq. (37) in the spectral method are defined over the entire domain, but the basis functions of the finite-element method are piecewise polynomials. These polynomials are local so that they are nonzero over only a small finite element. For example, piecewise linear "roof" functions have been used as basis functions. Coefficients $C_j(t)$ are then determined by application of the Galerkin approximation. Because the basis functions are nonzero over only a small domain, the expression for $W(x, t)$ resembles a finite-difference form.

One can combine the high accuracy of spectral calculations and the geometrical flexibility of finite-element representation. In this approach, called the *spectral finite-element method*, the computational domain is broken up into rectangular regions called elements, and within each of these elements the dependent variables are approximated by spectral functions. The temporal discrete equations are derived by using the Galerkin formulation in the same way as in the spectral method. This scheme may be particularly suited for regional prediction models which require flexibility and ease in the treatment of complex geometry and the convenience of mesh refinement in regions of small-scale dynamical variability.

D. Semi-Lagrangian Method

Normally, we adopt geometrically fixed (Eulerian) coordinates to perform the time integration. One problem with this approach is that the time integration schemes often require overly restricted time steps due to computation stability conditions. There is an alternative approach, called the Lagrangian method, in fluid dynamic calculations. In this approach, the properties of fluid are described in terms of the coordinates in space of the same fluid parcels following the movement of every fluid parcel. Although this approach is free from the computational stability condition, a disadvantage is that a regularly spaced set of parcels initially will often evolve to a highly irregularly spaced set at later times and observation of flow will be distorted. The idea of the so-called *semi-Lagrangian*

method is to approximate the conservation of a physical property along the particle trajectory within the framework of Eulerian coordinates. This can be achieved by using a different set of particles at each time step.

We present the basic idea of semi-Lagrangian method using a simple one-dimensional conservation (or advection) equation

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0 \quad (38)$$

and

$$\frac{dx}{dt} = u(x, t) \quad (39)$$

to demonstrate how to perform the advection calculation for a prognostic variable $u(x, t)$ as a function of x and t (Staniforth and Cote, 1991).

Let us assume that we know $u(x, t)$ at all mesh points x_i 's at times $t_n - \Delta t$ and t_n (see Fig. 1). Since the value of u does not change along a fluid trajectory, the value of u at x_i and $t_n + \Delta t$, denoted by point A in Fig. 1, can be approximated by

$$u(x_i, t_n + \Delta t) = u(x_i - 2\alpha, t_n - \Delta t) \quad (40)$$

where α is the distance BD in Fig. 1, which is the trajectory length of the particle traveled in x and Δt following an approximate space-time trajectory denoted by the line C-B-A. Thus, if we can find the distance α , then the value of u at the arrival point A at x_i and $t_n + \Delta t$ is just its value at the upstream point C at $x_i - 2\alpha$ and $t_n - \Delta t$. By definition, α can be expressed by

$$\alpha = u(x_i - \alpha, t_n) \Delta t \quad (41)$$

Since this is an implicit equation for α , an iterative method must be used to solve for α . To evaluate u between mesh points, spatial interpolation is necessary.

In principle, one can take the time step as large as one wishes. However, the accuracy of calculation will deteriorate rapidly if too large a time step is selected, because interpolations in time and space are involved. Nevertheless,

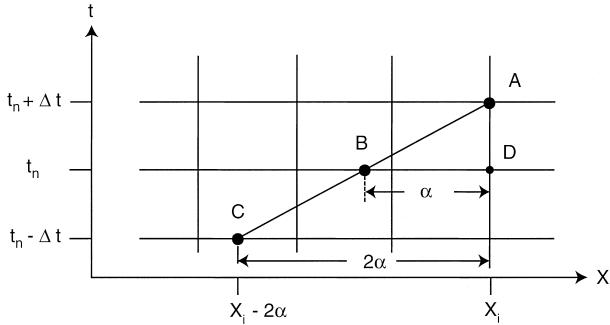


FIGURE 1 Three-time-level grid structure for the semi-Lagrangian scheme applied to the one-dimensional advection equation (38). See text for details.

the judicious use of the semi-Lagrangian method may provide a significant saving in computational resources (see Section IX).

Lastly, let us describe schematically how various integration methods can be put together to formulate an economical and practical numerical prediction algorithm. Let us write the atmospheric prediction equations in the form

$$\frac{dF(x, t)}{dt} + G(x, t) = R(x, t) \quad (42)$$

where

$$\frac{dF}{dt} = \frac{\partial F}{\partial t} + V(x, t) \cdot \nabla F \quad (43)$$

and

$$\frac{dx}{dt} = V(x, t) \quad (44)$$

Here, the coordinate variable x represents the vector involving the three space coordinates and F is the vector involving the prognostic variables. Similarly, G and R are the discretized expressions for the fast motions—such as acoustic and gravity waves—and the slow motions, such as planetary Rossby waves, respectively. The last term of (43) is a vector representation of the advection terms in Eq. (13).

Let us use the superscripts $+$, 0 and $-$, respectively, to indicate the value of a variable at the arrival point $(x, t + \Delta t)$, the midpoint of the trajectory $(x - \alpha, t)$ and the departure point $(x - 2\alpha, t - \Delta t)$, where α is defined as (see Fig. 1)

$$\alpha = V(x - \alpha, t) \Delta t \quad (45)$$

A semi-Lagrangian and semiimplicit approximation to Eq. (42) can be written as

$$\frac{F^+ - F^-}{2\Delta t} + \frac{1}{2}[G^+ + G^-] = R^0 \quad (46)$$

The advection terms are calculated by the semi-Lagrangian method by iteratively solving (45) for the vector displacements α in a similar way as done to (41). On the other hand, calculations of the first-motion terms G adopt the implicit method by evaluating the time average of its values at the end points of the trajectory. Calculations for the slow-motion terms R use the explicit method by evaluating at $x - \alpha$ and t . This algorithm gives a centered $O(\Delta t^2)$ approximation and is free from a restricted computational time step constraint. In practice, a time step of about 20 min is chosen for large-scale global weather prediction.

E. Special Problems in Atmospheric Modeling

There are many numerical problems unique to atmospheric modeling. For example, the longitude–latitude grid with constant increments in longitude and latitude

has computational disadvantages. Poles are singular where wind components in spherical coordinates are undefined. Grid points near the poles become dense and inhomogeneous because of the convergence of meridians toward the poles. Many solutions to this problem have been proposed, ranging from skipping grid points near the poles to redefinition of the grid points that cover the globe more uniformly using various mapping techniques. One interesting idea is to design a quasi-homogeneous grid based on the icosahedron. The icosahedron is the highest degree regular polyhedron consisting of 20 equilateral-triangular faces constructed within a sphere with its 12 vertices on the sphere. The vertices of the triangles can be connected by great circles to form 20 congruent major spherical triangles, which are then divided into many grid triangles. The spectral finite-difference method (Section V.C) may be suitable to design numerical algorithms for this kind of quasi-equilateral triangle grid, inspired by Buckminster Fuller's geodesic dome.

We often encounter the need for detailed prediction of weather over a relatively small area of the globe with a fine computational resolution, such as prediction of tropical cyclones and severe storms. One approach is to limit the forecast domain of concern in a global model and use a finer grid over the limited area. The limited area can be moved to another location as need arises. This configuration is referred to as a movable limited area model. The boundary conditions to the limited-area model may be supplied from a global prediction model. In this approach, no feedback of the limited-area prediction is considered to be outside of the limited area.

Another idea is to use a variable computational grid over the area where a detailed calculation is desired. Although the calculation must be performed over the entire global domain, the overhead of carrying out the calculation outside of the detailed area can be made relatively small by choosing a coarse grid outside of the area of interest. One feature of this variable-resolution approach in comparison with the limited-area modeling is that a single global model code can be used for both global and limited-area forecasts by minor changes in the model parameters, thereby avoiding the need to maintain two separate models of global and mesoscale predictions. The advent of the semi-Lagrangian semiimplicit algorithm, which uses a relatively large time step (Section V.D), has made this unified approach practical.

In dealing with the atmosphere we must be concerned with the discretization of variables in the vertical as well as in the horizontal. There are two ways to represent the vertical structure of the atmosphere. One is to use discrete points in the vertical with vertical derivatives approximated by finite differences. The other is to represent variables as a finite series of differentiable functions. However, the grid-point approach dominates the spectral

approach in discretization in the vertical. In fact, the potential usefulness of the spectral approach in the vertical has not been fully explored.

One reason that the difference method is preferred over the spectral method in the vertical is due to the handling of the earth's topography. Large mountain barriers influence atmospheric motions at all scales, and their incorporation into numerical prediction models is important. There are essentially two approaches to handling dynamical effects of mountains in prediction models. One is to block physically the integration domain in any coordinate system occupied by mountains. The other is to use terrain-following coordinates in which the earth's surface coincides with a coordinate surface. This latter approach is now widely adopted in atmospheric modeling because of its simplicity in computational treatment at the earth's surface. However, the pressure gradient calculation requires a careful hydrostatic correction owing to the slope of the coordinate surfaces relative to pressure or height coordinates. The need of accurate hydrostatic correction becomes greater for steeply sloped mountains. Further research is necessary to improve handling of mountains in numerical prediction models.

Whereas the atmosphere extends upward without limit, the numerical model of atmosphere must limit its vertical extent to a finite height. And, boundary conditions are required at the top of the model. Because the upper boundary conditions are not necessarily realistic, computational errors near the model top can propagate back into the domain of interest. Therefore, it is desirable to place the top boundary as high as possible. The top of an advanced global prediction model is placed at a height of 63 km above the ground, equivalent to 0.1 hPa in pressure, having 50 vertical levels. One advantage for extending the model top to such a height with a good vertical resolution is that a vast amount of atmospheric observations carried out by measuring infrared radiance from satellites can be better utilized to define the atmospheric state. And, the effects of radiatively active atmospheric compounds, such as ozone in the stratosphere, can be accurately calculated in the model (see Section VI). Especially for extended-range (1 month or more) forecasts, the impacts of motions in the stratosphere are no longer negligible. Having an expansive model top accompanied by high vertical resolution is a first step toward achieving accurate predictions of unique stratospheric weather and climate phenomena, such as sudden warmings and quasi-biennial oscillations.

VI. OBSERVING SYSTEMS AND DATA COLLECTIONS

The data that provide initial conditions for atmospheric prediction models come from a variety of observing

systems. Although the prediction models normally require the fields of horizontal velocity components u and v , temperature T , mixing ratio of water vapor q , and surface pressure p_s to start time integrations, the variational analysis method described later allows us to use various remote satellite infrared and microwave observations and other measurements, such as are performed by global positioning systems, for analysis of the initial conditions.

Initial conditions require accurate depiction of the three-dimensional structure of the atmosphere. Thus, the most important observing system is the one that provides the meteorological fields in three dimensions. The data on a single level tend to be less effective to use and their impact on a forecast is not as large as multilevel data. However, some of the single-level data, such as surface observations, are valuable from the standpoint of synoptic analyses with long-term records. Another classification of observing system is a direct (*in situ*) measurement vs indirect (remote) measurement. Most surface-based observations directly measure atmospheric variables, whereas satellite observations of atmospheric radiance and reflectivity at different wavelengths are used to infer the profiles of meteorological variables. There are also land-based remote measurements, such as radar, that can measure precipitation as well as winds. In general, direct measurements are accurate and easy to use but with limited spatial coverage, while remote measurements tend to be harder to use but with greater spatial coverage.

The most important and accurate observations that provide three-dimensional data are the measurements taken by sounding balloons, called radiosondes, released at several hundred weather stations around the world at fixed times at least once a day, mostly twice a day at 00 and 12 UTC (Coordinated Universal Time). These balloons reach to the height of 30 km, measure *in situ* temperature and humidity, and send the signals back to ground stations by radio and sometimes through communication satellites. The horizontal wind components are measured by tracking the balloons using radio range- and direction-finding devices. A recent technique also allows us to use satellites for tracking the balloons. Although the data from sounding balloons are more accurate compared to other measurements, no radiosonde observations are without problems. For example, the sun's radiation effects on temperature measurement in the stratosphere must be corrected. The *in situ* measurement of moisture in a low-temperature-low-pressure environment is very difficult and the data are normally available up to about 300 hPa, equivalent to the height of approximately 9 km. There is a simplified version of the balloon sounding called pilot balloons, that provides only horizontal wind components. Other *in situ* observation that provide accurate three-dimensional data are done by aircraft called ACARS

(Aircraft Communication Addressing and Reporting System). The data come from commercial aircraft that carry special instruments to measure wind and temperature with very short time intervals. The coverage of ACARS is the largest over North America, but it is gradually expanding to other parts of the world. High-quality vertical sounding data of temperature and winds are available from ACARS in the vicinity of airports during frequent take-offs and landings. The profiler is a land-based remote observation system using radar technology and provides accurate vertical sounding of winds with very high vertical resolution. This system is based on a fairly recent technology and is available over the United States and a limited number of stations in the tropics and other regions. These new data sources may become economical enough in the future to replace the conventional balloon soundings.

There are very large numbers of surface observations for winds, temperature, humidity, precipitation, cloud types and covers, radiation, etc., over the continents and islands. Many of the surface stations report at least twice a day at 00 and 12 UTC, but some of them report every hour. Other important surface observations come from traveling commercial ships that provide atmospheric and oceanic data at synoptic times. In addition, 100 or more drifting buoys gather atmospheric and oceanic data.

Satellite remote measurements are a very important component of the atmospheric observing system, and their importance is increasing steadily. Currently, several meteorological satellites are making measurements continuously and their data coverage is global with high spatial resolutions. In fact, the number of satellite observations is nearly 100 times more than that of land-based observations. There are several different types of satellite observations. Wind can be derived by tracing cloud or water vapor cells identified by the visible, infrared, and microwave images from geostationary satellites. The microwave measurements of the state of seas also provide wind direction and speed of the air near the surface. Radiance measurements at various wavelengths are used to infer a three-dimensional coverage of temperature and humidity in the troposphere. Microwave measurements are suited to infer humidity and liquid water fields. Special satellite observation programs are aimed specifically at measuring meteorological variables in the stratosphere. In the past, radiance measurements from the meteorological satellites were converted to temperature and humidity profiles using various retrieval methods, and the converted data were used in atmospheric data analysis. The retrieval process includes removal of observations contaminated by clouds and obtaining temperature and humidity profiles using statistical and/or physical methods. In the current three- and four-dimensional variational data analysis framework, these retrieval processes can be

TABLE III Number of Meteorological Data Processed at National Centers for Environmental Prediction at 00 UTC, 1 October 1999

Measurement platform	Number of data or profiles
Surface land observations	9,035
Surface ship observations	2,458
Other surface land/marine data	10,041
Ocean buoys	37
Radiosonde temperatures/winds ^a	650
Aircrafts	15,406
Pilot balloons	137
Wind profilers ^a	155
Surface SSM/I winds	17,079
Surface scatterometer winds	43,889
Satellite radiances ^b	81,158
Satellite cloud drift winds	9,957
Satellite water vapor winds	5,989
Total	195,991

^a Number of profiles.

^b Number of profiles from channels 11 and 14 of HIRS, MSU, GOES, and AMSU-A. (Acronyms: AMSU-A, Advanced Microwave Scanning Unit—A (NOAA); SSM/I, Special Sensor Microwave/Imager (Defense Meteorological Satellite Program); GOES, Geostationary Operational Environmental Satellite (NOAA); HIRS, High-Resolution Infrared Sounder; MSU, Microwave Sounding Unit.)

merged into the analysis scheme itself as described in Section VII.C.

Table III shows the number of meteorological data processed at the National Centers for Environmental Prediction (NCEP), Washington, D.C., at 00 UTC October 1, 1999. The data are transmitted via the Global Telecommunication System (GTS), which is a main trunk communication line connecting Bracknell (United Kingdom), Tokyo (Japan), Offenbach (Germany), Moscow (Russia), Melbourne (Australia), New Delhi (India), and other major cities around the world. Once data are received at operational forecasting centers, they are decoded and quality controlled before begin used in the objective analysis (Section VII). Normally, most of the data arrive within 5 to 6 hours after the map time.

The current global observing system is the outgrowth of networks established in 1979 at the epoch-making Global Weather Experiment (GWE) under the auspices of the Global Atmospheric Research Program (GARP). GARP was an international scientific endeavor to improve weather prediction, as well as to increase our understanding of the transient behaviors of global atmospheric circulation that control changes of weather. In fact, GWE was

the largest scientific experiment ever attempted to observe the atmosphere, utilizing every means of the observing systems. Unfortunately, there have been movements in these days to curtail the expenditure of maintaining the radiosonde stations and to replace atmospheric observations by using other means, notably meteorological satellites. Clearly, careful assessment is necessary to evaluate the impact of removal from, addition to, and replacement of data in the existing global observation networks. Such an assessment can be conducted by using the prediction model itself through the means of a so-called Observing System Simulation Experiment (OSSE) by constructing simulated “observed data” taken from a long-term time integration with a prediction model that is different from the one used for the OSSE. It is obvious that the simulated observations are only as realistic as the prediction model that produced the data, but data are self-consistent meteorologically.

One bright aspect in the transition of a global meteorological observing system from manpower-intensive to automated operations is the prospect of having a series of geophysical and meteorological satellites year after year for measurements of not only traditional meteorological field variables, but also a wide range of atmospheric and ocean model parameters. In advanced prediction models, additional initial and boundary conditions are necessary to specify the values of parameters involved in the physical processes (i.e., parameterizations; Section IV.A). For example, the prediction model used for operational global medium-range prediction at the NCEP/NOAA requires initial values for soil wetness and temperature at two levels in the ground, water equivalent snow depth, surface skin temperature, sea ice fraction, sea surface temperature, and ozone concentration. Furthermore, fixed constant or seasonally variable climatological values of surface albedo, vegetation fraction, vegetation type, surface roughness, and carbon dioxide concentration are needed. Currently, information on some of the parameter values is not easily available. Various new meteorological satellites are specifically aimed at collecting those data that have been totally missing in the traditional meteorological observations. An earth probe satellite named the Tropical Rainfall Measuring Mission (TRMM) was launched in 1997. This is the first radar to be flown in space to measure the precipitation rate accompanied by a microwave sensor to provide information on the integrated column precipitation content, and a visible/infrared radiometer to observe cloud coverage, type, and top temperatures. Precipitation rate as an initial condition plays a crucial role in improving the quality of weather prediction even in the midlatitudes (Section VII.B). The success of the proof-of-concept experiment, Global Positioning System/Meteorology (GPS/MET) in 1995–97, suggests that

the radio occultation technique may provide additional and independent information on the atmospheric temperature, pressure, and water vapor fields in the future. Thus, these new satellite observations will begin to provide needed information on the design of prediction models, as well as the present state of the atmosphere for accurate initial conditions, and eventually contribute to produce much improved weather forecasts.

VII. ATMOSPHERIC INPUT DATA FOR PREDICTION MODELS

We require initial conditions to start the time integration of prediction models. Initial conditions to the prediction system of Eqs. (7) to (12) and (19) are the three-dimensional fields of velocity components u , v , and w , pressure p , temperature T , or density ρ , and water vapor mixing ratio q (Section IV.C). For primitive equation models, vertical velocity w is a diagnostic variable (Section III.B). Also, the hydrostatic equation, Eq. (16), gives a relationship between p and ρ , or T . Therefore, for primitive equation models, it suffices to have the initial conditions of only u , v , T , q , and surface pressure, p_s . In this section, we discuss how to prepare the initial conditions from observed data.

A. Objective Analysis

The automatic process of transforming observed data from geographic, but irregularly distributed, locations at various times to numerical values at regularly spaced grid points at fixed times is called *objective analysis*. In the earlier days of numerical weather prediction, the two-dimensional fitting of polynomial was applied to observational data in an area surrounding a grid point at which the value of analysis is required. Polynomial fitting is suitable for processing relatively dense and redundant data, but it often gives unreasonable results in data sparse areas.

More recent procedures of objective analysis practiced at most of the operational forecasting centers consist of the steps schematically described in Fig. 2. Most observed data are collected worldwide every 6 hours, say 00, 06, . . . UTC (Coordinated Universal Time), and are blended into forecast values appropriate to the map time by the process referred to as *data assimilation*. The basic process of data assimilation, indicated by the box enclosed by dashed lines in Fig. 2, consists of two steps: One is objective analysis and the other is called *initialization*, which will be detailed in the next subsection.

The step of objective analysis is carried out by statistical interpolation by taking into account all of the available observations (Section VI) plus other prior information.

Short-term forecasts from the previous analysis cycle are used as prior information and are referred to as background fields.

Let us denote a collection of variables to be analyzed at the model grid by X , which is a vector of dimension M , and a collection of observed data given at irregularly distributed stations by y , which is a vector of dimension N . Normally, the number of observed data N is expected to be smaller than the number of model variables M , i.e., $N < M$. Therefore, we must blend the observed data into the background fields to produce the “best” estimate of the analysis X . We denote the background field by X_b . Since the locations of observations do not necessarily coincide with those of the model grid points where the background data exist, we need a relationship between y and X through some sort of interpolation in the form

$$y = HX + \epsilon_o \quad (47)$$

where H is the interpolation operator. The form of operator H can be fairly complex and nonlinear in the case of observed data given by passive sounders from satellites (Section VI). Therefore, H can be generally called the data conversion operator. In Eq. (47), the quantity ϵ_o denotes the observational error vector.

The blending of the observation data y with the background field X_b to obtain the analysis field X is done by

$$X = X_b + K(y - HX_b) \quad (48)$$

where K denotes the weight to the observation increment ($y - HX_b$). Of course, X is not the true analysis field because of the presence of errors in the observations, background fields X_b , and the operator H , as well as in the form of the weight K . The form of K , which is often referred to as gain matrix in estimation theory, is determined by requiring that the expected mean-square estimation error over the entire model grid points be minimized. In practice, we normalize the departures of all analyzed, observed, and background values from “true” values by the respective error estimates. Various error statistics and the form of operator H are reflected in the determination of the weight K . Also, the estimation formula of Eq. (48) is applicable for many variables, and multivariate statistical relationships among different variables can be incorporated. However, the analyzed prognostic variables thus obtained are not necessarily suitable as the input data to the prediction models.

B. Concept of Initialization

The solutions of primitive equation models correspond to two distinct types of motion. One type has low frequency. Its motion is quasi-geostrophic and meteorologically dominant. The other corresponds to high-frequency

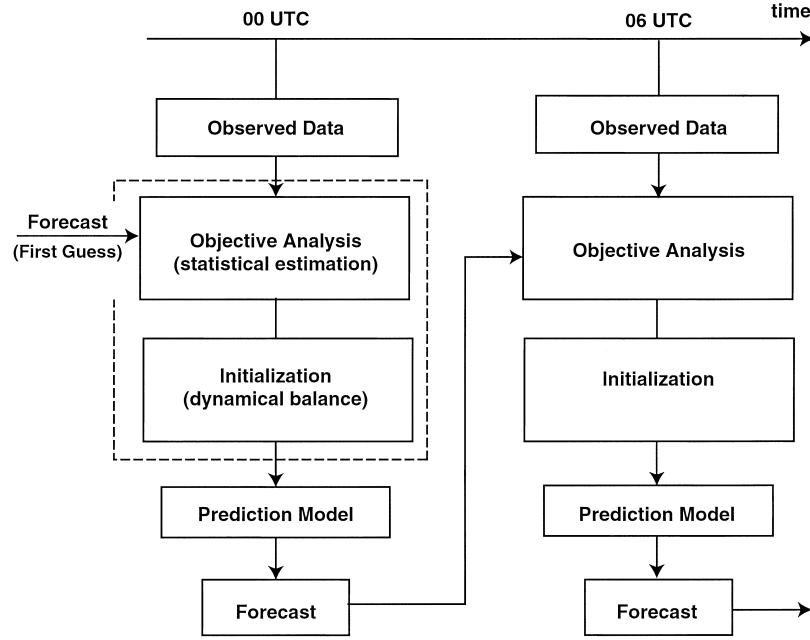


FIGURE 2 Schematic diagram for atmospheric forecast–analysis system. The two processes in the square enclosed by dashed lines can now be combined in the four-dimensional variational data assimilation using a continuous stream of observed data. See text for details.

gravity-inertia modes. The amplitude of the latter type of motion is small in the atmosphere. Hence, it is important to ensure that the amplitudes of high-frequency motions are small initially and remain small during the time integration when the primitive equations are solved as an initial value problem. The process of adjusting the input data to ensure this dynamical balance is called *initialization*.

Solution to the initialization problem was central to the successful transition in forecast practice from the use of quasi-geostrophic models to primitive equation systems during the 1970s. Since gravity-inertial motions are filtered out in quasi-geostrophic models, no special procedure was necessary and the objectively analyzed data could be used immediately as the input data to quasi-geostrophic models. Actually, there is an interesting history that led to solution of the initialization problem that is equivalent to the quest of understanding what the primitive equation forecast system really is. A promising break came with the advent of so-called *nonlinear normal mode initialization* (NNMI) during the latter part of the 1970s.

To illustrate the essential features of NNMI, we first point out that the characteristics of global atmospheric motions can be examined from the solutions of a linearized system of the basic hydrostatic prediction equations discussed in Section III.B. A simple perturbation system is the one linearized around the atmosphere at rest with a basic temperature that depends only on height. Solutions of such a linear system with appropriate boundary conditions are called normal modes.

If we represent the solutions of a primitive equation model in terms of its normal modes, we can symbolically express the evolution of the normal mode expansion coefficients using the following two spectral equations:

$$dW_R/dt + i\sigma_R W_R = N_R \quad (49)$$

for low-frequency meteorologically dominant motions, and

$$dW_G/dt + i\sigma_G W_G = N_G \quad (50)$$

for high-frequency gravity-inertia modes, which have small amplitudes. Here, W_R and W_G are the vector spectral coefficients for low- and high-frequency motions, respectively, and σ_R and σ_G denote corresponding frequencies. Also, N_R and N_G represent the nonlinear contributions to the time changes of W_R and W_G , respectively.

One way to suppress high-frequency motions initially is to set $W_G = 0$. However, this distorts the flow and leads to generation of high-frequency waves due to the nonlinear term N_G in Eq. (50). The essence of NNMI is to keep a small amount of W_G initially so that the initial W_G will not grow. For example, if we set $W_G = -i\sigma_G^{-1} N_G$ initially, dW_G/dt becomes zero initially as seen from Eq. (50). One can develop more sophisticated procedures to control the generation of excessive high-frequency motions.

Although much progress has been made on the procedure of initialization to control the generation of high-frequency motions, a basic question still remains: how to initialize the atmospheric motions that are subjected to

diabatic forcing such as heating due to the release of latent heat by condensation of water vapor. This question arises from attempting to initialize the tropical flows using an adiabatic initialization scheme without consideration of diabatic forcing. In that case, the initialization scheme often alters objectively analyzed flows beyond observational errors even over data-dense areas. The tropical motions are essentially driven by condensation heating, unlike the mid-latitude disturbances, which are controlled mostly by the dynamics due to meridional temperature contrast. Therefore, the effect of diabatic forcing must be included in the term N_G on the right-hand side of Eq. (50). Unfortunately, the evaluation of diabatic forcing term is not straightforward, because its evaluation requires the knowledge of model physics owing to the lack of direct measurements.

Another problem associated with the initialization of diabatic flow is that the moisture field must also be properly initialized in such a way that the initial precipitation rate calculated from the prediction model agrees with observed. The problem is often compounded by the fact that the precipitation rates on global scale are not easily available through conventional rainfall measurements. However, progress has been made recently in the moisture initialization owing to the advent of satellite observations, such as the TRMM (Section VI), and the development of variational data assimilation described in the next subsection.

C. Variational Data Assimilation

In the previous two subsections, we described the preparation of initial conditions in two separate steps of objective analysis and initialization. Also, we made an implicit assumption that all observational data are available as analysis variables (u , v , T , q , and p_s) at analysis times. However, as described in Section VI, the global earth observing system is a combination of various observing subsystems that have different capabilities concerning the type, measurement, quantity, frequency, and location of observations. For example, the satellite instruments measure vertically integrated temperature and water vapor content with certain weights as functions of height that are predetermined by the spectral windows of measuring atmospheric radiance. Since the radiometers observe only vertically integrated quantities, an inversion process is necessary to retrieve the vertical profiles of temperature and humidity from radiance measurements. This process is referred to as *satellite retrieval*. Since the number of spectral windows used to measure radiance is normally much less than the number of vertical analysis levels, some reference profiles of temperature and humidity are necessary to optimize the satellite retrieval process. In the past, the satellite retrieval process was performed separately from objective analysis

using climatological data as the reference profiles. Since short-range prediction provides much more accurate reference profiles, the functionality of the objective analysis is now extended to incorporate the retrieval processes. Thus, the prediction model is used as an integrator, whereby the three steps of retrieval, statistical estimation, and dynamical balance can be combined.

The method of analyzing the state of the atmosphere by utilizing a continuous stream of different kinds of observations to produce the evolution of flow that best fits all the observations during a period of assimilation window of 6 to 12 (or more) hours is referred to as *four-dimensional variational (4DVAR) data assimilation*.

Before describing the 4DVAR data assimilation, we present the three-dimensional case, which is a recast of the procedure of objective analysis (Section VII.A) to incorporate various constraints explicitly for achieving the “best” estimate of the analysis field X . The best estimate of X is said to be achieved if the background field X_b and the observation data y are combined in such a way to minimize the cost function J defined by

$$J = J_1 + J_2 \quad (51)$$

where

$$J_1 = \frac{1}{2}(X - X_b)^T B^{-1}(X - X_b) \quad (52)$$

and

$$J_2 = \frac{1}{2}(y - HX)^T O^{-1}(y - HX) \quad (53)$$

Here, B denotes the background error covariance matrix and O the observational error covariance matrix, representing the respective degree of confidence and used to weight the components of cost function. The superscript T represents the transpose of a matrix. The operator H denotes the data conversion operator defined by Eq. (47) and represents all the operations that are required to convert from model variables to observed variables. For instance, H includes radiation transfer calculations to convert model temperature and humidity to corresponding radiances, so that model-generated radiances can be compared directly with satellite radiances. The representation of Eq. (53) has the advantage of combining analysis and retrievals. Namely, very different types of observations, e.g., winds from radiosonde and radiance observations, can be treated simultaneously.

If one carries out calculations of the derivatives of J with respect to X to seek the minimum condition of J , the formal expression of the best estimate of X becomes similar to Eq. (48), but it is not the same because approximations used to derive the two systems are different. In the case of multidimensional problems it is more efficient to obtain the best estimate of X by an iterative method that

leads to reduction of J successively toward convergence. This approach is called the *three-dimensional variational (3DVAR) assimilation*.

In the grand scheme of 4DAR, all observations during the time window $t = 0$ to τ are assimilated into the forecast field X , which evolves according to the forecast model

$$\frac{\partial X}{\partial t} = L(X) + \epsilon_f \quad (54)$$

where L denotes the prediction model operator and ϵ_f represents prediction errors arising from model deficiencies.

The task now is to set up the optimization algorithm which produces the best estimate of the analysis X at the map time $t = 0$, denoted by X_o , using the series of observations given from $t = 0$ to τ . This is achieved by seeking the field X_o that minimizes the cost function J

$$J = J'_1 + J'_2 + J_3 + \text{other constraints} \quad (55)$$

where

$$J'_1 = \frac{1}{2}(X_o - X_b)^T B^{-1}(X_o - X_b) \quad (56)$$

$$J'_2 = \frac{1}{2\tau} \int_0^\tau (y - HX)^T O^{-1}(y - HX) dt \quad (57)$$

$$J_3 = \frac{1}{2\tau} \int_0^\tau \left[\frac{\partial X}{\partial t} - L(X) \right]^T P^{-1} \left[\frac{\partial X}{\partial t} - L(X) \right] dt \quad (58)$$

The quantity J'_1 is identical to J_1 of Eq. (52) in 3DVAR, except that the analysis field at $t = 0$ is denoted by X_o . Because all observations during $t = 0$ to τ are used, the constraint J'_2 is expressed as the integral form of J_2 of Eq. (53) with respect to time. The observational error covariance matrix O depends on time t . The quantity J_3 can be interpreted in the same way as J'_2 , except that the prediction model errors are involved instead of the observational errors. The quantity P in Eq. (58) denotes a covariance matrix that describes prediction error statistics.

The formulation of an iterative method to achieve the reduction of J successively toward convergence is no longer a simple optimization process as in 3DVAR because of the involvement of the prediction equation (54). The value of the gradient of the cost function J with respect to X_o is calculated by the backward time integration from $t = \tau$ to 0 of the transposed version of a linearized prediction model, referred to as the *adjoint model*, for the forecast field X . During the backward time integration of the adjoint model, observation $(y - HX)$ increments are assimilated as forcing at each time step. The practical aspects of the iteration process are highly technical. Although the 4DVAR works very well for adiabatic flows, there are many problems in implementing this procedure for actual atmospheric states in which diabatic forcings are involved. For example, it is

not straightforward to conduct the adjoint model calculations based on the prediction equations involving complex hydrological processes, such as cloud and precipitation. The estimation of the model error covariance matrix P is very difficult. In fact, the practical application of 4DVAR generally assumes no model errors, i.e., $J_3 = 0$. Nevertheless, the approach of 4DVAR is conceptually elegant and offers a great deal of promise to produce the optimal initial conditions. For example, the requirement of dynamical balance discussed in Section VII.B can be built into the 4DVAR through appropriate specification of other constraints in the cost function (55).

VIII. APPLICATIONS OF NUMERICAL PREDICTION

A. Predictability of the Atmosphere

One of the fundamental questions in weather prediction concerns the *predictability* of the atmosphere. There are two parts to this question. One is how well can we predict weather changes, assuming that we have a perfect forecasting model. The second is, not having the perfect model, how close we are to achieving this ideal state. Why can't we forecast perfectly, having the perfect model? This is because we can never determine the initial states perfectly because of uncertainties involved in observing the atmosphere and analyzing the data for the initial conditions. No matter how accurately instruments are designed to observe, the turbulent character of the fluid system gives rise to uncertainty in the "true" representation of fluid motions to be predicted. In addition to instrument and analysis errors, there are representativeness errors due to limited number of observations in the continuum.

Astronomical events such as eclipses can be predicted many years in advance with great precision. The laws governing celestial motions are precisely known from the dynamics of discrete bodies moving under gravitational fields, whereas the models used for predicting the motions of the atmosphere and oceans are approximate and nonlinear, involving very complex forcings.

In the early days of numerical weather prediction, Philip Duncan Thompson (1922–1994) put forward these questions on atmospheric predictability and investigated from the scientific grounds of nonlinear fluid dynamics. A dramatic turn on the theory of atmospheric predictability came in the early 1960s when Edward N. Lorenz quantitatively demonstrated how much tiny initial errors can grow in time to the extent that prediction becomes totally unskillful. In fact, his analysis contributed to blossoming of a physical concept called *chaos* in nonlinear dynamics.

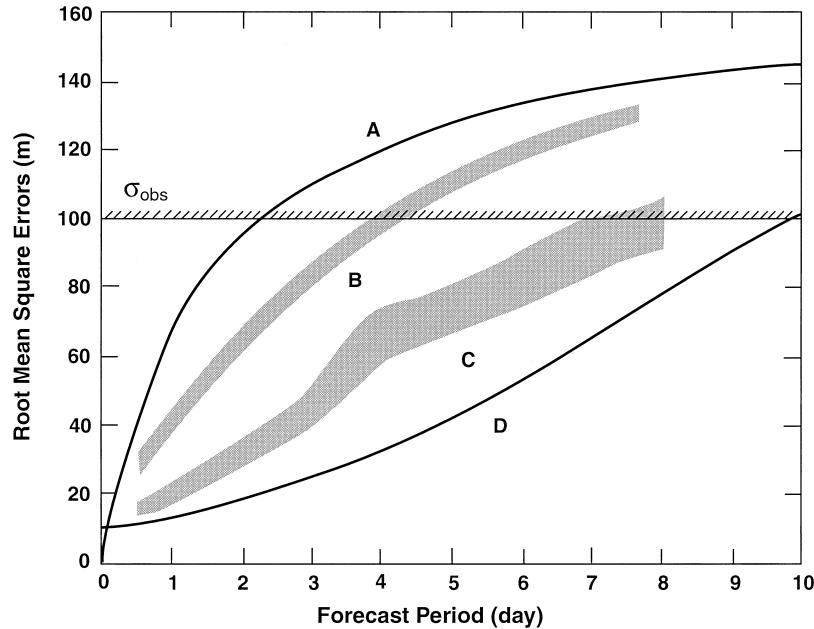


FIGURE 3 The growth of rms forecast errors as a function of time in days. The rms forecast errors are defined as the rms errors of the predicted height field of isobaric surface against its verification, averaged over the Northern Hemisphere north of 20° . Curve A represents the average error growth of persistence forecast. Bands B and C show average scores for the 1970s and 1990s operational forecasts, respectively. Curve D illustrates the rms error growth of an idealized atmospheric model, which has an error doubling time of 2 days, starting from the initial rms error of 10 m.

Let us consider how predictable the atmosphere is by numerical models. If we have no means of forecasting, all we can do is to use climatology or to assume that today's weather will continue, namely to use *persistence*. As one measure of verifying forecasts, we often use the root-mean-square (rms) error of the predicted height field of an isobaric surface. Figure 3 shows the growth of the average rms errors versus forecasting period in days during a boreal winter. Curve A represents the error growth of persistence forecasts. The time period of useful forecast may be defined as the time required for the forecast rms height error to reach the value of 100 m. Hence, the persistence forecasts are useful for only 2 days on average, which was comparable to the average score of weather forecasting prior to the dawn of the numerical weather prediction era. Average scores of operational forecasts during the 1970s are shown by band B, which indicates that the forecasts were useful up to 4 days on average. By contrast, the scores of the 1990s are shown by band C, which indicates that global prediction skill has now been extended to almost 7 days.

Now the question is, how far can we improve numerical prediction? The answer depends on the theoretical predictability of the atmosphere under the inevitable limit in our ability to prepare reliable initial conditions. It is therefore meaningful to ask how the uncertainty in initial state will grow with time during the course of numeri-

cal forecast. The predictability is said to be lost when the growth of initial error reaches the level of climatological variance denoted by σ_{obs} on Fig. 3.

The growth of initial error in atmospheric models occurs as a result of instabilities of the atmosphere and the nonlinear character of fluid motions. The transfer of error between various scales of motion and the growth of error are explained by statistical theories of turbulence. We can also determine the growth of error by calculating the separation of pairs of numerical solutions starting from slightly different initial states. The calculations using operational global models of ECMWF and NCEP indicate that the error growth rate is in the range of 1.0 to 2.0 days in error doubling time, depending on the phenomena being forecast.

Curve D in Fig. 3 shows how the rms error would grow with time if we assume that the initial rms error of 10 m grows with the doubling time of 2 days, and the rms will approach the asymptotic value of 145 m as seen from persistence forecasts. There are uncertainties as to the values of both the error growth rate and the level of error saturation, as well as the value of initial error. In fact, the use of 2-day error growth rate may be too optimistic since the quoted value tends to decrease in advanced models, which use high spatial grid resolutions and incorporate the physical processes of smaller scale motions that have much

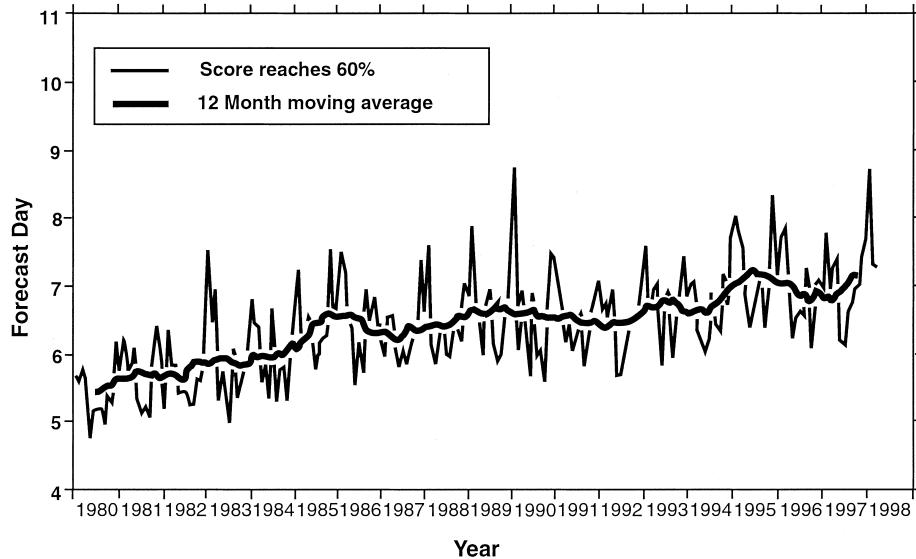


FIGURE 4 Monthly values of ECMWF medium-range forecast skill over the Northern Hemisphere during 1980–1997. The heavy smoothed solid line represents the 12-month moving average. The forecast skill here is defined by the forecast day at which the mean anomaly correlation (AC) of height change of 500 hPa surface between forecast and analysis reaches 60%. Forecasts are considered to be useful as long as the AC value is higher than the threshold value of 60%. Note the trend of increasing the useful forecast day year after year. Source: Courtesy of ECMWF.

shorter lifetime than the large-scale motions (Table I). Therefore, the theoretical predictability limit with an idealized model could be 9 days or even less rather than 10 days as curve D indicates. Now, what are the prospects of long-range prediction beyond the daily predictability limit, such as seasonal forecasts? These issues will be addressed in the remainder of this section.

To be more specific about the skill of forecasts, we show in Fig. 4 a long-term verification of ECMWF global medium-range (up to 10 days) forecasts since 1980 in terms of the useful forecast days against years of operation. Here, the definition of the useful forecast day is slightly different from the day that band B and C in Fig. 3 cross the climate rms value of 100 m. Instead of the rms height errors, the verification score uses the mean anomaly pattern correlation (AC) of height changes of 500 hPa surface in the midtroposphere between forecasts and analyses averaged over the Northern Hemisphere north of 20 degrees. The AC value is 100% at day 0, but the score drops gradually with forecast period in days. The forecast is said to be useful as long as the AC value stays above 60%, which is based on subjective evaluation of the forecasts by experienced forecasters. Thus, the day that the AC score crosses the 60% line defines another useful forecast day, which is used as the ordinate of Fig. 4. Note that the monthly averaged AC values fluctuate from month to month, but the smoothed curve for a 12-month running mean indicates an increasing trend of forecast score year after year. Notice that the score is generally higher during winter than summer.

We should note that all the arguments presented are for average of a large number of cases. We often encounter cases in which the model can make an excellent forecast even beyond 10 days. Predictability is a function of flow regimes, and finding the conditions for extended predictability is another challenging problem.

B. Prediction of Forecast Reliability

So far we have been concerned with the predictability of deterministic forecasts. Here the adjective of “deterministic” is used to indicate a single run from the “best” initial state available. However, recognizing the inevitable limit of deterministic predictability, the usefulness of a forecast will be enhanced significantly if the degree of forecast reliability is known. As average forecast skill scores vary month by month, daily forecast skill scores fluctuate day by day. It happens often that the forecast score suddenly drops a few days after having consistently high scores for 1 to 2 weeks. The public sometimes remembers such a forecast fiasco and discounts the usefulness of forecast in return. This strange behavior of numerical forecast may be caused by occasional large analysis errors in the initial conditions due to the lack of appropriate observations in crucial areas and/or the model’s inability to capture the transition of weather regimes from one type to another, such as the beginning of monsoons and the ending of blocking highs (prolonged fair-weather conditions). Clearly, any effort to elucidate these phenomena will contribute to add value to deterministic forecasts. The

first step to investigate these phenomena is to examine the sensitivity of forecasts on the uncertainty of initial states by running an ensemble of forecasts starting from slightly different initial conditions. Obviously, we expect a divergence of forecasts around the deterministic forecast. If such a divergence is small, we may place a high degree of confidence on its deterministic forecast, although all the ensemble forecasts are equally probable as long as the initial errors are within the analysis errors.

Clearly, what was stated makes sense conceptually, but a practical question will arise on the selection of various initial state around the “best” analysis for ensemble forecasts. Moreover, since the prediction models are not perfect, a dispersion of forecasts measured by a spread of the probability density function (pdf) in the N -dimensional phase space of model variables will also result from slight changes in the physical parameters of the models, even if the best analysis were true. To simplify our consideration for a moment, let us assume that the model is perfect, but there is uncertainty in the initial state.

A simple approach of ensemble forecasting is called *Monte Carlo forecasting*, in which random small errors are added to the best analysis in each ensemble run. The problem is that it requires a very large number, on the order of thousands, of runs in order to obtain a meaningful distribution of pdf after some forecast period. This is because the number of a model’s freedom in the phase space is on the order of millions, and a random selection of initial uncertainty is not an effective way to determine the likelihood of extreme situations in the forecasts. Since the divergence of forecasts in the early stage occurs because of the intrinsic instability of fluid flows, more effective selection of initial errors may be achieved by perturbing the flow patterns that have a tendency to grow rapidly.

Two approaches are mainly used for the selection of rapidly growing initial errors. One is called the *singular vector method*, which calculates most rapidly growing components by numerically solving the eigenvalues of a linearized version of the prediction model around the basic state obtained from the best analysis. The other is called the *breeding method*, which finds the most rapidly growing modes by repeated 1-day time integrations of the model itself, starting from small initial errors. At the end of each 1-day forecast, the forecast errors are rescaled to reduce their magnitudes to the level of initial errors for the subsequent “breeding” cycle of most growing modes.

At operational centers, as many as 50 model integrations are carried out usually up to a forecast period of 15 days, starting from different initial states that include a variety of rapidly growing error modes. Now, having as many as 50 forecasts in comparison with a single deterministic

forecast, the problem is to digest all useful information contained in the ensemble runs to infer the reliability of the forecast. One obvious thing to do is to measure the spread within the ensemble in terms of a specific norm such as rms height errors of the 500-hPa pressure surface. One might expect that a small spread of the ensemble is related to high skill of the forecast. However, this expectation does not always hold because of the presence of model errors. Nevertheless, the benefits of ensemble forecasts will become profound in the future, once scientists begin to learn how to utilize the statistical information of ensemble forecasts not only in probabilistic forecasts, but also in improving the data assimilation system.

One interesting result from validating ensemble forecasts is that the mean of the ensemble is generally better than the single deterministic forecast. Therefore, it is possible to improve the skill of deterministic forecasts through ensemble averaging, which in effect filters out unpredictable components. By averaging, the predictability error growth can be reduced, but the measure of useful information may also be diminished. Thus, it is important to find a particular process of averaging ensemble forecasts that can reduce the error growth without much affecting the strength of what we are looking for, i.e., signals. The prospect of long-range forecasts for the period of, say, 1 month is hinged upon our ability to extract these signals from the ensemble forecasts. How to tackle the problem of long-range forecasts will be discussed in the next subsection.

Up to this point, we have assumed that the prediction models are perfect and the divergence of ensemble forecasts occurs as a result of uncertainty in the initial states. This is certainly not the case in reality. The errors of prediction models are determined by verifying the forecasts against the corresponding analyses. When the difference between forecast and verification is averaged for a certain period, we find the model’s systematic error, which is often not negligible. In fact, the presence of systematic model errors is hurting our ability to succeed in long-range forecasts. Thus, the task of modelers is to reduce the systematic errors as much as possible, but it is still an elusive task. There are numerous tunable parameters in the physical processes of the model (Section IV), the values of which are chosen based on field and laboratory experiments. However, there are sometimes large uncertainties in their values and further tuning is needed to improve the performance of the model. One strategy of ensemble forecasts is to perform many runs using a set of model codes in which the values of some physical parameters are varied. Extending this step further, one can create a set of different initial states through the data assimilation system in which parameter values in the prediction model are perturbed slightly. By doing so, we now have a set of

perturbed initial states, as well as a set of perturbed prediction models to perform ensemble forecasts through a combination of two perturbed sets. The strategy of ensemble forecasts is currently a topic of active research along with better utilization of outputs from the ensemble forecasts.

C. Seasonal Prediction and Ocean–Atmosphere Coupled Modeling

The predictability limit mentioned earlier concerns the ability of models to predict the day-to-day evolution of the atmosphere, starting from a given initial state. The existence of such a deterministic limit does not imply that the models are incapable of longer range prediction of an averaged state of the atmosphere, called *climate*. In the long-term simulation of climate by the prediction models, the direct influence of initial state is minor in comparison with external forcings that evolve during the time period of simulation. The task of using the atmospheric models for seasonal prediction is to capture the dynamical response of external forcings, i.e., *signals*, by filtering out meteorological noise in the system produced by the day-to-day variations of weather.

The capability of atmospheric models to capture such signals has been demonstrated by numerical experiments conducted by the Atmospheric Model Intercomparison Project (AMIP) sponsored by the World Climate Research Program. The atmospheric models are integrated for a very long period of 20 or more years in which the distribution of sea surface temperature (SST) is prescribed as observed as a lower boundary condition. Many atmospheric models have succeeded in reproducing the so-called El Niño–Southern Oscillation (ENSO) phenomena. Southern Oscillation (SO) is a large-scale seesaw of atmospheric mass between the Pacific and Indian Oceans in the tropics and subtropics with an irregular period of 2 to 7 years. El Niño refers to an anomalously large warming of coastal waters off South America, a phenomenon often affecting the local fishing industry. These two phenomena, one in the atmosphere and the other in the Pacific Ocean, interact with each other and create the largest signal in interseasonal global climate variability. The ENSO as manifested by variations of warm and cold SST over the Equatorial Pacific influences the climate of the midlatitudes through the mechanism of atmospheric teleconnections. For example, the weather condition called Pacific North American (PNA) pattern (i.e., high pressure over the eastern equatorial Pacific, low pressure over the northern Pacific, high over North America, and low over the southeast coast of the United States) controls typical winter-season weather over the United States. The success of AMIP experiments suggests a prospect of dynamical seasonal forecasts, if we can predict SST several months in advance.

It is obvious, then, that we must couple two models of atmosphere and ocean for prediction of seasonal time scale. Physically, these two models interact with each other in such a way that the atmospheric model, uses SST predicted by the ocean model, which requires the surface wind stress, the net surface energy flux (i.e., the sum of sensible and latent heat fluxes), and the net influx of fresh water (i.e., precipitation minus evaporation). The atmospheric model requires information on surface energy exchange as well as the SST.

One problem of the coupled model is how to start the time integration of the model. There is not much problem for the atmospheric model as we know the state of the atmosphere through the analysis of observations including the SST. However, the three-dimensional state of oceans is only partially known because of sparse ocean observation networks. When the ocean model is integrated forward in time starting from an unrealistic initial state, it takes a very long time to come to an equilibrium with the atmosphere, which provides forcings to the ocean model. For example, it takes on the order of 10 years or less for the upper ocean, about 100 years for the midlevel ocean, and about 1000 years for the deep ocean. The phenomenon of a dynamical system that takes time to reach its equilibrium state starting from an unbalanced initial state is called a *spin-up problem*. Thus, the ocean part of the coupled model for seasonal forecast must be first spun up to the present-day state, which is in equilibrium under the present-day atmospheric forcings. Since the spin-up of the ocean model consumes large computer resources, many strategies have been formulated to accelerate the spin-up process.

Another notorious problem of the coupled model is its tendency to drift to the states of atmosphere and ocean that are unlike the present states during a long time integration. In the noncoupled system, the boundary conditions act as an anchor to the long-term trend of the model and the system cannot drift too far away from reality. By contrast, in the coupled system, because of the imperfection of the system, both the atmospheric and ocean models tend to drift toward their equilibrium states, which are sometimes far away from reality. This is called *climate drift* and is a cousin of the spin-up problem. In order to deal with this climate drift problem, techniques have been developed. They are based on the practice of *flux correction*, which makes *a posteriori* adjustments of various fluxes at the air-sea interface in order to prevent climate drift. Unfortunately, the correction terms are not necessarily small compared with the fluxes themselves. Therefore, the need of flux correction is indicative of the shortcoming in handling the interface conditions.

One promising approach to combat with climate drift is to apply the so-called *flux coupler*, which is the agent to specially handle the exchanges of various fluxes at the

earth's surface, including both air-sea and air-land interfaces, and make sure that conservative properties for momentum, heat, and fresh water are maintained as they are exchanged between various model components of the coupled model. One example of the need for special care is that the interfacial fluxes must be calculated on the finest of the model grids, which is usually the ocean model grid. The interfacial atmospheric fluxes on the coarse grid mesh are formulated to match with those on the finer ocean grid. Although the flux coupler does not eliminate the systematic errors in the component models, it does eliminate any inconsistency in exchanging the flux information at the interface of component models.

The coupled model prediction requires ocean initial conditions as well as atmospheric initial conditions. The ocean initial conditions are prepared by ocean data assimilation using observed ocean data in much the same way as is done for the atmospheric initial conditions, although the number of ocean observations is far fewer than that of atmospheric observations. In order to catch the signal by seasonal prediction, the model needs to be run in an ensemble mode. In the seasonal prediction, disturbances having a period less than 10 days are regarded as *meteorological noise*, since day-to-day variations are not predictable beyond the predictability limit of about 10 days. Main components of these disturbances are lows and highs traveling in the extratropics, which are major players in the general circulation of the atmosphere. We clearly need those players, but their noisy behaviors must be filtered out by appropriate averaging to capture the slowly varying trend of the signal.

Major operational centers have begun to experiment with coupled models to aid in issuing 1-month forecasts. The coupled models are run for a period of 3 to 6 months in an ensemble mode, having as many as 31 members. Unlike medium-range predictions, the choice of initial perturbations for ensemble forecasts is not very crucial and, in fact, initial conditions are taken from daily analyses 12 or 24 hours apart. Although a 5- to 10-member ensemble is often used in practice, a larger number of ensemble members is needed by a factor of 10 or more, if the forecast statistics of higher moments (such as variance and skewness) are required. Mostly for economic reasons, seasonal forecasts are made using relatively low-resolution (200–400 km grid) models. Although successful El Niño forecasts with coupled models are sometimes reported, the models still have systematic errors that may arise from shortcomings in coupling not only between air and sea, but also between air and land. There are studies suggesting the importance of land soil wetness, snow cover, and sea ice distribution for seasonal prediction. Because a long time integration is required, some of the physical processes that are considered to be minor in medium-range forecasts

become no longer negligible. Thus, in order to exploit the potential of dynamical seasonal prediction, the coupling should be extended to hydrology, biosphere, and sea-ice models. The dynamical seasonal prediction is still in an early stage of practical application, but the benefit from improvement of the coupled models will be enormous.

D. Severe-Storm Forecasts

Short-range forecasts up to 4 days aim at prediction of weather-making disturbances, such as fronts, thunderstorms, tornados, severe storms, and tropical storms, in contrast to medium-range forecasts that handle prediction of weather-carrying global scale planetary (Rossby) waves. However, traditionally the same model formulation based on the hydrostatic assumption (Section III.B) has been used for both short- and medium-range predictions, except that short- range prediction models adopt finer spatial and temporal resolutions of grids and have more detailed physics in limited domain than the medium-range prediction models in global domain. A typical limited-area model uses the horizontal grid of 10–30 km, while a global model usually uses 100 km.

With the increased capability of computers in both speed and memory, there is a trend to use higher grid resolution models. It appears difficult to successfully emulate subgrid-scale motions, those phenomena occurring within the grid increment of numerical model, through physical parameterization (Section IV.G). Even for medium-scale disturbances, such as fronts and tropical storms on the order of 1000 km in scale, the source of energy to drive the systems often resides in clusters of mesoscale convection on the order of 10 to 100 km in scale. Therefore, it is more straightforward to adopt finer grids of 10 km or less and to include explicit microphysics for dealing with cloud and precipitation processes.

For prediction models with a grid of 10 km or less, the hydrostatic assumption [Eq. (16)] is no longer justified and the vertical acceleration term in the vertical equation of motion must be retained as in Eq. (11). A rule of thumb for the importance of the nonhydrostatic term can be indicated by the square of the aspect ratio (depth over width) of motion. Since the depth of weather-making systems is about 10 km, the hydrostatic assumption still holds at a 30-km grid. However, the hydrostatic assumption deteriorates rapidly as the grid becomes finer. Once the acceleration term is retained, the dynamical model permits the propagation of acoustic waves that travel faster than the speed of sound in the vertical direction as well as in the horizontal. Therefore, it is impractical to use an explicit time integration scheme for the nonhydrostatic models. Instead, we must resort to the semiimplicit and, possibly, semi-Lagrangian integration approach (Section V.D). One

advantage of the nonhydrostatic formulation is the relative ease of handling the model's lateral boundary conditions due to the fact that the model is mathematically well posed in comparison with the limited-area hydrostatic model formulation.

Even only a decade ago, one of the obstacles of severe-storm numerical forecasts was the lack of observations for preparation of initial conditions. Also, appropriate data assimilation techniques with the nonhydrostatic models for analysis of unconventional data and remote sensing measurements mentioned in Section VI did not exist. There have been significant advances in both aspects in the past 10 years.

As far as the observational aspect is concerned, it is now possible, at least over the United States, to utilize a variety of instrument platforms and data, including rain gauge, Next Generation Doppler Weather Radar (NEXRAD), wind-profiler, and Aircraft Communication Addressing and Reporting System (ACARS) observations, in addition to traditional upper-air and synoptic records and various satellite measurements. This enables us to prepare initial conditions for a grid of 10 km with the aid of 4DVAR, adopting a nonhydrostatic model itself as the prediction model in data assimilation systems (Section VII.C). However, it is essential to include in the prediction model detailed cloud microphysics to deal with the mechanism of cloud and precipitation formation. It turned out that the analyses and forecasts are rather sensitive to such parameterizations. Therefore, the use of nonhydrostatic models does not solve all the problems of regional scale forecasting without further research and sufficient computer resources.

IX. COMPUTATIONAL ISSUES

The computational requirement for the time integration of prediction equations depends on the resolution of numerical model. Here, the *resolution* is a measure of model's spatial or temporal precision. For the spatial resolution, two measures are customarily used. One is the incremental distance of a finite difference grid. The other is the highest wavenumber used to represent a global field in terms of the spherical harmonics using a triangular truncation. For example, the T170 resolution means that there are 170 wave components along latitude circles with comparable meridional components. For this representation, we need a longitude and latitude lattice of at least 340 by 170 grid points. Actually, physical processes of a T170 model are ordinarily calculated at a much higher resolution of 512 by 256 points in a longitude and latitude lattice, called a Gaussian grid, to allow accurate nonlinear term evaluations. We are not too far off by saying that the T170 resolution corresponds to a grid resolution of 80 km.

The spatial resolution of a model in the vertical direction is normally designated by the number (integer L) of vertical levels or layers necessary to cover from the earth's surface to the model top. Thus, L42 means that the model has 42 vertical layers. In order to estimate the computer speed necessary to perform a 10-day medium-range forecast, let us take a T170L42 model as an example. For this model the number of grid points that covers the entire globe and depth of the atmosphere is about 2.4 million ($=340 \times 170 \times 42$). The number of prognostic variables is four (i.e., u , v , T , and q) for the hydrostatic model. Thus, the total number of discretized dependent variables is about 10 million. Now, a typical time step for this model would be 5 min. This means that nearly 3000 repeated calculations are needed to make one 10-day forecast ($=10 \text{ d} \times 24 \text{ h} \times 60 \text{ min}/5 \text{ min}$).

To perform one time step integration, we need to evaluate every term in Eqs. (8)–(12). The computation involves evaluation of the dynamical terms (pressure gradient, Coriolis and advection terms, etc.) and so-called physics Q and F terms (Section IV). Actually, the physics part takes more time to calculate than the dynamics part. This is the case for an advanced weather model, and this trend is increasing. As a very rough estimate, these parts require about 2500 floating point (FLOP) calculations per dependent variable per grid point. Thus, a total of 7.5×10^{13} FLOP calculations ($=2500 \times 3000 \times 10^7$) are required to produce one 10-day forecast. For operational forecasts, because it takes several hours to assemble, record, and analyze global observations, it is necessary to complete a 10-day forecast run within 1 h to meet the deadline for distribution of forecast products to weather services. This time constraint dictates that the computer speed be on the order of 2×10^{10} FLOP per second ($=7.5 \times 10^{13} \text{ FLOP}/(60 \text{ min} \times 60 \text{ s})$) or 2×10^{10} FLOP per second ($=7.5 \times 10^{13} \text{ FLOP}/(60 \text{ min} \times 60 \text{ s})$) or 20 gigaFLOPS.

Late in the 1950s, a typical horizontal grid resolution was 500 km and the speed of computers available in those days, such as the IBM 704, was roughly 0.04 MFLOPS (megaFLOPS). The resolution of operational models was increased to 400 km in the 1960s, to 300 km in the 1970s, to 200 km in the 1980s, and to 100 km in the 1990s. The increase of model resolution was made possible owing to a continued increase in the speed of supercomputers, which were developed one after another incessantly, such as the IBM 701, 704, 7090, 7094, and 360/195, followed by the CDC 6600, 7600 and a series of scientific computers developed by Seymour Cray (1925–1996) and his associates, such as the Cray 1 and Cray YMP. Use of a higher model resolution enables us not only to resolve more detail of motions, but also to incorporate more physical processes and to refine data analysis procedures. This has led to increased skill of numerical prediction

(Section VIII.A), and the computer speed became synonymous with an index of forecast skill. As a rule of thumb, when the grid sizes are reduced by one-half, the computational requirement increases by a factor of $2^4 = 16$, because the number of grid points is doubled in three space directions and the number of time steps also has to be doubled.

The order of magnitude increase of computer speed per decade during the past half-century was achieved by various technological inventions. At the beginning, the speed-up of computers was done by increasing the speed of processors. The invention of a vector processor contributed significantly to further increase the effective processor speed. The next step to increase the speed was to use multiple processors concurrently. This increased the speed by a factor of 10 or more, but a further speed-up was found difficult because of limitations resulting from all processors sharing the same memories. The latest approach reduces this limitation by using distributed memory attached to each processor while each memory communicates only when it is necessary. Some distributed memory machines use a very large number of processors, on the order of 1000 or more, to compensate for slow processor speed (Massively Parallel Processor machines). There is another type of distributed memory machine that utilizes a relatively small number of processors, but each processor is made of a very fast vector-processing unit. The last type has been demonstrated to be much more efficient and easy to use than the former, but is known to be costly.

The methods of achieving speed-up of supercomputers are beginning to put a heavy burden on users. As the machines evolve from vector processor, to shared-memory machine, to distributed-memory machine, the design and coding of prediction models becomes time consuming and manpower intensive. In order to make the model run efficiently on a distributed-memory machine, it is necessary to decompose the entire model domain into smaller subdomains and assign each subdomain to a different processor. Since communication is generally slow, the decomposition is designed to minimize communications between the processors. For the finite-difference models, the domain is divided into volumes with nearly equal numbers of grid points. Communication between the processors is necessary when finite-difference calculations are performed near the boundaries of subdomains (halo regions). Thus, the strategy for finite-difference models is relatively straightforward. Spectral models can also be adapted to distributed-memory machines, but a different strategy is needed. The spectral method requires conversion of data from spectral space to grid-point space and vice versa. This process involves summation of data along all latitude and longitude circles. The decomposition is performed in such a way that the summation can be done within one pro-

sor without communication. Communications are needed to transform the arrays of data from one configuration to the other to make such computation possible.

Unless computers were designed specifically to solve weather and climate problems, as happened to be done during the past four decades, the emergence of distributed-memory machines will have a profound impact on the way weather prediction models are designed and coded. We must take into account the architecture of a particular computer in designing the numerical methods for efficient arithmetic and logical operations. This is why the idea of using various polygonal grids that cover the globe with quasi-equal areas (Section V.E) has been brought back recently, because polygonal grids are more adaptable to distributed-memory machines. However, we must rethink not only the numerical methods for fluid dynamics, but also the way in which parameterizations of physical processes, such as radiation and moist convection, are formulated and coded.

The resolutions of medium-range forecast models in various operational centers in the world vary from country to country and are dependent on the computer power available. Currently, the U.S. National Weather Service is running a T170L42 model at the National Centers for Environmental Prediction (NCEP) using a massively parallel 384 processor IBM-SP machine with a peak speed of 690 gigaFLOPS. The peak speed is a theoretical limit and the model actually runs with a speed of 5–10% of the peak. This effective speed is referred to as sustained speed, and 30 gigaFLOPS is quoted for this machine. By the fall of 2000, the NCEP global model resolution will be increased to T254 when the computer will be upgraded to a peak speed of 2.5 teraFLOPS and deliver sustained 150 gigaFLOPS. One teraFLOPS is equivalent to 1000 gigaFLOPS (=1 trillion FLOPS).

At the European Centre for Medium-Range Weather Forecasts (ECMWF), the resolution of the medium-range forecast model in early 1999 is T319L50. In order to run this model in routine, a computer is required that delivers a sustained speed of at least 160 gigaFLOPS [$=((319/170)^3 \times (50/42) \times 20)$]. ECMWF uses a 116-vector parallel-processor Fujitsu VPP 700 computer with a sustained speed of 100 gigaFLOPS, which will have its augmentation to yield 400 gigaFLOPS by August 2000. In order to run such a high-resolution model routinely, various numerical expediences, including reduction of grid points near the poles and application of a semi-Lagrangian advection scheme (Section V.D), are adopted to speed up the model run. ECMWF is currently testing a T639 model as a future operational model. Such a model requires a sustained speed of 1 teraFLOPS for routine use.

Table IV shows the resolution of medium-range forecast models and the type of mainframe computers used

TABLE IV Resolution of Medium-Range Forecast Models and Type of Mainframe Computers Used at Operational Centers of Various Countries in the World in Early 1999

Country	Model resolution	Computer
Australia	T239L29	NEC SX-4 16CPU
Canada	Variable resolution min. 22 km L28	NEC SX-4 32CPU
ECMWF	T319L50	Fujitsu VPP 700 116CPU
France	T199L31	Fujitsu VPP 700E 26CPU
Germany	60 km L31	Cray T3E 1200CPU
Japan	T213L30	Hitachi S-3800/480 4CPU
USA (NCEP)	T170L42	IBM-SP 384CPU
United Kingdom	60 km L30	Cray T3E 840CPU

T170 resolution roughly corresponds to 80-km grid resolution. Increase in the T number corresponds to a proportional decrease in grid distance.

[Data source: WMO Technical Document Nos. 964 and 968.]

at operational centers of various countries in the world as of early 1999. Since each operational center has a plan to increase the model resolution accompanied by the augmentation of computing power in the future, this table should be regarded only as a snapshot.

So far, we presented the computational requirement for numerical prediction based on the computer time necessary for a 10-day forecast with a global model. In reality, this is only a portion of the operational use of computers. Comparable or even greater computational resources are needed to perform the analysis of input data with data assimilation systems (Section VII). Although the 4-D variational data assimilation has been demonstrated to produce a superior analysis, it is very time consuming to run the system and full implementation of the 4DVAR system awaits further augmentation of computer power. Beside medium- and short-range forecasts, many operational centers are producing ensemble forecasts and the computer time required is proportional to the number of ensemble members. Model development and improvements require running a large number of real data cases to see their impacts and performances before the model is put into operation. Thus, the science and the future of numerical weather prediction depend entirely on available computer power and resources.

X. FUTURE OUTLOOK

A. Utilization of Forecast Products

Outputs from numerical prediction models are the 3-D distributions of model-dependent variables at future times. In order for this information to be useful to users, many

more steps are required. To begin with, it is necessary to translate model-dependent variables into information that users can understand and that allows them to make use of the model output. For example, for general users the final product is the statement of future weather conditions, sunny, cloudy, rainy, etc., preferably with their probabilities. Some quantitative information, such as the amount of rain or snow, maximum and minimum temperatures, and winds, is also important. For more specific users, the predicted variables must be tailored to their needs. Farmers, forest managers, health care organizations, manufacturers, armed forces, and transportation, utility, and insurance industries all have their own specific requirements of which the producers of the forecast and the users must be aware. One notable example is that the issuance of evacuation orders to the inhabitants of the projected landfall location of a hurricane or typhoon requires accurate probabilistic prediction of the trajectory, wind field, and intensity of the storm.

The most commonly used method to translate the model output to more user-friendly products is called Model Output Statistics (MOS). This method utilizes statistical relationships between some of the predicted variables and observed weather elements at verifying stations. The statistical technique is intended to correct the model's systematic errors and enhance the value of forecasts by incorporating certain features of weather unique to specific geographic locations. Weather services want to automate to produce forecast products without human intervention, and MOS is suited for this purpose. However, MOS requires a very large sample, of the order of years, to obtain stable statistical relationships. The statistical method also requires that the model be unchanged during the period of application. This conflicts with the interest of modelers, who want to improve the models whenever they can. Therefore, there is a continued struggle between modelers and MOS developers.

Although the MOS technique is usually applied to enhance the value of output from a single model, the same idea can be extended to the outputs from the prediction models of different operational centers, i.e., multimodel output statistics. Because each prediction model has a certain bias and produces unique model errors, it is possible to correct the forecast by combining various multimodel outputs. Optimal weights to the outputs of different prediction models are determined in such a way as to minimize the sum of root-mean-square differences between model predictions and verifying analyses. Again, a large sample of outputs from each prediction model is necessary for the period of at least 6 months to obtain stable statistics. However, it has been reported that even a small number (fewer than six) of model ensembles can produce an enhanced forecast that is superior to the performance

of any member model in both short- and medium-range forecasts.

In addition to the statistical translation of model output, the forecasters in weather services apply subjective judgments to the model forecasts. In the early days of numerical prediction, subjectively corrected forecasts often beat the score of pure model products. The subjective judgments were based on past performance of the model and comparison with other model forecasts. However, the involvement of forecasters in the daily forecast process is decreasing steadily with time. This is due to improvement in the short- to medium-range prediction systems and the need for more research and development for long-range and seasonal forecasts as well as for severe storm forecasts.

B. Linkage to Climate Prediction

As John von Neumann stated at a meeting in 1955, it is instructive to divide the problem of atmospheric prediction into three categories depending on the time scale of prediction. In the first category, the prediction is primarily determined by the initial conditions. This is the case of short- to medium-range forecasts, which are the main topic of this article.

In the second category, we have the opposite situation, namely the case in which the prediction does not depend very much on the initial conditions. Motions in this very long time scale are determined mainly by the balance of external forcings, resulting from the differential imbalance between incoming solar radiation and outgoing terrestrial infrared radiation. Of course, no motion is ever completely independent of initial conditions, but what is meaningful is the ensemble of deterministic runs, called the climate, resulting from averaging of the weather in time and/or ensemble. Actually, there is little distinction in the way two different categories of prediction problem are formulated, since both problems are based on thermohydrodynamic principles. There are important differences, however, in the emphasis of modeling, such as more detailed physics considerations and long-term integration in the second category, whereas more attention is paid to the preparation of initial conditions in the first category.

The potential of numerical weather prediction technique to study the climate of the atmosphere was recognized in early days of numerical weather prediction when Norman A. Phillips succeeded in 1956 an extended time integration of a two-level quasi-geostrophic model, starting from an atmosphere at rest. He demonstrated that many features of the general circulation of the atmosphere could be simulated by the model. Since then many general circulation models (GCMs) have been started to develop based on the primitive-equation model formulation (Sec-

tion III.B). This activity was stimulated partly by the need for reliable global prediction models for planning to conduct the Global Weather Experiment in 1979 under GARP (Section VI) and partly by the concern for studying the impacts of anthropogenic activities, such as the possibility of atmospheric contamination by nuclear debris and the emission of greenhouse gases, on the atmospheric climate.

In parallel with the atmospheric GCM (AGCM) activity, modeling of the oceanic general circulation models (OGCMs) has been developed in varying degrees of complexity. Although the strategy of coupling AGCM with OGCM is still under scrutiny (Section VIII.C), the capability of coupled atmosphere–ocean–land–sea ice models to study the long-term variability of atmospheric climate has been well demonstrated. The impacts of minute changes in the atmospheric concentration of greenhouse gases on atmospheric climate can now be detected through controlled numerical experiments with the coupled models. One interesting feedback from operational weather prediction practice to the climate modeling community is the recognition of the importance of ensemble climate “predictions” to measure the reliability of climate impact projection.

Now, between the two extremes, as von Neumann said, there is the third category of prediction problems. In this category, we are sufficiently far from the initial state that details of the initial conditions have not much impact on what has developed. Nevertheless, certain features of the initial conditions bear considerable influence on the form which the circulation takes. The seasonal forecasting discussed in Section VIII.C over a period of 30 to 180 days falls into this category. The prediction problem at this intermediate time scale is most difficult to solve, because we must pay attention on both major features of the initial state and details of the physical processes in the earth environment system.

C. Future Opportunities

Before 1955 when the first operational forecast was produced at the U.S. Weather Bureau (later the National Weather Service), weather forecasts were done subjectively. Progress has been made steadily since then as measured by a continued improvement in the skill of deterministic medium-range forecasts, which now approaches the predictability limit. Efforts are underway to tackle the most difficult problem of seasonal predictions as well as to refine the prediction of regional scale disturbances, such as severe storms and tropical cyclones.

Meteorology is an old science, and the mathematics of atmospheric motions have been known for many centuries, but our ability to perform the time integration of

atmospheric equations did not exist until the middle of the 20th century. Clearly, the advent of electronic computers changed the face of weather prediction from an empirical art to a modern quantitative science. Unlike traditional physics and chemistry, in which major discoveries are made through laboratory experiments, observations were the only means in the study of meteorology. During the early 20th century, there was hope that the general circulation of the atmosphere could be understood through laboratory experiments using a rotating dishpan filled with water that is heated at the outer rim and cooled at the center. Although these experiments have helped in understanding the nature of thermal convection, dissimilarities in fluid, size, thermal stability, and rotation between the apparatus and the atmosphere, without mentioning the important role of water vapor, have prevented this approach from being the main road to studying the atmospheric general circulation.

The first weather satellite, TIROS I, was launched in 1960 and revolutionized the way we observe the atmosphere. Prior to the satellite era, atmospheric observations were limited and there were large gaps in our view of global circulation systems. However, the meteorological observations from satellites are so different from the traditional synoptic and radiosonde observations that the potential of satellite data have not been fully exploited until recently, when 3-D and 4-D variational data assimilation techniques became operational (Section VII.C). Data assimilation requires an accurate short-term forecast as the background in which observations are blended. Moreover, a global prediction model requires fine resolution in both horizontal and vertical directions to translate what the satellites "see" in terms of model variables. There is a synergy between improvement in the quality of data analysis and refinement of the prediction model that is used in data assimilation. This synergy has contributed to produce today's improved operational center forecasts.

As the data assimilation techniques improve, the resulting atmospheric analysis became more and more useful for meteorological applications beyond numerical weather prediction. Particularly, the importance of the global data has been recognized by the climate research community. The dynamically consistent and complete four-dimensional atmospheric global data at regularly spaced grid points are found to be essential in performing diagnostic and statistical studies. However, those users who started to use long series of analyses quickly realized that the analyses produced by operational centers have a major problem. Namely, changes in the modeling system from time to time introduced undesirable temporal discontinuities in the resulting analyses.

Recognizing the need for uniform quality historical analyses for climate studies, efforts have been under-

taken at major operational centers to produce historical analyses using a fixed data assimilation system to eliminate discontinuities. The reanalysis was first performed at NCEP, jointly with NCAR, for 50 years of data with their T62L28 model in 1998, and ECMWF also completed a 20+-year reanalysis with their T106L31 model. ECMWF is currently preparing a 40+-year reanalysis with an improved assimilation system. Although these efforts removed discontinuities in the analyses due to changes in their modeling systems, it became clear that changes in the observational systems due to introduction of meteorological satellites must also be accounted for in the reanalysis system, and this problem remains as a future challenge.

The reanalysis provided an excellent opportunity to examine the progress of numerical prediction and the relative role played in improvement of the observational system and the forecast system (modeling and data assimilation) during the past 40 years. A study suggests that about 25% of the gain in forecast skill during this time period comes from improvement in the observational system, while the remaining 75% is contributed by upgrading the forecast system by improving analysis techniques, model resolution, numerics, physics, etc. In other words, if the current technique of numerical prediction and the present computer power had been available in the 1960s, it would have been possible to produce weather forecasts almost as accurate as the forecasts made in the 1990s.

By now it is clear that this spectacular progress in our ability to predict weather and climate would not have materialized without the support of electronic industries, which provided not only number-crunching mainframe supercomputers, but also data storage and faster communications. Further progress in the science of predicting weather and climate depends critically on the future availability of supercomputers that can deliver a sustained speed of 20 TFLOPS. There are some prospects on the horizon that may lead to development of computers with this capacity. In the United States, the Accelerated Strategic Computing Initiative (ASCI) is leading an effort to construct two 40-TFLOPS (peak) machines by 2003. In Japan, the Earth Simulator Project supported by the Science and Technology Agency is developing a computer system that will deliver a peak speed of 40 TFLOPS and is scheduled to be completed by NEC in 2002. The system consists of a cluster of 640 nodes, and each node has 8 processors. The peak speed per node is 64 GFLOPS. These projects will, it is hoped, further encourage scientists to solve complex problems in many other physical science disciplines, thereby stimulating commercial development of supercomputers. As the science of weather and climate prediction enters the 21st century, it is gratifying to expect that future advances are still forthcoming in our quest to understand and predict the environment of the earth where we live.

ACKNOWLEDGMENTS

We conducted this research at the National Center for Atmospheric Research (NCAR), which is sponsored by the National Science Foundation, and at the National Centers for Environmental Prediction (NCEP), National Oceanic and Atmospheric Administration. Partial support was provided by the Central Research Institute of Electric Power Industry in Japan. We thank Roger Newson of the World Meteorological Organization, who provided materials for international comparison of models and computers used for medium-range operational forecasting; John Derber of NCEP, who supplied the data counts of NCEP operations; and David Burridge and Anthony Hollingsworth of the European Centre for Medium-Range Weather Forecasts for their permission to use Fig. 4 in the text. We also thank Joseph Tribbia of NCAR for his stimulating discussion on this subject, and Ronald Errico and David Baumhefner, both of NCAR, and Richard Anthes, University Corporation for Atmospheric Research (UCAR), for their useful comments on an earlier version of this manuscript. Prof. George W. Platzman, University of Chicago, who is one of the pioneers in this field and participated in the Meteorology Group at the Institute for Advanced Study in Princeton, kindly read our manuscript and gave us valuable comments for which we are grateful. The manuscript was typed by Barbara Ballard.

SEE ALSO THE FOLLOWING ARTICLES

ATMOSPHERIC DIFFUSION MODELING • CLOUD PHYSICS
 • MESOSCALE ATMOSPHERIC MODELING • OCEAN-
 ATMOSPHERIC EXCHANGE • PLANETARY WAVES (ATMO-
 SPHERIC DYNAMICS) • RADIATION, ATMOSPHERIC • SO-
 LAR TERRESTRIAL PHYSICS • THERMODYNAMICS • TIME
 AND FREQUENCY

BIBLIOGRAPHY

- Bengtsson, L. (1999). "From short-range barotropic modelling to extended-range global weather prediction: a 40-year perspective," *Tellus* **51AB**, 13–32.
- Daley, R. (1991). "Atmospheric Data Analysis," Cambridge University Press, New York.
- Ehrendorfer, M. (1997). "Predicting the uncertainty of numerical weather forecasts: a review," *Meteorol. Zeitschrift, N.F.* **6**, 147–183.
- Ghil, M., and Ide, K., eds. (1997). "Data Assimilation in Meteorology and Oceanography: Theory and Practice," *Meteorol. Soc. Japan* **75**(1B) (special issue).
- Hollingsworth, A., Capaldo, M., and Simmons, A. J. (1999). "The scientific and technical foundation of the ECMWF strategy 1999–2008," European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK.
- Kalnay, E., Lord, S. J., and McPherson, R. D. (1998). "Maturity of operational numerical weather prediction," *Bull. Am. Meteorol. Soc.* **79**, 2753–2769.
- Katz, R. W., and Murphy, A. H., eds. (1997). "Economic Value of Weather and Climate Forecasts," Cambridge University Press, New York.
- Lin, C. A., Laprise, R., and Ritchie, H., eds. (1997). "Numerical Methods in Atmospheric and Oceanic Modelling: The Andre J. Robert Memorial Volume," Canadian Meteorol. and Oceanograph. Soc., NRC Press, Ottawa.
- Randall, D. A., ed. (2000). "General Circulation Model Development," Academic Press, New York.
- Staniforth, A., and Cote, J. (1991). "Semi-Lagrangian integration schemes for atmospheric models—A review," *Mon. Wea. Rev.* **119**, 2206–2223.
- Trenberth, K. E., ed. (1992). "Climate System Modeling," Cambridge University Press, New York.
- Wiin-Nielsen, A. (1991). "The birth of numerical weather prediction," *Tellus* **43AB**, 36–52.