Appendix III.

Atmospheric Dynamics: Some Fundamentals, with Observational Implications

M. E. McIntyre

This Appendix to the GEFD Summer School lectures on "Fundamental Concepts and Processes" is a slightly expanded version of a tutorial article published in Proc. Internat. School Phys. "Enrico Fermi" CXV Course, *The Use of EOS for Studies of Atmospheric Physics,* edited by J. C. Gille and G. Visconti. North-Holland (Amsterdam, Oxford, New York, Toronto), pp. 313–386 (1992). ISBN 0-444-89896-4. The published version omits the background material about stratospheric ozone in section 0. A more thorough discussion of stratosphere–troposphere exchange, including the 'sideways–downward' control of tropical upwelling by extratropical wave driving — the mechanism discussed in §5 below, which we now tend to think of as a kind of '**gyroscopic pumping**' — is available in the review by Holton, J. R., Haynes, P. H., McIntyre, M. E., Douglass, A. R., Rood, R. B. and Pfister, L., 1995: Stratosphere–troposphere exchange, *Revs. Geophys. Space Phys.*, **33**, 403–439. Further information and updates are available from my home page, http://www.atm.damtp.cam.ac.uk/people/mem/

0. – Introduction: the stratosphere and the ozone layer.

This Appendix surveys some fluid-dynamical fundamentals and the scientific reasons for being interested in them, with chemical transport in the middle atmosphere particularly in mind, and stratosphere–troposphere exchange of chemical constituents; but many of the principles involved give insight into the workings of the troposphere as well, and into its links with the middle atmosphere. Some of the material has already appeared elsewhere (*e.g.*, ref. [1–5]), but it may be useful to pull the threads together in one place. First let us note, here and in the next section, a few observed facts of key importance.

The stratosphere, lying roughly between 10 and 50 kilometres altitude, is far more important to us than might be suggested by its relatively small mass, of the order of twenty percent of the total atmospheric mass. It partakes in the complicated ultraviolet, visible and infrared radiative transfer that contributes to shaping our global environment; and we know that its effects on climate, for instance, are not only significant now, but have been drastic from time to time in the past, following massive injections of debris from volcanic eruptions or meteoric impacts. The ozone layer, the bulk of which lies in the stratosphere, allows the biosphere as we know it to exist, by protecting it from photochemically powerful, and hence biologically devastating, solar ultraviolet at wavelengths below about 310 nanometers. Ozone is also one of the greenhouse gases that directly affect climate, by virtue of its infrared radiative properties.

The total amount of stratospheric ozone seems at first sight remarkably small in relation to its biological and human significance. Concentrations in the stratosphere are of the order of a few parts per million only, when expressed as mixing ratios. 'Mixing ratio' means the proportion of ozone in a given air sample, usually measured by volume, or equivalently by number of molecules. (Carbon dioxide mixing ratios, by comparison, are two orders of magnitude greater at a few hundred parts per million by volume, close to 350 at present.) The total amount of ozone in the stratosphere would occupy a layer only a few millimetres thick if isolated and brought to sea-level pressure and temperature. It is the small amount of ozone, together with the possibility of its catalytic destruction by man-made chemicals in still tinier amounts, a few parts per *billion*, i.e. per 10^9, that make the ozone layer potentially far more vulnerable than might once have been suggested by a naive contemplation of the vastness of the sky. A single atom of chlorine, for instance, can be re-cycled again and again to destroy large numbers of ozone molecules, perhaps thousands or tens of thousands of molecules. Just how many, and under exactly what circumstances, is one of the critical questions being researched today. The theoretical possibility of catalytic ozone destruction was first pointed out in the early 1970s; and we now know, as a result of the discovery of the Antarctic 'ozone hole' in the mid-1980s, that it is more than just a theoretical possibility. In at least some circumstances it is now a directly observable reality [6].

What is involved in understanding the behaviour of the ozone layer? Ozone production depends on a highly complicated set of photochemical reactions, requiring for their mathematical description a large set of nonlinear chemical-kinetics equations, already fifty or more in the state-of-the-art

models of a decade or more ago, with well over a hundred chemical species. It depends also, crucially, on the nonlinear fluid dynamics of the atmosphere. The distribution, total amount, and rate of production of ozone predicted from radiation and photochemistry alone differ greatly from the observed distribution and total amount and our best estimates of the actual rate of production. It has long been recognized that the explanation has to be sought in the transport of chemical constituents by fluid motion. Indeed the very fact that ozone is produced at a positive rate in the stratosphere depends upon the existence of fluid motion.

What happens is that ozone is produced in some regions, and transported to others where it has a relatively long photochemical lifetime. This interplay between radiation, chemistry and fluid dynamics depends on the nature of ozone photochemistry. Ozone mixing ratios have a tendency to relax photochemically toward a finite equilibrium value that depends on location; this is to be sharply distinguished from the behaviour of certain other chemical species that depend mainly on total exposure to solar ultraviolet. The photochemical timescales can be either much faster or much slower than fluid-dynamical timescales, also depending on location. Photochemistry is faster at higher altitudes and in more sunlit latitudes. Continuity considerations tell us therefore that transition regions must exist at some altitudes and latitudes, where the chemical and fluid-dynamical timescales are comparable. The rate at which ozone is produced and the ozone layer replenished must depend on the fluid motion in and near such transition regions, particularly on the time taken to cross them. This dependence is modelled only crudely, as yet, in global atmospheric models of the type being used today to make long-term assessments of ozone depletion, a point to which we shall return.

The formation of the Antarctic ozone hole demonstrated the importance of a rather different type of interplay between radiation, chemistry and fluid dynamics, to be mentioned in section 10 below. This too is inadequately represented in present-day ozone assessment models.

These examples illustrate a point that I believe to be relevant in a wider context. It is arguable that an improved scientific understanding of almost any aspect of the behaviour of the atmosphere, all the way down to the thin ($\simeq$ 1–2 km deep) boundary layer in which we live, will require not just supercomputer power but also a radically new and improved set of mathematical modelling *concepts*, not the least of which will be those required to understand and represent quantitatively the roles of fluid-dynamical processes. The atmosphere confronts us with a highly inhomogeneous 'wave-turbulence jigsaw puzzle' inaccessible to conventional turbulence theories. The implied need for radically new modelling concepts poses one of the greatest scientific challenges facing us today; and it is unlikely that such concepts will be developed quickly. In what follows, I shall try to sketch within space limitations a little more of what is already known and understood, and to show more precisely the nature of the challenge. In order to meet it, a sustained research effort involving mathematical, physical and chemical thinking – and lateral thinking – will be required.

1. – Temperatures and stable stratification.

Attempts to observe the stratosphere go back to the discovery of the tropopause by balloon-borne thermometers in the nineteenth century, and to studies of long-distance acoustic wave propagation in the First World War. The latter revealed the existence of a temperature maximum at altitudes near 50 km, now called the stratopause. Today there are a variety of highly sophisticated terrestrial and space-based remote sensing devices and *in situ* balloon, aircraft and rocket techniques. Thanks largely to the global coverage from space-based remote sensors, we have a reliable broad-brush view of the typical global-scale temperature structure, and of the global-scale distributions of several of the important chemical constituents. With the help of insights from fluid-dynamical theory and modelling, and from special observations including those from the recent airborne polar stratospheric expeditions that confirmed the causes of the ozone hole, we also have a fair number of clues about some aspects of the fluid-dynamical behaviour, and how this

affects the chemical behaviour. Refs. [18] and [32] provide useful compendia of the information that was available prior to, and after, the airborne expeditions.

Fig. 1, from a compilation by Barnett and Corney [7] of temperatures retrieved from satellite-borne radiometers, shows a typical mean January temperature structure. The mean involves time averaging over the Januarys of two or three seasons (for precise dataset information see ref. [7,8]) combined with averaging around latitude circles at constant pressure-altitude, this latter being the "Eulerian" zonal mean in the usual meteorological sense. The pressure altitude is shown on the left using a logarithmic scale in hectopascal or millibar (*i.e.* in units of 100 $\mathrm{N/m^2}$), and on the right in e-folding pressure scale heights, corresponding to geometric altitude increments of about 7 km in the stratosphere (one scale height $H_p = RT/g$ being almost exactly 7 km at temperature $T = 239\,\mathrm{K}$ in a hydrostatic, isothermal, perfect-gas atmosphere with specific gas constant $R = 287\,\mathrm{m^2 s^{-2}\,K^{-1}}$ when gravity g is $9.8\,\mathrm{ms^{-2}}$). Thus fig. 1 shows altitudes up to about 85 km, taking in most of the middle atmosphere (stratosphere and mesosphere) as well as the far more massive troposphere. The picture for July, not shown, is roughly a mirror image of fig. 1, the main difference being that the temperature extrema in the wintertime Antarctic stratosphere and lower mesosphere are more pronounced than their Arctic counterparts in fig. 1, reaching 180 K and 270 K at 4 and 8 scale heights respectively.

The corresponding mean zonal winds $\overline{u}$ (not reproduced here; see, for instance, ref. [8]) can be estimated from the mean temperatures $\overline{T}$ shown, extending the wind field upward from the troposphere. The troposphere is relatively well observed by the operational meteorological network, which provides direct wind observations. The upward extension can be computed, for instance, on the assumption that the middle atmosphere is a perfect gas with specific gas constant $R = 287\,\mathrm{m^2 s^{-2}\,K^{-1}}$, and that the mean fields are in approximate hydrostatic and cyclostrophic balance (equivalent to a balance between the centripetal acceleration in a sidereal reference frame and the pressure gradient and gravitational forces). Reference [8] (*q.v.* for a thorough discussion and bibliography) uses a more refined approximation, and shows that the refinement does not greatly affect the results. The idea of "balanced motion" and its generalizations is of great importance in atmosphere-ocean dynamics; we shall encounter it again in sect. 8. In the case corresponding to fig. 1, the zonal winds relative to the Earth show broad, planetary-scale jets peaking at magnitudes of about $\overline{u} \simeq 60\,\mathrm{ms^{-1}}$ to $70\,\mathrm{ms^{-1}}$ in middle latitudes at altitudes of 9 to 10 scale heights, the sense being westerly in the northern, winter hemisphere (meaning eastward or prograde, by convention), and easterly in the summer hemisphere (meaning westward or retrograde). Zonal wind speeds such as $70\,\mathrm{ms^{-1}}$ may be compared with the Earth's rotational speed, for instance $327\,\mathrm{ms^{-1}}$ at 45° latitude. A useful exercise in order-of-magnitude analysis is to check that these magnitudes are indeed roughly consistent with fig. 1 and with hydrostatic and cyclostrophic balance on the rotating Earth.

One of the most conspicuous features in fig. 1 is the stratopause temperature maximum (picked out by the light shading near 7 scale heights or 50 km). This feature is always strongest at the summer pole (the sign of the associated horizontal temperature gradient accounting for the signs of the zonal-wind jets) and is largely due to the absorption of solar UV (ultraviolet) radiation by ozone (O_3), whose partial pressure is sufficient at these altitudes for it to begin to intercept a substantial fraction of the incoming UV photons (*e.g.* ref. [9], fig. 4.3, 4.5). Temperatures $\overline{T}$ near the summer polar stratopause are thought to be fairly close to the temperature T_{rad} that would be calculated from radiative transfer alone (*e.g.* fig. 2 of ref. [10] in this volume; also *e.g.* ref. [11–14]). Just how close is still a matter for debate, but an educated guess would probably put it at 10 K or closer.

It should be noted incidentally that the maximum with respect to latitude of the diurnally averaged solar irradiance $\bar{S}$ does, in fact, occur at the pole in midsummer. It is a straightforward exercise in vector analysis to show this, approximating the Earth as spherical and using the fact that the tilt α of the Earth's axis is about 23.6° or 0.41 radian. The main steps are first to show that the fractional length of day λ, *i.e.* the proportion of the 24-hour day for which the sun is above

the horizon, depends upon latitude ϕ, in the northern midsummer case, according to

$$\lambda(\phi) = \pi^{-1} \arccos[\max\{-1, \min(1, -\tan\alpha \tan\phi)\}]. \tag{1}$$

and then to show that the diurnally averaged vertical component of solar irradiance is the full solar irradiance ($S_\perp = 1368 \pm 2\,\mathrm{Wm}^{-2}$ [15]) multiplied by

$$\bar{S}/S_\perp = \lambda(\phi)\sin\alpha\sin\phi + \pi^{-1}\sin\{\pi\lambda(\phi)\}\cos\alpha\cos\phi\ . \tag{2}$$

This function is plotted in fig. 2a. It has 3/2 power behaviour at the Arctic and Antarctic circles ($\pm$ 66.4°) and hence a continuous first derivative. The secondary maximum would disappear if the Earth's tilt α were made greater than about 25°. A plot of the complete seasonal evolution of the curve in fig. 2a is given in fig. 2b, from ref. [16]. This evolution can be computed by replacing α in eqs. (1) and (2) by the appropriate solar declination or 'effective tilt angle' α_{eff} ($|\alpha_{\mathrm{eff}}| \leq \alpha$). This is such that $(\frac{1}{2}\pi - \alpha_{\mathrm{eff}})$ is the angle between two vectors one of which gives the direction of the Earth's axis, i.e., points toward the pole star, and the other points from the Earth toward the Sun at the time of interest. A simple approximation to α_{eff}, idealizing the Earth's orbit as circular, would be $\alpha_{\mathrm{eff}} = -\arcsin\{\sin\alpha\cos\Delta t\}$ where Δt is the time after the December solstice measured in units of (1 year)/(2π).

A dynamically important consequence of the vertical temperature distribution is the well-known stable stratification, or static stability, of the atmosphere. Most of the atmosphere is stably stratified, *i.e.* exhibits stability to adiabatic, quasi-static (nonacoustic) vertical displacements. A natural measure of this static stability is the buoyancy frequency (Brunt-Väisälä frequency) N, defined by

$$N^2 = g\,\partial(\ln\theta)/\partial z\ , \tag{3}$$

where z is the geometric altitude, g the gravitational force per unit mass, and θ the potential temperature. The potential temperature θ is the temperature that an air parcel would have if compressed adiabatically to a nominal sea-level pressure p_0, and is, therefore, a Lagrangian or material invariant for dry adiabatic motion; p_0 is usually taken as 1000 hPa, but sometimes, for historical reasons, as 1013 hPa. It is a straightforward exercise to show from the diatomic-perfect-gas laws, again taking R =287 $\mathrm{m^2 s^{-2}\,K^{-1}}$ and taking specific heats in the ratio $\gamma = 7/5$, that θ depends on T and pressure p according to

$$\theta = T(p_0/p)^\kappa\ , \tag{4}$$

where $\kappa = R/c_p = (\gamma - 1)/\gamma = 2/7 = 0.286$, c_p being the specific heat at constant pressure. It follows that in a hydrostatic atmosphere the quantity N^2 is positive, *i.e.* the atmosphere is stably stratified, whenever T either increases with height, or decreases more gently than the "adiabatic lapse rate" g/c_p = 9.8 K/km. To show that $\partial T/\partial z = -g/c_p$ corresponds to $N^2 = 0$, use the hydrostatic relation in the form

$$\partial(\ln p)/\partial z = -g/RT \tag{5}$$

to deduce from (3) and (4) that

$$N^2 = \frac{g}{T}\left(\frac{\partial T}{\partial z} + \frac{g}{c_p}\right)\ . \tag{6}$$

N^2 can be shown to be just the Archimedean restoring force per unit mass per unit displacement of a vertically displaced, thermally isolated air parcel in pressure equilibrium with its surroundings. N^2

is positive everywhere in fig. 1, by a good margin. In fact most of the atmosphere is stably stratified nearly all the time, by this criterion. It is noted in passing, however, that the stable stratification of the tropical troposphere is tightly controlled by deep tropical cumulonimbus convection, so that the tropical troposphere is also close to neutral *moist* static stability [17].

The stratosphere, as its name suggests, has the strongest stable stratification found anywhere in the atmosphere, in this coarse-grain, global-scale view. As fig. 1 shows, the temperature increases upwards for the most part, making $\mathbf{N}^2$ larger than anywhere else. The constant-θ or isentropic surfaces are very close to being horizontal, with slopes $\lesssim 10^{-3}$. The term "isentropic" arises from the fact that a constant-θ surface is also a surface of constant specific entropy $s(\theta) = c_p \ln \theta + \text{const.}$ Typical buoyancy periods $2\pi/N$ work out to be four or five minutes in the stratosphere. This means that the Archimedean restoring force must be thought of as an immensely strong constraint on adiabatic vertical motion, whenever we are dealing with air motions on time scales of days, weeks or longer. This includes most types of motion having direct relevance to the global-scale advective transport of chemical constituents.

An important dimensionless measure of the stratification constraint is the Richardson number

$$\mathrm{Ri} = N^2 (\partial \overline{u}/\partial z)^{-2} \ , \tag{7}$$

which in the stratosphere has typical values of tens or hundreds. It can be shown that this implies that three-dimensional turbulent mixing across isentropic surfaces — for instance, the mixing of chemical constituents — is likely to be a very sporadic and intermittent process, and that, at a given instant, large portions of the stratosphere are probably not undergoing such mixing at all. Such a picture is consistent with observation, and is not altogether surprising when one recalls the smoothness of passenger jet flight in the lower stratosphere, punctuated only occasionally by encounters with turbulence. It represents an important difference between the real atmosphere and most numerical models of it, which for reasons of numerical stability assume a fictitious, all-pervasive eddy diffusivity.

2. – Temperature anomalies and CFC lifetimes.

Some of the features in fig. 1 are less easy to explain than the stratopause temperature maximum — indeed some are quite surprising. For instance the coldest place in the atmosphere — indeed the coldest place on Earth, by far — is found in the summer polar cap (see also [18–21]). It occurs at the summer mesopause, at altitudes near 85 to 90 km, as suggested by the dark shading indicating temperatures as low as 150 K at the top left of fig. 1. Even 150 K is nowhere near the lowest temperature observed, being only a statistical average about which there are quite large fluctuations. Individual rocket soundings have occasionally recorded temperatures as low as 110 K or –163°C (*e.g.* [20,21]). The extreme cold is one of the reasons for the formation at altitudes near 83 to 85 km of the world's highest clouds, the "noctilucent clouds" that can be seen after midnight on some clear nights in July and August, from latitudes north of about 50°, as an electric blue glow above the northern horizon (*e.g.* [22,23]), and also from space as "polar mesospheric clouds" extending over the polar cap (*e.g.* [24]). These tenuous, stratiform clouds are believed to be made of ice crystals. The conditions that lead to their formation appear also to lead to an anomalous radar reflectivity that has often been observed at nearby altitudes, probably indicative, in one way or another (*e.g.* [25,26] and references therein), of the presence of large, heavily hydrated "cluster ions".

One does not need to be an expert on atmospheric radiative heat transfer to see that these low mesopause temperatures are remarkable in themselves. Indeed, they used to be regarded as one of the major enigmas in atmospheric physics [27]. They occur despite the large midsummer solar irradiances implied by fig. 2, and despite the additional heating by infrared radiation from the

warm and relatively dense stratopause below (see, *e.g.*, ref. [28], fig. 2.19). In fact, a temperature $T = 110\,\mathrm{K}$ is probably of the order of a hundred degrees lower than T_{rad}. Despite some uncertainties in the quantitative calculation of T_{rad}, in part because of departures from local thermodynamic equilibrium, there is little doubt that T_{rad} does behave qualitatively as one might guess, being lowest at the winter pole and highest at the summer pole. Calculations of T_{rad} at 12 scale heights predict a latitude distribution somewhat like that shown at the top of fig. 1, but with the sign of the latitude reversed. (Again see fig. 2 of ref. [10] in this volume, or the corresponding information on pp. 280, 617, 281 of ref. [11,14,18] respectively.) Thus the winter pole, also, is the site of a major temperature anomaly at 12 scale heights, with $\overline{T}$ departing significantly from T_{rad} in the opposite sense, $\overline{T} > T_{\mathrm{rad}}$. The winter polar temperature anomaly extends much further down, into the stratosphere; and in the case of the Arctic middle atmosphere, at least, this winter anomaly is very substantial at all altitudes. Even the coldest stratospheric $\overline{T}$ values shaded at the right of fig. 1 are several tens of degrees warmer than T_{rad}; and the anomaly increases with altitude to an order of magnitude approaching, again, values of the order of a hundred degrees warmer than T_{rad} in the mesosphere. In the Antarctic winter in July, the pattern is similar except that the mesospheric warm anomaly is stronger and the stratospheric warm anomaly weaker.

The tropical lower stratosphere is a more subtle case. Here it is less useful to think of T_{rad} as given, as we have so far been tacitly doing, if only because the effects of radiative heat transfer cannot be separated, even approximately, from the powerful constraint imposed by deep moist tropical-cumulonimbus convection upon vertical temperature profiles in the troposphere just below, and the consequent direct effect upon tropopause height [17], along with the strong radiative effects of upper-tropospheric cloudiness. But it will be seen in what follows that the observed mean stratospheric temperatures must, in fact, be somewhat lower than their radiative values for given tropospheric conditions.

Let us note one other observed fact, at first sight unconnected with the foregoing, but actually, as we shall see, very closely connected. This is the now-notorious longevity of man-made chlorofluorocarbons, or CFCs, in the troposphere. CFCs are the highly stable, non-toxic, industrially and domestically useful chlorine-containing substances now implicated as a major factor in the formation of the Antarctic ozone hole, and as a possible threat to the entire ozone layer. The tropospheric mixing ratios of two of these CFCs in particular, known as CFC-11 and CFC-12 (with respective chemical formulae $CFCl_3$ and CF_2Cl_2), have been building up over the past two decades at rates between about 4% and 6% per year to values of the order of a fraction of a part per billion by volume each, which is at least two decimal orders of magnitude above the detection limits of the global monitoring system [29,30]. In addition, we have a rough idea, from the chemical industry's statistics, of the rates at which the CFC-11 and CFC-12 must have been leaking into the troposphere [18,31–33]; and we know from laboratory and field measurements that only a tiny fraction is disappearing into the oceans — probably 0.2% per year or less (*e.g.*, [34,35]). The implication is that almost all the CFC-11 and CFC-12 released into the troposphere is still there. If one summarizes the information just referred to in terms of a notional atmospheric e-folding "lifetime" τ for each substance, then the numbers that come out are so large, in comparison with a decade, that their precise magnitudes are rather ill defined especially for CFC-12. The τ values are estimated to be very roughly of the order of a century, somewhat less for CFC-11 and somewhat more for CFC-12 (*e.g.*, [18–32]). Because the lifetimes τ are defined in an e-folding sense, this means that it would take something of the order of *several* centuries to get rid of most of the present atmospheric burden of CFC-11 and CFC-12, even if the leakage of these substances into the troposphere could somehow be stopped immediately.

3. – The mean circulation, and "total-exposure constituents" *vs.* "local-relaxation constituents".

How are the long atmospheric lifetimes of CFC-11 and CFC-12 connected with the observed pattern of middle-atmospheric temperature anomalies $\overline{T} - T_{\rm rad}$? It is here that the fluid dynamics begins to enter the picture.

For reasons to emerge, there is a systematic, highly persistent global-scale mean mass circulation in the stratosphere and mesosphere. Figure 3 shows a modern estimate [36] of the circulation, for January 1979 but thought to be broadly typical of other northern-hemispheric winters. Along with similar estimates given in ref. [37,38], it is based largely on a uniform dataset, in this case data from the LIMS infrared radiometer on board the Nimbus 7 satellite. Other datasets and combinations thereof have been used in recent years to produce independent estimates [14,39,42]; all the estimates differ from each other in points of detail, but show much the same large-scale pattern. References [14,38] give the most recent discussions of the differences and the sources of uncertainty, and the assumptions involved in the calculations. The heavy dashed line extending above fig. 3 is schematic only, and indicates the qualitative sense of the mesospheric mean circulation, deduced from observational and theoretical evidence other than that in [36] (see, *e.g.*, [18,28,43]). Although observations are much sparser in the mesosphere, the gross sense of the circulation seems to be in no real doubt, taking all the evidence into consideration, and is qualitatively in accord with the sense deduced in a pioneering study by Murgatroyd and Singleton [43]. On a more quantitative and detailed level, there is considerable interannual variability in the strength and shape of the circulation, especially in the winter stratosphere. In the case of January 1979, for instance, the stratospheric circulation was somewhat stronger than average, probably by a factor of order 1.5 to 2.

It is largely the mean circulation that connects the long atmospheric lifetimes of CFC-11 and CFC-12 with the observed global-scale departures from $T_{\rm rad}$. Indeed the latter form the starting point for all the estimates cited above, although the circulation can also be estimated in other ways [10], and is a feature of the behaviour of global general-circulation models [18,44] as well as that of the real atmosphere. Where the air is descending on average, as in the polar night, it is being powerfully compressed and is trying to warm adiabatically, pulling $\overline{T}$ above $T_{\rm rad}$. Where the air is rising, as in the tropical lower stratosphere and in the summer polar mesophere, the associated expansion and adiabatic cooling pulls $\overline{T}$ below $T_{\rm rad}$. It makes sense to talk about "pulling" $\overline{T}$ away from $T_{\rm rad}$ because, broadly speaking, radiative heat transfer behaves somewhat like a relaxation or Newtonian cooling process in which $\overline{T}$ would, under radiation alone, relax exponentially towards $T_{\rm rad}$. As Fels [11] puts it, radiation in the middle atmosphere acts somewhat like a heavily-damped "spring". The calculations from which fig. 3 and similar estimates were obtained are based on this idea, but use quantitatively accurate radiation models rather than simple Newtonian cooling. The relevant e-folding relaxation time scale $\tau_{\rm rad}$ is of the order of weeks in the lower stratosphere and days near the stratopause, increasing again towards the mesopause. The effectiveness of the vertical motion in pulling $\overline{T}$ away from $T_{\rm rad}$ is related to the fact, already mentioned, that the atmosphere is stably stratified.

The same mean circulation is a crucial factor in the vertical advective transport of chemical substances. For instance, the noctilucent and polar mesospheric clouds mentioned earlier, and probably the anomalous radar echoes also, depend not only on the adiabatic cooling of the summer mesopause by the upwelling branch of the mean circulation, but also on a supply of water vapour from below, carried by the same upwelling (*e.g.*, [22,45,46]). Similarly, the upward mean motion in the tropics is the main route by which tropospheric air, and its chemical constituents including CFC-11 and CFC-12, are carried high into the stratosphere. The chemical stability that makes substances like CFC-11 and CFC-12 so convenient for industrial and domestic use also means that the main place where they are destroyed at a significant rate is at middle-stratospheric or higher altitudes, above 25 km or so, where solar UV photons are sufficiently energetic to break the tight molecular bonds photolytically. (It is the resulting molecular fragments, containing chemically active forms of chlorine, that can catalytically destroy ozone.) Absorption into the ocean is distinctly slower,

probably by a factor of five or more [35]. Thus it is mainly the strength of the mean circulation that controls the rate at which the atmospheric burdens of CFC-11 and CFC-12 are being reduced.

We note in passing that, contrary to what has sometimes been suggested, the chemistry of the Antarctic ozone hole itself [31,32,47] exerts no direct control over CFC-11 and CFC-12 destruction rates. This is because the ozone hole occurs below 25 km and in polar latitudes, well out of reach of CFC-destroying UV photons in winter and spring when ozone-hole chemistry operates. Under these circumstances CFC-11 and CFC-12 are chemically inert. Indeed we would expect nearly all the unphotolyzed CFC-11 and CFC-12 molecules that reach the ozone hole to be carried intact back down into the troposphere since, as fig. 3 illustrates, the mean circulation has a persistently poleward–downward sense in the extratropical lower stratosphere. As we shall see in sect. 9, there are strong theoretical as well as observational reasons to expect a lower-stratospheric circulation in this sense, connected with the properties of a basic fluid-dynamical quantity known as *potential vorticity* and an associated low-frequency wave phenomenon known as *Rossby-wave motion*.

The time scale for a notional air parcel to rise through the tropical stratosphere is typically of the order of two years (see, for instance, ref. [39]). This is entirely consistent with the observed CFC-11 and CFC-12 e-folding lifetimes of the order of a century since, although a typical journey through the stratosphere may take only several years (see also [48]), the typical altitude difference between stratosphere and troposphere is of the order of three to four scale heights, corresponding to density ratios of the order of e^{-4} to e^{-3} or two to five times 10^{-2}. The upshot is that, for this purpose, the troposphere acts approximately like a large, well mixed reservoir, of whose total mass only about one percent per year circulates high enough for the CFC-11 and CFC-12 to be photolyzed. Exactly how high that needs to be varies with the chemical stability of the substance in question; the altitudes required are a little higher, for instance, for CFC-12 than for CFC-11, consistent with the somewhat longer lifetime of CFC-12.

There are other long-lived trace chemical constituents whose behaviour in the stratosphere is essentially similar. One important example is the naturally occurring constituent nitrous oxide (N_2O). Along with CFC-11 and CFC-12 it enters the atmosphere at the Earth's surface, and is well mixed in the troposphere, with a mixing ratio close to 307 p.p.b.v. [49], as measured in 1989. Nitrous oxide is destroyed irreversibly in the middle and upper stratosphere by two UV-related processes: direct photolysis, and reaction with excited atomic oxygen, itself produced by photolysis. The stratospheric destruction rate corresponds to a tropospheric e-folding lifetime of roughly one and a half centuries [18,31], even longer, probably, than that of CFC-12.

Nitrous oxide, CFC-11 and CFC-12 are all examples of what might be broadly characterized, to a first approximation, as "total-exposure constituents". The word "exposure" is used here in the same sense as in photography. The mixing ratios of such constituents in a given air parcel depend mainly on the total past exposure of the air parcel* to sufficiently energetic UV photons, and relatively little on details of the history of the exposure. In the case of nitrous oxide, CFC-11 and CFC-12, mixing ratios decrease with increasing exposure, relaxing toward zero just as the emulsion on a photographic plate relaxes toward its fully exposed state. This photochemical behaviour is to be sharply distinguished from that of ozone, which might be called a "local-relaxation constituent" inasmuch as it tends to relax toward a finite, nonzero equilibrium mixing ratio that depends strongly on location, especially height and latitude. The relaxation time scales are themselves strongly location-dependent [9,50], varying (in the case of ozone) from an hour or less near the sunlit stratopause to a year or more in much of the lower stratosphere, except when ozone hole chemistry

* Note added in proof: Strictly speaking, the notions of "air parcel" and "air parcel history" should be replaced by the notions of "typical molecule" and "transport history". Here "transport" denotes the way in which molecules are carried around in the atmosphere not only by large-scale advection, but also by small-scale mixing, whose combined effects can destroy the identity of a macroscopic air parcel. See ref. [57,58], and the cautionary remarks on p. 10154 of ref. [188].

operates. Such characterizations are very highly idealized, and can claim no more than rough qualitative validity; for more accurate yet simplified modelling of standard (non-ozone-hole) ozone chemistry, for instance, the reader may consult, for instance, ref. [51], and for the state of the art of sophisticated chemical modelling, taking into account the perturbations due to active forms of chlorine and other species, ref. [9,18,31,32,52]. But the simplistic "total-exposure" *vs.* "local-relaxation" idealizations correctly point to some potentially important consequences for the way in which the air motion affects the chemistry.

One consequence is that, unlike the mixing ratios of nitrous oxide, CFC-11 and CFC-12, ozone mixing ratios depend hardly at all on total exposure. Another is that they are likely to depend more strongly on air parcel histories (more correctly, transport histories), particularly on what happens during a typical molecule's last journey, if any, across a region of comparable photochemical and fluid-dynamical time scales. This has implications, not yet well understood, for the magnitude and interannual variability of the rate at which ozone is produced naturally (*e.g.*, [53]). It likewise has implications for the legitimacy of using comparisons between measurements of ozone and nitrous oxide, for instance, as evidence of non-standard chemistry leading to ozone depletion. These points will come up again, in sect. 10, after discussing the phenomenon of *Rossby-wave breaking*.

4. – Definitions of "mean", and the three-dimensional diabatic mass circulation.

Despite these caveats about the consequences of local-relaxation behaviour, the idea of vertical advective transport by a mean circulation like that in fig. 3 appears to go a long way towards explaining the gross, global-scale features of many observed constituent distributions, the tropospheric lifetimes of CFCs being merely one aspect. The observations range from those of pioneers like Brewer and Dobson in the first half of this century, who first hypothesized a lower-stratospheric circulation like that in fig. 3 (*e.g.*, [27] and references therein), all the way to modern remote and *in situ* observations using increasingly ingenious and sophisticated instrumentation. A notable recent example is the work of Fahey, Loewenstein and their co-workers on the measurement of nitrous oxide and its photolysis products from the NASA ER-2 aircraft at lower-stratospheric altitudes $\lesssim 18\,\mathrm{km}$ (*e.g.*, [54,55] and references therein). Many other examples, including data from high-altitude balloons and from terrestrial and space based remote sensors including the LIMS instrument already mentioned, are reviewed in refs. [18,31,32]. The mean-circulation picture is especially persuasive in the case of total-exposure constituents, because it seems reasonable to hypothesize that, by following the mean motion, one obtains a good first approximation to the actual cumulative exposure of air parcels even though this simplistically ignores eddy motions hidden from view by the averaging used in fig. 3. The hypothesis receives support, for instance, from a comparison presented in ref. [36] between SAMS satellite observations and photochemical modeling results for two total-exposure constituents, nitrous oxide and methane, using the circulation estimate shown in fig. 3 and its counterparts for other months of the year, and largely ignoring eddy motions. Indeed it is clear from the LIMS and SAMS data alone that the tropical upwelling suggested by fig. 3 has a major role in constituent transport; one can see the plumes of several total-exposure constituents ascending in the tropics, just as suggested by fig. 3.

It is important to note, on the other hand, first that not all aspects of the observed constituent distributions can be thus explained, even in the case of total-exposure constituents (*e.g.*, [18,36,41,56–58]) — another point to which we shall return when eddy motions are discussed in sect. 9 and 10 — and second that the foregoing picture is not even qualitatively correct unless a suitable definition of "mean" is used for the global-scale mass circulation. If, for instance, the traditional Eulerian zonal mean at constant pressure-altitude is used, then the wintertime circulation pattern looks entirely different from that shown in fig. 3, and the overall picture far less simple, even as regards total-exposure constituents. Eddy advective fluxes associated with departures from zonal symmetry then become as important as mean advective fluxes, both in the thermal and in

the chemical budgets, and there is much cancellation between the mean and eddy contributions — implying that both must be accurately estimated even for a qualitatively correct first approximation to the net effect.

This and related points are touched on in ref. [10] in this volume, and further information can be found in reviews such as those in ref. [28] and in chapt. 6 and 12 of ref. [18]. The definition used as the conceptual basis for the calculation in fig. 3 is the isobaric "transformed Eulerian mean" or "residual" circulation (see ref. [10] in this volume and the other references just cited). Under the conditions of interest here, which include the fact that $\tau_{\rm rad}$ values are typically smaller than the time scale for seasonal change [14], this residual circulation is closely related to various definitions of the two-dimensional "diabatic circulation", such as the "isentropic Eulerian mean" described, *e.g.*, in ref. [28,38,59,60], and also related to the Eulerian "transport circulation" of ref. [44,61] (see also [18], chapt. 6). The technicalities involved in these definitions, their approximate interrelation, and their relation to the concepts of Lagrangian mean and Stokes drift — which latter concepts are not always directly applicable for reasons discussed, *e.g.*, in ref. [18,61,62] — are fascinating to the theoretician but beyond the scope of the present discussion. The interested reader may consult the above-mentioned reviews and, for recent insights into the differences inherent in the traditional Eulerian, isentropic Eulerian, and Lagrangian viewpoints, ref. [38], sect. 8 of ref. [4], and ref. [63] in this volume. This last (also [64–66]) describes a new and important application of Lagrangian mean and related concepts to the detailed analysis of polar-stratospheric chemical data, which promises among many other things to help clarify some currently controversial questions (*e.g.*, [46,67]) regarding mean descent within, outside and near the edge of the wintertime stratospheric polar vortex.

Notwithstanding the technicalities, however, the leading requirement for a relevant definition of the global-scale mean mass circulation is to some extent conceptually simple, and easy to state. The two most critical considerations are first the strong altitude dependence of solar UV fluxes, and second the powerful dynamical constraint on vertical motion represented by the stable stratification. Arguably the most important thing, therefore, is to define the mean vertical motion such that it relates as closely as possible to the exact *three-dimensional diabatic mass circulation*, or rate at which air parcels and their chemical constituents ascend or descend diabatically relative to the isentropic or constant-θ surfaces marking the stable stratification — regardless of how the latter surfaces may move up or down as a result of eddy motion that may well be relatively rapid and involve relatively large, but temporary, adiabatic vertical displacements. (It is these adiabatic displacements that underlie the near-cancellation of mean and eddy fluxes in the traditional Eulerian-mean description, *e.g.*, [18,41,62,68].) Some first estimates of three-dimensional diabatic circulations from observations and from a general-circulation model have recently been obtained by, respectively, Pawson and Harwood [38] and Thuburn [69]. Devising and acquiring good quantitative estimates of the three-dimensional diabatic mass circulation, and relevant mean representations of it — along with other statistical properties relevant to "total-exposure", "local-relaxation", and more complicated types of chemical behaviour — will be one of the most critical challenges to the next generation of three-dimensional numerical modeling and data assimilation studies in the runup to EOS.

5. – Mean circulation dynamics, "downward control" and "stratosphere-troposphere exchange".

What maintains the mean circulation illustrated in fig. 3? In particular, what causes the persistent, constituent-transporting, cross-isentropic motion we have been talking about, approximately summarized by the mean vertical motion in fig. 3? One sometimes hears the answer "diabatic heating, of course." But does it make sense to regard the diabatic heating as the *cause*? The balance of terms in an equation is not generally the same thing as a causal link (and it is particularly dangerous to argue for causality from only one out of the complete set of governing equations of

a system). In this connection it will prove instructive to consider, in sect. 6, a "perpetual-winter" thought-experiment in which some extra diabatic cooling is artificially applied to the winter polar stratosphere. We shall see that (other things being equal, in a sense to be explained) the extra diabatic cooling does not cause any permanent increase in mean descent rates [40]. Indeed, in this particular case, the opposite tends to occur. Under reasonable assumptions, one finds that mean descent rates ultimately settle down to slightly *smaller* values than before.

In order to understand this, and other aspects of the circulation dynamics, it is essential to consider first the fact that the Earth is a rapidly rotating planet. The rotation is rapid in the sense that the angular momentum of the atmosphere is dominated by the contribution from the Earth's angular velocity, and is relatively little affected by the air motion relative to the Earth. (Recall the relative zonal-wind magnitudes mentioned in sect. 1.) It follows that, except for a region within the tropics, the mean angular momentum $\overline{m}$ per unit mass always has a strong latitudinal gradient. This is a robust feature, only minutely affected by changing the definition of "mean" and, what is more important, practically certain to survive any foreseeable global change.

Figure 4 shows the latitude-height distribution of $\overline{m}$, defined as the isobaric Eulerian mean, for the January conditions corresponding to fig. 1. The contribution from zonal winds relative to the Earth bends and displaces the isopleths of $\overline{m}$ to some extent, but not nearly enough to change their quasi-vertical orientation in the extratropics, seen in a vertically stretched coordinate system. The implication is that, in any zonally averaged description, a mean circulation such as that of fig. 3 can persist in the extratropics if, and only if, the extratropical mean state feels a persistent torque, Γ say, about the Earth's axis. Otherwise the horizontal branches of the mean flow, required by mass conservation, cannot continue to cross the isopleths of $\overline{m}$.

The existence and robustness of the implied causal link between this mean torque Γ and the mean circulation is clearly seen in analytical and numerical model experiments ([40] and references therein) that solve the equations of zonally symmetric fluid motion on the rotating Earth, with Γ prescribed as a function of latitude ϕ, altitude z and time t. An equation describing the thermal relaxation of mean temperatures $\overline{T}$ towards $\overline{T}_{\rm rad} = \overline{T}_{\rm rad}(\phi, z)$ has to be included. For any realistic initial $\overline{m}$ distribution the result is always the same: as long as Γ persists, the mean circulation persists and the distributions of $\overline{T}$ and $\overline{m}$ remain realistic. Note incidentally that, except in the summer mesosphere, the required sense of Γ is negative, *i.e.* against the Earth's rotation. If Γ is turned off, the circulation begins to die away. $\overline{T}$ then begins to approach $\overline{T}_{\rm rad}$ on a time scale related to the radiative relaxation time scale $\tau_{\rm rad}$, while the $\overline{m}$ distribution begins to look increasingly unrealistic as unrealistically strong zonal winds develop (see also [70] and references therein). This underlines the point, already evident from fig. 3 and 4, that something must, on average, be exerting the required torque Γ on the real atmosphere.

Before turning to the question of what that something might be — although one can easily guess that zonally averaged eddy effects must be involved — we need to say a little more about the extratropical response to a prescribed Γ distribution. Calculating that response poses an interesting problem in partial differential equations involving both elliptic and hyperbolic behaviour, in the following sense brought out by an idealized example. Figures $5a, b$ show schematically the early and late stages of the incremental response to a change $\Delta\Gamma$ in Γ made at, say, time $t = 0$, as predicted by "quasi-geostrophic theory" [28,71]. $\Delta\Gamma$ is taken to be nonzero within the shaded region only. Quasi-geostrophic theory approximates the extratropical $\overline{m}$ distribution somewhat crudely, but qualitatively correctly, by the contribution to it from the Earth's angular velocity alone. (Considerations of balance show that this approximation is consistent with the fact, mentioned earlier, that Richardson numbers Ri are typically large.) The change in the mass streamfunction satisfies an elliptic equation in the limit $t \to 0$ and, as sketched in fig. $5a$, spreads out somewhat like a dipolar electrostatic potential from the forcing region. This is the classical "Eliassen problem". The analogy with electrostatics is imperfect, but qualitatively reasonable when viewed in a vertically stretched coordinate system in which the aspect ratio of the stretching is of the order of Prandtl's

ratio

$$N/f \gtrsim 10^2 \ , \tag{8}$$

where N as before is the buoyancy frequency characterizing the stable stratification, and $f = 2\Omega \sin\phi$, the Coriolis parameter, here evaluated at a representative middle latitude ϕ, with $\Omega = 0.7292 \times 10^{-4}\ \mathrm{s}^{-1}$, the Earth's angular velocity. There is a tendency for the response in the stream-function to be stronger below the forcing region, especially when the horizontal scale on which $\Delta\Gamma$ varies $\gtrsim NH_p/f \gtrsim 10^2 H_p$, in round numbers 10^3 km or more [40,72,73].

Fig. 5*b* shows schematically the limit $t \to \infty$. Except in a frictional boundary layer near the Earth's surface, this limiting case is governed by a hyperbolic equation, whose characteristics are approximately vertical when viewed in the stretched coordinate system. The streamlines follow the characteristics down into the frictional boundary layer, which (*a*) closes the circulation, and (*b*) permits a steady-state angular momentum balance to be attained.

The mathematical theory, including a full discussion of the time-dependent problem describing the transition from fig. 5*a* to fig. 5*b*, is described in ref. [40], with explicit illustrations both from analytical theory and from numerical models. As well as depending on τ_{rad}, the evolution again depends on the horizontal scale, being slower for the smaller scales $\lesssim 10^3$ km. In the limit $t \to \infty$, the domain of influence of the forcing region extends entirely downward, as suggested by fig. 5*b*. In a theoreticians' bottomless atmosphere, the change in circulation burrows downward for ever!

In a description more accurate than quasi-geostrophic, the long-time behaviour is similar in important respects, the limit $t \to \infty$ again being governed by a hyperbolic equation along whose characteristics the streamlines extend downward from the forcing region. In an exact description the characteristics are just the $\overline{m}$ isopleths. The only new feature is that, during the time-dependent adjustment, the bending and displacement of the $\overline{m}$ isopleths changes slightly as the angular-momentum distribution changes, an effect neglected in the quasi-geostrophic theory and involving upward as well as downward propagation of information. The final, steady-state relation between the Γ distribution and the mean circulation is still simple, however, provided that we write it in terms of the total torque Γ rather than the change $\Delta\Gamma$. It can be expressed in the isobaric transformed Eulerian-mean (TEM) formulation, or equally well in the isentropic Eulerian-mean formulation, as follows. Denoting the mean vertical velocity as $\bar{w}^*$ in either case, we can show that in the steady state

$$\bar{w}^*(\phi, z) = \frac{1}{\bar{\rho}(\phi, z)\cos\phi} \frac{\partial}{\partial\phi} \left\{ \int_z^\infty \left\{ \frac{\bar{\rho}(\phi', z')\Gamma(\phi', z')\cos\phi'}{\overline{m}_\phi(\phi', z')} \right\} \Bigg|_{\phi' = \Phi(z'; \phi, z)} dz' \right\} \tag{9}$$

anywhere above the frictional boundary layer, in the extratropics ([40], sect. 2). The function $\Phi(z'; \phi, z)$ gives the latitude, at altitude z', of the (nearly vertical) $\overline{m}$ isopleth or characteristic passing through latitude ϕ and altitude z. Thus the integration is carried out along the characteristics, rather than exactly vertically. The vertical coordinate z is chosen in a manner appropriate to whichever formulation, TEM or isentropic Eulerian mean, is being used, and $\bar{\rho}$ is the mass density in the chosen coordinate system; in the TEM system $\bar{\rho}$ depends on z alone. For convenience we may take $z = H_{\mathrm{s}} \ln(p_0/p)$ in the TEM formulation (as in fig. 3) and $z = \kappa^{-1} H_{\mathrm{s}} \ln(\theta/\theta_0)$ in the isentropic Eulerian-mean formulation, H_{s} being a standard, representative value of the pressure scale height, usually taken as 7 km exactly, and p_0 and θ_0 nominal sea-level values of pressure p and potential temperature θ. This makes z roughly equal to geometrical altitude in each case; exact equality would hold in an isothermal, hydrostatic atmosphere at temperature gH_{s}/R ($= 239$ K when $H_{\mathrm{s}} = 7$ km). The symbol $\overline{m}_\phi$ means the latitudinal derivative of $\overline{m}$ at constant z, representing the latitudinal angular-momentum gradient, and $\overline{m}_\phi(\phi', z')$ means the same quantity evaluated at

latitude ϕ' and altitude z'. Note that $\overline{m}_\phi$ is negative in the northern hemisphere, the dominating contribution from the Earth's rotation being $-a^2 f \cos\phi$, with $f = 2\Omega \sin\phi$ at the relevant latitude ϕ. Like $\overline{m}$, Γ is reckoned per unit mass. Γ can be taken as the Earth's mean radius $a = 6371\,\mathrm{km}$ times $\cos\phi$ times the mean zonal force per unit mass at latitude ϕ and altitude z. This entails the inessential, but traditional, meteorological approximation (*e.g.*, [74], p.17) of using $a\cos\phi$ in place of the exact distance to the Earth's axis, incurring a minor error of order 1%. The mean mass density $\bar{\rho}$ falls off exponentially with z, by a factor of order e for each pressure scale height, which turns out to be enough to ensure the convergence of the integral at its upper limit; this point is further discussed in ref. [2,46]. Given that the integral converges and that the steady state is reached, relation (9) follows directly from the angular-momentum equation together with a steady-state mass conservation equation of the form

$$\frac{1}{a\cos\phi}\,\frac{\partial(\bar{\rho}\bar{v}^*\cos\phi)}{\partial\phi} + \frac{\partial(\bar{\rho}\bar{w}^*)}{\partial z} = 0\ , \tag{10}$$

as shown in [40]. The adiabatic heating or cooling implied by (9) pulls mean temperatures $\overline{T}$ away from $\overline{T}_{\mathrm{rad}}$ in the manner already described. It is the main contribution, as seen in both the TEM and the isentropic Eulerian-mean formulations, to what is sometimes called the 'dynamical heating' of the mean atmosphere [11,75].

To the extent that the $\overline{m}$ and $\bar{\rho}$ distributions are regarded as given, (9) tells us that the $\bar{w}^*$ distribution at a given altitude depends on the Γ distribution only above that altitude, and not below it, always assuming that the time scale is long enough — say seasonal or longer, depending on precisely what horizontal scales and other aspects are of interest [40]. This does not, of course, say that all causal linkages are downward. It is only the long-time-average control of the extratropical mean circulation by the torque Γ that is exerted downward, in the sense just described. Other causal linkages are also of interest: besides the weak upward feedback on $\overline{m}$ itself there remains, for instance, the question of what causes Γ. We shall see that the latter involves a very strong upward linkage on all time scales of interest, related to the upward propagation of certain wave motions.

Nevertheless, (9) already conveys useful insight into what is involved in determining the strength of the circulation through the stratosphere, and hence, for instance, in determining CFC and nitrous oxide lifetimes and the rate at which photolyzed stratospheric material is supplied to the troposphere. Such questions are often discussed under the heading "stratosphere-troposphere exchange". It is sometimes thought that rates of exchange of air and chemicals between stratosphere and troposphere are bound up mainly with what goes on near the tropopause; but the relevance of Γ values well above the tropopause, made evident by the relation (9), warns us that such a viewpoint would be misleading. It would be almost like supposing that the mean rate of outflow of water from a garden tap or faucet, dripping through a damp cloth interposed between the tap and the ground, can be changed by moving the cloth around. One can certainly change the locations where water drips onto the ground, and temporarily change the rate at which this happens — and both things may be of interest in themselves — but one cannot thus control the total, long-time-mean outflow rate.

Reference [40] further discusses these issues, and notes the implications for the vertical extent (see also ref. [76]) of numerical models whose global circulation, and stratosphere-troposphere exchange rates, are required to be causally as well as numerically correct. Included in [40] is an attempt to estimate from observational data the number of scale heights that contribute significantly to the right-hand side of (9) in the real atmosphere. For instance, to approximate (9) to within 20% of its true value at $z \simeq 18\,\mathrm{km}$, it appears from the data that the vertical integration starting at $z \simeq 18\,\mathrm{km}$ must often be taken up to the stratopause, and sometimes even higher (*e.g.* in the southern polar winter). To obtain CFC-11 destruction rates correct to 20%, it would be more relevant to take the lower limit of integration around $z \simeq 25\,\mathrm{km}$, implying an even higher upper limit,

especially when it is recalled that the relevant UV fluxes increase with altitude. The corresponding requirements for CFC-12, nitrous oxide and certain other total-exposure stratospheric constituents, such as methane, would be likely to be even more stringent, increasingly so in the order just stated. Such requirements are related to the vertical excursions of streamlines in plots like fig. 3. For these chemical purposes it is not the total stratosphere-troposphere mass exchange rate that is relevant, but rather the part of it that circulates high enough. It would be interesting to see the sensitivity of results from a sophisticated two-dimensional photochemical modelling package to various choices of the upper limit of integration in (9).

6. – The "polar cooling" thought-experiment.

Relation (9) also helps us to discern what happens in the perpetual-winter thought-experiment mentioned at the beginning of the last section, in which extra diabatic cooling is applied to the winter polar stratosphere, "other things being equal". We take the latter phrase to mean (*a*) that the comparison is to be made at a given pressure altitude z, *i.e.* with a fixed mass of air above the location at which descent rates are to be compared, and (*b*) that the Γ field is to be held fixed. It has been argued (*e.g.* [11,77,78]) that the magnitude of Γ might well, in fact, diminish in such a situation; but holding it fixed is enough to make the point.

We shall focus on the northern winter and content ourselves with a comparison of mass-weighted mean descent rates integrated over an Arctic polar cap bounded by latitude $\phi = \phi_{\rm c}$, say, in other words downward mean mass fluxes in the polar cap $\phi \geq \phi_{\rm c}$, at a given altitude $z = z_{\rm c}$. We assume furthermore that the cooling is applied over a large depth and that it has the effect of making $\overline{T}_{\rm rad}$ decrease at all altitudes in the polar cap. Then the time-dependent theory shows that, not surprisingly, $\overline{T}$ follows $\overline{T}_{\rm rad}$ and decreases also. The westerly circumpolar zonal winds then become stronger than before, since hydrostatic and cyclostrophic balance and a colder polar cap can be shown to imply that the westerlies increase more strongly with altitude from the surface, where friction keeps wind speeds relatively small. The strengthening of the circumpolar westerlies means in turn that the isopleths of $\overline{m}$ crowd slightly closer together in some range of circumpolar latitudes, one of the small, finite-Richardson-number effects neglected in quasi-geostrophic theory, making $|\overline{m}_\phi|$ slightly larger than before. We assume that $\phi_{\rm c}$ can be chosen such that the $\overline{m}$ isopleth $\phi = \Phi(z; \phi_{\rm c}, z_{\rm c})$ passing through latitude $\phi = \phi_{\rm c}$ at altitude $z = z_{\rm c}$ lies within the region where $|\overline{m}_\phi|$ has increased.

Since $z_{\rm c}$ is to represent a given pressure altitude in (9), it is simplest to use the version of (9) based on log-pressure coordinates and the TEM description. In the numerator of (9) we then have $\bar{\rho}(\phi', z') = \bar{\rho}(z')$, as already remarked; in fact, the function $\bar{\rho}(z)$ that appears in (10) and, therefore, in (9) is by definition proportional to $\exp(-z/H_{\rm s})$ in this case. Multiplication of (9) by $\bar{\rho}$ and integration over the polar cap $\phi \geq \phi_{\rm c}$, with respect to the area element $dA = 2\pi a^2 \cos\phi \, d\phi$, gives the vertical mass flux,

$$\iint\limits_{\text{polar cap}} \bar{\rho}(z_{\rm c}) \bar{w}^*(\phi, z_{\rm c}) \, dA = 2\pi a^2 \int\limits_{\phi_{\rm c}}^{\pi/2} \bar{\rho}(z_{\rm c}) \bar{w}^*(\phi, z_{\rm c}) \cos\phi \, d\phi$$

$$= -2\pi a^2 \int\limits_{z_{\rm c}}^{\infty} \left\{ \frac{\bar{\rho}(z') \Gamma(\phi', z') \cos\phi'}{\overline{m}_\phi(\phi', z')} \right\} \Bigg|_{\phi' = \Phi(z'; \phi_{\rm c}, z_{\rm c})} dz' \ . \tag{11}$$

We note again that $\overline{m}_\phi$ and Γ are both negative, in the northern winter hemisphere. Thus the last integral is positive, consistent with mean descent in the polar cap. Now in the thought-experiment

the only change in that integral is the slight increase in $|\overline{m}_\phi|$ already noted. Hence, as was asserted, the result of applying the extra cooling is to make the mass-weighted mean descent rate settle down to values slightly smaller than before.

There is nothing surprising about this, except from the viewpoint that diabatic cooling "causes" diabatic descent. The example emphasizes that such a viewpoint would be misleading (as has long been recognized by dynamicists [11,18,28,79]) not least when trying to understand long-term circulation trends on a rapidly rotating planet. In this particular thought-experiment, it happens that the effect of applying the cooling is to push the system into a state in which TEM descent rates, hence local dynamical heating rates, are slightly reduced. In other words the dynamical feedback is slightly positive, under the assumed conditions, with the consequence that actual mean temperatures $\overline{T}$ diminish even further toward $\overline{T}_{\rm rad}$ than would have been the case with fixed dynamical heating.

7. – The origin of the torque Γ.

In the Earth's atmosphere the torque Γ that sustains the mean mass circulation arises from the systematic, irreversible transport of angular momentum by certain fluid-dynamical wave motions that disturb the mean state illustrated in fig. 1, 3 and 4. Here our understanding is far less complete, for reasons that will emerge; but it appears that for the most part the waves in question propagate or diffract upwards from the troposphere, and dissipate in the middle atmosphere [18,28]. This is the predominantly *upward* causal link already mentioned.

It is a basic rule in physics that wave propagation and diffraction are generally accompanied by a flux or transport of momentum, and hence of angular momentum. The essential effect is the setting-up of a "radiation stress" that acts to transport momentum between regions of wave generation and regions of wave dissipation [68,80–82]. The last stage of the process is sometimes loosely referred to as "momentum deposition", as if the waves themselves possessed momentum that is deposited at the location where they dissipate; but for waves in a material medium it is more generally correct ([83] and references therein) to speak of "momentum flux deposition". As L. Brillouin [80] wrote long ago, "it is not ultimately the density of momentum that matters, but rather the *flux of momentum*. This latter may well differ from zero even when the density of momentum is zero." [My translation, but his italics.]

There are many beautiful experimental demonstrations of this type of wave-induced momentum transport, including some that require no special apparatus. For instance, anyone can easily do the experiment shown in fig. 6*a*, taken from ref. [4] (see the caption to fig. 6 for more detail). The cylinder, or any other anisotropic wavemaker, is oscillated fairly rapidly, say at 5 to 6 Hz, so as to radiate capillary-gravity waves more strongly in some directions than in others. The resulting wave-induced momentum transport gives rise, in this case, to a mean flow (indicated by the arrows) that becomes conspicuous as soon as one sprinkles a little powder, such as chalk dust or white pepper, on the surface of the water. Carefully stopping the wavemaker and observing the persistence of the mean flow demonstrates that it is not merely a "Stokes drift" dependent on the continued presence of the waves.

The mathematical theory needed to describe such phenomena lies beyond the scope of this lecture; but it can be seen at once that the theory needs to be more accurate than the standard linearized wave theory, which is correct to the first order in wave amplitude. The linearized theory, within its regime of validity, describes only the temporary or reversible, back-and-forth oscillation of fluid parcels that comprises the basic phenomenon of wave motion in a material medium. To describe the mean flow one needs a self-consistent description of the fluid motion correct to the second order in wave amplitude or better. It is at this same level of accuracy, incidentally, that the distinctions mentioned earlier between the various Eulerian, Lagrangian and other definitions

of mean flow first become important (*e.g.*, [62]), making these distinctions one of the inevitable concerns of wave-mean interaction theory.

The experiment shown in fig. 6*a* works robustly when the cylinder, or a wavemaker of almost any shape, is oscillated in almost any direction. But using a vertically oscillating cylinder, as suggested in fig. 6*a*, proves that the observed mean flow is predominantly wave-induced, and not boundary-layer streaming from the surface of the wavemaker. The latter streaming then has the opposite sense (*e.g.*, ref. [82], eq. (215); ref. [84], fig. 31). By using a longer, curved wavemaker, such as that sketched in fig. 6*b*, in a larger body of water [4], one can focus the waves on a spot well away from the wavemaker and demonstrate wave-induced momentum transport over a longer distance. In that case the result is an easily observable mean flow concentrated near the focus, especially if the waves can be induced to break slightly, generating turbulent flow and localizing the dissipation more strongly. The air entrainment that often makes wave breaking conspicuous when it occurs on a larger scale, as on ocean beaches, is unimportant for present purposes, as has been clearly recognized by water wave researchers [85]. It is the turbulence and the consequent extra wave dissipation that matters.

It turns out that, in order to characterize the kind of wave dissipation processes that give rise to the mean torque Γ in the Earth's atmosphere, one is forced into generalizing the notion of "wave breaking" in a certain way, consistent with its established meaning for water waves and its fundamental significance for the momentum transport phenomena under discussion. The generalization, to be described in sect. 9, is dictated by the way in which the Lagrangian description of fluid motion enters theories of wave propagation and wave-mean interaction in stratified, rotating fluid systems [62,68,86–88]; the need for such a generalization can also be seen from the role of the potential vorticity to be discussed below.

For the global circulation the two most important types of waves are those known as internal gravity waves and Rossby waves. Both depend, in contrasting ways, upon the atmosphere's strong stable stratification. Some of the characteristics of gravity waves and Rossby waves are summarized in table I, which will be a focal point for the discussion to follow.

Internal gravity waves are the waves that arise directly from the Archimedean or buoyancy restoring force associated with the stable stratification. This means that most internal gravity waves are, in effect, fast oscillations, from the viewpoint suggested at the end of sect. 1. The theory of internal gravity wave propagation is an interesting topic whose details would again take us too far afield (and are readily available elsewhere [28,82]). It begins by noting that the intrinsic frequency $\hat{\omega}$ of plane waves ($\hat{\omega}$ meaning the radian frequency Doppler-shifted to the reference frame of an observer moving with the local mean flow) is independent of the magnitude $|\mathbf{k}|$ of the wave vector $\mathbf{k}$ when $|\mathbf{k}|H_p \gg 1$, implying that the intrinsic group velocity $\partial\hat{\omega}/\partial\mathbf{k}$ is at right angles to the wave vector. This can be seen at once from dimensional analysis, assuming that the buoyancy frequency N and Coriolis parameter f are the only relevant properties of the basic state (*i.e.* that H_p is unimportant). It can further be shown that the vertical phase and group velocities always have opposite signs (a rule that applies, as it happens, to all atmospheric wave types capable of vertical propagation apart from sound waves, whose effects are negligible for present purposes). If $\hat{\omega}^2 \gg f^2$ then $\hat{\omega}$ is given simply by N times the sine of the angle between $\mathbf{k}$ and the vertical. These fast internal gravity waves are generally more efficient at transporting momentum and angular momentum than the slower "inertio-gravity waves" with $\hat{\omega}^2$ not much greater than f^2, which feel a significant restoring force from the Coriolis as well as from the buoyancy effects.

In the upper mesosphere, several lines of evidence point clearly to the conclusion that Γ is dominated by breaking internal gravity waves incident from below, with intrinsic frequencies $\hat{\omega}$ satisfying $\hat{\omega}^2 \gg f^2$, wavelengths $2\pi/|\mathbf{k}| \gtrsim 10\,\mathrm{km}$, and group velocity vectors that appear close to vertical in a stretched coordinate system (but are closer to horizontal in reality) giving fast propagation times of hours from the troposphere to the mesosphere (*e.g.*, [18,27,91–96]). There

is a smaller contribution to Γ from wave dissipation by infrared radiative heat transfer; ref. [11] gives an authoritative discussion. It is mainly the irreversible angular-momentum transport due to gravity waves, then, that is believed to drive the upper-mesospheric branch of the circulation sketched schematically in fig. 3. In other words, it is these waves that are believed to cause the remarkable coldness of the summer polar mesopause, and the advective transport of water vapour from below, that together give rise to noctilucent clouds and related phenomena. The same goes for the equally remarkable warmth of the winter polar mesosphere. The applicability of eq. (9) to this problem has been discussed in ref. [40,46].

The breaking of the dominant, fast gravity waves arises naturally from what might be called the "ocean-beach effect" as the waves, initially with small amplitudes, propagate upwards (meaning, of course, that the vertical *group velocity* component is upwards) through many density scale heights (*e.g.* [18,28,92] and references therein). The evanescence of the density puts the waves into a situation resembling that of ocean surface gravity waves approaching a beach. The notional "beach" steepens where N increases, as, for instance, at the mesopause [46,93,97]; it also steepens wherever the mean shear Doppler-shifts the intrinsic frequency toward zero, the celebrated "critical level" effect [98]. The critical level is the linear-theoretic location of zero intrinsic frequency, and corresponds to the edge of the beach, in that the waves break, or otherwise dissipate, before reaching it, under conditions typical of the middle atmosphere with coarse-grain Richardson numbers not too small. That is, real upward-propagating gravity waves, under realistic conditions, dissipate below, not "at", their critical levels ([18] fig. 6-32, [92] fig. 9).

The hastening, and vertical confinement, of wave breaking by the critical-level effect is important for understanding the selective transmission or "filtering" of waves with different phase speeds through the westerlies or easterlies [99]. This in turn is a key to understanding why Γ tends to be positive in the summer mesosphere, and negative in the winter mesosphere, as implied by the heavy dashed line above fig. 3.

Whereas internal gravity waves involve small-scale, relatively fast, vertical displacements of the isentropic surfaces of the stable stratification, on time scales of minutes to hours, Rossby waves are characterized by large-scale, relatively slow, nearly horizontal displacements of air parcels along isentropic surfaces, on time scales of days. Such motions may aptly be described as "stratification-constrained", signifying an *absence* of gravity waves. They are sometimes also called "layerwise-two-dimensional" since, although they involve some vertical motion, the vertical velocities w' are relatively tiny, and the horizontal velocity field (u', v') is two-dimensionally nondivergent to a crude first approximation. "Large scale" most often means horizontal scales L of the order of 10^3 km or more; and in the extratropics "relatively slow" can generally be taken to mean $\hat{\omega}^2 \ll f^2$, for a wave of intrinsic frequency $\hat{\omega}$. The order of magnitude of w' is $v'H_p/L$ times $\hat{\omega}/f$, which for horizontal disturbance velocities v' of tens of meters per second will typically imply w' values of a a few centimetres per second.

The restoring mechanism to which Rossby waves owe their existence, and their ability to transport angular momentum, will be discussed in sect. 9. The restoring mechanism depends on the presence of a gradient of potential vorticity along some or all of the isentropic surfaces of the stable stratification. This gradient produces a kind of "sideways stratification" that is fundamental to very many large-scale atmospheric dynamical processes [1]. In the context of the Antarctic ozone hole, for instance, this sideways stratification is a key factor in inhibiting latitudinal eddy advective chemical transport into the ozone hole region (sect. 10). The same inhibition of eddy transport is also what makes possible the important new quasi-Lagrangian methods of data analysis described in ref. [63] in this volume.

Rossby waves are believed on good evidence to dominate Γ in much of the wintertime stratosphere. Related types of motion, whose ability to transport angular momentum likewise depends on the existence of isentropic gradients of potential vorticity, are believed to dominate Γ in the summertime lower stratosphere also. In the latter case, however, as was first suggested by Charney and

Drazin in a classic paper [100], Rossby-wave propagation to greater altitudes is largely suppressed by the summertime mean-flow configuration. This permits only a more restricted, quasi-diffractive upward penetration from the troposphere [101], akin to quantum-mechanical tunnelling, on vertical scales of the order of the Rossby height $fL/N \lesssim 10^{-2}L$ (cf. eq. (8)). For a more precise statement about penetration height scales see ref. [1], p. 902. The more restricted upward penetration of Rossby-type disturbance motions is believed to be part of what underlies the summer-winter asymmetry seen in fig. 3.

In what follows we shall look more closely at the most essential aspects of linear and nonlinear Rossby-wave dynamics, after recalling some basic properties of the potential vorticity — to be precise, the Rossby–Ertel potential vorticity [102–104], hereafter "PV". We need some knowledge of these properties not only to understand the dynamics of the Rossby waves themselves, but also to understand more clearly the circumstances in which both they and internal gravity waves can give rise to irreversible momentum and angular-momentum transport ([4] and references therein). This is because the dynamics of the mean or basic state itself, on which the waves are conceived to propagate, can be described most succinctly in terms of the PV, as a consequence of cyclostrophic balance. In fact, the most important aspects of the mean evolution can be described in terms of the irreversible "rearrangement", in a generalized sense to be explained later, of PV on isentropic surfaces — an idea closely related to the generalized concepts of wave-mean interaction and wave breaking already mentioned.

8. – PV fundamentals: Ertel's theorem and PV invertibility.

The most accurate and general version of the potential-vorticity concept, for a continuously stratified fluid such as the middle atmosphere, is that associated with the exact set of definitions given by Ertel [104]. A hydrostatic version was given earlier by Rossby [102,103]. The difference between the exact and hydrostatic versions is usually unimportant in atmospheric dynamics, the main exception being when one seeks to understand in detail how the PV is affected by three-dimensional turbulent mixing (sect. 11, [5,105]). The exact (Ertel) definition of the PV, now using true geometric coordinates in all three dimensions, including the vertical, may be taken as

$$Q = \rho^{-1}\boldsymbol{\zeta}_{\mathrm{a}}\cdot\nabla\theta \ , \tag{12}$$

where from now on ρ denotes the mass per unit volume, in the ordinary geometric sense of the word "volume", and where $\boldsymbol{\zeta}_{\mathrm{a}} = 2\boldsymbol{\Omega} + \boldsymbol{\zeta}$, the absolute vorticity vector, obtained by adding twice the Earth's angular-velocity vector $\boldsymbol{\Omega}$ to the relative vorticity vector $\boldsymbol{\zeta} = \nabla \times \mathbf{u}$, with $\mathbf{u}$ the air velocity relative to the Earth; ∇ is the three-dimensional gradient operator with respect to geometric position $\mathbf{x}$, and θ is the potential temperature as before.

Because of the atmosphere's stable stratification, $\nabla\theta$ is usually directed nearly vertically. Roughly speaking, therefore, Q is a measure of the component of absolute spin about the vertical, including the vertical component of the Earth's rotation. More precisely, Q can be regarded, for reasons to emerge shortly, as measuring the intrinsic "cyclonicity" of an air parcel, in a sense that is highly relevant to the stratification-constrained, layerwise-two-dimensional motion. This is related to the fact that, on each isentropic surface of the stable stratification, Q is proportional to the component of absolute vorticity precisely normal to that surface, together with the fact that this component tends to increase or "spin up" when $|\nabla\theta|$ is decreased by adiabatic vertical motion, and *vice versa*, tending to keep the Q value of an air parcel constant. Extratropical stratospheric air has a very high intrinsic cyclonicity, in this sense, in comparison with tropospheric air; and this is a key factor, for instance, in extratropical explosive cyclogenesis, one of the commonest causes of severe maritime surface weather, in which extratropical lower-stratospheric air descends along a sloping isentropic surface and interacts dynamically with warm, moist lower-tropospheric air (*e.g.*, [1,106,107] and references therein).

It is no more than a matter of convention, incidentally, that θ is used in (12) and not, for instance, the specific entropy $s(\theta) = c_p \ln\theta + \text{constant}$. As Ertel pointed out, it would be equally valid in principle to adopt any of an infinite number of other definitions of the form $Q_S = \rho^{-1}\boldsymbol{\zeta}_\mathrm{a}\cdot\nabla\{S(\theta)\}$, where $S(\theta)$ is any monotonic, differentiable function of θ. For convenience we shall nevertheless follow meteorological convention and refer to Q, as defined by (12), as "the" PV. There are four fundamental points about it that will be emphasized here.

The first has already been hinted at: on short enough time scales, of the order of days in most parts of the atmosphere, the PV tends to be *materially conserved*, or *advected*, to a useful approximation. In other words, its value following an air parcel remains constant, like the mixing ratio of a chemically inert tracer substance that is simply being carried with the flow, and not diffusing, sedimenting, or reacting. As already mentioned, this is part of why it is useful to think of the PV as measuring the intrinsic cyclonicity *of an air parcel*. It neatly takes account of phenomena such as the often-temporary spinup or spindown caused by adiabatic vertical motion, allowing a correct description of important aspects of the dynamics without explicit reference to the vertical motion.* Sufficient, albeit not necessary, conditions for the material conservation of Q (see sect. 11) are first that θ is materially conserved (the motion is adiabatic), and second that no frictional or other nonconservative forces act. This is the classical result known as *Ertel's theorem*. Under the conditions assumed by the theorem, isentropic distributions of PV evolve in a way that is relatively straightforward to visualize and comprehend, namely by quasi-horizontal, two-dimensional advection.

The second fundamental point about PV is the idea of "invertibility", which in one form or another goes back to Charney [71] and Kleinschmidt [108]. Further historical notes can be found in ref. [1], sect. 1, and [109], chapt. 11. The idea is that, besides being especially suitable for visualization purposes, like the isentropic distributions of certain chemical mixing ratios, isentropic distributions of PV contain, in fact, *nearly all the dynamical information* that is relevant to the stratification-constrained, layerwise-two-dimensional, non-gravity-wave part of the motion. More precisely, there is an "invertibility principle", to the effect that if

(a) a suitable balance condition is imposed, to eliminate gravity and inertio-gravity waves from consideration, and if

(b) a suitable reference state specified, for instance by specifying the mass under each isentropic surface, as is done in the theory of "available potential energy" in stratified atmospheres [74], then

(c) a knowledge of the distribution of Q on each isentropic surface, and of θ at the lower boundary, is sufficient to deduce, diagnostically, all the other dynamical fields such as winds, temperatures, pressures, and the altitudes of the isentropic surfaces [1].

For brevity this diagnostic process may be called "PV inversion". The word "diagnostic" implies the use of information at a single instant only. To the extent that PV inversion can be done accurately, it enormously simplifies the prognostic or timestepping aspects of the problem — which otherwise involve all three velocity components since, although vertical motions are small, they are still dynamically significant because of their effect on $\nabla\theta$ together with the large contribution to Q from the Earth's rotation. The prognostic part of the problem now reduces to timestepping the distribution of Q on each isentropic surface, the distribution of θ at the lower boundary, and the mass under each isentropic surface if diabatic heating is important. The distribution of θ at the lower boundary (more precisely, just above the planetary boundary layer†) is crucial to tropospheric dynamical processes including cyclogenesis [1,109–111], but relatively unimportant for most aspects of middle-atmospheric dynamics [112]. So in the latter case we may say that the single scalar field

* which, incidentally, is usually too small to measure directly

† roughly the first kilometre or two above the Earth's surface

Q, or more precisely its distributions on isentropic surfaces, describes nearly everything about the dynamics.

The balance condition, sometimes loosely referred to as a "slow-manifold" condition (cf. [113] and references therein), includes, and generalizes, the notions of stratification-constrainedness and cyclostrophic balance. It can be thought of as an assumption that gravity and inertio-gravity waves are either absent altogether or can be averaged out and, for the purposes of PV inversion, ignored. It is known that in reality such waves are spontaneously emitted by unsteady layerwise-two-dimensional motions ([113] and references therein), and so must almost always be present to some extent, implying that balance and invertibility are inherently approximate concepts (and that the entity called the "slow manifold", therefore, cannot be a true manifold, in the mathematical sense of the term). Hence the phrase "nearly all the dynamical information". But the spontaneous emission is often surprisingly weak, hence the approximations involved often surprisingly good — far better than might be supposed from the usual theories of balanced motion based on "filtered equations", which include quasi-geostrophic theory and its refinements. Exactly how good, under what circumstances, is still a topic of current research and will be touched on again in the next section.

Note incidentally that, since $Q_S = S'(\theta)Q$, the information content of isentropic distributions of Q is exactly the same as when Q is replaced by any of the other potential vorticities Q_S defined by Ertel, provided only that the arbitrary function $S(\theta)$ is monotonic and differentiable so that $S'(\theta) \neq 0$. In particular, it is evident that whenever there is an isentropic gradient of Q, *i.e.* a gradient of Q on a constant-θ surface, there must always be a proportionate isentropic gradient of Q_S. By the same token, vertical gradients of PV have no particular significance, and are largely a matter of the choice of $S'(\theta)$. In this respect PV is fundamentally *unlike* the mixing ratio of a chemical constituent like nitrous oxide, a point to which we return in sect. 11.

The ability to characterize the layerwise-two-dimensional motion largely in terms of the two scalar fields Q and θ, in the manner just described, with an evolution that is straightforward to visualize and comprehend, is clearly an important simplifying principle in atmosphere-ocean dynamics. It is a further important reason for regarding the PV as the most fundamental measure of inherent "cyclonicity". The invertibility principle encapsulates and summarizes an enormous number of results and insights scattered throughout a vast theoretical research literature, on both linear and nonlinear aspects of the dynamics, going back to pioneering work in the 1940s (*e.g.*, [1,109,114]) and including many classic process studies, such as those on Rossby-wave propagation and on related processes like baroclinic instability [1,109]. Much of this theoretical literature has a mathematically formidable appearance, even when use is made of quasi-geostrophic theory, the theoretical tool of maximum simplicity and least accuracy, and even when the idealized conditions demanded by Ertel's theorem are fulfilled. The mathematical details can divert attention from the conceptually simple way in which the problem of layerwise-two-dimensional motion can be separated into its diagnostic and prognostic parts, and from the demonstrable fact that the same conceptual separation applies, also, at higher accuracy and in a much wider range of circumstances than the circumstances in which quasi-geostrophic theory is valid. It is almost entirely in the diagnostic part of the problem, corresponding implicitly or explicitly to the PV inversion operation, that the mathematical technicalities arise. A simple but telling illustration of this point is given in the review of "nonlinear Rossby-wave critical-layer theory" in sect. 2 of ref. [115] (which theory, contrary to popular belief, is now, for the reasons just indicated, a very-well-understood theory; see also ref. [116], and sect. 5 of ref. [117]), and sect. 5.6 of ref. [28]).

A study of typical examples, such as those discussed in [1,110,111,113,118,119], soon gives a qualitative feel for the nature of the inversion operation at quasi-geostrophic and at higher accuracies; and modern computing technology can be harnessed to make this quantitative when necessary, the price for higher accuracies being, for the reasons just stated, more a computational than a conceptual price. In the next section we shall discuss a few of the simplest possible examples;

the reader interested in going further may consult the references just cited. This will make the foregoing ideas more concrete and explicit, at the same time showing how the Rossby restoring mechanism works. Discussion of the third and fourth fundamental points about the PV, which have recently been the subject of some controversy, is postponed to sect. 11. They are relevant to understanding the diabatic and frictional evolution of PV and the concept of "generalized PV rearrangement" on isentropic surfaces, which leads to a clearer understanding of wave, mean-flow interaction theory and the origin of the torque Γ.

9. – Basic examples: Rossby wave propagation and breaking, and the effect on Γ.

The dynamical system now to be considered, nondivergent barotropic vortex dynamics, is the simplest system having the two basic characteristics just described: a scalar field, Q, that is materially conserved in the absence of diabatic and frictional effects, and an invertibility principle saying that all the other dynamical variables can be diagnosed from a knowledge of Q. Besides illustrating the Rossby-wave mechanism, it will also illustrate very clearly the way in which Rossby-wave breaking leads to the persistently negative Γ values already referred to in connection with the lower part of fig. 3, and how this phenomenon, as well as the wave propagation itself, depends on the existence of isentropic gradients of PV.

The symbol Q will now be used to represent not the Rossby–Ertel PV of the stratification-constrained, layerwise-two-dimensional motion, but simply the absolute vorticity of a strictly non-divergent, strictly two-dimensional motion

$$\mathbf{u} = (u, v) \ , \qquad u = -\partial\psi/\partial y \ , \qquad v = \partial\psi/\partial x \ , \tag{13}$$

on an approximately flat model Earth (Rossby's original model, the so-called "beta plane"). Here (x, y) are eastward and northward Cartesian coordinates and (u, v) the corresponding components of the wind velocity vector $\mathbf{u}(x, y, t)$. The function $\psi(x, y, t)$ is called the streamfunction. Q is here taken simply as $f + \nabla^2\psi$, where the Laplacian is two-dimensional, $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$, the simplest elliptic partial differential operator. As before, f is the Coriolis parameter, possibly a function of y. It is only through the y-dependence of f that Rossby's model Earth is not completely flat. The evolution equations are

$$DQ/Dt = \text{frictional terms} \ , \tag{14a}$$

$$\psi = \nabla^{-2}(Q - f) \ , \tag{14b}$$

where D/Dt is the two-dimensional material derivative, defined by

$$D/Dt = \partial/\partial t + \mathbf{u}\cdot\nabla = \partial/\partial t + u\partial/\partial x + v\partial/\partial y \ . \tag{15}$$

Together with the boundary conditions implicit in the inverse Laplacian operator ∇^{-2}, eqs. (14a) and (14b) define the dynamical system to be considered. If the right-hand side of (14a) is zero, the equation states that Q is materially conserved. Note that, if we were to watch a moving picture of the Q field (whether or not it is materially conserved), then we would be following everything about the dynamics since, by (14b), knowledge of Q implies knowledge of ψ and hence, by (13), of the wind field. This is the relevant invertibility principle, and three aspects of it are emphasized here:

(1) Local knowledge of Q does not imply local knowledge of ψ or $\mathbf{u}$; the inversion is a global process. In particular, as already mentioned, it depends on specifying suitable boundary conditions to make the inverse Laplacian ∇^{-2} unambiguous.

(2) In this system the balance condition, on which invertibility depends, corresponds simply to the absence of sound or external gravity waves. They have been filtered out by the assumption of incompressible, nondivergent motion.

(3) There is a scale effect, whereby small-scale features in the Q field have a relatively weak effect on the ψ and $\mathbf{u}$ fields, while large-scale features have a relatively strong effect. In particular, ψ and $\mathbf{u}$ are to varying degrees insensitive to fine-grain structure in the Q field. The inverse Laplacian ∇^{-2} in (14b) is a smoothing operator, and some of the smoothing survives even when followed by the single differentiations in (13).

Equations (14a) and (14b) summarize, with remarkable succinctness, the peculiar way in which fluid elements push each other around. The nonlocalness of the inversion operator and the implied action at a distance (aspect (1)) are related of course to aspect (2). The invertibility principle holds exactly in this case, because the waves representing departures from balance have been assumed to propagate infinitely fast and to have infinitely stiff restoring mechanisms.

The succinctness of (14a), (14b) is to be compared with what is involved in thinking directly in terms of Newton's second law. Newton's law applied to every fluid parcel is mathematically equivalent to (14), but requires us to think explicitly about subtle aspects of the pressure field. The formulation (14) represents an economy of thinking analogous to, but greater than, the economy that results from treating normal reaction forces as constraints when discussing the dynamics of a roller coaster on a rigid track. This treatment reduces a three-dimensional problem to a one-dimensional one, making the problem easier than explicitly using the two normal components of Newton's second law. In the fluid-dynamical system the pressure field and boundaries likewise have the nature of constraints: to this extent they play a role analogous to the normal reaction forces in the roller-coaster problem. The power of the viewpoint represented by (14) has long been recognized, and made use of, by aerodynamicists as well as by theoretical meteorologists. Indeed the idea of invertibility, as exemplified by (14b), is built into classical low-Mach-number aerodynamical language, in such phrases as the velocity field "induced by" a given vorticity field, and in such ideas as the idea that a strong vorticity anomaly can roll "itself" up into a nearly circular vortex.

It is pertinent to note from (14b) that diagnosing the ψ field from the Q field is almost the same thing, mathematically, as calculating the electrostatic potential induced by a given charge distribution [120], or the static displacement of a stretched membrane induced by a given pressure distribution on it. Thus strong local anomalies in Q tend to induce strong circulations around them in the corresponding sense. Such a Q anomaly, together with its induced velocity field, is nothing other than the coherent structure that fluid dynamicists call a "vortex". Examples of the corresponding layerwise-two-dimensional coherent structures in a stratified, rotating atmosphere, long familiar to synoptic meteorologists as "cyclones" and "anticyclones", may be found in ref. [1,110], and [121]. The term "vortex" is used interchangeably to mean either the entire coherent structure, or simply the PV anomaly that induces it.

For layerwise-two-dimensional flow the PV inversion operation at quasi-geostrophic accuracy is very like a three-dimensional version of (14b) [1,122]. This incidentally is where the problem of fig. 5a gets its elliptic character, and is also related to the quasi-diffractive behaviour mentioned earlier (sect. 7) in connection with the asymmetry of fig. 3. More accurate inversion operators are nonlinear, albeit often only weakly so (the superposition principle remaining qualitatively valid [110,111]). One may perhaps draw a rough analogy with electrostatics in a strange kind of nonlinear dielectric. Such accurate inversion operators have been formulated and studied in ref. [113,119] for the next simplest system, a single-layer "shallow-water" rotating system in which the balance-disturbing gravity waves have finite phase velocities c_{GW}, and in ref. [111] for a three-dimensional rotating, stratified atmosphere. In the shallow-water system the evolution equation still takes the exact form (14a), provided that Q is now interpreted as the relevant "shallow-water potential

vorticity" [102, 103],

$$Q = \frac{(f + \nabla^2 \psi)}{h} \ , \tag{16}$$

where $h(x, y, t)$ is the layer depth, and ψ now describes the rotational part of a velocity field that is also slightly divergent (Helmholtz decomposition, *e.g.*, [74], p. 20). The results of these studies show that balance conditions and corresponding inversion operators can be defined that are not only much more accurate than quasi-geostrophic (at the price of being computationally more elaborate, and to some extent nonlinear) but also surprisingly accurate even in cases where $c_{\rm GW}$ values are not at all large numerically in comparison with typical relative flow speeds $|\mathbf{u}|$. We possess examples in which the local Froude number $|\mathbf{u}|/c_{\rm GW}$ reaches values as high as 0.7, while the accuracy remains much better than typical observational and numerical forecast accuracies [113]. This is very surprising indeed. The explanation appears to be the weakness of the nonlinear coupling from vortical motions to gravity waves, precisely analogous to the well known weakness of "aerodynamic sound generation" [82] pointed out by Lighthill in the 1950s. Rotation probably makes this coupling still weaker, although not drastically so in these examples since the domain is hemispherical, implying f values that are arbitrarily small near the equator. It is this same weak nonlinear coupling that underlies the fact that balance and invertibility are indeed inherently approximate, albeit often remarkably accurate, concepts. Whether the ultimate limitations on inversion accuracy have yet been reached is still an open question.

In the shallow-water system the PV inversion operators involved are qualitatively not unlike the simple inverse Laplacian (14*b*), except that the implied action-at-a-distance has a short-range character related to the finiteness of $c_{\rm GW}$. This short-range character is another part of why PV inversion can remain surprisingly accurate as Froude numbers $|\mathbf{u}|/c_{\rm GW}$ increase.

If we put such distinctions aside for a moment longer, contenting ourselves with a crude qualitative level of description, then the simple model system (14) provides an adequate illustration of the Rossby-wave propagation mechanism, as will now be shown.

Imagine a basic state of relative rest ($\psi = 0$ everywhere) in which f, and therefore Q, has a large-scale northward gradient $\beta = df/dy > 0$. If the right-hand side of (14*a*) is neglected, then the contours of constant Q are also material contours. If a disturbance makes these contours undulate as suggested in fig. 7, then the right-hand side of (14*b*) will be alternately positive and negative as indicated by the plus and minus signs in fig. 7. To see what the induced ψ field must look like, one can solve (14*b*), or simply picture the contours of ψ as the equipotentials of the electrostatic field due to a pattern of alternating positive and negative charges (with the sign changed), or as the topographical contours giving the displacement of a stretched elastic membrane that is pushed (+) and pulled (−) alternately in the same pattern. It can be seen that ψ will have hills and valleys centred respectively on the minus and the plus signs, and (13) then shows that the strongest north-south velocities will occur at intermediate positions, a quarter wavelength out of phase with the displacement, and in the sense shown by the heavy, dashed arrows in the figure. If one now makes a moving picture in one's mind's eye of what this induced velocity field will do to the contours, one can see at once that the undulations propagate westward.

The scale effect, aspect (3) above, implies that the westward phase speed $c_{\rm RW}$ increases with wavelength. This is another robust and well-known property of Rossby waves and, when extended to the stratified case, helps explain why, for instance, it is mainly the longest, planetary-scale Rossby waves that reach the wintertime middle stratosphere and higher [28,100]. It is easy to see in the simplest case in which the dynamical system is (14) and β is constant. We then have simple solutions of the type $\psi \propto \cos \ell y \cos\{k(x - c_{\rm RW}t)\}$, which can be seen at once to satisfy (14*a*) and (14*b*) — in this case without linearization, as it happens — provided that the phase speed $c_{\rm RW}$ and the wavenumber components k, ℓ satisfy the celebrated Rossby dispersion relation

$$c_{\rm RW} = -\beta/(k^2 + \ell^2) \ . \tag{17}$$

The scale-dependent factor $-(k^2+\ell^2)^{-1}$ corresponds precisely to the inverse Laplacian operator ∇^{-2} in (14*b*), explicitly showing the scale effect upon phase speed.

In crude qualitative terms fig. 7 applies equally well to the shallow-water case also, and to the real stratified, layerwise-two-dimensional case. The only differences lie in the computational and quantitative details of the different PV inversion operators involved, and hence in the strength and degree of scale dependence of the induced velocity fields. In the shallow-water case, for instance, the scale dependence of the phase speed becomes very weak at large wavelengths, because of the short-range character of the inversion operator. What is common to all cases is that, whenever air parcels are displaced back and forth across an isentropic gradient of PV (assuming small displacements that make the PV contours undulate gently), the resulting PV anomalies induce velocities a quarter wavelength out of phase with displacements, giving oscillatory behaviour and hence wave propagation, one manifestation of which is phase propagation along the undulating PV contours. The sense of this phase propagation is always such that high PV is on the right, in a right-handed coordinate description; in coordinate-independent terms we may call this phase propagation "retrograde". We are talking of course about the intrinsic phase propagation, that is to say the propagation relative to any mean flow that may be present. The group velocity (discussed in terms of PV inversion in sect. 6*c* of ref. [1]) may in the real, layerwise-two-dimensional case, if isentropic gradients of PV are strong enough, have vertical as well as latitudinal components, proportional to minus the effective wave-induced momentum flux components shown in Table I, also known as "Eliassen-Palm flux" components [28,89,90]. The minus sign is related to the retrograde sense of the intrinsic phase speed.

What is involved in Rossby-wave "breaking", and how does it contribute to the torque Γ? The answer exposes in another way the origin of the minus sign just mentioned, and the negativeness of the corresponding Γ contribution that accounts for much of what is seen in fig. 3.

We shall look at some specific examples shortly, but the essential effect involved in Rossby wave breaking is to rearrange PV advectively along isentropic surfaces in a more or less irreversible way, in contradistinction to the reversible undulations of PV contours depicted in fig. 7. As suggested in Table I, breaking Rossby waves may be compared and contrasted with breaking gravity waves [2,4,87]. Whereas the latter tend (i) to generate three-dimensional turbulence, (ii) to deform isentropic surfaces irreversibly, and (iii) to rearrange entropy and chemicals downgradient in the vertical, breaking Rossby waves tend

(i) to generate layerwise-two-dimensional, stratification-constrained (so-called "geostrophic") turbulence,

(ii) to cause the irreversible deformation of otherwise wavy PV contours on isentropic surfaces, and

(iii) to rearrange PV and chemicals downgradient along the same isentropic surfaces.

For reasons of general applicability, and relevance to wave–mean interaction theory (and also because the notion of "turbulence" is not itself always clearly defined), it seems best to take the *defining* property of wave breaking to be the rapid and irreversible deformation of those material contours that would otherwise be undulated reversibly by the waves' restoring mechanism, under the conditions assumed by linear, nondissipative wave theory, as illustrated in fig. 7. "Rapid" means on time scales not much greater than $\hat{\omega}^{-1}$, where $\hat{\omega}$ is intrinsic wave frequency as before [86]. In the case of Rossby waves the relevant material contours are the PV contours. This is the generalized concept of wave breaking mentioned earlier; its motivation and theoretical justification has been argued very carefully in ref. [86,87], and is implicit in [62,68,88].

Note carefully that this generalized concept does not concern itself with case-dependent details, such as particular wave-breaking morphologies, nor with the issue of whether or not dynamical instabilities are involved [86]. Instabilities are, as it happens, directly involved in many cases of

internal-gravity-wave breaking — to be precise, those for which $\hat{\omega}^2/N^2$ is not too close to unity (*e.g.*, [99,123]) — but they are not essential, for instance, to all cases of ordinary ocean wave breaking. So it is necessary to avoid defining wave breaking in terms of instabilities, if the ordinary ocean beach case is to be included.

Consider now a very simple thought-experiment, using the dynamical system (14), in which the initial Q distribution has a uniform gradient, shown by the dashed profile in fig. 8*a*. Suppose that Rossby-wave breaking, in the sense just referred to, takes place in a narrow zone near the y-origin, with the end result that the Q distribution becomes perfectly mixed within that zone and unaffected outside it, as suggested by the solid profile in fig. 8*a*. Real Rossby-wave breaking does not necessarily produce anything like perfect mixing [86,124,125], but fig. 8*a* is arguably relevant as an idealization. In fact, it is not always a very severe idealization, a case in point being the class of solutions of eqs. (14) that can be found in the literature under the heading of "nonlinear Rossby-wave critical-layer theory". Figure 8*b* shows the zonally or x-averaged Q profile evaluated from an accurate, x-periodic nonlinear solution of (14) with the right hand side of (14*a*) zero, obtained by the methods of critical layer theory as in ref. [116] (P. H. Haynes, personal communication). These methods involve some mathematical technicalities (partly associated with inverting ∇^2 in a particular geometry [115]), which need not concern us here except to note that the nonlinear solutions thus obtained provide accurate, dynamically self-consistent descriptions of Rossby-wave breaking in certain situations where the wave-breaking region or "surf zone" — otherwise known in this context as the "critical layer" — is narrow as compared with the other horizontal length scales in the problem. The theory also provides valuable insights into the way in which a Rossby-wave surf zone can interact dynamically with the incident wave field, including the net absorption and back-reflection of wave activity. There tends to be more reflection than in ocean beach surf zones [89,115–117]; this fact has recently been taken into account in a new and potentially important parametrization of Rossby-wave breaking, for use in computationally economical models of middle-atmospheric circulation and chemistry [126].

The narrow surf zones for which critical-layer theory is valid occur when wave amplitudes are not too large, near locations where mean shear Doppler-shifts intrinsic phase speeds to locally small values. As with internal gravity waves, small intrinsic phase speeds are conducive to wave breaking, although the morphology of the critical layer is entirely different. In the case of monochromatic Rossby waves the wave-breaking region tends to straddle the critical line where linear theory predicts zero intrinsic frequency and phase speed (in three dimensions, the critical surface), whereas, as mentioned earlier, gravity waves tend to break before reaching their critical levels.

Now the inversion operator (14*b*), adapted to the x-periodic case, implies that a change $\delta\bar{Q}(y)$ in the zonal mean Q field induces a change

$$\delta\bar{u}(y) = \int_y^\infty \delta\bar{Q}(y')\,dy' \tag{18}$$

in the zonal mean zonal velocity component $\bar{u}(y)$. For the case of fig. 8*b* the resulting $\delta\bar{u}(y)$ profile is the profile shown in fig. 8*c*; for the case of fig. 8*a* the $\delta\bar{u}(y)$ profile (not shown) looks very similar to the central part of fig. 8*c* except that the shape is exactly parabolic, as is evident from a moment's consideration of (18). Because the right-hand side of eq. (14*a*) has been taken to be zero, corresponding to free, frictionless fluid motion, and because there is no zonal mean pressure gradient, it is clear from eq. (18) and fig. 8*c* that the effect of rearranging Q has been the same as that of exerting a mean force in the x-direction. It is also clear that the sign of this effective mean force is negative.

From what has already been said about PV inversion operators one would expect these two qualitative results to carry over to stratified atmospheres; and it will be seen in another way that this must, indeed, be true, when we come to the general results of sect. 11. Robinson [118,127] has

performed numerical experiments for stratified atmospheres that explicitly show the quantitative behaviour in some examples of interest. The dynamical system used is a realistic model of the wintertime stratosphere on a spherical Earth. In these experiments, PV is rearranged along isentropic surfaces in surf zones of varying widths confined to the middle stratosphere, in the manner of fig. 8. As expected, the effect is just the same as that of exerting, at latitudes and altitudes corresponding to the location of the surf zone, a negative mean torque Γ. The total angular momentum transferred, as measured by the time integral of Γ during the formation of the surf zone, scales roughly as the cube of the width of the zone. As can easily be seen from (18), this cubic scaling law is exact for the idealized case of fig. 8*a*.

It is worth noting that the foregoing ideas — which date back to G. I. Taylor's pioneering attempts in the early part of the century to construct a "vorticity transfer theory" of turbulence (see bibliography in [4]) — have occasionally been regarded with suspicion, or even stated authoritatively to be plain wrong [128]. The argument invokes some fluid-dynamical subtleties, but essentially says that such things cannot happen because they would violate momentum conservation. But this is unequivocally disproved, for instance, by the accurate, fully nonlinear solutions of (14) already mentioned, including that of fig. 8*b*, *c*. What the argument from momentum conservation overlooks is the possibility, and physical reality, of wave-induced momentum transport from outside the surf zone. Both the "wavelike" and the "turbulent" aspects of the problem, concerning the exterior and interior of the surf zone and the way they interact dynamically, and the implications for the amount of momentum transported, are beautifully illustrated by the solutions from critical-layer theory.

Also well illustrated by the critical-layer theory is a very basic point that is equally clear from the general theory of wave–mean interaction, particularly from the so-called "nonacceleration theorem", a consequence of Kelvin's circulation theorem (ref. [68], below eq. (5.9)). The point is that the wave "dissipation" that leads to irreversible wave-induced momentum transport must be understood, for this purpose, to include wave breaking in the generalized sense envisaged here, involving the rapid and irreversible deformation of the relevant material contours, here identified with the PV contours in fig. 7. This reflects a purely fluid-dynamical irreversibility, which is familiar, for instance, from many studies of two-dimensional turbulence, including a classic laboratory study of material-contour behaviour by Welander [129]. In real fluid media this fluid-dynamical irreversibility may well interact with overtly dissipative effects, such as friction and diabatic heating. But its existence does not depend on the latter. In the case of the two-dimensional system (14), at least, it is clear that the irreversible, two-dimensionally turbulent behaviour, sometimes loosely referred to as the "enstrophy cascade", is actually a property of *frictionless* two-dimensional motion. The idea of frictionless two-dimensional motion is known to correspond to a well-defined thought-experiment since (14), with zero friction, has been proven, in the rigorously mathematical sense of the word, to have a unique, smooth solution for all time if initialized with a smooth Q distribution (*e.g.*, [130] and references therein).

What then do breaking Rossby waves look like? There are many ways of irreversibly deforming the PV contours in fig. 7; and breaking Rossby waves, like breaking gravity waves, come in a great variety of shapes and sizes. As with other kinds of wave motion, there is no unique and automatically recognizable "signature" of wave breaking, though some morphologies are commoner than others. Nonlinear critical-layer theory provides one (historically important) set of examples based on (14), and used in references cited earlier including [28,86]; but the morphological details will not be emphasized here because different parameter regimes, involving larger wave amplitudes, have more direct relevance to atmospheric observations.

Detailed examples for a wide range of parameter regimes are becoming increasingly available from various kinds of high-resolution numerical experiment (*e.g.*, [125, 131–137]). Figures 9*a*, *b* present two snapshots from one such high-resolution numerical experiment by Norton [132], using a single-layer model winter stratosphere, a shallow-water pseudospectral numerical model [131] with a

4 km mean equivalent depth and a horizontal numerical resolution better than a degree of latitude. The Q patterns, containing nearly all the dynamical information, are shown in gray scale. The darkest shades denote the most highly cyclonic air masses, except that, for reasons to be explained, an intermediate band of Q values is shown in white, as indicated at the side of the figure. The highest cyclonicities, hence the darkest shades, tend to occur in the polar cap. The region shown in fig. 9 is exactly a hemisphere, and the map projection polar stereographic. The hemisphere is the entire model domain in this case; to save computational expense an artificial, stress-free boundary was introduced at the equator.

The parameter regime is at an opposite extreme to that for which Rossby-wave critical-layer theory is valid, in some ways far closer to realism for the wintertime stratosphere, although the corresponding surf zone — which here consists of the entire region outside the white band — is now unrealistically broad, extending all the way to the model's equatorial boundary. This broad surf zone is two-dimensionally turbulent, in striking contrast with the undular behaviour of the white band suggested by fig. 9*a*. The corresponding region in the real wintertime stratosphere (*e.g.*, [3,18,138]) and in lower-resolution but fully three-dimensional models (*e.g.*, [139–141]) usually reaches as far as the subtropics only, even in the most strongly disturbed winters; see also the remarks on the quasi-biennial oscillation at the end of sect. 10. Thus the real situation is to this extent more like the situation idealized by fig. 8*a*, albeit with a surf zone that is still far broader, usually (cf. [138]), than the very narrow surf zones that are quantitatively describable by nonlinear critical-layer theory.

Haynes and Norton (personal communication) have recently taken single-layer modelling to a higher level of sophistication, including a promising technique for linking ultra-high-resolution single-layer model runs more closely, perhaps even semi-quantitatively, to coarse-grain three-dimensional model runs and, it is hoped, ultimately to three-dimensional observational data. One result has been the discovery of single-layer cases [142] that appear to have a much more realistic tropics than the case of fig. 9, the surf zone being confined more to middle latitudes in the manner suggested schematically by fig. 8*a*. It appears that two factors in attaining greater realism were a somewhat smaller polar vortex and, for more subtle reasons, the use of a global rather than a hemispherical domain. See also ref. [134].

To see the relationship between single-layer experiments and reality, one can imagine the real stratosphere as made up of hundreds of layers marked by isentropic surfaces, dynamically coupled to each other through an appropriate PV inversion operator, over vertical scales related to horizontal scales *via* the fundamental dynamical aspect ratio $f/N \lesssim 10^{-2}$. Again for a more precise statement see ref. [1], p. 902, also ref. [112]. Along each such "θ-layer" the air moves almost frictionlessly, exhibiting a combination of layerwise-two-dimensional undular and turbulent behaviour like that in the single layer illustrated by fig. 9 and in the more realistic single-layer experiments just mentioned. To the layerwise-two-dimensional motion must be added the relatively weak diabatic, cross-isentropic motion comprising the three-dimensional diabatic circulation. The small-scale features in fig. 9 should be imagined to be the intersections, with a given isentropic surface, of sloping, three-dimensional material structures. Thus, for example, the white band in fig. 9 would correspond in reality to a sloping polar-vortex wall (see, for instance, the three-dimensional visualizations of the real polar vortex in ref. [143]).

If one had the observational resolution, one would probably see, in the real stratosphere, especially the lower stratosphere, a more marked, finer-grained small-scale horizontal structure on each isentropic surface than even the finest structure in fig. 9. In the case of fig. 9 the thin filaments in the surf zone are being unrealistically attenuated by the model's artificial hyperdiffusivity, even at the very high numerical resolution used here, and the very scale-selective (triharmonic) hyperdiffusivity. The filaments would correspond in reality to thin, sloping sheets of air with distinct chemical properties (and would appear as horizontal layers to a balloon or other vertical sounder). There is a strong tendency to form such sloping sheets in layerwise-two-dimensional turbulence because of

the combined effects of horizontal and vertical shear [125,133,144], the most probable slopes being of the order of the same fundamental aspect ratio $f/N \lesssim 10^{-2}$. Such structures tend to be stable while they are still being stretched and sheared, contrary to what might be concluded from naive application of a local shear instability analysis [145,146].

These characteristics of layerwise-two-dimensional turbulence may prove significant for surf zone chemistry, inasmuch as they imply an exponentially rapid shrinking of vertical scales as horizontal scales shrink, equally rapidly, by differential horizontal advection, on e-folding times of the order of days [125, 142]. This in turn points to a highly anistropic, multiscale interplay between the continuous layerwise-two-dimensional horizontal mixing, on the one hand, and intermittent three-dimensional vertical mixing, on the other, all of which has to take place before previously separate chemical constituents can be brought into molecular contact (see remarks about "mixdown" below). The net effect cannot be described by the usual eddy-diffusivity assumption, even with an anisotropic eddy diffusivity, a point to which we shall return in the next section.

Fig. 9*a* exhibits only a weak form of Rossby-wave breaking, in which thin filaments of cyclonic material are being eroded from the edge of the polar vortex and mixed into the surf zone. The corresponding thin sloping sheets in the real atmosphere would be likely to be observationally invisible except in high-resolution balloon and aircraft measurements of chemical constituents (*e.g.*, [18,147,148]). Figure 9*b*, from a later time in the same model run, shows a larger amplitude Rossby-wave-breaking event (more precisely, the state of affairs immediately following the occurrence of two such events). Note the self-evidently irreversible deformation of some parts of the (previously undular) white band.

Figure 10 shows a case from the real middle stratosphere, comprising a clear example at still larger amplitude (note, however, that the outermost latitude circle in fig. 10 is 15°N, not the equator). The figure shows an isentropic map of PV from a very careful case study [124] in which the observational-data analysis, from operational meteorological data including satellite infrared radiances, was cross-checked in several ways. There is little doubt that the PV structures shown in this coarse-grain view are real, despite the well-known difficulty of evaluating the PV from observational data. The map shows three secondary vortices that resulted from a single, massive Rossby-wave-breaking event of planetary-scale dimensions. As already mentioned, the scale effect whose most elementary illustration is eq. (17) favours very large horizontal, and also vertical, scales, for Rossby-wave propagation up into the wintertime middle stratosphere [100]. The vortices seen in fig. 10 consist of large amounts of cyclonic air ejected from the main polar vortex on the left, a process that can clearly be seen taking place in earlier maps over the previous few days. If the observational resolution were an order of magnitude better, one would almost certainly see, in addition, filamentary debris somewhat like that evident in fig. 9*b*, although different in detail.

A number of other observational case studies showing direct evidence for Rossby-wave breaking and consequent polar-vortex erosion include, for instance, those reported in ref. [86,148,149]. The last two concern the lower stratosphere, and directly reinforce the presumption that the combined effects of polar-vortex erosion and polar-vortex chemistry may have significant consequences for present and future mid-latitude ozone depletion. The first case cited occurred in January 1979, corresponding to fig. 3, and was another case of PV rearrangement in the middle stratosphere on a grand scale, comparable to that seen in fig. 10. This appears to have produced a contribution to Γ that played a substantial part in maintaining the strong poleward-downward mean circulation seen in the right-hand half of fig. 3. The only examples exhibiting still more drastic PV rearrangement than this case or the one in fig. 10 are the most extreme examples of all, the spectacular northern-hemispheric dynamical events known as major midwinter stratospheric warmings. In typical examples, the polar vortex splits apart and irreversible PV rearrangement may occur over the entire extratropical winter hemisphere [86,117]. PV inversion then implies a strong warming of the polar regions by many tens of K [112], just as observed. Such events can contribute substantially to the mean dynamical heating (sect. 5) in some northern winters. In the less strongly disturbed

southern hemisphere (*e.g.*, [143]), major midwinter warmings have not been observed, although the "final spring warming" and the precursory vortex erosion may in some years involve very substantial PV rearrangement and diabatic attenuation spread over a longer time, with possible implications for the interannual variability of the Antarctic ozone hole ([150,151]).

Mean-circulation estimates like fig. 3 must be understood, of course, as statistical residuals from ensembles of wave-breaking events occurring during the time of averaging, together with any effects of gravity-wave breaking, its interaction with Rossby waves, and of other, *e.g.* purely diabatic, mechanisms of Rossby-wave and gravity-wave dissipation — all of which may contribute to Γ by amounts whose relative importance in the stratosphere is not clear today (*e.g.*, compare [11,126]), and indeed in ways that probably involve significant interactions between the different processes. It is unlikely, for instance, that one can think of Rossby-wave dissipation by diabatic effects as if it were something entirely separate from dissipation by Rossby-wave breaking: the interaction between the two is arguably of direct relevance to the statistical behaviour of isentropic distributions of PV [152–154], and hence to Γ. Late winter in the southern hemisphere (*e.g.*, [143]) seems to provide a very clear example in which vortex erosion cooperates with diabatic heating. References [4,18,153] discuss different aspects of how global-scale Eulerian-mean circulations might be viewed as statistical residuals of the behaviour of diabatically attenuating PV anomalies; these issues bear, again, on the question of interannual variability (*e.g.*, [48,150,151]).

When using an Eulerian-mean description one probably needs to average over periods longer than the single month of fig. 3, perhaps over most of a season, if all the irreversible aspects are to become statistically distinct from the temporary effects such as reversible, bodily displacements of the polar vortex. The latter can make large but successively opposite-signed contributions to Γ, with a net effect that would be exactly zero if the vortex were to return to its original position without being eroded or Ertel's theorem being violated. That the net effect would, indeed, then be zero is obvious from considerations of PV invertibility, even though very far from obvious from the equations of the traditional Eulerian-mean description. By the same token, the net effects of Rossby-wave breaking may themselves be reduced if some of the ejected cyclonic air re-merges with the main vortex [124] before its PV values are affected diabatically.

It should be noted, finally, that the ideas illustrated by fig. 8 and eq. (18) apply also to the summertime lower stratosphere, notwithstanding the more restricted, quasi-diffractive upward penetration of Rossby-type disturbances mentioned earlier, e.g. above eq. 16. To get a persistently negative Γ all that is necessary is a persistently downgradient rearrangement of PV, where downgradient refers to the large-scale latitudinal gradient arising from the Earth's rotation. "Persistently" again need be true only in a statistical, long-time-mean sense. The quasi-diffractive (upward evanescent) behaviour is partly due to the fact that the intrinsic phase speeds of most tropospheric disturbances tend to be eastward relative to the zonal flow in the summertime stratosphere, which is the wrong sign for free Rossby propagation in the stratosphere [100]. This leads to a situation dominated to a first approximation by the elliptic character of PV inversions applied to strong, tropopause-related PV anomalies. For this reason the quasi-diffractive behaviour is robust, in reality, even though linear Rossby-wave theory is very far from being valid for these lower-stratospheric disturbances. The tropopause-related PV anomalies are associated with vigorous, large-scale weather systems, and the disturbances are still very strongly nonlinear — the waves may be "diffracting", as far as their vertical structure is concerned, but they are also breaking — especially in the few kilometres above the tropopause jet altitudes where, again, intrinsic phase speeds tend to be small. This is indeed why irreversible PV rearrangement does, therefore, typically take place over a great range of latitudes in the extratropical lower stratosphere (*e.g.*, chapt. 5 of ref. [18], and fig. 6 of ref. [155]). It seems likely that this is the primary cause of the persistently poleward-downward mean circulation seen on the left, as well as on the right, of the lower part of fig. 3. Breaking gravity waves may also make a noticeable contribution to Γ in the lower stratosphere, although exactly how much, and in what locations and seasons, is currently controversial (*e.g.*, [156,157]).

Certain other consequences of Rossby wave breaking are worth noting. For example, it explains why the amplitude growth that might arise as Rossby waves propagate upwards through many density scale heights, towards the mesopause and above, cannot lead to an atmospheric corona like the solar corona. The interesting question of why there is no atmospheric corona was first posed in ref. [100], starting from a consideration of the the relatively enormous energies of planetary-scale tropospheric motions on Rossby time scales (for further historical remarks see [109]). Wave breaking provides a robust "saturation mechanism" for limiting, in order of magnitude, the amplitudes of upward-propagating waves, including Rossby waves, long before the associated wave energy densities could become comparable to thermal-energy densities as the mean mass density decreases. What it means in practice is that, for all the significant types of atmospheric waves, horizontal disturbance velocities are highly unlikely to exceed intrinsic horizontal phase speeds by more than a modest factor. For all the significant wave types these speeds are always far smaller than acoustic and thermal molecular speeds. In the Rossby-wave case the amplitude order-of-magnitude limitation imposed by wave breaking was deduced in another way in ref. [158]; in the gravity-wave case the existence of the order-of-magnitude limitation was explicitly recognized much earlier ([95] and references therein). These considerations show, among other things, why one can be confident that the integral in eq. (9) converges at its upper limit.

10. – Chemical constituent behaviour, mixdown, PV eddy-transport barriers, and the failure of the eddy-diffusivity hypothesis.

Let us now return to the questions about chemical-constituent behaviour that were raised at the end of sect. 3. These concerned the effects of dynamical processes upon different types of chemical constituents, of which the idealized categories of "total-exposure constituents" and "local-relaxation constituents" represented opposite extremes chosen to give insight into some of the possibilities.

It was pointed out, for instance, that, unlike nitrous oxide mixing ratios, ozone mixing ratios in the stratosphere depend hardly at all on total exposure to UV, and are likely to depend more strongly on air parcel histories (more correctly, transport histories), particularly on what happens during the parcel's last journey, if any, across a region of comparable photochemical and fluid-dynamical time scales.* The extreme case would be a parcel containing a local-relaxation constituent that travels rapidly and irreversibly from a region with fast photochemistry and a high equilibrium mixing ratio, such as the subtropical upper stratosphere, to a region with slow photochemistry and a low equilibrium mixing ratio, such as the polar night. A moment's reflection shows that such a parcel will end its journey with a higher mixing ratio of the constituent than another parcel making the same journey more slowly and hence staying closer to local equilibrium for a greater part of the journey. A simple example of this effect has been analyzed in ref. [159], and argued to be relevant to understanding aspects of the seasonal evolution of the real ozone layer. With more complicated trajectories, such as might be expected in wintertime surf zones, and involving both layerwise-two-dimensional turbulent motion and diabatic, cross-isentropic motion, there may likewise be significant transport history effects. These might be especially strong, for instance, in the mesosphere, above the zonal wind jet maximum, where the Rossby-wave "surf" may extend all the way across the pole [160] and diabatic descent rates may be intermittently large, or in major stratospheric warmings, in which the same can occur over a deeper layer including the middle stratosphere.

It seems clear also that different types of chemical constituent might be differently affected by the mixing processes themselves, especially where interacting catalytic cycles are involved [9,18,31,32,52,163]. There is, for example, the potentially important question of the time delays

* More correctly, *typical molecule's* last journey; again see footnote near the end of sect. 3.

before previously separate air parcels come into thermal and chemical contact. Hardly anything is known about these "mixdown times", but typical surf zone values seem likely to be of the order of one to three weeks [3]. Mixdown is likely to involve the interplay, already referred to, between relatively large-scale, layerwise-two-dimensional eddy motion, on the one hand, and small-scale three-dimensional vertical mixing, on the other. It appears that, as expected for large Ri, the latter occurs sporadically in thin layers of small-scale, three-dimensional turbulence ($\lesssim 10^2$ m thick), commonly seen in radar and balloon data (*e.g.*, [161,162]); but there is disagreement over typical vertical mixing rates (*e.g.*, [3] and references therein, [28] sect. 12.5.2, p.459, [166]). Mixdown times must depend among other things on how frequently such layers occur, and on the probability distribution of their altitudes and thicknesses [166]. Such statistics may well be highly variable in space and time. It is thought that many of the turbulent layers are caused by the breaking of inertio-gravity waves with intrinsic frequencies $\hat{\omega}$ not far above f. The breaking of these waves involves shear (Kelvin-Helmholtz) instability, and if the waves are monochromatic and the group velocity is upward, then the layers move downward, as is observed on occasion in the radar data. Mixdown times of the order of weeks are comparable to some of the relevant chemical time scales; and the possible implications for the chemistry have only very recently begun to be considered [148,163].

It is emphasized that neither the dynamics nor the chemical transports we have been discussing can be represented, in a completely correct way, by models based on quasi-horizontal eddy diffusivities [3,164–165], although aspects of the vertical mixing can legitimately be represented by a vertical eddy diffusivity, as was apparently first recognized by Dewan [166]. For quasi-horizontal eddy motions such as those illustrated in fig. 9, however, the eddy-diffusivity or flux gradient hypothesis leads to deeply paradoxical conclusions even if applied to coarse-grain, long-time-mean quantities, let alone used to answer more detailed chemical questions. The mixdown time delay is not the only problem. Some of the reasons will be explained shortly, and are further discussed in ref. [3].

For all these reasons it remains very much an open question as to why the coarse-grain eddy-diffusivity models currently being used for long-term global chemical assessments (*e.g.*, [18,31,32]) perform as well as they appear to for some purposes. Part of the answer may be that, if simple mean circulations like that of fig. 3 represent a usable first approximation to chemical transport, in conjunction with a quasi-horizontal eddy diffusivity, then there must be something very robust about the chemistry. Total-exposure behaviour, at least, would appear likely to be robust in this sense, as already hinted in sect. 3. For long-lived total-exposure constituents ref. [36] provides some evidence that the main effect of the wintertime stratospheric surf zone (apart from its role in the dynamics of the mean circulation itself) is no more than the obvious one of flattening the chemical mixing ratio isopleths along the isentropes in the locality of the surf zone. But the underlying question is difficult to answer convincingly for local-relaxation constituents, let alone realistic model constituents with more complicated chemical behaviour. On top of the profound conceptual, let alone quantitative, uncertainties about how to model transport, including mixing, in a way that has some claim to quantitative validity, there are significant quantitative uncertainties about the modelling of the photochemistry itself (*e.g.*, [47,52]). This makes it difficult to rule out the possibility of cancelling errors, especially when models are "tuned" to fit coarse-grain observations as well as possible.

The difficulties are well illustrated by one particular issue of current concern — public as well as scientific — to which the idealized distinction between "total-exposure constituents" and "local-relaxation constituents" may have immediate relevance. A recently proposed argument interprets simultaneous measurements of lower-stratospheric ozone and nitrous oxide from the NASA ER-2 aircraft as evidence of nonstandard chemistry leading to ozone depletion outside the Antarctic and Arctic polar vortices ([167,168] and references therein). The starting point is to say that, according to standard gas phase chemistry, there should be a functional relationship between the nitrous oxide

(N_2O) and ozone (O_3) mixing ratios observed in high latitudes in the wintertime lower stratosphere — somewhat like the observed (and very striking) functional relationship between the mixing ratios of N_2O and "total odd nitrogen" NO_y, which has attracted much interest in the recent literature (*e.g.*, [55,169,170]). The reason given [167] is that "standard gas-phase photochemistry identifies the primary source region for O_3 and the loss region for N_2O as the tropical middle stratosphere and also predicts long photochemical lifetimes for N_2O and O_3 (more than a year) for high-latitude winter." Reference [167] continues "From this, we deduce that air parcels at high latitudes with the same N_2O mixing ratios should have very similar O_3 mixing ratios. Deviations from this behaviour thus indicate chemical ozone loss occurring by processes not included in standard gas phase photochemistry models".

However, the possibility of parcel history sensitivity, or transport history sensitivity, is relevant here. Although stratospheric NO_y has a total-exposure character, stratospheric O_3, according to standard (gas-phase) photochemistry, does not. Stratospheric NO_y has a total-exposure character because it arises mainly from one of the reactions that destroy N_2O in the tropical middle stratosphere. Indeed the mixing ratios of NO_y and N_2O, and the observed functional relation between them, all appear to be good examples of things that might be expected to be transport-history-insensitive, since (*a*) they all depend mainly on total exposure, and (*b*) the functional relationship implied by simple total-exposure photochemistry is a linear relationship, and hence preserved under mixing. ("Simple total-exposure photochemistry" means here that a fixed proportion of the N_2O destroyed becomes NO_y; then mixing two air parcels evidently has the same effect as changing the exposure to an intermediate value.) The observed functional relationship is approximately linear, suggesting that the actual photochemistry may not be far from this ideal, for the relevant air parcel ensembles. But there is no corresponding basis for expecting a functional relationship between N_2O and O_3. On the contrary, as already noted, O_3 might well be parcel-history-sensitive.* One cannot, therefore, discount the possibility that the ER-2 data in question could merely be showing the effects of natural fluctuations in ozone production rates due to varying parcel history statistics — including the statistics of the highly anistropic, multiscale process simplistically referred to as "mixing". On the other hand, the conclusion of ref. [167,168] could still turn out to be correct — see also ref. [148] — since nonstandard chemistry and transport history effects could both be occurring in reality. Our observational knowledge, and modelling capability, both fall very far short of being able to say which is the most important.

One way forward might be to use high-resolution, single-layer models like that of fig. 9, and then perhaps several layers, to perform idealized experiments, at first using simple total-exposure and local-relaxation constituents, to begin to sharpen our understanding of how nonlinear eddy transport, including mixing, might affect chemistry. There have been analytical "chemical eddy" theories dealing, in effect, with the chemical fluctuations arising in an undular situation like that within the white band in fig. 9*a*; but the need now is to understand the *nonlinear* chemical eddy or surf zone effects, *i.e.* the effects of ensembles of reversible and irreversible parcel excursions upon the chemistry, including local-relaxation-type chemistry. Such experiments need to be made synergistic with theoretical work on possibly-relevant nonlinear processes like vortex merging and with current efforts to model chemistry and transport in three dimensions. The three-dimensional models are, for the time being, restricted to what in fluid-dynamical terms are extremely low resolutions, far lower even than fig. 9*b*, which itself has insufficient resolution to describe the smallest scales being

* Note added in proof: All these expectations receive confirmation from the results of a new and interesting model study by Plumb and Ko [188] who, however, propose a quite different explanation for the approximate linearity of the observed functional relationship between N_2O and NO_y. This relies not on special chemistry but on assuming a three-way separation of time scales in the lower stratosphere, between global-scale quasi-horizontal mixing (fastest), mean vertical circulation (intermediate), and chemical relaxation (slowest).

generated by the two-dimensional turbulence, let alone those directly involved in the interplay with vertical mixing. So the three-dimensional models should be thought of as models in which mixdown times are artificially short. Nor are they likely to represent polar-vortex erosion realistically.

It is the inhibition of large-scale, latitudinal eddy advective transport by the Rossby-wave mechanism that gives rise to what probably qualifies as the most conspicuous failure of the eddy-diffusivity hypothesis. As already mentioned, this inhibition is believed to be a significant factor in the formation of the Antarctic ozone hole and its possible future Arctic counterpart [47,170,171]. An understanding of this point is essential, moreover, to resolving some of the controversial questions mentioned in sect. 4, regarding mean descent near the edge of the wintertime stratospheric polar vortex. One of these controversies appears to arise from assuming not only that the eddy-diffusivity hypothesis is justifiable, but also that there is, in reality, an all-pervasive eddy diffusivity in the atmosphere that tends continually to smooth steep gradients everywhere (ref. [172], fig. 11 and pp. 16806–9), so that a mean flow converging laterally on the vortex edge is necessary to account for the steep PV gradients there. However, the assumption of an all-pervasive smoothing by eddy effects has been clearly shown, by high-resolution numerical experiments such as those cited above, to be incorrect, which in turn reduces the plausibility of the hypothesized, laterally convergent mean flow.

Consider again the white band in fig. 9*a*. It has already been remarked that, in striking contrast with the turbulent behaviour of the model's very extensive Rossby surf zone, the behaviour of the white band is largely undular. This is especially plain to see from animated versions of fig. 9, such as can be produced on modern graphics workstations (*e.g.*, [173]). The white band, when not being disrupted as in fig. 9*b*, displays a peculiar resilience or quasi-elasticity. This is, in part, a visible manifestation of the Rossby restoring mechanism. As can be deduced from the scale of relative Q values on the left, the white band is a region of very steep Q gradients, where a substantial fraction of the total range of Q values, hence the Rossby mechanism, is concentrated. The white band is also a material entity, to the extent that Ertel's theorem holds, which is true here to good approximation. The upshot is that it tends to act as a flexible, material "barrier" to eddy advective transport. This state of things is a direct result of the Rossby-wave-breaking process and the consequent two-dimensional rearrangement of the Q field, including the erosion of the main vortex that takes place as a result of a variety of Rossby-wave-breaking events — ranging from the observationally invisible fine-scale erosion seen in fig. 9*a* to the massive mid-stratospheric breaking events exemplified by fig. 10. The situation is somewhat like a sideways, global-scale, Rossby-wave counterpart of the notorious "atmospheric boundary-layer inversion", a thin layer of large N^2 that can trap smog over cities by means of the *gravity-wave* restoring mechanism.

Other model experiments like that of fig. 9, but representing passive chemical tracers as well as Q, show clearly that the tracers form steep gradients at the same locations as does Q (*e.g.*, [131,139,142,174]). They also provide a wealth of evidence that the formation of the sharp edge is entirely a consequence of the wave-breaking and erosion processes, as in the idealized scenario of fig. 7*a*. No converging mean flows are necessary. In the experiment of fig. 9, for instance, the initial conditions comprised a smooth Q distribution with no steep gradients anywhere; another such experiment is reported in ref. [131]. There is no doubt that, in this and similar experiments, the sharp edge results from the two-dimensionally turbulent PV rearrangement. The formation of exceedingly sharp edges in frictionless models, comprising (14) with exactly zero right-hand side, has been demonstrated repeatedly (*e.g.*, [136]). Like the thinning of material filaments and sheets, it is an exponentially rapid process on the time scale of the large-scale horizontal velocity-gradient field. It seems that the problem is often not one of explaining the observed sharpness of polar-vortex PV edges (scales typically of order 10^2 km), but rather of explaining why they are not observed to be sharper still.

There can be no real doubt that a corresponding subpolar "PV barrier" to quasi-horizontal eddy transport typically occurs in the real wintertime stratosphere and that, together with certain

temperature-sensitive processes such as the formation of "polar stratospheric clouds" [47,170,171], it plays a significant part in the formation of the Antarctic ozone hole by chemically isolating the interior of the polar vortex to a considerable extent, in the sense of strongly inhibiting layerwise-two-dimensional eddy transport into the vortex from the surrounding surf zone. There is, indeed, a longstanding body of purely observational evidence in favour of some such vortex isolation effect in the real wintertime stratosphere, (*e.g.*, [155,175,176] and references therein). This is consistent with what one might expect from the apparently robust behaviour of models like that of fig. 9, its laboratory counterparts [177], and its numerical counterparts for the simplest relevant dynamical system (14) [125], together with the family resemblance between PV inversion operators in all the different dynamical systems discussed earlier.

To all this evidence is now added that from the new data analysis of ref. [63] in this volume. The method used depends on the undular behaviour of the real vortex edge as an approximate material entity. In effect, it assumes that the isentropic distributions of PV within the edge — the real-stratospheric counterpart of the white band in fig. 9*a* — are to a first approximation not being irreversibly rearranged, so that PV values can be used as a quasi-Lagrangian latitudinal coordinate there. The consistency of the results using independent data subsets strongly suggests that the real situation conforms to this assumption, to within the limitations on the analysis imposed by the use of coarse-grain operational meteorological analyses of isentropic gradients of PV, and that the Rossby-wave mechanism really is operating there.

The outstanding questions now concern the precise sharpness of the real edge (as opposed to its coarse-grain representation in operational meteorological analyses) under different circumstances, and the extent, quantitatively speaking, of the vortex isolation. That is, how leaky is the vortex edge? The numerical experiments suggest that the two questions are closely related. The second question is currently controversial (*e.g.*, [63–67,148,172,174,178]), and will remain a key issue both for data analysts and for modellers. On the question of what limits the sharpness of the vortex edge in the real atmosphere, one candidate is scale-selective radiative transfer [179], the scale-selectivity arising from CO_2 infrared opacity. Another is the intermittent vertical turbulent mixing already mentioned [166], although it should be cautioned that the effect of this on the PV field is quite unlike its effect on the distributions of chemical constituents (sect. 11 below). Of course, the sharpness of the edge will be sensitive to the degree of large-scale disturbance of the vortex: it will tend to be sharper, for instance, if the vortex has been more violently disturbed during the preceding weeks, and the numerical experiments illustrate this.

Why are the wave-breaking, erosion and edge formation phenomena under discussion irreconcilable with the standard eddy-diffusivity or flux gradient hypothesis? The question has recently been discussed in ref. [3,165], and in earlier papers they refer to; but in view of the widespread acceptance, or at least use, of the hypothesis let us briefly put down a few of the relevant considerations.

As usually applied, the eddy-diffusivity hypothesis supposes, among other things, that eddy transport is greatest where and when mean gradients are greatest. The situation highlighted by the white band in fig. 9, and its counterpart in the real atmosphere, is at an opposite extreme: the transport, meaning the layerwise-two-dimensional eddy advective transport, is smallest where and when the gradients are greatest. The standard eddy-diffusivity hypothesis also carries the presumption that more eddy activity (in practice measured by fluctuating velocity variances) should give more transport. More eddy activity does often mean more vortex erosion; but this also means more edge-sharpening, resulting in the creation of a stronger PV contrast across the surviving vortex edge. This implies a more resilient PV barrier, and a stronger tendency to inhibit further eddy advective transport in the immediate neighbourhood. One can to some extent mimic this situation by hypothesizing an eddy diffusivity that is a strong, and time-variable, function of horizontal position. This, however, is at best an unnatural (and numerically ill-conditioned) expression of the situation, including the asymmetry, or one-sidedness, often exhibited by the vortex-erosion

phenomenon [125,135,137], which might, among other things, be relevant to the question of mid-latitude ozone depletion [148]. Worse still, the idea of an eddy diffusivity having strong spatial variability is an inherently self-contradictory idea.

Why is this? The classical turbulence-theoretic justification for the eddy-diffusivity hypothesis relies on an assumption of "nearly homogeneous turbulence" [165], implying a *scale separation* between the largest eddies and still larger scales of variation of mean quantities. Hypothesizing that mean quantities, including the postulated eddy diffusivity itself, are strong functions of horizontal position evidently precludes, on the other hand, the possibility of such a scale separation. This is the contradiction. Further paradoxes deeply embedded in the eddy-diffusivity hypothesis (and seemingly most acute for layerwise-two-dimensional turbulence) are described or referred to in ref. [3,165]; see also [2], p. 286, and [115] for another eddy viscosity paradox, albeit of a more superficial character, that has sometimes caused confusion in the context of nonlinear Rossby-wave critical-layer theory. It is sometimes claimed (adding to the confusion) that one can always define "the" eddy diffusivity as minus the flux of a trace constituent divided by its mean gradient. What this argument overlooks is that the result, even if nowhere negative or infinite, will usually depend sensitively on the tracer distribution even for a chemically inert tracer. That is, the resulting eddy diffusivity will imply a description of the "transport" that is not independent of the thing transported.

The essential fact that confronts us is that the particular situation illustrated by fig. 9, and also, indeed, that of fig. 8 described by Rossby-wave critical-layer theory, are both at opposite extremes to the scale-separated situations envisaged by classical, near-homogeneous turbulence theory. In reality we have to deal with closely adjacent wavelike and turbulent regions that interact in an essentially nonlinear manner. Since this type of situation is neither accessible to the classical turbulence theories nor amenable to the alternative approach of Dewan [166] (based on random location of mixing events and a scale separation associated with mixing-layer thinness), its theoretical description and adequate numerical modelling comprise a central challenge in atmospheric dynamics today.

The inhibition of eddy advective transport by a flexible "PV barrier" like the white band in fig. 9*a* is clearly related to the linear Rossby-wave restoring mechanism, as has been said, but it should also be admitted that this is to some extent an oversimplification [155]. As can be seen from fig. 9, a wide range of scales is involved; and linear Rossby-wave theory is not strictly applicable. It does not even begin to describe everything that is happening at the smallest scales, including finer-scale aspects of vortex erosion such as are evident in fig. 9*a*. Nevertheless, high-resolution numerical experiments and, in addition, laboratory experiments using dye tracers [177], have consistently and repeatedly shown the physical reality of the PV barrier effect. As pointed out in ref. [125], the strong shear near the edge plays an important role, in nonlinear reality, in preventing the barrier being breached at small scales.

We have been concentrating on effects that correspond to the top half, as it were, of fig. 8*a*, while thinking of the whole mixed region as an idealized conceptual model for a mid-latitude stratospheric surf zone (while remembering that the mixing may be far from perfect in reality). The bottom half of fig. 8*a* prompts the question of whether *subtropical PV barriers* exist in the real stratosphere. It appears that the Rossby mechanism must certainly be effective in inhibiting cross-equatorial eddy transport in the lower stratosphere, if only because wave-mean interaction theory would otherwise predict, for the reasons illustrated in fig. 8c, a much faster easterly acceleration stage than in the observed quasi-biennial oscillation of $\overline{u}$ in the equatorial lower stratosphere [3]. But after seeing the new and more realistic numerical experiments by Norton mentioned earlier [142], I have become less convinced that one can often expect to see subtropical PV barriers that are nearly as sharp as the subpolar one. The much vaster area of the subtropics seems to dilute the effects of Rossby-wave breaking and make less probable the concentrated erosion that can so effectively sharpen the polar-vortex edge. It appears that if one can speak of a subtropical PV

barrier, then it is on average a more diffuse, ill-defined one. But it remains true that sharp edges may appear intermittently and that the general inhibition of cross-equatorial eddy transport — more precisely, layerwise-two-dimensional eddy advective transport — by the Rossby mechanism is likely to be another significant factor that must be taken into account in global chemical modeling.

11. – PV fundamentals: the conservation and impermeability properties and the concept of "generalized PV rearrangement".

We come finally to the third and fourth fundamental points about that undoubtedly strange, yet very basic, quantity, the PV. They concern PV behaviour under the conditions in which Ertel's theorem is violated, a topic whose importance is an inevitable consequence of everything that has been said up until now. As already mentioned, the two points to be made have attracted some controversy; moreover one of them appears to have gone altogether unrecognized until a few years ago [153]. They suggest a very general way of thinking simultaneously about general-circulation dynamics and chemical transport that may prove to have far-reaching significance. This involves replacing the notion of the mean torque Γ by the notion of "PV flux" or "PV transport", in an appropriately precise and general sense [4,5,153], consistent with the *general* use of the words "flux" and "transport" in mainstream physics and chemistry. This includes, but is not restricted to, advective transport. In various approximate forms the idea has a long history, going back to the classical work by Taylor and others on the "vorticity transport" theory of turbulence already mentioned in sect. 9, and already illustrated in fig. 8.

Ertel's theorem is a particular case of the general result

$$DQ/Dt = -\rho^{-1}\nabla\cdot\mathbf{N}_Q \tag{19}$$

(*e.g.* [104,120]), in which the three-dimensional material derivative, and the nonadvective flux or transport, are defined respectively by

$$D/Dt = \partial/\partial t + \mathbf{u}\cdot\nabla \ , \qquad \mathbf{N}_Q = -\mathcal{H}\boldsymbol{\zeta}_{\mathrm{a}} - \mathbf{F}\times\nabla\theta \ , \tag{20a,b}$$

$\mathbf{u}$ being the three-dimensional velocity field, $\mathbf{F}$ the viscous or other nonconservative body force per unit mass, and $\mathcal{H}$ the diabatic heating rate expressed as the material rate of change of θ, $\mathcal{H} = D\theta/Dt$. We continue to use ordinary geometric coordinates as in sect. 8.

The flux form of (19), from which (19) itself can be recovered using the mass conservation equation $\partial\rho/\partial t + \nabla\cdot\{\rho\mathbf{u}\} = 0$, is

$$\frac{\partial}{\partial t}(\rho Q) + \nabla\cdot\mathbf{J} = 0 \ , \qquad \text{where} \qquad \mathbf{J} = \mathbf{u}\rho Q + \mathbf{N}_Q \ . \tag{21a,b}$$

This equation expresses exact conservation. Here one does not, of course, mean conservation in the material or Lagrangian sense of Ertel's theorem, but in the traditional, *general* sense used in theoretical physics and chemistry and as applied, for instance, to basic conservable quantities like energy, and angular momentum. Material conservation is the special case $\mathbf{N}_Q = 0$. The exact conservation form of (21a) is a direct consequence of the mathematical form of (12) and the fact that div curl of any vector field vanishes identically.

The flux form (21a) expresses conservation in the most general possible sense; for instance it is indifferent to whether the system is also mass-conserving [5]. It is the form (19), not the form (21a), that depends on assuming also that $\partial\rho/\partial t + \nabla\cdot\{\rho\mathbf{u}\} = 0$. Equations in the form (21a), stating the vanishing of a four-dimensional divergence, are the most general possible way of expressing exact conservation of the additive, extensive quantity whose amount per unit volume appears in parentheses in (21a), consistent with the most fundamental principles of physics [180].

What does this imply for the the analogy between PV and chemical mixing ratios? We can now make this precise. The general notion of conservation corresponds, with certain provisos, to the notion of an indestructible chemical substance, that is to say a chemical constituent that has *zero source*. Equations (19) and (21a) can, therefore, be read as saying that the PV behaves like the mixing ratio of a peculiar chemical "substance", or generalized trace constituent, that has zero source away from boundaries. The word "source" is being used here in its standard chemical sense; see below for remarks on other senses of the word. A chemical substance with zero source means a chemical substance whose molecules are neither created nor destroyed. The *mixing ratio* of such a substance can of course change — for instance by dilution — and so, likewise, of course, can the PV, as we know already from (19).

One of the peculiarities of the generalized tracer substance whose mixing ratio is the PV — let us call it "PV-substance", or "PVS" — is that one can have both positive and negative amounts of it, like electric charge [120]. This means that the notion of "indestructibility" has to be understood in the same generalized sense as it is for electric charge in fundamental physics. One can have "pair production" and "mutual annihilation". Note again that ρQ in (21a) is the amount of PVS per unit volume, and Q the amount of PVS per unit mass.

Another peculiarity of PVS behaviour is what might be called the "impermeability property". This expresses a strikingly simple fact about the flux or transport $\mathbf{J}$ in relation to isentropic surfaces that promises corresponding simplifications in our thinking about the global circulation. It is the basis for the new view of general-circulation dynamics and chemical transport already mentioned. It shows that the Taylor relationship between wave-induced forces and sideways PV transport in the simple system (14), illustrated by the example of fig. 8, has a formally exact counterpart, applicable to the real atmosphere in all its wavelike and turbulent complexity, whose dynamical significance is limited only by the accuracy of the PV invertibility principle [4].

The impermeability property says that the notional "molecules" of PVS — or perhaps we should call them notional "charged particles" — behave as if they cannot cross isentropic surfaces. In this sense the flux or transport of PV(S) is always exactly, and not just approximately, along isentropic surfaces. In other words, isentropic surfaces behave exactly as if they were impermeable to PVS, even when diabatic heating or cooling makes them permeable to mass and chemical substances ($\mathcal{H} \neq 0$); in this respect isentropic surfaces behave like semi-permeable membranes.

The impermeability property is another direct, and simple, consequence of the mathematical form of the definition (12) of Q or, equally, of any of its variants Q_S. One can show quite straightforwardly [5], by manipulating the expression (21b) giving the total, advective plus nonadvective, flux or transport $\mathbf{J}$, that

$$\mathbf{J} = \rho \mathbf{u}_{\theta\perp}\, Q + \rho \mathbf{u}_{||}\, Q - \mathcal{H}\boldsymbol{\zeta}_{\mathrm{a}||} - \mathbf{F} \times \nabla\theta \ , \tag{22}$$

where the subscript $_{||}$ denotes projections parallel to the local isentropic surface:

$$\mathbf{u}_{||} = \mathbf{u} - \frac{\mathbf{u}\cdot\nabla\theta}{|\nabla\theta|^2}\nabla\theta \ , \qquad \boldsymbol{\zeta}_{\mathrm{a}||} = \boldsymbol{\zeta}_{\mathrm{a}} - \frac{\boldsymbol{\zeta}_{\mathrm{a}}\cdot\nabla\theta}{|\nabla\theta|^2}\nabla\theta \ , \tag{23a,b}$$

and where

$$\mathbf{u}_{\theta\perp} = -\frac{\partial\theta/\partial t}{|\nabla\theta|^2}\nabla\theta \ , \tag{23c}$$

which is just the velocity of the isentropic surface normal to itself. If Q is to be replaced by Q_S, the only change required is to replace θ by $S(\theta)$ and $\mathcal{H}$ by $\mathcal{H}S'(\theta)$ everywhere.

Since the last three terms in (22) all represent vectors lying parallel to the local isentropic surface, while the first is just ρQ times the normal velocity $\mathbf{u}_{\theta\perp}$ of that surface, it follows that *a point moving with velocity* $\mathbf{J}/(\rho Q)$ *always remains on exactly the same isentropic surface,* whether or not the air is moving across that surface as occurs when the diabatic heating $\mathcal{H} \neq 0$. The velocity $\mathbf{J}/(\rho Q)$ can be pictured as the velocity with which PVS molecules or particles would move, discounting notional thermal motions.

It is difficult, incidentally, to make any sensible remarks about the history of these ideas. The conservation property is mathematically almost trivial, as already indicated, amounting to little more than the identical vanishing of div curl. Not surprisingly, the mathematical aspects were noticed long ago by many authors including Ertel himself, albeit usually leaving the conservation implications unstated. Ertel [104] noted, for instance, the divergence form of the $\mathcal{H}$ contribution to the right hand side of (19). It is not clear why he considered only $\mathcal{H}$ and not $\mathbf{F}$. Truesdell [181] did much the same thing with $\mathbf{F}$ and not $\mathcal{H}$. The interested reader may wish to consult ref. [120,153,182] for more history. Once one does start thinking about the conservation implications, and the PV-chemical analogy, one soon notices the impermeability property as well. But as far as I have been able to gather, after extensive correspondence, the impermeability property was not pointed out until Haynes and I noticed it simultaneously and independently [153]. We then quite unexpectedly found ourselves enmeshed in controversy, over the *conservation* as well as the impermeability property. This was compounded by an interdisciplinary language barrier problem of which we were unaware at the time.

It appears that, despite the analogy between PV and chemical mixing ratios, a separate convention has grown up in which *PV behaviour* is thought of in a manner not directly related to the traditional, general notions of conservation relation and conservable quantity — implicit, for example, in all discussions of *chemical behaviour.* Along with this has grown up a separate usage of words like flux, transport, source, sink, creation, destruction, etc., when used in connection with the PV. I have not seen a systematic account of this separate convention, nor a set of explicit definitions of the associated vocabulary, but the convention appears to define the words "flux" and "transport" to mean the quantity $\mathbf{u}\rho\chi$, in traditional language the *advective contribution* to the total transport of any quantity whose mixing ratio is χ. There is also a third convention in which the word "transport" is used to mean the quantity $\mathbf{u}\cdot\nabla\chi$, in traditional language the *advection.* It should be noted that each of the latter two conventions prohibit the use of a traditional physico-chemical idea such as "molecular-diffusive transport" since they render such an idea self-contradictory. The same goes for the idea of "angular momentum transport" that is so basic, as we saw earlier, to understanding the mean circulation, stratosphere-troposphere exchange rates, CFC lifetimes, and so on.

The word "source" is also problematical. It is very widely used in a nonconservational sense, as in statements like "charge is the source of the electric field", or "$Q - f$ is the source of the streamfunction" in eq. (14*b*). Here the word "source" means simply the cause of something, with no implication that we are dealing with a conservable quantity. It might be useful to speak of a "source in the second sense", if there is a danger of confusion. Thus in phrases like "the source of the ozone" or "the source of the energy" we usually mean source in the first sense, whereas in a phrase like "the source of the difficulty" the word is clearly being used in the second sense.

The commonplace occurrence of the second sense has led to the word "source" being used to mean almost anything on the right-hand side of an equation. This includes not only (14*b*) but also, unfortunately for present purposes, (19). Since it seems inevitable that we shall want to continue to think of PVS and chemical substances together — and ref. [64] in this volume adds to the case for this — it is arguably worth some effort to use a consistent terminology. The effort need not be too onerous; for instance we could avoid confusion by using for the right-hand side of (19) a self-explanatory phrase like "material tendency", rather than "source".

The other reason for the controversy is more substantial, however. There appears to be a genuine mistake — actually three fundamental mistakes — that have become entrenched to some

extent in parts of the atmospheric-science literature, and in parts of the atmospheric-science folklore (see bibliography in [5]). The first is an incorrect assumption that PV behaves like the mixing ratio of a chemical substance in the presence of three-dimensional turbulence. This is equivalent to the statement that PVS, the notional substance whose mixing ratio is the PV, behaves as though it is mixed like a chemical by the turbulence. Let us call this the "mixing fallacy".

The impermeability property tells us at once that no such behaviour is possible, because although there is nothing to stop chemical substances being turbulently mixed across isentropic surfaces, the same surfaces act as if they are impermeable to PVS. This implies an actual, in principle measurable, difference between the behaviour of PV and chemical mixing ratios, quite independent of theoretical issues like the ambiguity of flux vectors [5,153]. The same conclusion can be drawn, albeit slightly less directly, from other well-known principles like the Kelvin-Bjerknes circulation theorem, if one devises suitable thought-experiments. The mixing fallacy appears to underlie the idea, encountered in the literature, that one can quite generally think of PV as the mixing ratio of something that always behaves rather like stratospheric ozone, with a source in the stratosphere and a sink in the troposphere. This last idea is profoundly wrong. It may also come, in part, from forgetting which of the two senses of the word "source" is in question.

The actual mistake, which is not a matter of semantics, may well have originated in a tacit neglect of the strong diabatic heating and cooling that occurs on the Kolmogorov turbulent microscale during turbulent mixing. This microscale heating and cooling is, of course, crucially important, first for seeing how air and trace chemicals can cross isentropic surfaces, and second for seeing how (19) and (21a) are satisfied in the presence of three-dimensional turbulence (whose microscale aspects are intriguing, and understandable [5], but which will be left aside here). The point is relevant irrespective of whether one regards the ρ, θ, $\mathbf{u}$ and $\boldsymbol{\zeta}_{\rm a}$ fields as explicitly representing the detailed, fine-grain reality including the Kolomogorov microscales — in which case the diabatic heating field $\mathcal{H}$ represents the effects of molecular conduction only — or whether one is taking the coarse-grain view necessary in practical observational work, in which the ρ, θ, $\mathbf{u}$ and $\boldsymbol{\zeta}_{\rm a}$ fields are considered to be some kind of averaged approximations to reality, with corresponding adjustments to the fields $\mathcal{H}$ and $\mathbf{F}$ to include the eddy flux convergences from unresolved scales. The interested reader may consult ref. [4,5,105] for further discussion, and some more history.

The other two most serious mistakes in the literature might be called the "friction fallacy" and the "correlation fallacy". Again the reader is referred to the references just cited. In brief, the friction fallacy says that in atmospheric dynamics $\mathbf{F}$ can always be neglected in (20b). Fundamentally, this asserts that we can ignore molecular viscosity at the turbulent microscale, together with its macroscopic consequences, as far as the evolution of PV distributions is concerned. Here the underlying mistake appears to have been a misidentification of two distinct and unequal quantities, torque and force curl. The correlation fallacy consists of ignoring the difference between the coarse-grain average of (12), on the one hand, and the result of substituting the coarse-grain ρ, θ, $\mathbf{u}$ and $\boldsymbol{\zeta}_{\rm a}$ fields into (12), on the other (the latter being inevitably what we mean in practice by the "observed PV"). This neglects the unresolved correlations in the averaged triple product comprising the average of (12); the point is discussed in ref. [5,105], again with some history.

An example that is relevant to the global-circulation problem, and in which the fundamental difference between PV and chemical behaviour shows up especially strikingly, is the the effect of gravity-wave breaking noted in Table I. The effects on PV of this and other modes of gravity-wave dissipation are well established both theoretically and experimentally, and provide unmistakable counterexamples to the first two fallacies just described, with which they are in conspicuous disagreement [4,5]. The issue is one of some scientific importance: if the mixing and friction fallacies were to be upheld and shown not to be fallacious, after all, then the entire edifice of wave-mean interaction theory, from Lord Rayleigh's contributions onwards, together with a vast array of supporting experiments and field observations, would be overthrown.

In summary, then, and setting controversy aside, a pursuit of the PV–chemical analogy to its logical conclusion leads to the expression (22) for the flux **J**, showing that the PV behaves as if particles of signed "PV-substance" or "PVS" are transported exactly along isentropic surfaces, but not across them, and created or destroyed (apart from "pair production" and "mutual annihilation"*) only where isentropic surfaces meet boundaries. These properties are the counterparts for PV of the well known conservation and transport properties of vorticity (vortex lines created or destroyed only at boundaries, and total vorticity transport tensor identically antisymmetric [4,5,153]). Being completely general, they apply to the real atmosphere in all its complexity.

They show, in particular, that the effects of breaking Rossby and gravity waves upon the global distribution of PV, and hence (via PV invertibility) the way in which they control the mean circulation, can be thought of in terms of the total transport of PVS exactly along the isentropic surfaces of the atmosphere's stable stratification, no matter how complicated the details. This suggests that the mathematical description of the middle-atmospheric circulation that admits the most far-reaching generalizations, and which is conceptually the most succinct, involves replacing the concepts of "force" and "torque" by the concept of "PVS flux" or "PVS transport" — or "generalized rearrangement" of PVS — in the manner suggested by the example of fig. 8 and by the form of the last term of (20*b*), and first noted by Taylor long ago for the simplest relevant dynamical system (14). The words "flux" and "transport" are used interchangeably and are to be understood in the general sense of mainstream physics. The phrase "generalized rearrangement" is meant to suggest the horizontal migration of PVS particles confined to each θ-layer, allowing for pair production and mutual annihilation processes, and dilution and concentration effects as mass enters and leaves the layer. The idea makes sense because of the conservation and impermeability theorems. Different aspects of this picture, which complements, and does not try to supersede, the meteorological "air mass transformation" or "material tendency" viewpoint associated with (19), are further discussed in ref. [4,153].

12. – Concluding remarks: the challenge for EOS.

One of the things we have learnt from modern space-based instruments, and in an even more sharply focused way from high-resolution numerical experiments using modern supercomputers, is that it is very far from the mark to think of the global-scale atmosphere as "simply turbulent", in the sense of being permeated by a quasi-uniform eddy diffusivity that continually tries to smooth all large-scale gradients. Turbulence can, and seemingly often does, have the opposite effect. Thus, for instance, it is very clear that laterally convergent mean circulations are not required to maintain the sharpness of the PV gradients bordering the edge of the ozone hole; quasi-horizontal eddy advective transport can do this very efficiently, on their own, by the process of "vortex erosion".

Numerical experiments like that shown in fig. 9, and the many others cited, including an increasing number of multilayer experiments, show such erosion and edge-sharpening phenomena again and again. These phenomena seem robustly indifferent to the numerical discretization method used; the sharpness of the edges produced seem to be determined mainly by the model resolution that can be afforded. Far from seeing smoothing by the eddies, then, one usually has the impression of dealing with closely adjacent, strongly interacting regions of wavelike motion and turbulent motion, a kind of highly inhomogeneous "wave-turbulence jigsaw puzzle" in which the waves strongly modify, indeed often give rise to, the turbulence (as on ocean beaches), and in which the turbulence, in turn, modifies the local spatial distribution of the wave restoring mechanism, and also, after a propagation delay, the wave field at greater distances. The waves and turbulence

* Pair production and mutual annihilation could be important in the tropics where the PV changes sign, usually at locations not too far from the equator.

in the atmosphere seem even more closely interlinked than they are in the case of ordinary ocean-beach waves. The word 'turbulence' is being used here, as above, in the broad sense that includes layerwise-two-dimensional as well as ordinary three-dimensional turbulence.

Such situations lie beyond the reach of classical turbulence theories, which assume a scale separation between mean and eddy quantities, leading to the concept of eddy diffusivity. It appears that in the real atmosphere such a scale separation may often be one of the worst modelling assumptions one can make — fig. 9 being a case in point — although there are some exceptions to this rule, the main one of interest here being vertical mixing by breaking inertio-gravity waves, where a scale separation having nothing to do with classical turbulence theory arises from the thinness and random altitudes of the turbulent layers (vitiating, incidentally, some widely used classical relations between eddy diffusivities and turbulent dissipation rates) [46,166].

Some of these difficulties are now beginning to be alleviated by increased computer power, so that the dimensionless artificial diffusivities and hyperdiffusivities of simplified atmospheric models are now getting down towards the level of, perhaps, the real fluid diffusivities in small-scale laboratory experiments. By the time EOS is launched, we might expect to have a three-dimensional capability routinely and affordably capable of simulating the entire middle atmosphere at horizontal resolutions comparable to that of fig. 9, and compatible vertical resolutions conforming to the fundamental aspect ratio (8) at least in extratropical latitudes. This is still far too diffusive in comparison with the real atmosphere, but decisively better than what is possible currently. If model development goes hand in hand with the effort to improve our understanding — indeed a major reason for being involved in numerical modelling is that the models can be used for thought-experiments and hypothesis-testing, not just to fit observations by tuning parameters — then we can hope for real progress in answering many of the questions raised in this survey, and in beginning to make some of the qualitative insights we now have more quantitative.

I think we can, indeed, claim to have come a long way in terms of qualitative understanding, over the past two decades. A case in point, not mentioned so far, is what used to be called the enigma of the atmosphere's large-scale "negative viscosity", or upgradient large-scale eddy angular-momentum transport seen in coarse-grain observational "global-circulation statistics" (*e.g.*, [74], pp. 85, 150). Today this is no longer an enigma, since it can now be seen as a natural consequence of the general properties of wave-induced momentum transport illustrated in fig. 6. Rossby waves generated at and near the Earth's surface in middle and high latitudes have a tendency to propagate upwards and equatorwards, to some extent guided along the steep PV gradients at the tropopause and to some extent simply feeling the Earth's spherical geometry. These waves tend to break in the subtropics. Figure 8 represents the simplest of a hierarchy of conceptual and numerical models for this process; it is too simple in some ways since here the "surf zone" is the tropical upper troposphere and in the real atmosphere is profoundly affected, also, by massive injections of zero-PV air from the main centres of deep cumulonimbus convection (*e.g.*, [184,185]). Nevertheless, the robust one-signedess of irreversible Rossby-wave-induced angular momentum transport shows through in all the models, and seems to explain very neatly the persistent phenomenon that used to be thought of as "negative viscosity".

The same picture leads to a unified view of a whole range of other phenomena. One is the origin of the tropical upper-tropospheric cold vortices long studied as isolated entities by synoptic meteorologists: from the general perspective summarized here these are variants of the wave-breaking morphologies illustrated in fig. 9*b* and 10 (fig. 2 and 10 of ref. [1]). Another is the upper-tropospheric "mid-Pacific trough", which is the statistical signature of a preferred site for Rossby-wave breaking, or subtropical jet erosion, related to a large-scale anticyclonic circulation due to zero-PV injection at its own preferred site in and near the West Pacific (*e.g.*, [185]).

It has been a triumph of remote-sensing data acquisition and analysis that we have already moved so far from the position in 1982, when in a review of our understanding of major stratospheric

warmings I referred to "the near-impossibility of drawing isentropic maps of potential vorticity from even the best data analyses and thus seeing directly what is going on [dynamically]" [117]. That was what may people thought of as the conventional wisdom at the time. Now, within the lifetimes of most present here, there is a good chance that we shall, indeed, see and understand what is going on dynamically with enormously greater clarity, and begin to make our description of parts of the global atmospheric wave-turbulence jigsaw puzzle more quantitative. For instance, I have said hardly anything about how best to quantify the all-important fluxes or transports of Rossby- and gravity-wave activity, *e.g.* from the troposphere to the middle atmosphere, that drive the global mean circulation [186], although a fleeting mention was made of the Eliassen–Palm flux in connection with Table I, and of another celebrated manifestation of wave-mean interaction, the tropical quasi-biennial or "26-month" oscillation, at the end of sect. 10. These are all topics of current research activity. A largely unresolved, yet central, issue is quantification of the reflectivity of real Rossby-wave surf zones [117,126].

Perhaps I shall have done enough, in this survey, if I have succeeded in showing the central importance — indeed the practical necessity, from a dynamics and chemical-transport viewpoint — of getting improved observational estimates of isentropic distributions of that strange quantity, the Rossby-Ertel potential vorticity or PV. The invertibility principle tells us that we cannot claim to have a quantitative handle on the dynamics and chemical transport until we have a quantitative handle on isentropic distributions of PV — despite the very severe theoretical, observational and modelling challenge posed by the spatial inhomogeneity of such distributions, and their intertwined wavelike and turbulent aspects. Modelling with supercomputers has already given us a better idea of their likely qualitative appearance, when observed at higher spatial resolution than has hitherto been possible for the real atmosphere.

Observational information about chemical constituents, also, will be important from this viewpoint, as well as in connection with all the purely chemical questions that confront us, and the even more difficult questions of interactive dynamics and chemistry, some of which were raised in sect. 10. Perhaps one of the greatest milestones for the analysis of future satellite data will be the simultaneous retrieval of PV and chemicals. It is clearly necessary to attempt this — tall order though it may be — in order to get the best value from the data. It will presumably involve a fully integrated retrieval and data assimilation, in which simulated radiances from a sufficiently accurate chemical-dynamical model are matched to the observed radiances using modern "four-dimensional" methods based on the model and its adjoint [187], giving a chemically and dynamically consistent solution to the radiance inversion problem. This could take us far further than present-day, climatologically-based, one-at-a-time retrieval methods.

The simultaneous retrieval problem could well be said to be a Holy Grail in the quest for a better understanding of our planet's atmosphere. It should beckon the brightest young atmospheric scientists involved in modelling and data analyis, as computers become increasingly powerful. What might be a comparable Holy Grail for the brightest theoreticians and model-builders? Perhaps it could be said to be the discovery of radically new modelling techniques that somehow reduce our dependence on the eddy-diffusivity hypothesis, in future models of coupled radiation, dynamics and chemistry.

* * *

Many of the ideas that are central to this survey, and to the fluid dynamics of wave-mean interaction in general, grew out of material originally developed for an essay that shared the 1981 Adams Prize in the University of Cambridge. The development of these ideas has been influenced, in countless ways, by a number of scientific co-authors with whom it has been a privilege to work closely — and by other friends and colleagues too numerous to mention, except to say that an early interest in wave-mean interaction was stimulated by my doctoral thesis supervisor F. P. Bretherton, and in

different ways by D. O. Gough and E. A. Spiegel. The Atmospheric Dynamics research group at Cambridge has received generous support from the Natural Environment Research Council (through the UK Universities' Global Atmospheric Modelling Project and through the British Antarctic Survey), the Science and Engineering Research Council, the UK Meteorological Office, the US Office of Naval Research, the Nuffield Foundation, the UK Department of the Environment (in connection with the 1989 Airborne Arctic Stratospheric Expedition), and all three UK national supercomputer centres: the Atlas Laboratory at the Rutherford Appleton Laboratory, the University of London Computing Centre, and the Manchester Computing Centre. A shortened version of this review is to appear in *Science Progress*, ref. [6]. P. H. Haynes and W. A. Norton, who have made important contributions to the more recent phases of the research, have generously allowed me to use the unpublished material reproduced in fig. 8 and 9. I am also indebted to D. G. Andrews, D. G Dritschel, J. C. Farman, D. W. Fahey, R. R. Garcia, W. L. Grose, J. R. Holton, B. J. Hoskins, M. N. Juckes, D. Keyser, D. J. Lary, D. M. Murphy, A. O'Neill, R. A. Plumb, R. Rotunno, M. R. Schoeberl, T. G. Shepherd, S. Solomon, A. F. Tuck and T. E. VanZandt for helpful comments or correspondence regarding the material in several draft versions of the manuscript. Finally, I thank especially Guido Visconti without whom this survey would never have been written.

REFERENCES

[1] B. J. Hoskins, M. E. McIntyre and A. W. Robertson: On the use and significance of isentropic potential-vorticity maps. Q. J. Roy. Meteorol. Soc. **111**, 877–946. Also **113**, 402–404 (1987).

[2] M. E. McIntyre: Dynamics and tracer transport in the middle atmosphere: an overview of some recent developments. In: *Transport Processes in the Middle Atmosphere*, edited by G. Visconti and R. R. Garcia. Dordrecht, Reidel, 267–296 (1987). (Proc. NATO workshop, Erice.)

[3] M. E. McIntyre: Middle atmospheric dynamics and transport: some current challenges to our understanding. In *Dynamics, Transport and Photochemistry in the Middle Atmosphere of the Southern Hemisphere* (Proc. San Francisco NATO Workshop), edited by A. O'Neill, 1–18 (Dordrecht, Kluwer, 1990).

[4] M. E. McIntyre and W. A. Norton: Dissipative wave–mean interactions and the transport of vorticity or potential vorticity. J. Fluid Mech. **212** (G. K. Batchelor Festschrift Issue), 403–435 (1990); see also *Corrigendum*, **220**, 693 (1990).

[5] P. H. Haynes and M. E. McIntyre: On the conservation and impermeability theorems for potential vorticity. J. Atmos. Sci., **47**, 2021–2031 (1990).

[6] J. C. Farman, B. G. Gardiner and J. D. Shanklin: Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interactions. Nature, **315**, 207–210 (1985).

[7] J. J. Barnett and M. Corney: Middle atmosphere reference model derived from satellite data. *Handbook for The Middle Atmospheric Program, vol, 16: Atmospheric structure and its variations in the region 20–120 km: Draft of a new reference middle atmosphere*, edited by K. Labitzke, J. J. Barnett and B. Edwards, 47–143. Available from SCOSTEP secretariat, University of Illinois, 1406 W. Green St, Urbana, Ill. 61801, U.S.A. See also reference [18], chapter 6 (1985).

[8] C. J. Marks: Some features of the climatology of the middle atmosphere revealed by Nimbus 5 and 6. J. Atmos. Sci., **46**, 2485–2508 (1989).

[9] G. Brasseur and S. Solomon: *Aeronomy of the middle atmosphere.* Dordrecht, Reidel, 441pp. (1984).

[10] J. R. Holton: Dynamics of the middle atmosphere: its role in transport and troposphere–stratosphere coupling. In Proc. Internat. School Phys. "Enrico Fermi", CXV Course, edited by J. C. Gille and G. Visconti, 387–405. (North-Holland, Amsterdam, Oxford, New York, Toronto, 1992).

[11] S. B. Fels: Radiative–dynamical interactions in the middle atmosphere: in *Issues in Atmospheric and Oceanic modeling*, Advances in Geophysics (ed. S. Manabe) **28A**, 277–300, Academic Press, Orlando, Florida (1985).

[12] J. C. Gille and L. V. Lyjak: Radiative heating and cooling rates in the middle atmosphere. J. Atmos. Sci., **43**, 2215–2229 (1986).

[13] J. T. Kiehl and S. Solomon: On the radiative balance of the stratosphere. J. Atmos. Sci., **43**, 1525–1534 (1986).

[14] K. P. Shine: The middle atmosphere in the absence of dynamical heat fluxes. Q. J. Roy. Meteorol. Soc., **113**, 603–633 (1987). See also E. P. Olaguer, H. Yang and K. K. Tung: *A reexamination of the radiative balance of the stratosphere, J. Atmos. Sci.* **49**, 1242 (1992).

[15] R. C. Willson and H. S. Hudson: Solar luminosity variations in solar cycle 21. Nature, **332**, 810–812 (1988).

[16] A. C. Cogley and W. J. Borucki: Exponential approximation of daily average solar heating or photolysis. J. Atmos. Sci., **33**, 1347–1356 (1976). See also, for instance, the textbook by J. M. Wallace and P. V. Hobbs: *Atmospheric Science*, New York, Academic, p.346 (1977); also R. J. List (ed.): *Meteorological Tables*, 6th edn., Washington DC, Smithsonian Inst., table 132, p. 417 (1958).

[17] K-M. Xu and K. A. Emanuel: Is the tropical atmosphere conditionally unstable? Mon. Wea. Rev., **117**, 1471–1479 (1989).

[18] *Atmospheric ozone 1985: Assessment of our understanding of the processes controlling its present distribution and change: World Meteorological Organization Global Ozone Research and Monitoring Project Report No. 16,* Geneva: World Meteorol. Org. Available from: NASA Earth Science and Applications Division, code EEC National Aeronautics and Space Administration, Washington D.C. 20546, U. S. A.; In 3 volumes, 1095pp + 86pp refs (1985).

[19] F.-J. Lübken, U. von Zahn, A. Manson, C. Meek, U.-P. Hoppe, F. J. Schmidlin, J. Stegman, D. P. Murtagh, R. Rüster, G. Schmidt and H.-U. Widdel: Mean state densities, temperatures and winds during the MAC/SINE and MAC/EPSILON campaigns. J. Atmos. Terr. Phys., **52**, 955–970 (1990). Also J. Geophys. Res., **96**, 20841 (1991).

[20] E. Kopp, F. Bertin, L. G. Björn, P. H. G. Dickinson, C. R. Philbrick and G. Witt: The "CAMP" campaign 1982. Proc. 7th ESA Symp. on European Rocket and Balloon Programmes and Related Res., (Europ. Space Agency SP-229, July 1985) p. 117. See also *Adv. Space Res.*, **4**, 153 (1984). See also C. R. Philbrick, J. J. Barnett, R. Gerndt, D. Offermann, W. R. Pendleton, P. Schlyter, J. F. Schmidlin and G. Witt: Temperature measurements during the CAMP campaign, Adv. Space Res., **4**, 153 (1984).

[21] O. A. Avaste, A. V. Fedynsky, G. M. Grechko, V. I. Sevastyanov and Ch. I. Willmann: Advances in noctilucent cloud research in the space era. Pure Appl. Geophys. **118**, 528–580 (1980). (See p.543.)

[22] G. E. Thomas, J. J. Olivero, E. J. Jensen, W. Schröder and O. B. Toon: Relation between increasing methane and the presence of ice clouds at the mesopause. Nature, **338**, 490–492 (1989).

[23] M. Gadsden and W. Schröder: *Noctilucent clouds.* (Heidelberg, Springer, 1989).

[24] E. J. Jensen, G. E. Thomas and O. B. Toon: On the diurnal variation of noctilucent clouds. J. Geophys. Res., **94**, 14693-14702 (1989).

[25] U.-P. Hoppe, D. C. Fritts, I. M. Reid, P. Czechowsky, C. M. Hall and T. L. Hansen: Multiple-frequency studies of the high-latitude summer mesosphere: implications for scattering processes. J. Atmos. Terr. Phys., **52**, 907–926 (1990).

[26] E. J. Jensen, G. E. Thomas and B. B. Balsley: On the statistical correlation between polar mesospheric cloud occurrence and enhanced mesospheric radar echoes. Geophys. Res. Lett., **15**, 315–318 (1988).

[27] J. T. Houghton: The stratosphere and the mesosphere. Q. J. Roy. Meteorol. Soc. (Presidential Address), **104**, 1–29 (1978).

[28] D. G. Andrews, J. R. Holton and C. B. Leovy: *Middle Atmosphere Dynamics.* Academic Press, 489pp (1987).

[29] D. M. Cunnold, R. G. Prinn, R. A. Rasmussen, P. G. Simmonds, F. N. Alyea, C. A. Cardelino, A. J. Crawford, P. J. Fraser and R. D. Rosen: The Atmospheric Lifetime Experiment. 3. Lifetime methodology and application to three years of $CFCl_3$ data. J. Geophys. Res., **88**, 8379–8400 (1983).

[30] D. M. Cunnold, R. G. Prinn, R. A. Rasmussen, P. G. Simmonds, F. N. Alyea, C. A. Cardelino and A. J. Crawford: The atmospheric lifetime experiment. 4. Results for CF_2Cl_2 based on three years' data. J. Geophys. Res., **88**, 8401–8414 (1983).

[31] *Report of the International Ozone Trends Panel, 1988: World Meteorological Organization Global Ozone Research and Monitoring Project Report No. 18.* Geneva: World Meteorol. Org. Available from: NASA Earth Science and Applications Division, code EEC National Aeronautics and Space Administration, Washington D.C. 20546, U. S. A.; In 2 volumes, 828pp + 70pp refs (1988).

[32] *Scientific assessment of stratospheric ozone, 1989: World Meteorological Organization Global Ozone Research and Monitoring Project Report No. 20.* Geneva: World Meteorol. Org. Available from: NASA Earth Science and Applications Division, code EEC National Aeronautics and Space Administration, Washington D.C. 20546, U. S. A.; Volume I, 486pp (1989).

[33] L. R. Ember, P. L. Layman, W. Lepkowski and P. S. Zurer: Tending the global commons. Chemical and Engineering News, **64** (47), 14–64 (1986).

[34] P. S. Liss and P. G. Slater: Flux of gases across the air-sea interface. Nature, **247**, 181–184 (1974).

[35] C. Junge: The role of the oceans as a sink for chlorofluoromethanes and similar compounds. Z. Naturforsch., **31a**, 482–487 (1976).

[36] S. Solomon, J. T. Kiehl, R. R. Garcia and W. L. Grose: Tracer transport by the diabatic circulation deduced from satellite observations. J. Atmos. Sci. **43**, 1603–1617 (1986).

[37] J. C. Gille, L. V. Lyjak and A. K. Smith: The global residual mean circulation in the middle atmosphere for the northern winter period. J. Atmos. Sci., **44**, 1437–1452 (1987).

[38] S. Pawson and R. S. Harwood: Monthly mean diabatic circulations in the stratosphere. Q. J. Roy. Meteorol. Soc., **115**, 807–840 (1989).

[39] L. B. Callis, R. E. Boughner and J. D. Lambeth: The stratosphere: climatologies of the

radiative heating and cooling rates and the diabatically diagnosed net circulation fields. J. Geophys. Res., **92**, 5585–5607 (1987).

[40] P. H. Haynes, C. J. Marks, M. E. McIntyre, T. G. Shepherd and K. P. Shine: On the "downward control" of extratropical diabatic circulations by eddy-induced mean zonal forces. J. Atmos. Sci., **48**, 651 (1991).

[41] J. D. Mahlman, D. G. Andrews, D. L. Hartmann, T. Matsuno and R. G. Murgatroyd: Transport of Trace Constituents in the Stratosphere, MAP Handbook 3 (1981), p. 14. Also in Dynamics of the Middle Atmosphere, edited by J. R. Holton and T. Matsuno (Terrapub, Tokyo, and Reidel, Dordrecht, 1984), p. 387.

[42] J. E. Rosenfield, M. R. Schoeberl and M. A. Geller: A computation of the stratospheric diabatic circulation using an accurate radiative transfer model. J. Atmos. Sci., **44**, 859–876 (1987).

[43] R. J. Murgatroyd and F. Singleton: Possible meridional circulations in the stratosphere and mesosphere. Q. J. R. Meteorol. Soc., **87**, 125–135 (1961).

[44] R. A. Plumb and J. D. Mahlman: The zonally averaged transport characteristics of the GFDL general circulation/transport model. J. Atmos. Sci., **44**, 298–327 (1987).

[45] R. R. Garcia: Dynamics, radiation and photochemistry in the mesosphere: implications for the formation of noctilucent clouds. J. Geophys. Res., **94**, 14605–14615 (1989). [Special Issue for the International Workshop on Noctilucent Clouds.]

[46] M. E. McIntyre: On dynamics and transport near the polar mesopause in summer. J. Geophys. Res., **94**, 14617–14628 (1989). [Special Issue for the International Workshop on Noctilucent Clouds.]

[47] S. Solomon: Progress towards a quantitative understanding of Antarctic ozone depletion. Nature, **347**, 347 (1990).

[48] U. Schmidt and A. Khedim: *In situ measurements of carbon dioxide in the winter Arctic vortex and at midlatitudes: an indicator of the age of stratospheric air*, Geophys. Res. Lett., **18**, 763–766 (1991). [The technique is to exploit the well known trend in tropospheric CO_2 mixing ratio to estimate how long the air observed in the vortex has been in the stratosphere. The results suggest that this "CO_2 age" has average value 5.6 ± 1.1 year, but may fluctuate systematically by almost a factor of two on a timescale of 3 to 5 years.] See also W. H. Pollock, L. E. Heidt, R. A. Lueb, J. F. Vedder, M. J. Mills and S. Solomon: *On the age of stratospheric air and ozone depletion potentials in polar regions*, J. Geophys. Res., **97**, 12993–12999 (1992). [This paper suggests CO_2 ages of order "3–5y" at pressure altitudes of order 17–21km.] Also L. Heidt, S. Hovde, A. F. Tuck, J. F. Vedder and R. Weiss: *The age of the air in the Arctic lower stratospheric vortex during late winter 1988/9*, J. Geophys. Res., submitted Nov 1992. [Nominal CO_2 age $\simeq$ 4y but highly variable; probably consistent with Schmidt and Khedim.] Also, for background discussion and further bibliography, T. M. Hall and M. J. Prather: *Simulations of the trend and annual cycle in stratospheric CO_2*, J. Geophys. Res., **98**, 10573–10581 (1993).

[49] M. Loewenstein, J. R. Podolske and S. E. Strahan: ATLAS instrument characterization: accuracy of the AASE and AAOE nitrous oxide data sets. Geophys. Res. Lett., **17**, 480–484 (1990).

[50] S. Solomon, R. R. Garcia and F. Stordal: Transport processes and ozone perturbations. J. Geophys. Res. **90 D**, 12981–12989 (1985).

[51] D. Cariolle and M. Déqué: Southern hemisphere medium-scale waves and total ozone

disturbances in a spectral general circulation model. J. Geophys. Res. **91 D**, 10825–10846 (1986).

[52] R. Toumi, B. J. Kerridge and J. A. Pyle: Highly vibrationally excited oxygen as a potential source of ozone in the upper stratosphere and mesosphere. Nature, **351**, 217–219 (1991).

[53] A. W. Brewer and A. W. Wilson: The regions of formation of atmospheric ozone. Q. J. Roy. Meteorol. Soc., **94**, 249–265 (1968). (See also references therein.)

[54] M. Loewenstein, J. R. Podolske, K. R. Chan and S. E. Strahan: N_2O as a dynamical tracer in the Arctic vortex. Geophys. Res. Lett., **17**, 477–480 (1990).

[55] D. W. Fahey, D. M. Murphy, K. K. Kelly, M. K. W. Ko, M. H. Proffitt, C. S. Eubank, G. V. Ferry, M. Loewenstein and K. R Chan: Measurements of nitric oxide and total reactive nitrogen in the Antarctic stratosphere: observations and chemical implications. J. Geophys. Res., **94**, 16665–16681 (1989).

[56] H. W. Feely and J. Spar: Tungsten-185 from nuclear bomb tests as a tracer for stratospheric meteorology. Nature, **188**, 1062–1064 (1960).

[57] J. R. Holton, 1986: Meridional distribution of stratospheric trace constituents. J. Atmos. Sci., **43**, 1238–1242.

[58] J. D. Mahlman: Mechanistic interpretation of stratospheric tracer transport, in *Issues in Atmospheric and Oceanic Modelling, Advances in Geophysics* (ed. S. Manabe), vol. **28A**, Academic Press, Orlando, Florida, pp.301–323 (1985).

[59] D. R. Johnson and W. K. Downey: Azimuthally averaged transport and budget equations for storms: quasi-Lagrangian diagnostics 1. Mon. Wea. Rev., **103**, 956–979 (1975).

[60] K. K. Tung: On the two-dimensional transport of stratospheric trace gases in isentropic coordinates. J. Atmos. Sci., **39**, 2330–2355 (1982).

[61] H. Kida: General circulation of air parcels and transport characteristics derived from a hemispheric GCM. Part I: a determination of advective mass flow in the lower stratosphere. J. Meteorol. Soc. Japan, **61**, 171–188 (1983).

[62] M. E. McIntyre: Towards a Lagrangian-mean description of stratospheric circulations and chemical transports. Phil. Trans. Roy. Soc. **A 296**, 129–148 (1980). (Special Middle Atmosphere issue.)

[63] M. R. Schoeberl and L. R. Lait: Conservative coordinate transformations for atmospheric measurements. In Proc. Internat. School Phys. "Enrico Fermi", CXV Course, edited by J. C. Gille and G. Visconti, 419–431. (North-Holland, Amsterdam, Oxford, New York, Toronto, 1992).

[64] M. R. Schoeberl, L. R. Lait, P. A. Newman, R. L. Martin, M. H. Proffitt, D. L. Hartmann, M. Loewenstein, J. Podolske, S. E. Strahan, J. Anderson, Chan, K. R. and B. Gary: Reconstruction of the constituent distribution and trends in the Antarctic polar vortex from ER-2 flight observations. J. Geophys. Res., **94**, 16815–16846 (1989). (Airborne Antarctic Ozone Experiment special issue.)

[65] L. R. Lait, M. R. Schoeberl, P. A. Newman, M. H. Proffitt, K. K. Kelly, M. Loewenstein, J. R. Podolske, S. E. Strahan, K. R. Chan, B. Gary, E. Browell, M. P. McCormick and A. Torres: Reconstruction of O_3 and N_2O fields from ER-2, DC-8, and balloon observations. Geophys. Res. Lett., **17**, 521–524 (1990). (Airborne Arctic Stratospheric Expedition special issue.)

[66] M. R. Schoeberl, L. R. Lait, P. A. Newman and J. E. Rosenfield: The structure of the polar vortex. J. Geophys. Res., **97**, 7859, special Polar Ozone Issue (1992).

[67] R. A. Plumb: Ozone depletion in the Arctic. Nature, **347**, 20–21 (1990).

[68] D. G. Andrews and M. E. McIntyre: An exact theory of nonlinear waves on a Lagrangian-mean flow. J. Fluid Mech., **89**, 609-646 (1978).

[69] J. Thuburn: Baroclinic wave life cycles, climate simulations and cross-isentropic mass flow in a hybrid isentropic coordinate general circulation model. Q. J. R. Meteorol. Soc., submitted (1992).

[70] T. J. Dunkerton: Nonlinear propagation of zonal winds in an atmosphere with Newtonian cooling and equatorial wavedriving. J. Atmos. Sci., **48,** 236–263 (1991) (1986).

[71] J. G. Charney: On the scale of atmospheric motions. Geofysiske Publ. **17** (2), 3-17 (1948). Reprinted in ref. [109].

[72] R. R. Garcia: On the mean meridional circulation of the middle atmosphere. J. Atmos. Sci., **44**, 3599–3609 (1987).

[73] T. J. Dunkerton: Body force circulations in a compressible atmosphere: key concepts. Pure Appl. Geophys., **130**, 243-262 (1989).

[74] E. N. Lorenz: *The Nature and Theory of the General Circulation of the Atmosphere.* Geneva: World Meteorol. Org., 161pp. (1967).

[75] S. B. Fels, J. D. Mahlman, M. D. Schwarzkopf, and R. W. Sinclair: Stratospheric sensitivity to perturbations in ozone and carbon dioxide: radiative and dynamical response. J. Atmos. Sci., **37**, 2265–2297 (1980).

[76] B. A. Boville and D. P. Baumhefner: Simulated forecast error and climate drift resulting from the omission of the upper stratosphere in numerical models. Mon. Wea. Rev., **118**, 1517-1530 (1990).

[77] J. D. Mahlman and L. J. Umscheid: Dynamics of the middle atmosphere: successes and problems of the GFDL 'SKYHI" general circulation model. In *Dynamics of the Middle Atmosphere* (edited by J. R. Holton and T. Matsuno), 501–525, Tokyo, Terrapub, and Dordrecht, Reidel (1984).

[78] R. A. Plumb: On the seasonal cycle of stratospheric planetary waves. Pure Appl. Geophys., **130**, 233–242 (1989).

[79] R. E. Dickinson: Theory of planetary wave-zonal flow interaction. J. Atmos. Sci., **26**, 73-81 (1969).

[80] L. Brillouin: On radiation stresses (in French). Annales de Physique **4**, 528–586 (1925).

[81] M. S. Longuet-Higgins and R. W. Stewart: Radiation stress in water waves; a physical discussion, with applications. Deep-Sea Research, **11**, 529–562 (1964).

[82] M. J. Lighthill: *Waves in Fluids.* Cambridge University Press, 504 pp (1978). See also M. J. Lighthill: Emendations to a proof in the general three-dimensional theory of oscillating sources of waves. Proc. Roy. Soc. Lond. Ser. A, **427**, 31 (1990).

[83] M. E. McIntyre: On the "wave momentum" myth. J. Fluid Mech. **106**, 331–347 (1981). For some further relevant points, see §5 of M. E. McIntyre, 1993: *On the role of wave propagation and wave breaking in atmosphere–ocean dynamics*, Sectional Lecture, Proc. XVIII Int. Congr. Theoret. Appl. Mech., Haifa, ed. S. R. Bodner, J. Singer, A. Solan, Z.

Hashin. Elsevier, 281–304.

[84] M. Van Dyke: An Album of Fluid Motion. Stanford, Parabolic Press (1982).

[85] M. L. Banner and O. M. Phillips: On the incipient breaking of small scale waves. J. Fluid Mech., **65**, 647-656 (1974). Also **211**, 464 (1990).

[86] M. E. McIntyre and T. N. Palmer: The "surf zone" in the stratosphere. J. Atm. Terr. Phys., **46**, 825-849 (1984). See also J. R. Holton and M. P. Baldwin: Climatology of the stratospheric polar vortex and planetary wave breaking. J. Atmos. Sci., **45**, 1123 (1988). Also, e.g., J. Austin and N. Butchart: A study of air particle motions during a stratospheric warming and their influence on photochemistry. Q. J. Roy. Meteorol. Soc., **115**, 841 (1989).

[87] M. E. McIntyre and T. N. Palmer: A note on the general concept of wave breaking for Rossby and gravity waves. Pure Appl. Geophys., **123**, 964–975 (1985).

[88] M. E. McIntyre and T. G. Shepherd: An exact local conservation theorem for finite-amplitude disturbances to nonparallel shear flows, with remarks on Hamiltonian structure and on Arnol'd's stability theorems. J. Fluid Mech., **181**, 527-565 (1987).

[89] T. J. Dunkerton, C.-P. F. Hsu and M. E. McIntyre: Some Eulerian and Lagrangian diagnostics for a model stratospheric warming, J. Atmos. Sci., **38**, 819 (1981).

[90] H. J. Edmon, B. J. Hoskins and M. E. McIntyre: Eliassen-Palm cross-sections for the troposphere, J. Atmos. Sci., **37**, 2600 (1980). Also Corrigendum, J. Atmos. Sci., **38**, 1115 (1981), especially second last item.

[91] C. O. Hines: Gravity waves in the atmosphere. Nature, **239**, 73–78 (1972).

[92] D. C. Fritts: Gravity wave saturation in the middle atmosphere: A review of theory and observations. Revs. Geophys. Space Phys. **22**, 275–308 (1984).

[93] A. Hauchecorne, M. L. Chanin and R. Wilson: Mesospheric temperature inversion and gravity wave breaking. Geophys. Res. Lett., **14**, 933–936 (1987).

[94] I. M. Reid and R. A. Vincent: Measurements of mesopause/mesospheric gravity wave momentum fluxes and mean-flow accelerations at Adelaide, Australia. J. Atmos. Terrest. Phys., **49**, 443-460 (1987).

[95] D. C. Fritts: Gravity waves in the middle atmosphere of the Southern Hemisphere. In *Dynamics, Transport and Photochemistry in the Middle Atmosphere of the Southern Hemisphere* (Proc. San Francisco NATO Workshop), edited by A. O'Neill, 171–189. (Dordrecht, Kluwer, 1990).

[96] T. Tsuda, Y. Murayama, M. Yamamoto, S. Kato and S. Fukao: Seasonal variation of momentum flux in the mesosphere observed with the MU radar. Geophys. Res. Lett., **17**, 725 (1990).

[97] T. E. VanZandt and D. C. Fritts: A theory of enhanced saturation of the gravity wave spectrum due to increases in atmospheric stability. Pure Appl. Geophys., **130**, 399-420 (1989).

[98] J. R. Booker and F. P. Bretherton: The critical layer for internal gravity waves in a shear flow. J. Fluid Mech., **27**, 513-539 (1967).

re [99] NOTE: ref. [99] of Appendix III (Rossby 1936) has been re-numbered as ref. [102].

[99] R. S. Lindzen: Turbulence and stress owing to gravity wave amd tidal breakdown. J.

Geophys. Res., **86**, 9707-9714 (1981).

[100] J. G. Charney and P. G. Drazin: Propagation of planetary-scale disturbances from the lower into the upper atmosphere. J. Geophys. Res., **66**, 83-109 (1961). Reprinted in ref. [109].

[101] C. G. Rossby: On temperature changes in the stratosphere resulting from shrinking and stretching. Beitr. Phys. Freien Atmos., **24**, 53-59 (1938).

[102] C. G. Rossby: Dynamics of steady ocean currents in the light of experimental fluid mechanics. Papers in Physical Oceanography and Meteorology (Mass. Inst. of Technology and Woods Hole Oc. Inst.), **5** (1), 43pp (1936). See especially eq. (75).

[103] C. G. Rossby: Planetary flow patterns in the atmosphere. Q. J. R. Meteorol. Soc., **66**, Suppl., 68–97 (1940).

[104] H. Ertel: Ein Neuer hydrodynamischer Wirbelsatz. Met. Z., **59**, 277–281 (1942). (English translation in W. Schröder: *Geophysical Hydrodynamics and Ertel's potential vorticity (Selected Papers of Hans Ertel)*, pp.33–40. Interdivisional Commission of History, No.12, International Association of Geomagnetism and Aeronomy, Hechelstrasse 8, D-2820 Bremen-Rönnebeck, Germany, 218 pp (1991).

[105] D. Keyser and R. Rotunno: On the formation of potential-vorticity anomalies in upper-level jet-front systems. Mon. Wea. Rev., **118**, 1914–1921 (1990).

[106] L. W. Uccellini, D. Keyser, K. F. Brill and C. H. Wash: The Presidents' Day cyclone of 18-19 February 1979: Influence of upstream trough amplification and associated tropopause folding on rapid cyclogenesis. Mon. Wea. Rev., **113**, 962-988 (1985).

[107] B. J. Hoskins and P. Berrisford: A potential-vorticity perspective of the storm of 15–16 October 1987. Weather, **43**, 122–129 (1988).

[108] E. Kleinschmidt: Uber Aufbau und Entstehung von Zyklonen (1,2,3. Teil) Met. Runds., **3**, 1-6, 54-61, 89-96 (1950).

[109] R. S. Lindzen, E. N. Lorenz and G. W. Platzman, (eds.): *The Atmosphere: a Challenge (The Science of Jule Gregory Charney)*. Boston, Amer. Meteorol. Soc., 321pp (1990).

[110] A. J. Thorpe: Synoptic scale disturbances with circular symmetry. Mon. Wea. Rev. **114**, 1384–1389 (1986).

[111] C. A. Davis and K. A. Emanuel: Potential vorticity diagnostics of cyclogenesis. Mon. Wea. Rev., **119**, 1930 (1991). See also C. A. Davis: *Piecewise potential vorticity inversion*, J. Atmos. Sci., **49**, 1397–1411 (1992). Also D. J. Raymond: *Nonlinear balance and potential-vorticity thinking at large Rossby number*, Q. J. Roy. Meteorol. Soc., **118**, 987–1015 (1992). Also M. E. McIntyre: *Isentropic distributions of potential vorticity and their relevance to tropical cyclone dynamics.* Proc. ICSU/WMO International Symposium on Tropical Cyclone Disasters, ed. J. Lighthill, Z. Zheng, G. Holland and K. Emanuel. Peking [*sic*] University Press, Beijing, China, 143–156 (1993).

[112] W. A. Robinson: Analysis of LIMS data by potential vorticity inversion. J. Atmos. Sci., **45**, 2319–2342 (1988).

[113] M. E. McIntyre and W. A. Norton: Potential-vorticity inversion on a hemisphere. J. Atmos. Sci., to appear (1993).

[114] A. Eliassen: Geostrophy. Q. J. Roy. Meteorol. Soc., **110**, 1-12 (1984).

[115] P. D. Killworth and M. E. McIntyre: Do Rossby-wave critical layers absorb, reflect or

over-reflect? J. Fluid Mech., **161**, 449-492 (1985).

[116] P. H. Haynes: The effect of barotropic instability on the nonlinear evolution of a Rossby wave critical layer. J. Fluid Mech., **207**, 231–266 (1989).

[117] M. E. McIntyre: How well do we understand the dynamics of stratospheric warmings? J. Meteorol. Soc. Japan **60**, 37-65 (1982).

[118] W. A. Robinson: The behavior of planetary wave 2 in preconditioned zonal flows. J. Atmos. Sci., **43**, 3109–3121 (1986).

[119] J. Egger: Some aspects of potential vorticity inversion. J. Atmos. Sci., **47**, 1269-1275 (1990).

[120] A. M. Obukhov: On the dynamics of a stratified liquid. Dokl. Akad. Nauk SSSR, **145** (6), 1239–1242 (1962). English transl. in Soviet Physics - Doklady **7**, 682–684 (1962).

[121] W. A. Norton: *Balance and potential vorticity inversion in atmospheric dynamics.* PhD Thesis, University of Cambridge (1988), 167 pp.

[122] C. Mattocks and R. Bleck: Jet streak dynamics and geostrophic adjustment processes during the initial stages of lee cyclogenesis. Mon. Wea. Rev., **114**, 2033-2056 (1986).

[123] D. C. Fritts and P. K. Rastogi: Convective and dynamical instabilities due to gravity wave motions in the lower and middle atmosphere: theory and observations. Radio Science, **20**, 1247–1277 (1985).

[124] S. A. Clough, N. S. Grahame and A. O'Neill: Potential vorticity in the stratosphere derived using data from satellites. Q. J. Roy. Meteorol. Soc., **111**, 335–358 (1985).

[125] M. N. Juckes and M. E. McIntyre: A high resolution, one-layer model of breaking planetary waves in the stratosphere. Nature, **328**, 590–596 (1987).

[126] R. R. Garcia: Parameterization of planetary wave breaking in the middle atmosphere. J. Atmos. Sci., 48, 1405 (1991). See also R. R. Garcia, F. Stordal, S. Solomon and J. T. Kiehl: *A new numerical model of the middle atmosphere. 1. Dynamics and transport of tropospheric trace source gases*, J. Geophys. Res., **97**, 12967–12991 (1992).

[127] W. A. Robinson: Irreversible wave–mean flow interactions in a mechanistic model of the stratosphere. J. Atmos. Sci., **45**, 3413–3430 (1988).

[128] R. W. Stewart and R. E. Thomson: Re-examination of vorticity transfer theory. Proc. Roy. Soc. Lond., **A 354**, 1–8 (1977).

[129] P. Welander: Studies on the general development of motion in a two-dimensional ideal fluid. Tellus, **7**, 141–156 (1955).

[130] S. Childress, G. R. Ierley, E. A. Spiegel and W. R. Young: Blow-up of unsteady two-dimensional Euler and Navier-Stokes solutions having stagnation-point form. J. Fluid Mech., **203**, 1-22 (1989).

[131] M. N. Juckes: A shallow water model of the winter stratosphere. J. Atmos. Sci., **46**, 2934-2955 (1989).

[132] M. N. Juckes, M. E. McIntyre and W. A. Norton: High-resolution barotropic simulations of breaking stratospheric planetary waves and polar-vortex edge formation. Q. J. Roy. Meteorol. Soc., to be submitted (1992).

[133] P. H. Haynes, 1990: High-resolution three-dimensional modeling of stratsopheric flows: quasi-two-dimensional turbulence dominated by a single vortex. In: *Topological Fluid*

Mechanics, edited by H. K. Moffatt and A. Tsinober, pp.345–354. Cambridge University Press.

[134] M. L. Salby, D. O'Sullivan, R. R. Garcia and P. Callaghan: Air motions accompanying the development of a planetary wave critical layer. J. Atmos. Sci., **47**, 1179-1204 (1990).

[135] D. G. Dritschel: The repeated filamentation of two-dimensional vorticity interfaces. *J. Fluid Mech.* **194**, 511-547 (1988).

[136] D. G. Dritschel: Strain-induced vortex stripping, in *Mathematical Aspects of Vortex Dynamics*, edited by R. E. Caflisch, page 107 (Philadelphia, Society for Industrial and Applied Mathematics, 1989).

[137] L. M. Polvani and R. A. Plumb: Rossby wave breaking, microbreaking, filamentation and secondary vortex formation: the dynamics of a perturbed vortex. J. Atmos. Sci., **49**, 462 (1992).

[138] C. B. Leovy, C. -R. Sun, M. H. Hitchman, E. E. Remsberg, J. M. Russell, L. L. Gordley, J. C. Gille, L. V. Lyjak: Transport of ozone in the middle stratosphere: evidence for planetary wave breaking. J. Atmos. Sci., **42**, 230-244 (1985).

[139] W. L. Grose, J. E. Nealy, R. E. Turner and W. T. Blackshear: Modeling the transport of chemically active constituents in the stratosphere. In: *Transport Processes in the Middle Atmosphere*, edited by G. Visconti and R. R. Garcia. Dordrecht, Reidel, 229–250 (1987). (Proc. NATO workshop, Erice.)

[140] A. O'Neill and V. D. Pope: Simulations of linear and nonlinear disturbances in the stratosphere. Quart. J. Roy. Meteorol. Soc., **114**, 1063–1110 (1988).

[141] J. A. Kaye, R. B. Rood, D. J. Allen, E. M. Larson and C. H. Jackman: Three dimensional simulation of spatial and temporal variability of stratospheric hydrogen chloride. Geophys. Res. Lett., **16**, 1149-1152 (1989).

[142] W. A. Norton: *Breaking Rossby waves in a model stratosphere diagnosed by a vortex-following coordinate system and a technique for advecting material contours*, J. Atmos. Sci., **51**, 654–673 (1994). See also the independent work of D. W. Waugh and R. A. Plumb: *Contour advection with surgery: a technique for investigating fine scale structure in tracer transport.* J. Atmos. Sci., **51**, 530–540 (1993).

[143] C. R. Mechoso: The final warming of the stratosphere. In *Dynamics, Transport and Photochemistry in the Middle Atmosphere of the Southern Hemisphere* (Proc. San Francisco NATO Workshop), edited by A. O'Neill, 55–69 (Dordrecht, Kluwer, 1990).

[144] J. C. McWilliams: Statistical properties of decaying geostrophic turbulence. J. Fluid Mech., **198**, 199-230 (1989).

[145] D. G. Dritschel: On the stabilization of a two-dimensional vortex strip by adverse shear. J. Fluid Mech., **206**, 193–221 (1989).

[146] D. G. Dritschel, M. N. Juckes, P. H. Haynes, and T. G. Shepherd: The stability of a two-dimensional vorticity filament under uniform strain. J. Fluid Mech., **230**, 647 (1991).

[147] M. H. Proffitt and R. J. McLaughlin: Fast-response dual-beam UV-absorption ozone photometer suitable for use on stratospheric balloons. Rev. Sci. Instrum., **54**, 1719-1728 (1983).

[148] A. F. Tuck, T. Davies, S. J. Hovde, M. Noguer-Alba, D. W. Fahey, S. R. Kawa, K. K. Kelly, D. M. Murphy, M. H. Proffitt, J. J. Margitan, M. Loewenstein, J. R. Podolske, S. E.

Strahan and K. R. Chan: Polar stratospheric cloud processed air and potential vorticity in the Northern Hemisphere lower stratosphere at mid-latitudes during winter. J. Geophys. Res., **97**, 7883 (1992).

[149] R. J. Atkinson, W. A. Matthews, P. A. Newman, and R. A. Plumb: Evidence of the mid-latitude impact of Antarctic ozone depletion. Nature, **340**, 290–294 (1989).

[150] R. R. Garcia and S. Solomon: A possible relationship between interannual variability in Antarctic ozone and the quasi-binennial oscillation. Geophys. Res. Lett., **14**, 848 (1987).

[151] J. C. Farman, B. G. Gardiner and J. D. Shanklin: How deep is an "ozone hole"? Nature, London, **336**, 198 (1988).

[152] N. Butchart and E. E. Remsberg: Area of the stratospheric polar vortex as a diagnostic for tracer transport on an isentropic surface. J. Atmos. Sci., **43**, 1319-1339 (1986).

[153] P. H. Haynes and M. E. McIntyre: On the evolution of vorticity and potential vorticity in the presence of diabatic heating and frictional or other forces. J. Atmos. Sci., **44**, 828–841 (1987). See also ref. [5].

[154] A. O'Neill and V. D. Pope: The seasonal evolution of the extra-tropical stratosphere in the southern and northern hemispheres: systematic changes in potential vorticity and the non-conservative effects of radiation. In *Dynamics, Transport and Photochemistry in the Middle Atmosphere of the Southern Hemisphere* (Proc. San Francisco NATO Workshop), edited by A. O'Neill, 33–54 (Dordrecht, Kluwer, 1990).

[155] M. E. McIntyre: On the Antarctic ozone hole. J. Atmos. Terrest. Phys., **51**, 29–43 (1989). (Corrigendum: delete "and temperature changes" on page 31b, lines 4–5; see also reference [40].)

[156] T. N. Palmer, G. J. Shutts and R. Swinbank: Alleviation of a systematic westerly bias in general circulation and numerical weather prediction models through an orographic gravity wave drag parametrization, Q. J. R. Meteorol. Soc., 112, 1001 (1986). See also ref. [183].

[157] E. Klinker and P. Sardeshmukh: The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements, J. Atmos. Sci., **49**, 608 (1992).

[158] R. S. Lindzen and M. R. Schoeberl: A note on the limits of Rossby-wave amplitudes. J. Atmos. Sci., **39**, 1171-1174 (1982).

[159] J. C. Farman, R. J. Murgatroyd, A. M. Siklnickas and B. A. Thrush: Ozone photochemistry in the Antarctic stratosphere in summer, Q. J. R. Meteorol. Soc., **111**, 1013 (1985).

[160] T. J. Dunkerton and D. P. Delisi: The subtropical mesospheric jet observed by the Nimbus 7 Limb Infrared Monitor of the Stratosphere. J. Geophys. Res., **90**, 10681 (1985).

[161] T. Sato and R. F. Woodman: Fine altitude resolution observations of stratospheric turbulent layers by the Arecibo 430MHz radar. J. Atmos. Sci., **39**, 2546-2552 (1982).

[162] J. Barat: The fine structure of the stratospheric flow revealed by differential sounding. J. Geophys. Res., **88**, 5219–5228 (1983).

[163] M. Prather and A. H. Jaffe: Global impact of the Antarctic ozone hole: chemical propagation. J. Geophys. Res., **95**, 3473-3492 (1990).

[164] A. F. Tuck: A comparison of one-, two-, and three-dimensional model representations of stratospheric gases. Phil. Trans. Roy. Soc. London, **A 290**, 477-494 (1979).

[165] U. Frisch: Fully developed turbulence and intermittency. In Proc. Internat. School Phys. "Enrico Fermi", LXXXVIII Course, edited by M. Ghil, R. Benzi and G. Parisi, 71–88. North-Holland, Amsterdam, Oxford, New York, Toronto (1985).

[166] E. M. Dewan: Turbulent vertical transport due to thin intermittent mixing layers in the stratosphere and other stable fluids, Science, **211**, 1041 (1981). Also J. R. Ledwell and A. J. Watson: The Santa Monica Basin tracer experiment: a study of diapycnal and isopycnal mixing. J. Geophys. Res., **96**, 8695 (1991).

[167] M. H. Proffitt, J. J. Margitan, K. K. Kelly, M. Loewenstein, J. R. Podolske and K. R. Chan: Ozone loss in the Arctic polar vortex inferred from high-alititude aircraft measurements. Nature, **347**, 31–36 (1990).

[168] M. H. Proffitt, S. Solomon and M. Loewenstein: Comparison of 2-D model simulations of ozone and nitrous oxide at high latitudes with stratospheric measurements, J. Geophys. Res., **97**, 939 (1992).

[169] D. W. Fahey, S. Solomon, S. R. Kawa, M. Loewenstein, J. R. Podolske, S. E. Strahan and K. R. Chan: A diagnostic for denitrification in the winter polar stratospheres, Nature (London), **345**, 698 (1990).

[170] A. F. Tuck, R. T. Watson, E. P. Condon, J. J. Margitan and O. B. Toon: The planning and execution of ER-2 and DC-8 aircraft flights over Antarctica, August and September 1987. J. Geophys. Res., **94**, 11181–11222 (1989). [Special Issue on the Airborne Antarctic Ozone Experiment; see also the other papers therein.]

[171] R. Turco, A. Plumb and E. Condon: The Airborne Arctic Stratospheric Expedition: Prologue. Geophys. Res. Lett. **17**, 313–316 (1990). [Special issue on the AASE; see also the other papers therein.]

[172] M. H. Proffitt, K. K. Kelly, J. A. Powell, M. Loewenstein, J. R. Podolske, S. E. Strahan and K. R. Chan: Evidence for diabatic cooling and poleward transport within and around the 1987 Antarctic ozone hole. J. Geophys. Res., **94**, 16797–16813 (1989). (Special issue on the Airborne Antarctic Ozone Experiment.)

[173] S. P. Cooper and W. A. Norton: Towards an improved person-machine interface in atmospheric modelling. In: *Proc. NERC/IMA Symposium on the Use of Numerical Models in the Evironmental Sciences*, edited by D.G. Farmer and M.J. Rycroft, 189–197. Oxford University Press (1991).

[174] R. B. Rood, J. E. Nielsen, R. S. Stolarski, A. R. Douglass, J. A. Kaye and D. J. Allen: Episodic total ozone minima and associated effects on heterogeneous chemistry and lower stratospheric transport. J. Geophys. Res., **97**, 7979 (1992).

[175] M. P. McCormick, C. R. Trepte and G. S. Kent, : Spatial changes in the stratospheric aerosol associated with the north polar vortex. Geophys. Res. Lett., **10**, 941-944 (1983).

[176] J. C. Farman: Ozone measurements at British Antarctic Survey stations, Phil. Trans. Roy. Soc. Lond., **B279**, 261–271 (1977).

[177] J. Sommeria, S. D. Meyers and H. L. Swinney: Laboratory model of a planetary eastward jet, Nature (London), **337**, 58 (1989). Also J. Sommeria, S. D. Meyers and H. L. Swinney: Experiments on vortices and Rossby waves in eastward and westward jets, in *Nonlinear Topics in Ocean Physics*, edited by A. R. Osborne (North-Holland, Amsterdam, 1991) p. 227.

[178] A. F. Tuck: Synoptic and chemical evolution of the Antarctic vortex in late winter and

early spring, 1987. J. Geophys. Res., **94**, 11687–11737 (1989). (Airborne Antarctic Ozone Experiment Special Issue.)

[179] P. H. Haynes and W. E. Ward: The effect of realistic radiative transfer on potential vorticity structures, including the influence of background shear and strain. J. Atmos. Sci., to appear (1992).

[180] R. P. Feynman, R. B. Leighton and M. Sands: The Feynman Lectures on Physics: Vol. 2, Mainly Electromagnetism and Matter. (Addison-Wesley, Reading, Mass., 1964).

[181] C. Truesdell: Proof that Ertel's vorticity theorem holds on average for any medium suffering no tangential acceleration on the boudnary. Geofisica pura e applicata (Pure and Appl. Geophys.), **19**, 167 (1951).

[182] A. J. Thorpe and K. A. Emanuel; Frontogenesis in the presence of small stability to slantwise convection. J. Atmos. Sci., **42**, 1809 (1985).

[183] C. McLandress and N. A. McFarlane: Interactions between orographic gravity wave drag and forced stationary planetary waves in the winter northern hemisphere. J. Atmos. Sci., to appear (1992).

[184] P. D. Sardeshmukh and B. J. Hoskins: The generation of global rotational flow by steady idealised tropical divergence. J. Atmos. Sci., **45**, 1228 (1988).

[185] V. Magaña and M. Yanai: Tropical-midlatitude interaction on the time scale of 30 to 60 days during the northern summer of 1979. J. Climate, **4**, 180–201 (1991).

[186] P. H. Haynes: Forced, dissipative generalizations of finite-amplitude wave-activity conservation relations for zonal and non-zonal basic flows. J. Atmos. Sci., **45**, 2352–2362 (1988).

[187] P. Courtier and O. Talagrand: Variational assimilation of meteorological observations with the adjoint vorticity equation. II: Numerical results. Q. J. Roy. Meteorol. Soc., **113**, 1329–1347 (1987).

[188] R. A. Plumb and M. K. W. Ko: Interrelationships between mixing ratios of long-lived stratospheric constituents, J. Geophys. Res., **97**, 10145–10156 (1992).

FIGURE CAPTIONS

Fig. 1. Averaged observed temperature structure of the troposphere, stratosphere, and mesosphere, for January conditions; the light shading shows the warmest regions and the dark shading the coldest. The scale on the right is in nominal e-folding pressure scale heights of about 7 km, so that for instance the stratopause is at about 50 km. Temperatures are in degrees Kelvin (so that 273 K is freezing point at sea level). The equivalent pressure altitude shown on the left varies by a factor e^{12} or about five decimal orders of magnitude; the density varies over a comparable range since it is accurately given by the perfect gas law, and temperatures vary relatively little.

Fig. 2a. Diurnally averaged relative solar irradiance at midsummer in the northern hemisphere, as a function of latitude, according to eq. (1). Relative solar irradiance means incident energy per unit time per unit horizontal area, expressed as a fraction of the solar energy (flux) per unit time per unit area normal to the direction of the Sun (eq. (2)).

Fig. 2b. Diurnally averaged relative solar irradiance as a function of season, from ref. [16].

Fig. 3. Mass transport streamlines of the global-scale mean circulation for January 1979 (light curves), from reference [36], estimated using satellite data. The picture gives the typical latitude and height dependence of the longitude and time averaged mass circulation of the stratosphere, defined in a quasi-Lagrangian sense giving a simplified, but roughly correct, indication of the vertical advective transport of chemical constituents (see text, sect. 4). The heavy dashed streamline (schematic only) indicates the qualitative sense of the mesospheric "Murgatroyd–Singleton circulation", deduced from other observational and theoretical evidence (*e.g.*, [18,28]) and not from ref. [36]. The left-hand scale is pressure altitude $z = H_p \ln(p_0/p)$ in nominal kilometres, assuming a pressure e-folding scale height $H_p = 7$ km. Thus the vertical domain shown corresponds roughly to the middle third of fig. 1 (multiply the right-hand scale in fig. 1 by 7 km). The upward mean velocities are a smallish fraction of a millimetre per second, of the order of 0.2 $\mathrm{mm\,s^{-1}}$ [10,37,39,40], albeit somewhat larger in the case of January 1979 when the wintertime stratosphere was dynamically very active (sect. 9). The northward mean velocities at top right (not counting the heavy dashed streamline) are of the order of two or three $\mathrm{ms^{-1}}$. The time for a marked fluid element to rise from the tropopause to, say, 40 km is generally of the order of two years [39]. The lower pair of circulation cells is often collectively referred to as the "Brewer–Dobson circulation"; for historical reviews see [27,41].

Fig. 4. Zonally averaged absolute angular momentum $\overline{m}$ per unit mass, in units of $6.4 \times 10^7\ \mathrm{m^2 s^{-2}}$, shown for the middle atmosphere only, from ref. [40]. The zonal wind fields used are those calculated in ref. [8], corresponding to the January temperatures in fig. 1. The abscissa shows latitude on a linear scale with the south pole on the left, the north on the right, and the equator marked by the central tick. The left-hand scale is pressure altitude in scale heights; the right-hand scale is the same in millibars or hPa. Courtesy of Dr C. J. Marks.

Fig. 5. a) Transient and b) steady-state responses of the mean mass circulation to a change in the mean torque applied to the extratropical atmosphere (see text, sect. 5).

Fig. 6. Simple experiments with water waves, illustrating the wave-induced momentum transport — here manifested by the appearance of a strong mean flow — that results from the generation of waves in one place and their dissipation in another (see text, sect. 7). The mean flow can be made visible by sprinkling a little powder such as chalk dust on the surface of the water. I have often done experiment (a) as a lecture demonstration using a cylinder about 10 cm long and 4 cm in diameter, and experiment (b) using a curved wavemaker that is rather bigger, about 60 cm in arc

length. Good results are obtained with capillary-gravity waves having frequencies $\gtrsim 5$Hz. From ref. [4].

Fig. 7. The restoring mechanism or quasi-elasticity to which Rossby waves owe their existence (diagram taken from ref. [1]). The + and − signs, respectively, indicate the centres of the cyclonic and anticyclonic PV anomalies due to air parcel displacements across a basic positive isentropic y-gradient of PV. The heavy, dashed arrows indicate the sense and relative phase of the induced velocity field (see text), which causes the leftward propagation of the phase of the pattern, with individual parcels always accelerating against their displacements. The Rossby-wave mechanism is fundamental to many large-scale dynamical processes, including for instance baroclinic instability [1,109], and is a key factor in the chemical near-isolation of the wintertime stratospheric polar vortex and, to some extent, the equatorial lower stratosphere as well.

Fig. 8. The equivalence between potential-vorticity rearrangement rate and mean force or torque, in the simplest relevant model system (see text). Courtesy P. H. Haynes; see also references [116, 118, 127].

Fig. 9. Distribution of shallow-water potential vorticity $Q = (f + \nabla^2\psi)/h$ (see text, sect. 9) at two instants in a high-resolution pseudospectral numerical experiment (triangular truncation to total wavenumber 159, corresponding to a mesh size less than a degree of latitude) on a single-layer hemispheric shallow-water model stratosphere of 4 km mean equivalent depth (courtesy of Dr. W. A. Norton). Reference [132] gives further information about this experiment. Values of Q are shown in gray scale as indicated on the left, in units of $7.5\times10^{-8}\mathrm{m}^{-1}\mathrm{s}^{-1}$. The gray scale is monotonic except for the band of intermediate values shown white, marking a location where a substantial part of the equator-to-pole Q gradient, and hence the Rossby-wave restoring mechanism (cf. fig. 7), is concentrated. This produces a quasi-elastic resilience, acting as an undular "PV barrier" against horizontal eddy advective transport of mid-latitude air into the core of the polar vortex.

Fig. 10. Coarse-grain isentropic map of PV on the $\theta = 850$ K isentropic surface in the middle stratosphere (near 30 km altitude) on 9 December 1981, from ref. [124] (see text). Note that the outermost latitude circle is 15°N, not the equator. The arrows show the velocity field (scale at bottom left). The three cyclonic blobs on the right have been shown by careful data analysis and cross-checking [124] to be real features.

Table I. – *Some basic characteristics and typical magnitudes relating to (inertio-)gravity waves and Rossby waves. "PV" means the Rossby–Ertel potential vorticity (sect. 8). The wave-induced momentum fluxes shown are those appropriate, in a first approximation, to the TEM description of the mean state* [28,89,90]. *The northward and upward components of the effective flux of zonal or eastward momentum are shown in that order, where* u', v', w' *are, respectively, the eastward, northward and upward disturbance velocity components and subscripts denote derivatives. The quantity* $f\overline{v'\theta'}/\overline{\theta}_z$ *is negligible for fast gravity waves but significant for low-frequency inertio-gravity waves, tending to cancel the term* $\overline{u'w'}$. *For Rossby waves* $f\overline{v'\theta'}/\overline{\theta}_z$ *is dominant and* $\overline{u'w'}$ *negligible. The abbreviation "PVS" stands for "PV substance", the notional tracer "substance", or signed "charge" whose mixing ratio is the PV; see sect. 11.*

	Gravity waves	Rossby waves
Time scales:	minutes to hours	days
Basic restoring mechanism:	buoyancy (-Coriolis)	PV-induced horizontal motion
Activated by undulations of:	θ-surfaces, vertically	PV contours on θ-surfaces, horizontally
in the presence of:	a vertical gradient of θ	an isentropic gradient of PV
Kinematical character:	substantial vertical velocity and horizontal divergence ($u'_x \sim w'_z$ &/or $v'_y \sim w'_z$)	layerwise-2D: small vertical velocity and horizontal divergence ($u'_x \gg w'_z$ &/or $v'_y \gg w'_z$)
Wave-induced momentum transp. dominated by:	$\{0,\ \overline{u'w'} - f\overline{v'\theta'}/\overline{\theta}_z\}$	$\{\overline{u'v'},\ -f\overline{v'\theta'}/\overline{\theta}_z\}$
Consequences of wave "breaking":	• 3D turbulence	• layerwise-2D turbulence
	• θ-surfaces deform irreversibly, instead of undulating	• PV contours deform irreversibly along θ-surfaces, instead of undulating
	• entropy & chemicals mixed vertically, downgradient.	• PVS & chemicals mixed along θ surfaces, down their isentropic gradients
	• PVS transported nonadvectively *along* θ-surfaces, possibly upgradient	