Journal of Atmospheric and Terrestrial Physics, Vol. 46, No. 9, pp. 825–849, 1984.
Printed in Great Britain.

0021-9169/84 $3.00 + .00
Pergamon Press Ltd.

The 'surf zone' in the stratosphere

M. E. McIntyre
Department of Applied Mathematics and Theoretical Physics, University of Cambridge,
Cambridge CB3 9EW, U.K.

and

T. N. Palmer
Meteorological Office, Bracknell, Berks. RG12 2SZ, U.K.

(*Received in final form* 2 *April* 1984)

Abstract—Synoptic, coarse-grain, isentropic maps of Ertel's potential vorticity Q for the northern middle stratosphere, estimated using a large-Richardson-number approximation, are presented for a number of days in January–February 1979, together with some related isentropic trajectory calculations. The effects of substituting FGGE for NMC base data are noted, as well as some slight corrections to maps published earlier. The combined evidence from the observations and from dynamical models strongly indicates the existence of planetary-wave breaking, a process in which material contours are rapidly and irreversibly deformed. In the winter stratosphere this occurs most spectacularly in a gigantic 'nonlinear critical layer', or 'surf zone', which surrounds the main polar vortex, and which tends to erode the vortex when wave amplitudes become large. Some of the FGGE-based Q maps suggest that we may be seeing glimpses of local dynamical instabilities and vortex-rollup phenomena within breaking planetary waves. Related phenomena in the troposphere are discussed. An objective definition of the area $A(t)$ of the main vortex, as it appears on isentropic Q maps, is proposed. A smoothed time series of daily values of $A(t)$ should be a statistically powerful 'circulation index' for the state of the winter-time middle stratosphere, which avoids the loss of information incurred by Eulerian space and time averaging.

1. INTRODUCTION

In this Symposium and in a recent article in *Nature* (McIntyre and Palmer, 1983), we presented and discussed some synoptic maps of Ertel's potential vorticity Q on the 850 K isentropic surface, for the extratropical Northern hemisphere on several days during the warming events of January–February 1979. The 850 K isentropic surface lies in the middle stratosphere at altitudes around 30 km. Q is defined as

$$Q = \rho^{-1}(2\mathbf{\Omega}+\nabla\times\mathbf{u})\cdot\nabla\theta, \qquad (1)$$

where $\mathbf{\Omega}$ is the Earth's angular velocity, $\mathbf{u}$ is air velocity relative to the Earth, ρ is air density and θ may be taken as specific entropy or any function of it, such as potential temperature. For a discussion of the approximations and data-processing procedures involved in constructing these isentropic Q maps, the reader is referred to the *Nature* article (hereafter MP) and to the Appendix A of this paper, and also to a forthcoming paper by Clough *et al.* (1984).

The reason for thinking in terms of such maps, despite the well known difficulties in deriving them from real data, is that they are fundamentally the simplest and most useful way in which to visualize large-scale dynamical processes. This is not merely because of the simplicity of the Lagrangian behaviour of Q, which is approximately a material tracer, but also because the distribution of Q controls the time development of almost every large-scale dynamical process of meteorological interest, through its diagnostic connection with the mutually balanced pressure and velocity fields. This important fact was implicitly recognized by V. Bjerknes in his early emphasis on the use of 'circulation theorems', following concepts originally developed by Kelvin and Helmholtz (Gill, 1982; Pedlosky, 1979). The point was made particularly clear by the work of Charney and Stern (1962, equations 2.25b, 2.31); see also Bretherton (1966a) for the extension of the concept to include the lower boundary condition and Hoskins *et al.* (1984) for its extension from quasi-geostrophic to semi-geostrophic theory and further refinements.

The most familiar dynamical uses of the potential vorticity concept refer to adiabatic processes like Rossby wave propagation, the Rossby-wave restoring effect being related to the isentropic *gradient* of Q, as is well known. However, frictional and diabatic effects need not be excluded from consideration. Indeed, an example in which we shall be interested is the way in which diabatic processes form the stratospheric circumpolar vortex each autumn. For some purposes this, too, is simplest to think of directly in terms of the evolution of isentropic distributions of Q, rather than primarily in terms of the balanced wind, temperature and pressure fields and the ageostrophic circulations which couple them together.

The stratospheric Q maps presented in MP, being severely limited by data resolution and aliasing problems due *inter alia* to the spacing of satellite orbits (cf. Salby, 1982, but also Clough *et al.*, 1984), gave only a coarse-grained estimate of the large-scale Q distribution on one isentropic surface, "resembling a blurred view of reality seen through a pane of knobbly glass". Despite this, our experience in working with these maps has been one of seeing what is happening in the stratosphere much more clearly than ever before. In particular, the maps gave the first reasonably convincing direct view of the breaking of planetary-scale Rossby waves, a phenomenon whose existence, and importance for large-scale dynamical and tracer-transport processes in the stratosphere, had been anticipated from theoretical and numerical modelling work, including the theory of nonlinear critical layers (e.g. Stewartson, 1978; Warn and Warn, 1978; McIntyre, 1982, and references therein; Fig. 2 below). The possibility of making this phenomenon directly visible, albeit in blurred and distorted form, arose from the enormous spatial scale of some of the wave-breaking events in the middle stratosphere. They span a large portion of the hemisphere and can almost certainly be called the 'world's largest' breaking waves.

The Q maps immediately suggested another key fact about the anatomy of large-scale dynamical and tracer-transport processes in the stratosphere, namely that in winter, especially late winter, the isentropic surfaces of the extratropical middle stratosphere are divided into two sharply-defined, zonally asymmetric regions, a polar *main vortex*, characterized by steep gradients of Q at its edge, surrounded by a broad *surf zone* within which systematic, large-scale gradients of Q are comparatively weak. As its name is meant to suggest, the surf zone appears to be the main region of wave breaking, where the strongest quasi-horizontal mixing and irreversible tracer transport takes place. It is a real-life analogue of the theoretical nonlinear critical layer. In the late winter of 1979 it occupied more than a third of the area of the entire northern hemisphere. We argued that this 'main vortex, surf zone' structure, and its observed time evolution, in which the surf zone broadened and the main vortex shrank as the winter progressed, can be attributed in large part to erosion of the main vortex by the action of the breaking planetary waves in the surf zone. Independent corroboration of this picture, and further details of it, including a clearer view of the equatorward boundary of the surf zone, have emerged from recent analyses of LIMS ozone data from the Nimbus 7 satellite (Leovy *et al.*, 1984). The results have thrown new light on phenomena such as stratospheric sudden warmings, and have important implications for modelling the transport of pollutants in the ozone layer.

In this paper we present a fuller account of the original evidence for the surf-zone concept, including some isentropic trajectory calculations whose results were described qualitatively in MP but not presented for lack of space. We also give some further discussion of the implications of that evidence not only for the middle atmosphere, but also for analogous, weather-related phenomena in the troposphere, which have long been familiar to synoptic meteorologists but which can now be seen in a new and simpler light. Both the stratospheric and the tropospheric cases seem to involve a basically similar interplay between wave generation, propagation and breaking, which is proving to be an important part of the highly inhomogeneous "wave-turbulence jigsaw puzzle" which must be pieced together if we are to improve our understanding of large-scale atmospheric flows, and our ability to predict them.

We also take the opportunity to present some further information about the stratospheric Q maps, acquired since the publication of the earlier (*Nature*) article particularly the extent to which the maps are affected by substituting FGGE* for the original NMC† base analyses at 100 mbar while keeping the same stratospheric thickness estimates as before (from the Tiros-N Stratospheric Sounding Unit, hereafter SSU—see Pick and Brownscombe, 1981). In fact all the Q maps presented in this paper are FGGE-based, except those in Fig. 1. It seems reasonable to assume that the FGGE 100 mbar analyses are better on the whole than the NMC analyses, especially in the subtropics. The FGGE-based Q maps appear to show a little more detail about the nature of the wave-breaking process. In addition, we draw attention to a mistake in the computer code which produced the original, NMC-based Q maps, which regrettably was discovered after the original publication, MP, appeared in print. It does not affect the conclusions, having little effect on the qualitative features of interest, but it does, for instance, imply that for quantitative purposes the indicated maximum value $Q_{\max}$ of Q in the main vortex was overestimated in the original maps (lying nearer the value given by the geostrophic approximation, instead

* First GARP Global Experiment data, as analyzed at the European Centre for Medium Range Weather Forecasts.

† National Meteorological Center, Washington, DC.

of being given, as intended, by the gradient-wind approximation).

To illustrate the extent of the change, Fig. 1a presents the original, slightly incorrect, NMC-based Q map for 27 January 1979, some of whose qualitative features were emphasised in the discussion presented in MP, including the long tongue of high-Q air, shown lightly shaded, which extends around the Aleutian anticyclonic vortex (Fig. 1c), and which was taken to be one of the more spectacular visible manifestations of wave breaking. The main vortex, as we defined it, is approximately the heavily shaded region. We identified most of the remaining area as belonging to the surf zone. Fig. 1b is the same map with the coding error corrected, so that the resulting changes can readily be seen. The qualitative features of interest are much the same in both these maps, but the highest contour value in the correct map (Fig. 1b) corresponds to 1.84 rather than 2.46 times the maximum planetary vorticity 2Ω, in terms of the convention used in MP and in equation (A4) below, with 29K mbar^{-1} taken as the reference static stability. The situation is much the same in the case of the other maps. Figure 1c is the NMC-based 10 mbar height map for the same day.

2. WAVE BREAKING AND IRREVERSIBILITY

Before going further, it is necessary to say more precisely what we mean by wave breaking. Like the familiar breaking of surface gravity waves as they approach an ocean beach, the breaking of planetary or Rossby waves, as we conceive it, is characterized by a rapid and irreversible deformation of material contours. MP argued from wave-mean interaction theory that this is the most useful and fundamental sense in which the notion of wave-breaking can be generalized, from familiar instances, so as to cover all kinds of transverse (non-acoustic) wave motions in fluid media, including internal gravity waves and Rossby waves. It is whether or not the waves break, in the sense envisaged, that often dominates how effective they are at inducing quasi-permanent changes in the basic state on which they propagate, for instance in the distributions of angular momentum and of advected scalars.

The relevant material contours, for this purpose, are those which, under the conditions assumed by linear, nondissipative wave theory, would *not* deform irreversibly, but merely undulate back and forth* (cf. the perceptive remarks by DICKINSON, 1969, p. 78). In the case of Rossby waves, which in nondissipative wave theory satisfy the relations

$$\mathrm{D}\theta/\mathrm{D}t = 0 \qquad (2)$$

and

$$\mathrm{D}Q/\mathrm{D}t = 0, \qquad (3)$$

where $\mathrm{D}/\mathrm{D}t$ is the material derivative, the relevant contours are the contours of constant Q in each isentropic surface, which are indeed material contours if equations (2) and (3) hold. Thus the breaking of Rossby waves, in the sense envisaged, should in principle be visible in Q maps. Taken at face value, Fig. 1 seems to be a case where this was also true in practice. The gross shape of the Q contour marked '4', enclosing the lightly shaded area, is roughly but strikingly suggestive of the material contour deformation expected theoretically for the simplest kind of Rossby wave-breaking event. A theoretical example, taken from the theory of barotropic, nonlinear critical layers in a dissipationless fluid, is shown in Fig. 2 (see caption for details). This represents an analytical solution, with infinitely good spatial resolution. The family resemblance to the phenomenon suggested by Fig. 1 is emphasised by means of the shading. The kinematics of such contour deformations involves the formation of thin tongues of advected material, which tend to become ever longer and thinner, an irreversible process for all practical purposes (see below) and, incidentally, illustrative of the physical reality behind the so-called potential-enstrophy 'cascade' (e.g. WELANDER, 1955; BATCHELOR, 1969; BRETHERTON and HAIDVOGEL, 1976; RHINES, 1979).

It should be borne in mind that such tongues, because of their tendency to become longer and thinner as time goes on, may rapidly disappear from view in any approximate representation having limited horizontal resolution, even if they involve a substantial isentropic contrast in Q values. An excellent illustration of this point can be seen in the numerical simulations of two-dimensional turbulence at different resolutions shown in Fig. 8a, b of HERRING *et al.* (1974).† Such tongues may also be dynamically unstable (the more so, the greater the contrast in Q values), because of the change in the sign of the isentropic gradient of Q across the tongue (CHARNEY and STERN, 1962). Although Charney and Stern studied only zonally symmetric flows, we have in mind a local approximation in which the narrowness of the tongue in comparison with its length permits it to be treated locally as if it were part of a zonally symmetric flow. Such instabilities might be capable of causing a tongue to break up into eddies at some stage of its evolution. The analytical solution of Fig. 2 has, in fact, been shown to be unstable in this way (KILLWORTH and McINTYRE, 1984; HAYNES, 1984), but we are not yet sure how significant this might be for cases like that of Fig. 1, since the parameter regime is very different.

Whether or not such instabilities turn out to be

* - such undulations usually resulting from the restoring effect giving rise to the wave propagation (associated with the basic-state vertical density gradient in the case of gravity waves, and with the basic-state isentropic gradient of potential vorticity in the case of Rossby waves). For further discussion see PAGEOPH 123, 964-975. (1985)

† See also the numerical "contour-dynamics" simulations reported in, for instance, Deem, G.S. and Zabusky, N.J. (1978) Phys. Rev. Lett. 40, 859-862; Dritschel, D.G. (1987) J. Fluid Mech., 172, 157-182 (E.g. figure 13).

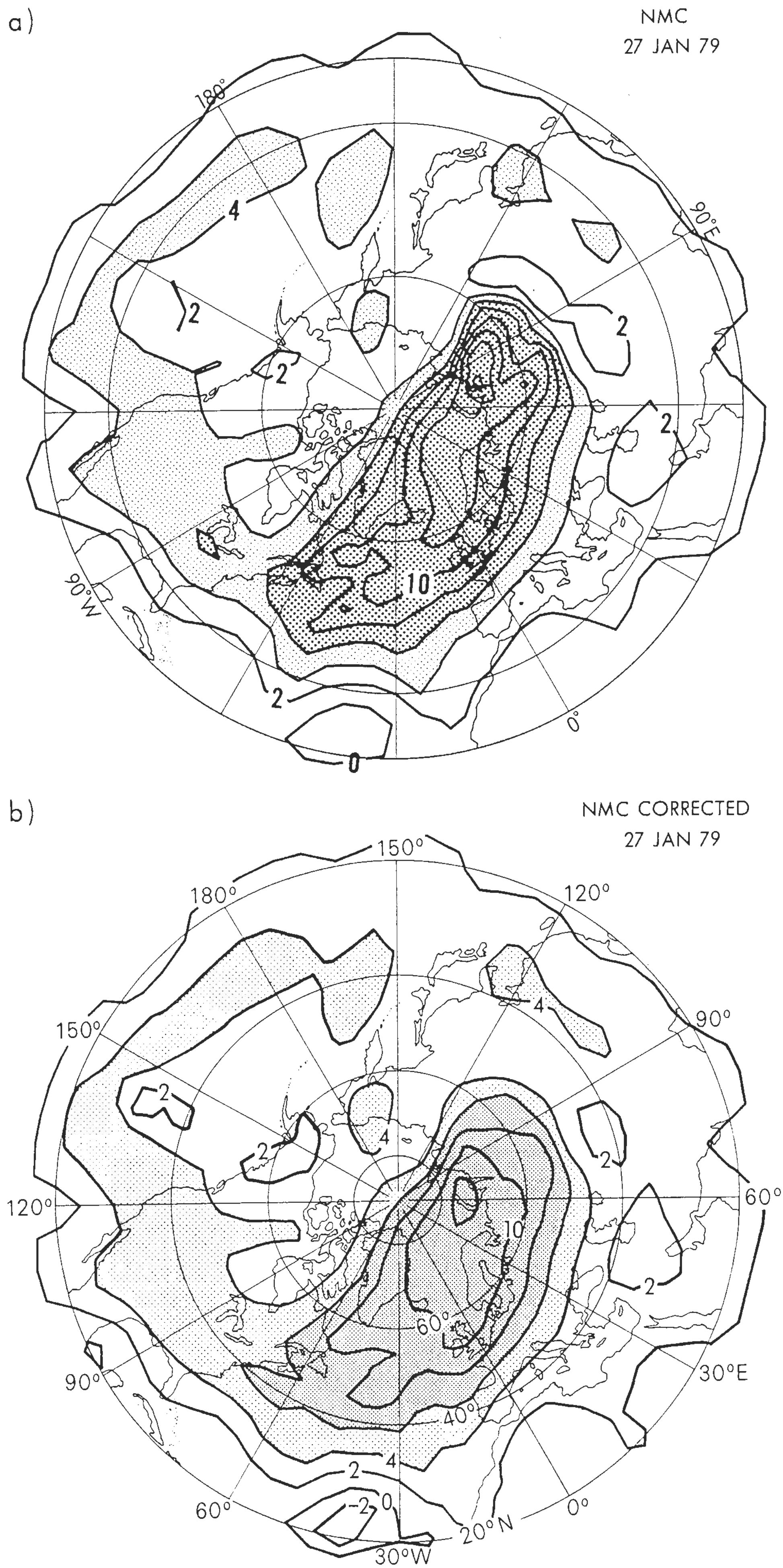


Fig. 1a, b
(see over)

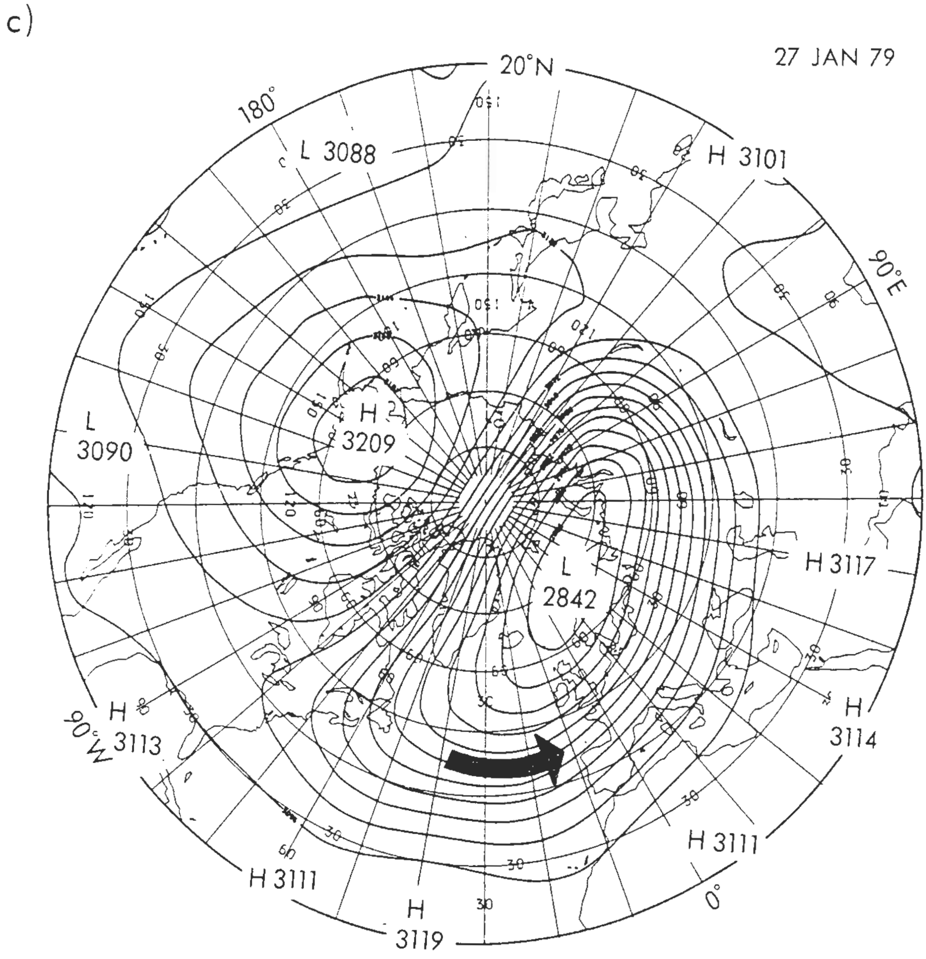


Fig. 1. (a), (b) Uncorrected and corrected coarse-grain NMC-based estimates of Ertel's potential vorticity Q (see text) on the 850 K isentropic surface, near the 10 mbar isobaric surface, on 27 January 1979 at 00Z. For units see Appendix A. Contour interval is 2 units. Values greater than 4 units are lightly shaded and greater than 6 units heavily shaded. (c) Corresponding NMC-based analysis of the geopotential height of the 10 mbar isobaric surface in dekametres. Contour interval is 24 dekametres. The lowest contour value is 28.56 km and the highest 31.92 km. Map projections are polar stereographic. The southernmost latitude circle shown is 20°N.

important, the length of time for which a wave breaking event can be followed in detail, if it is visible at all, will clearly be sensitive to the spatial scale of the event in relation to the horizontal resolution of the maps, as well as to the strength of the isentropic Q contrast across the tongue. Figure 1 corresponds to the wave event of largest amplitude and scale during the whole winter of 1978–79, and could well be a case where breaking was visible to an unusual extent despite the limited resolution. Caution should still be exercised, however, regarding the realism or otherwise of some details in the picture, especially near the far end of the tongue, which lies in the subtropical Pacific where the data quality may be worst. These questions will be taken up again in the next two sections.

Although the meaning of 'rapid and irreversible' deformation of material contours may seem almost self-evident as soon as one has some specific examples in mind, including the familiar one of surface gravity waves on an ocean beach, it may be useful to be more specific. The rest of this section, and Appendix B, examines more closely what is involved.

Take 'rapid' first. A reasonable judgement as to what should be meant in general by rapid seems to be that the time scale of the material deformation should not be much longer than a typical intrinsic wave period in the locality concerned, and should also be much shorter than dissipative time scales. Slower, overtly dissipative processes can also lead to irreversible material deformation, through their cumulative effects over many wave periods, but it would not be natural to call this wave breaking. The analytical example of Fig. 2 is an example of rapid deformation in the foregoing sense. From the solution given in STEWARTSON (1978), it can

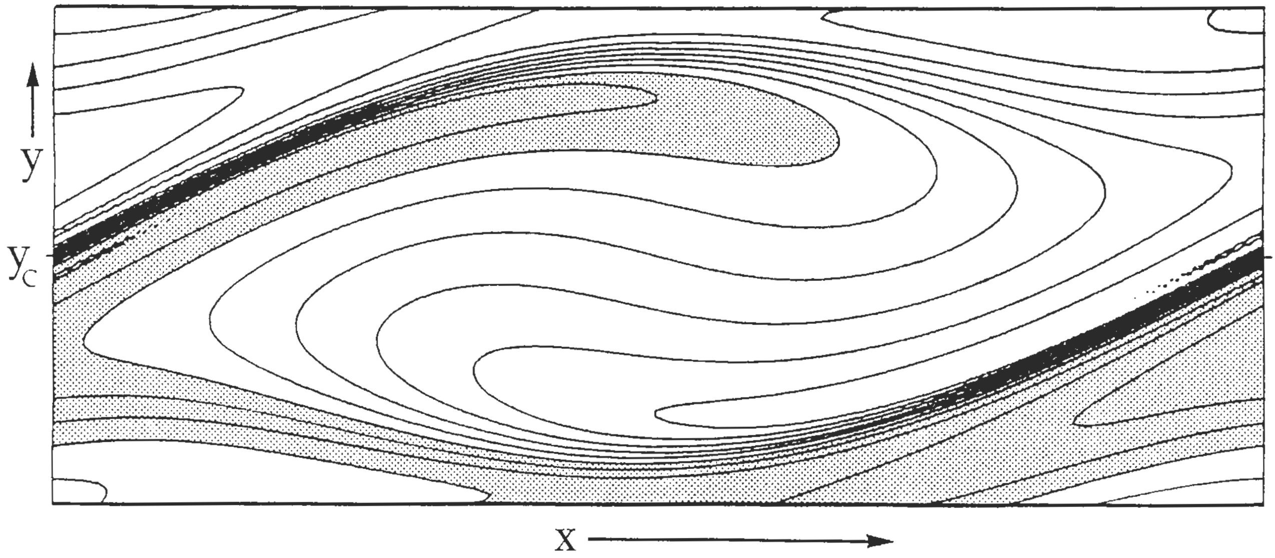


Fig. 2. Analytical solution from the theory of nonlinear critical layers (Stewartson, 1978; Warn and Warn, 1978), exhibiting the effectively irreversible deformation of material contours due to a two-dimensional (height-independent) 'Aleutian vortex' set up by a stationary Rossby wave on a shear flow. An equation of the form (3) holds exactly, and so the contours are both material contours and contours of Q, or equally (in this case) its two-dimensional counterpart, the vertical component of absolute vorticity. The shading picks out values of Q intermediate between the higher and lower values (unshaded) at bottom and at top/centre, respectively. The solution is periodic in x, which corresponds to longitude west for the purpose of a qualitative comparison with Fig. 1b. The y scale is exaggerated and corresponds to minus the latitude. Initially the contours lie parallel to the x axis. The initial flow is in the x direction, its velocity proportional to $(y-y_c)$ so that there is a 'critical line' at $y = y_c$. The time elapsed is 0.57 of the time for an air parcel to make one complete trip around the centre of the vortex or 'cat's eye'. The solution can be shown to apply whether or not the contours represent equally spaced values of Q, i.e. it applies when the initial Q distribution is a given, arbitrary function of y. For instance the initial gradient could be small outside the shaded region.

be shown that the time for a material contour near the centre of the picture to rotate clockwise through 360° is just one intrinsic wave period, when the intrinsic wave period is evaluated from linear theory at a value of y halfway between the centre and the edge of the model 'Aleutian vortex', or slightly less than halfway between the centre and bottom of the picture.

Second, what do we mean by 'irreversible'? The notion of irreversibility in use here is similar, at a *conceptual* level, to that employed in ordinary statistical mechanics (for a clear discussion see Peierls, 1979, pp. 76, 79). But it should be kept firmly in mind that the associated physical phenomena, such as the continual lengthening of material contours and the accompanying potential-enstrophy 'cascade', are very different. These phenomena are to be carefully distinguished from phenomena associated with irreversibility of the microscopic, molecular and radiative processes usually studied in statistical mechanics, a quite separate set of phenomena also present in the real atmosphere. The mechanics we are concerned with here is that of the bulk fluid motion (particularly the large-scale, quasi-two-dimensional fluid motion), not that of the individual air molecules and photons. The irreversibility in question would be present even in a fluid-dynamical model which, like the model from which Fig. 2 was obtained, had no overtly dissipative terms in its equations of motion, i.e. which utterly neglected the irreversibility due to microscopic processes. (Whether there exists a numerical method which could in any sense cope with such a model is an entirely separate question, to which the answer is no; but that does not destroy the legitimacy of the thought-experiments, and the resulting concepts, arising from a consideration of such models and their analytical solutions. The same can be said about the question of whether such models exhibit true statistical-mechanical *equilibria*—to which the answer is, again, almost certainly no. All we are concerned with here is the fact, for which the theoretical and empirical evidence is overwhelming, that they exhibit irreversibility. As is well known, this in no way contradicts the fact that the model equations are time-symmetric; see Appendix B.)

Figure 3, taken from Welander (1955), reminds us very effectively of the general nature of the fluid-

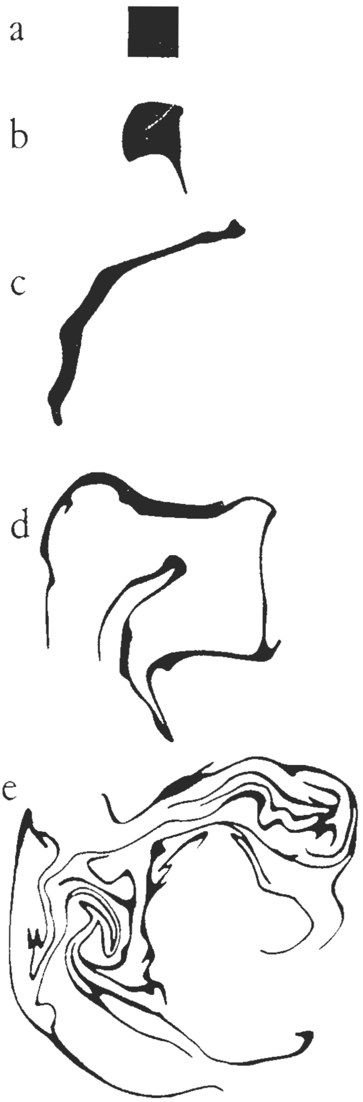


Fig. 3. A classical illustration of fluid-dynamical irreversibility (see text): deformation of a material contour marked by a passive tracer in a time-dependent, two-dimensional, approximately nondivergent velocity field. Time goes forward from (a) to (e). From Welander (1955).

dynamical irreversibility of interest here. It shows the evolution of a passive material tracer in a time-dependent, approximately two-dimensional velocity field, produced in a laboratory tank of water on a rotating table. Part of the fluid is marked with dye at a given instant (Fig. 3a) and then followed for subsequent times (Figs. 3b–e). A reversed evolution from (e) to (a) is never observed in practice, even though overt diffusion of dye plays no significant role. Notice the tendency for ever smaller scales to appear.

A similar, albeit not identical, behaviour has been found from numerical 'two-dimensional turbulence' experiments to be characteristic of the smaller scales in the Q distribution, when overt dissipation is small enough (e.g. Herring *et al.*, 1974; Bretherton and Haidvogel, 1976; Rhines, 1979; Holloway and Kristmannsson, 1984). This is to be expected since, although Q is not a passive tracer—a fact which is crucial to the Rossby restoring mechanism and hence, for instance, to the resistance to erosion of a *large*-scale Q-feature like the stratospheric main vortex—Q nevertheless behaves somewhat like a passive tracer on sufficiently small scales. The mathematical reason lies in the smoothing properties of the operators for finding diagnostically the balanced height and velocity fields corresponding to given isentropic Q distributions (and a given mass distribution of θ). The operator giving the height field has the approximate character of an inverse Laplacian (Charney and Stern, 1962, equation 2.25b; Hoskins *et al.*, 1984), a fact correctly suggested by the smoothness of Fig. 1c in comparison with Fig. 1b. The operator giving the velocity field is the same followed by only a single spatial differentiation. Consequently, the velocity field tends to become insensitive to the ever-decreasing scales in the isentropic Q distributions, once a situation locally resembling that in Fig. 3e develops for whatever reason. In such situations, a rapidly varying, small-scale Q field is advected by a smoother, larger-scale velocity field, the two being only weakly correlated on small scales. Figure 2 is actually an extreme example of this scale effect, in that the small-scale correlation is weak enough to be exploited mathematically, permitting the use of the method of matched asymptotic expansions to derive the solution (Stewartson, 1978; Warn and Warn, 1978).

Another way of viewing the scale effect is to note that it corresponds to the well known weakening of the Rossby wave restoring mechanism as spatial scales become smaller. As smaller and smaller scales develop in the shape of a Q contour, the restoring mechanism, which depends on the correlation between the velocity and Q fields, becomes less and less capable of opposing the deformation of the contour. What is more, even if Q is not quite passive on the smallest scales, i.e. even if a significant local Rossby restoring effect remains, then the likely result is to make the contour shape more convoluted, if anything, than it would be for a passive tracer. Once the Q contours have begun to fold over, the Rossby wave mechanism will tend to manifest itself, if at all, in the form of the dynamical instabilities mentioned earlier. This is because the folding implies sign changes in the local isentropic gradient of Q, on which the local Rossby restoring mechanism depends. As is well known, the possibility of such dynamical instabilities, whether barotropic, baroclinic, or a mixture of the two, is related to the existence of oppositely-directed Rossby wave propagation in horizontally or vertically adjacent regions (Fjørtoft, 1950; Lighthill, 1963, p. 93; Bretherton, 1966b).† An example amenable to

† & Hoskins et al, §6

analytical solution is again given by the special case of Fig. 2, where the instability analysis mentioned earlier provides us with a detailed picture. The upshot is that, in one way or another, the small-scale deformation of material contours coincident with contours of the Q distribution can be expected to be no less irreversible than the passive-tracer behaviour illustrated in Fig. 3, once the deformation is initiated. How important the averaged, large-scale dynamical effects of that deformation may be is a separate question, of course, the answer to which will depend very much on the detailed circumstances. The same goes for the related question of how complete or incomplete is the consequent mixing of the Q distribution.

An interesting discussion relevant to this is in Lindzen & West JAS 41, 3021.

Whether the word 'dissipation' should be used to refer to the effects of the fluid-dynamical irreversibility under discussion is philosophically a moot point. However, in a practical sense these quasi-two-dimensional fluid motions are, in effect, dissipative, and in a way that depends hardly at all on the presence or absence of small, overtly dissipative terms in the equations, provided that such dissipative terms only smear out the finest details in the Q field and have a correspondingly small effect on the velocity field. In formulating numerical models, the distinction between fluid-dynamical and molecular or radiative irreversibility tends to get blurred, of course, since all irreversible processes have to be lumped willy-nilly into artificial, overtly dissipative terms in order for the model equations to be tractable numerically at finite resolution. (The question then becomes how to choose the form of those terms so as to do the least damage to the basic processes being modelled, including the quasi-two-dimensional fluid-dynamical irreversibility.)

Finally, it is re-emphasised that the general concept of wave breaking, as developed here, in no way depends on the extent to which recognizable instabilities can be said to play a role in deforming the material contours. Instabilities are important in certain cases, but the crucial question, which is relevant to all cases, is whether or not the contour deformations are irreversible. Such deformations can take place with or without the help of recognizable instabilities. Cases in which instabilities are recognizable and important include quasi-hydrostatic internal gravity waves (e.g. LINDZEN 1967, 1981; HODGES, 1967) and internal gravity waves on layered basic states (PHILLIPS, 1966; WOODS, 1968). Instabilities have been shown theoretically to be important in the special Rossby wave example of Fig. 2, as already mentioned (see especially HAYNES, 1984). But there are other instances, including classical cases of surface gravity waves, where the concept of 'instability' is peripheral or irrelevant to the concept of wave breaking (e.g. BENJAMIN and OLVER,* 1982, p. 146). The analytical solution shown in Fig. 2, which represents an unstable situation in which the instabilities have been artificially suppressed, provides us with a dynamically consistent thought-experiment in which the material deformation is, nonetheless, effectively irreversible (and destroys the initial large-scale gradient of Q), without the help of any instabilities. As for the tongues of high-Q air which we believe must be eroded from time to time into the real stratospheric surf zone, whether they are sufficiently unstable to make a major difference to what happens in the stratosphere is still an open question, although we shall see in Section 4 that there is now some direct evidence pointing towards the actual occurrence of such instabilities. It appears moreover that their effect may be to make the mixing *less* complete than might otherwise be the case.

3. AIR PARCEL TRAJECTORIES AND WAVE-BREAKING SIGNATURES

As mentioned in the last section, and in McINTYRE and PALMER (1983), and as we shall also see from the FGGE-based Q maps, there is some doubt as to whether the data should really be showing us a tongue fully as long as that appearing in Fig. 1. Although the gross shape of the apparent tongue is extremely plausible by comparison with the simplest available theoretical models, we cannot be certain of the precise extent and strength of the real tongue, expecially since the quality of the NMC base data, in particular, is probably approaching its worst in the subtropical Pacific. We therefore decided to see first of all whether isentropic trajectory calculations could throw any light on the matter. Such calculations have their own problems but at least they involve a certain amount of integration as well as differentiation, somewhat mollifying the effects of noise in the data.

Figure 4 shows a computer-generated plot of three bundles of isentropic trajectories, integrated backwards from three sets of positions centred on the 32.5°N latitude circle, within or near the apparent tongue of 27 January 1979. These were kindly produced for us by Mr J. Austin, using FGGE-based SSU gradient-wind fields computed on the 850 K isentropic surface (for details of the method see AUSTIN and TUCK, 1984). The small circles are time markers for each tenth of a day. If equation (2) were exactly satisfied, then the isentropic trajectories would represent the paths of material parcels, apart from effects due to the errors in the wind fields. The latter include not only data errors and space–time interpolation errors, but also errors from the finite differencing scheme used to estimate the gradient winds from the height fields, probably worst near the pole, since a latitude–longitude grid was used.

* Also, e.g.: Longuet-Higgins & Cokelet 1976, Proc.Roy.Soc.London A 350, p.17
Baker, Meiron & Orszag 1982, J.Fluid Mech. 123, p.477
New 1983, J.Fluid Mech. 130, p.219.
Pullin 1982, J.Fluid Mech. 119, 507 (517).
New, McIver & Peregrine, J.Fluid Mech. 150, p 244, 248.

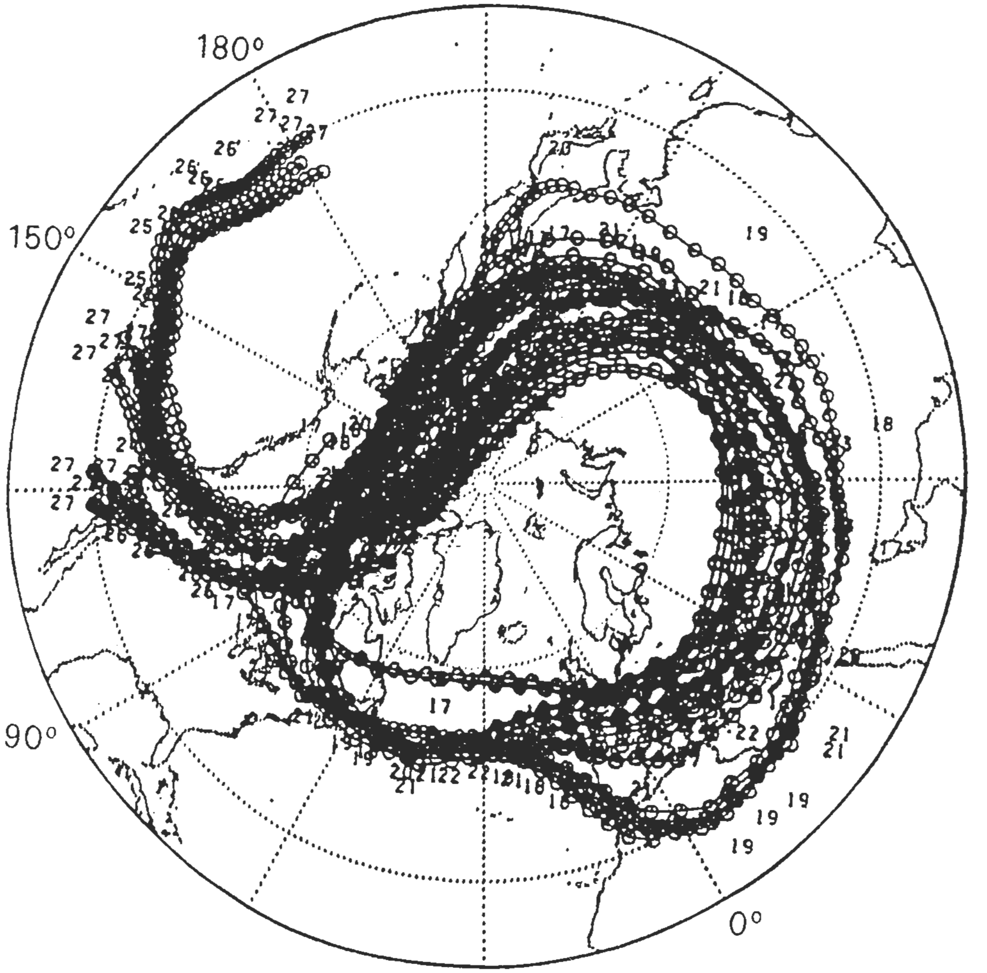


Fig. 4. Three bundles of machine-calculated 850 K isentropic trajectories ending on 27 January 1979. Time markers are shown every tenth of a day. Gradient winds were estimated from the Montgomery streamfunction on a latitude–longitude grid (see AUSTIN and TUCK, 1984) and the deceleration correction (A5) was applied. Dates shown by the computer graphics are at time 00Z.

A 'deceleration correction' was applied to allow for the error in the gradient-wind approximation itself, due to changes in the wind speed |**u**| along the trajectory. That correction, described in Appendix A, can be cumulatively quite significant for an air parcel which peels off the rapidly-moving edge of the main vortex. The correction arises from the well known fact, neglected in the gradient-wind approximation, that an air parcel has to head slightly 'uphill', or more precisely, into the isobaric height gradient, if it is to reduce its speed (e.g. PALMÉN and NEWTON, 1969, §8.2).

Figure 5 gives an idea of the sensitivity of the trajectory calculations to the deceleration correction, and also to the choice of spatial differencing scheme from which to estimate the gradient winds. Four different estimates are shown for the part of the isentropic trajectory beginning on 23 January and ending at 177.5°W, 35°N on 27 January, with time markers every half day. The main scatter in the results is between days 23 and 24, when the trajectories are emerging from the region of large horizontal wind shear near the edge of the main vortex.

The solid, heavy curve in Fig. 5 is the result of a hand computation ignoring the deceleration correction, done by careful graphical interpolation in space and time from machine-produced daily maps of FGGE-based SSU gradient-wind vectors. The gradient-wind vectors were computed from finite differences not on a latitude-longitude grid, but rather on a square grid on a polar stereographic projection, grid size about 580 km at 50°N, the same grid as was used for estimating Q (Appendix A). The heavy dotted curve shows the result of applying the deceleration correction from day 23 to 24, the main period of deceleration in this case, with |**u**| going from 62 m s^{-1} on day 23 up to a maximum of 72 m s^{-1} and then steeply down to 34 m s^{-1} on day 24. The correction brings the trajectory significantly nearer the pole. It would probably lie nearer still if the correc-

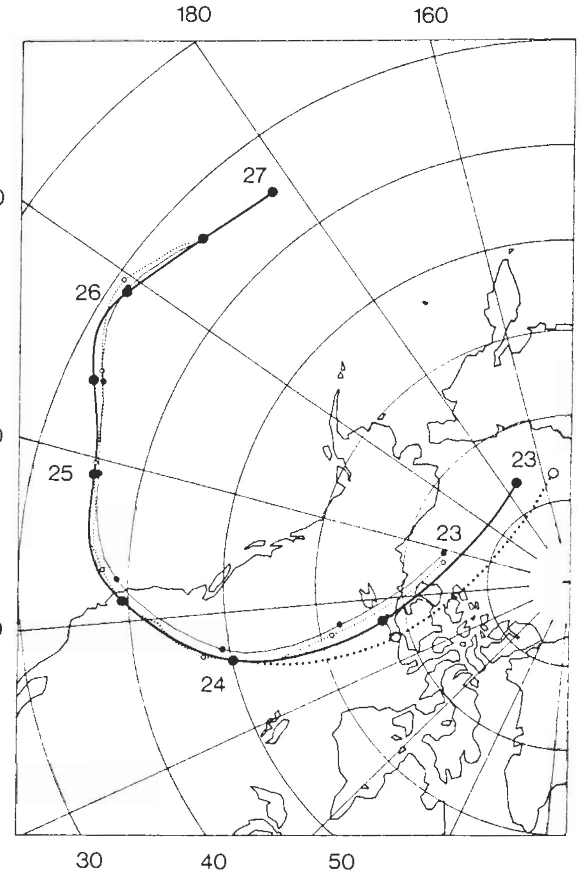


Fig. 5. Two hand-calculated (heavy curves) and two machine-calculated (light curves) 850 K isentropic trajectories, with time markers every half day, showing the effect of the deceleration correction (dotted curves with open time markers) and of different differencing schemes for computing the gradient wind (heavy versus light curves). See text.

tion had been applied instead from day 23 to 25, or over the whole trajectory, day 23 to 27, at the far end of which |**u**| is down to about 20 m s^{-1}, but we did not attempt this because of uncertainty as to how to handle the effects of an unusually severe local nonuniformity in the estimated gradient-wind field encountered between days 24.5 and 25. This prevented us from applying the correction over that section with any real confidence that it would be meaningful, i.e. not strongly dependent upon particular choices of interpolation methods. The computational effects of this nonuniformity of the wind field upon the various differencing and interpolation schemes may well account for the larger scatter in the time markers for day 24.5, as compared to those for day 25 onwards, where the overall agreement with the light curves is remarkably good.

The light curves are two machine-calculated trajectories produced in the same way as for Fig. 4 using the gradient winds from the latitude–longitude grid. Points are plotted every tenth of a day and joined by straight-line segments. The dotted curve reproduces the computation of Fig. 4, that is to say with the deceleration correction applied at each time step over the whole length of the trajectory. The time step was a hundredth of a day. The light, continuous curve is the corresponding machine computation without the deceleration correction. The cumulative effect of the deceleration correction can be seen to be much smaller in this case. This was traced to the fact that the gradient winds estimated from the latitude–longitude grid are considerably weaker than those from the polar stereographic grid in the computationally critical region near the edge of the main vortex. For instance, at day 23 on the corrected machine trajectory (beginning of the light dotted curve), the gradient-wind speed |**u**| seen by the machine computation was only 38.5 m s^{-1}, as compared to about 65 m s^{-1} interpolated to the same point from the polar stereographic grid values from which the heavy curves in Fig. 5 were constructed. This accounts for the shorter length of the machine trajectories, as well as for the smaller deceleration correction.

Notice how the deceleration correction has a tendency to exacerbate the computational sensitivity near the edge of the main vortex. If |**u**| at the beginning of a trajectory is large, then the acceleration correction is relatively large, bringing the trajectory further towards the pole and into still stronger winds, and so on. In the case of the heavy dotted curve in Fig. 5, it was found that one or two iterations were enough to settle its position at each time step. As already mentioned, the maximum value of |**u**| on the heavy dotted curve was estimated to be 72 m s^{-1}. This occurred in the sector lying just north of 80°N. The maximum gradient-wind speed indicated on the polar stereographic grid for day 23 was 82 m s^{-1}, at 84°N, 126°W.

The shape and position of the main vortex on day 23, defined in terms of the Q distribution on the 850 K isentropic surface, is indicated in Fig. 6, the FGGE-based Q map for 23 January. (Note that the contour interval is half that in Fig. 1). McINTYRE and PALMER (1983) defined the edge of the main vortex as the edge of the main region of steep Q gradients, which in the case of Fig. 6 can reasonably be placed somewhere within the lightly shaded, continuous band lying between the contour values 4 and 6. The edge of the main vortex, so defined, is very close to the starting points of the trajectories shown in Fig. 5. It is pointless to try to decide exactly how close, since, like the trajectories, the Q maps themselves are subject to finite-differencing errors on the polar-stereographic grid, as well as data errors. Moreover, diabatic corrections to the trajectories, and to equation (3) for the evolution of Q itself, have yet to be given a state-of-the-art assessment.

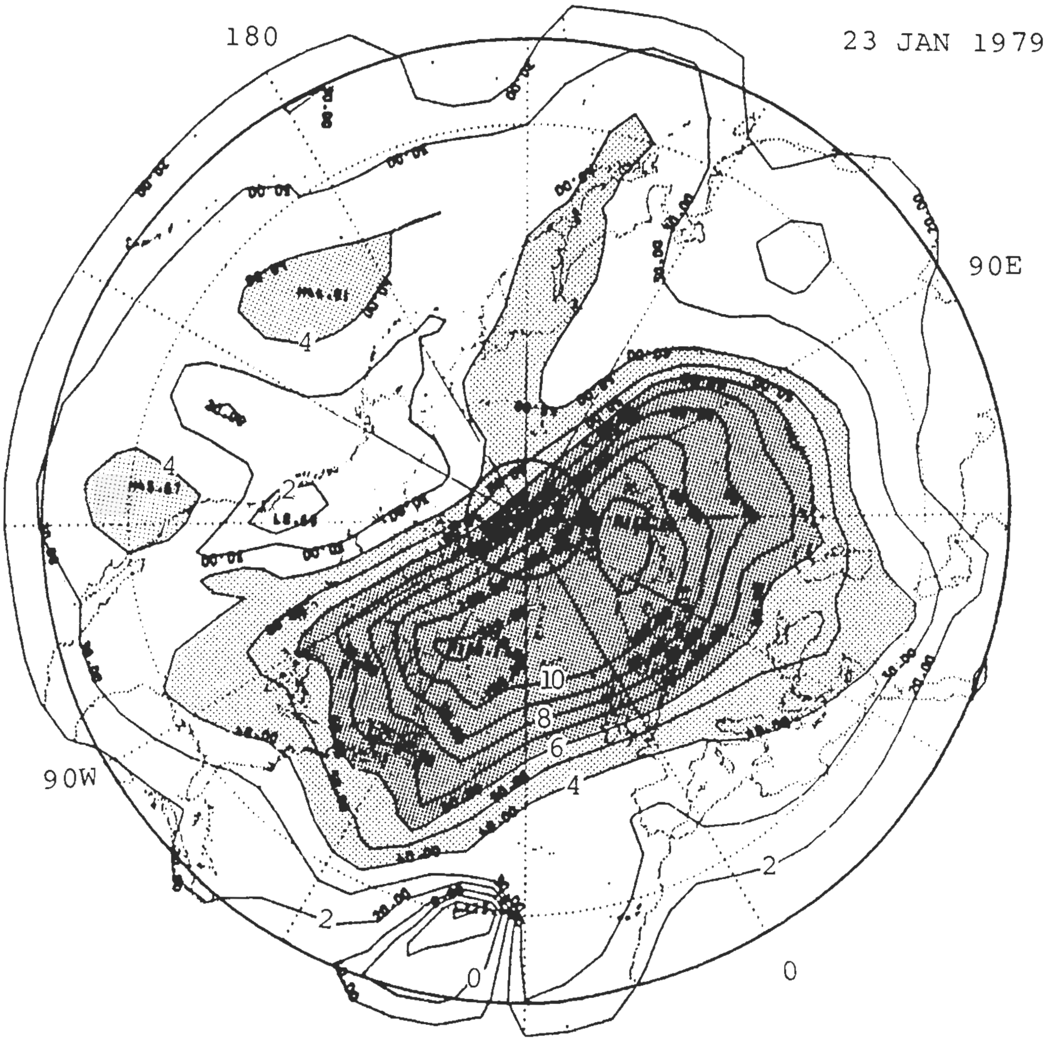


Fig. 6. Coarse-grain FGGE-based estimate of Ertel's potential vorticity on the 850 K isentropic surface on 23 January 1979 at 00Z (direct reproduction of the computer-generated plot). For units see Appendix A. Contour interval is 1 unit, twice as fine as in Figs. 1a, b. (The small computer-generated numbers are to be divided by 10, to give the units defined in Appendix A.) Values greater than 4 units are lightly shaded and those greater than 6 units heavily shaded, as in Figs. 1a, b. The 80°N latitude circle is shown (solid circle) in order to facilitate comparison with Fig. 5. Map projection is polar stereographic.

Among other things, this would require careful attention to the actual local distribution of ozone. However, it seems fair to say that all the evidence so far seems consistent, at least, with the general nature of the hypothesized wave-breaking and erosion mechanism, in which air parcels are removed from the edge of the main vortex and mixed quasi-horizontally into the surrounding surf zone, the whole process involving irreversible deformation of material contours.

It should be cautioned that the completeness or incompleteness of the mixing is not clear observationally; we cannot see the Q distribution in nearly enough detail to tell this directly. In addition, it should be kept in mind that other, overtly dissipative, processes may play a role in determining the fate of an air parcel eroded from the main vortex. An important task for the future will be to assess this. For instance, if the vertical structure of the day-27 tongue were sufficiently baroclinic, then Q values in the tongue might be quite vulnerable to diabatically induced changes in the tropics. However, none of this argues against the likely physical reality and importance of the erosion mechanism itself.

The hypothesized picture becomes even more plausible when we take the circumstantial evidence into account, particularly the fact that the Aleutian vortex was in the process of growing towards its maximum size for the whole winter during the period in question, 23–27 January. On kinematical grounds alone this would appear to make erosion of the main vortex almost inevitable, with the expanding Aleutian vortex 'eating its way into the potential-vorticity gradient at the edge of the main vortex', during that period. In particular, although it may never be able to be proven beyond doubt, it seems entirely reasonable to suppose that even the further part of the apparent tongue in Fig. 1 could be real, and that it could have originated, at least in part, from a thin band of high-Q air at the edge of the main vortex, just as theoretical analysis and modelling originally suggested.

It is also possible that, while the further part of the visible tongue may be a real feature of the isentropic Q distribution, some of it may have originated as a piece of high-Q 'debris' from a previous wave-breaking event, subsequently caught up in the growing Aleutian vortex. An example of this might be the piece of high-Q air indicated in Fig. 6 to be lying over Japan and Sakhalin on day 23. On the assumption that its northernmost end is a real feature of the Q distribution (and that the longest trajectories in Fig. 5 are the most realistic), this could have contributed to the Q contrast across the furthest part of the day-27 tongue and, therefore, to the strength and visibility of that tongue.

The corresponding questions are less delicate as regards the nearer part of the day-27 tongue, say east of about 150°W. The evidence from the trajectory computations and the daily sequence of Q maps leads us to believe that the air in the nearer part of the tongue originated from a little deeper within the high-Q air near the edge of the main vortex. The nearer part of the tongue is a robust feature exhibiting continuity between the Q maps for days 26 and 27, whether FGGE or NMC based. To summarize so far, taking all the lines of evidence into account, the physical reality of the hypothesized irreversible material contour deformation, and erosion of the main vortex, seem to be in very little doubt. It is only the finer details, such as the exact length, shape and strength of the day-27 tongue, and how complete or incomplete the eventual mixing is within the surf zone (how much of the Q distribution ends up looking like Fig. 3e), that remain uncertain.

To illustrate the point about robustness and

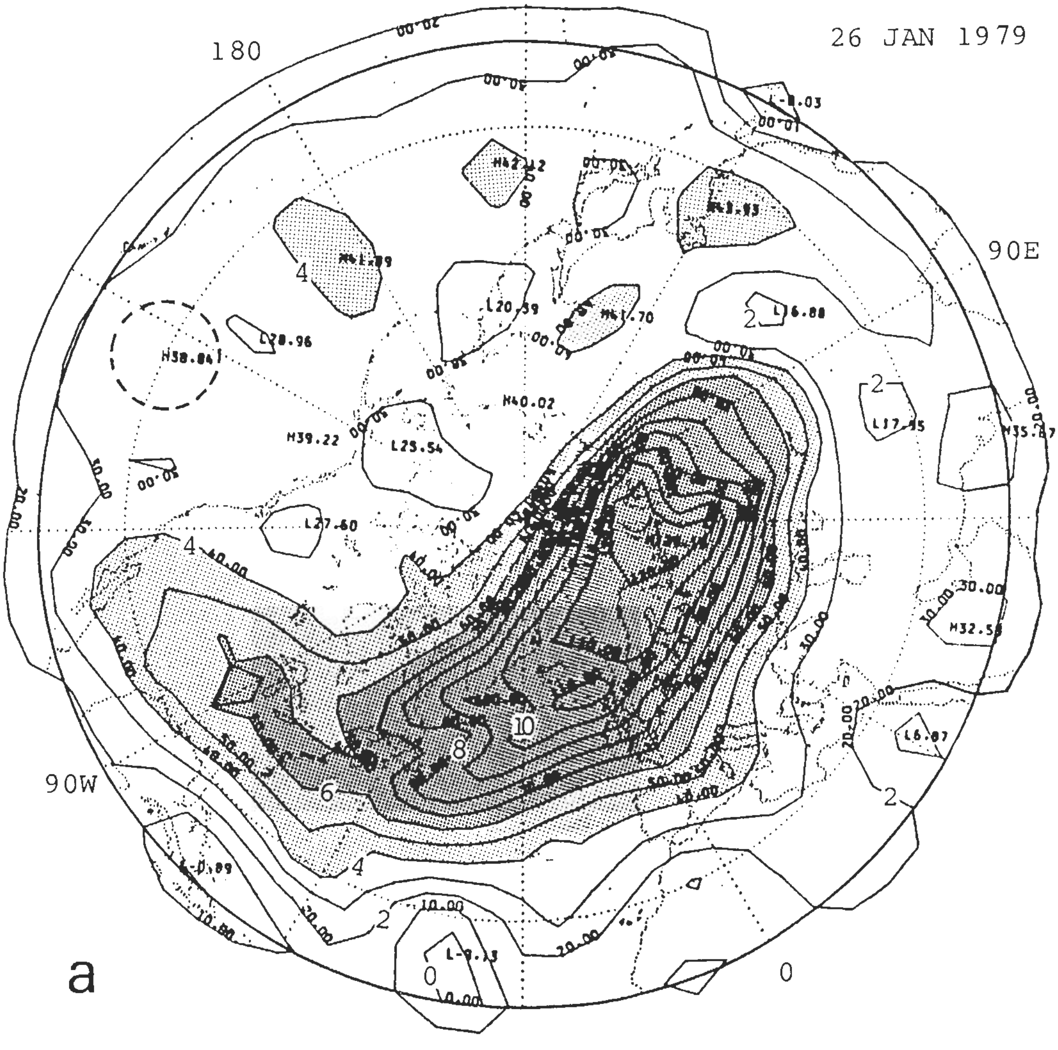


Fig. 7a
(see over)

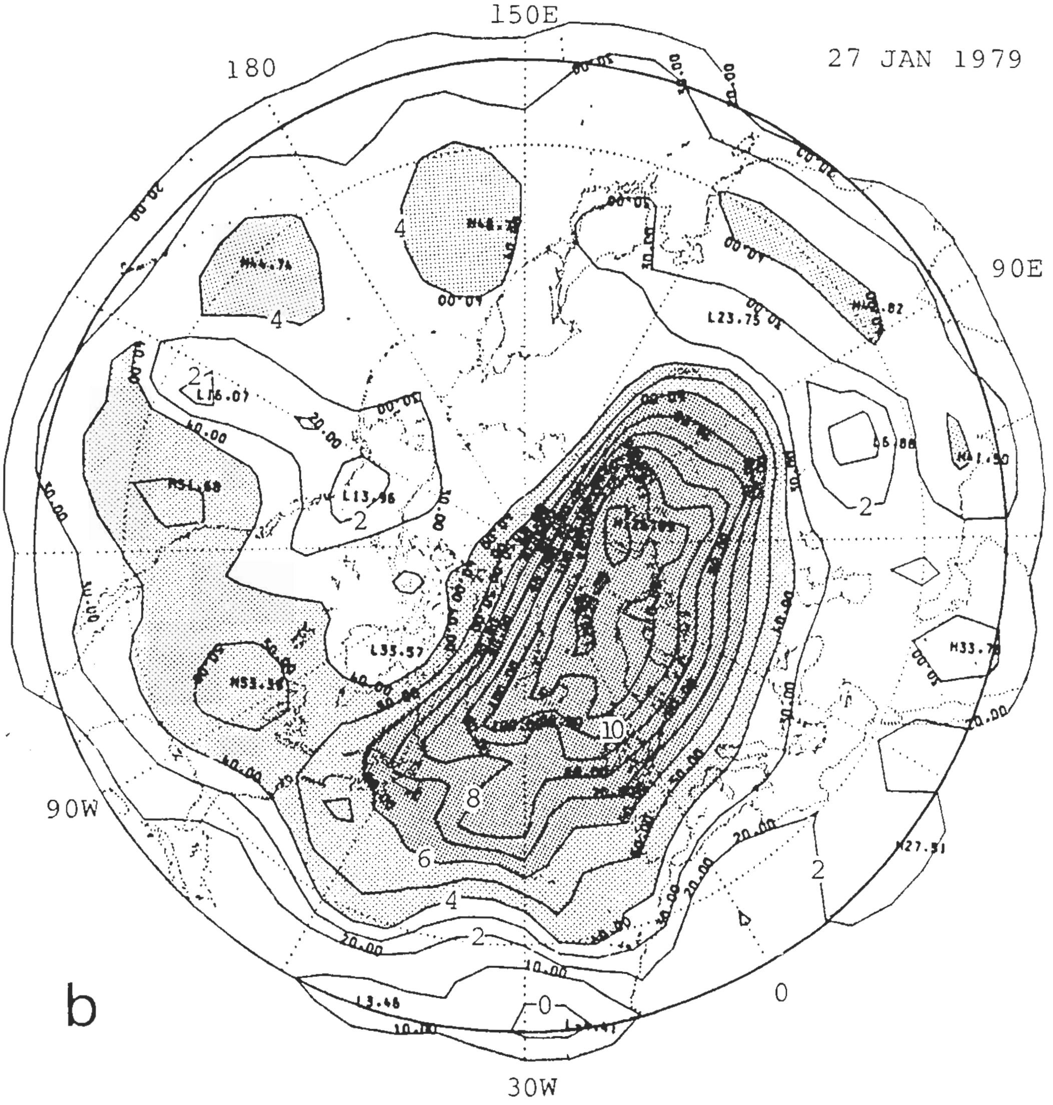


Fig. 7. FGGE-based potential vorticity as in Fig. 6, but (a) for 26 January and (b) for 27 January 1979. The shading refers to contour values greater than 4 and 6 units, as before. The contour interval is the same as in Fig. 6 (and twice as fine as in Figs. 1a, b). The small dashed circle in (a) marks the position, at its centre, of a local maximum value of 3.884 units, which only just misses being picked up by the contouring (see text, section 4).

continuity, Fig. 7 shows the FGGE-based Q maps for days 26 and 27. Notice the large chunk of high-Q air over North America on day 26 (Fig. 7a), which appears to be detaching itself from the main vortex to become the nearer part of the day-27 tongue (Fig. 7b). We say 'appears to', because some of this air actually goes around the south edge of the chunk and straight back into the edge of the main vortex. The trajectory calculations indicate that it is only the north edge of the chunk which is sheared out to form the nearer part of the day-27 tongue, supplemented by a further supply of high-Q air coming from the edge of the main vortex nearer the pole. The general appearance of Figs. 7a, b is quite typical of the wave-breaking signatures seen on the Q maps for the period, both NMC and FGGE-based. Another example is presented in Fig. 8, which shows the FGGE-based Q maps for 31 January and 1 February.

Figures 7a and 8a remind us of a feature commonly seen on middle-stratospheric geopotential *height* maps in winters during which large planetary-wave amplitudes develop. This is the transitory 'comma shape' of the outermost height contours surrounding the main vortex. Figure 9 is a typical example. It is, in fact, the 10 mbar height map for the same day, 26 January, as in Fig. 7a. It seems likely that this comma shape will prove to be generally indicative of very large scale wave-breaking events. The tilting of troughs associated with the upward and equatorward propagation of planetary-wave activity, an essentially linear phenomenon, merges imperceptibly into the nonlinear wave-breaking signature as the wave becomes 'surf'. Because of the inverse Laplacian (Charney and Stern, Hoskins *et al.*, *loc. cit. ante*), a time sequence of height maps gives an out-of-focus view of the process in comparison with the view given by a time sequence of Q maps, even coarse-grain ones like Figs. 7 and 8.

4. Local dynamical instability?

Figure 7b forces us to return briefly to the question of the further part of the day-27 tongue. If Fig. 7b gives a truer picture than Fig. 1b, as it may do in virtue of the better quality of the FGGE base data, then the strikingly regular blobbiness of the apparent tongue and its extension across the Pacific, which in Fig. 7b has the appearance of a tongue that has completely broken up into blobs, prompts the suggestion that we might actually be seeing the signature of a local dynamical instability of the sort discussed in Section 2. It could be a baroclinic or a barotropic instability, or more likely a mixture of the two. Order-of-magnitude estimates indicate tentatively that an instability with a large enough growth rate would be dynamically feasible if the Q contrast were on the high side of the range suggested by the evidence presented above. Quantitative instability calculations have yet to be done, and observational questions about vertical structure have yet to be answered. Note that the spacing of the blobs in Fig. 7b does not coincide with the spacing of satellite orbits. Looking along and just to the north of the 30°N latitude circle in Fig. 7b (shown dotted), one counts about five blobs, actual or incipient, between 150°E and 30°W, whereas there are seven satellite orbits.

Although there is no self-evident continuity of the blobs between Figs. 7a and 7b, it is noteworthy that the computer output comprising Fig. 7a does show a

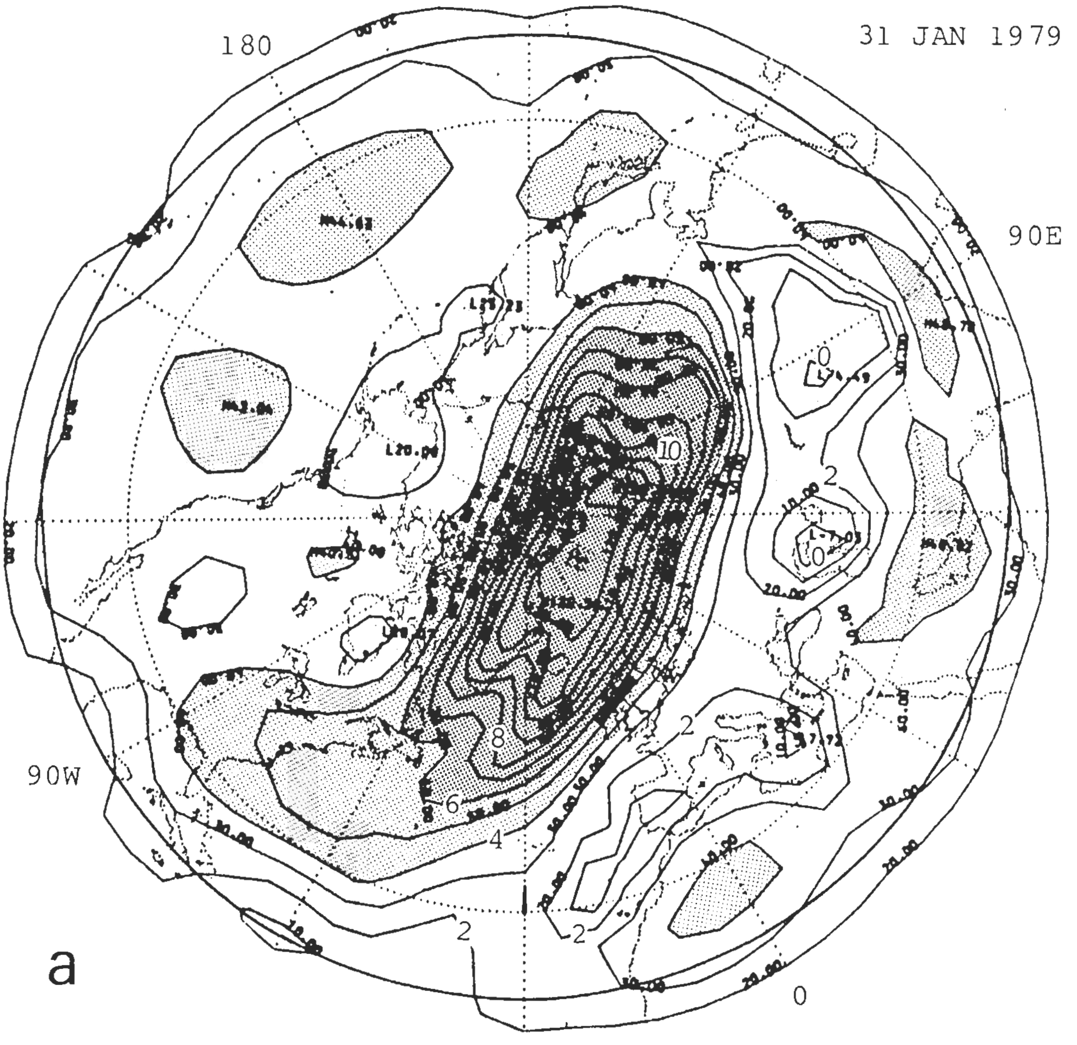


Fig. 8a
(see over)

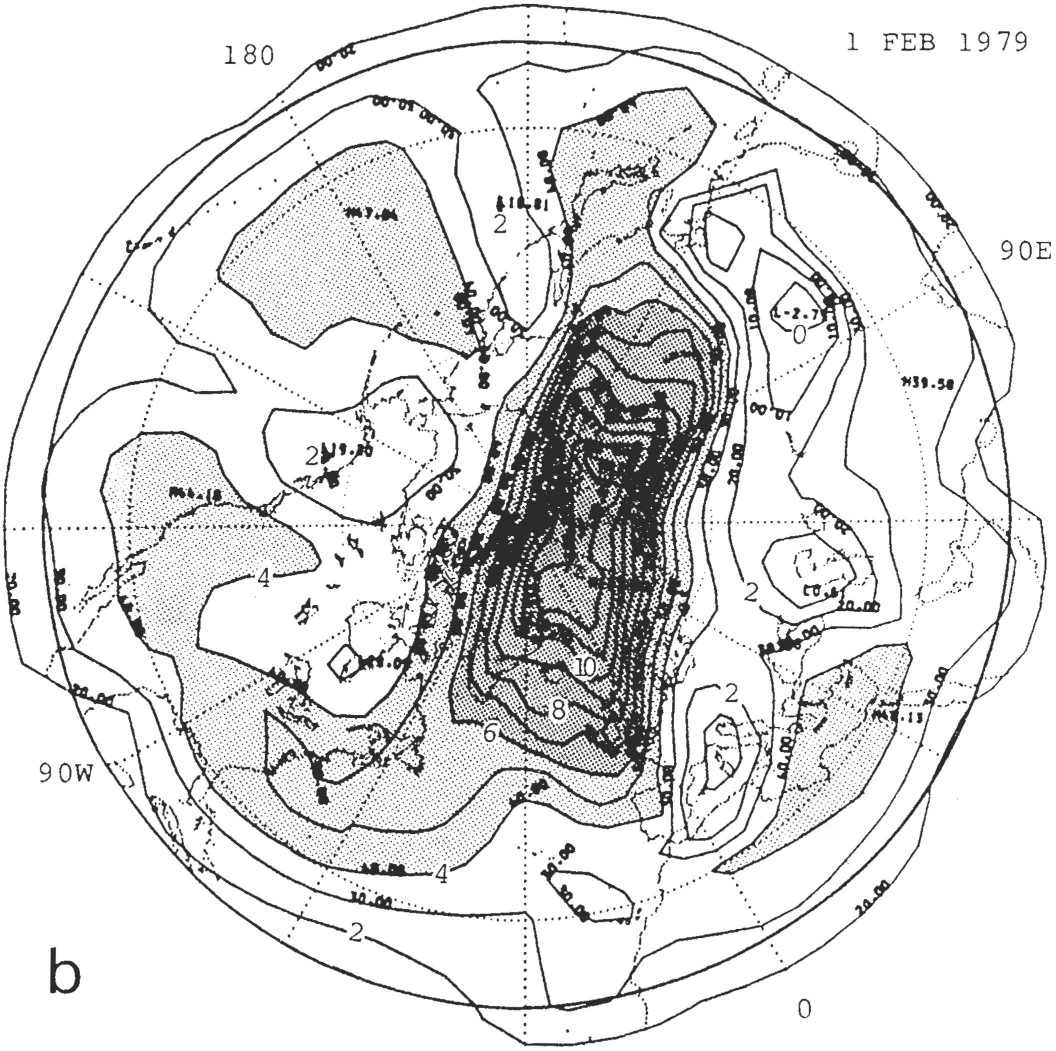


Fig. 8. FGGE-based potential vorticity as in Figs. 6 and 7, but (a) for 31 January and (b) for 1 February 1979. The shading refers to contour values greater than 4 and 6 units, as before.

maximum at 146°W, 30°N (at the centre of the dashed circle) which only just misses being picked up by the contouring, having the value 3.9 contour units (38.84 in the units used by the computer output, shown in small figures). It is in just the right position to correspond to the first fully detached blob at 165°W in Fig. 7b at advection speeds of the order indicated by Fig. 5. The apparent fluctuation in blob strength between Figs. 7a and 7b could be due to an interference or moiré effect between patterns due to real features and patterns due to orbital artifacts (PICK and BROWNSCOMBE, 1981), and this idea seems consistent with the order of magnitude of the apparent blob motion and the much slower drift of the orbits.

According to the local-instability hypothesis, the blobs, if resolved an order of magnitude better, would be expected to look like rolled-up billows or occluded cyclones, connected by very thin wisps of high-Q air, as suggested by the 'artist's impression' shown in Fig. 10. If the instability hypothesis is correct, one could use instability theory to put a rough lower bound on the Q contrast across the tongue, by requiring that the growth rate be sufficient to cause the hypothesized breakup in the time available. The concept of 'vortex rollup' (e.g. BATCHELOR 1967, p. 590) would also be relevant, especially to the furthermost vortex.

Regarding the balance of probabilities that the blobs reflect real features, we remark that although SSU orbital coverage in the subtropical Pacific was not perfect on days 26 and 27, there were always data from descending orbits wherever ascending ones were gappy, and vice versa, on both days. The ascending and descending orbits crossed near 20°N. It is pertinent to note in addition that case studies, in which similar blob-like features appear, have now been done for more recent winters during which data from more than one

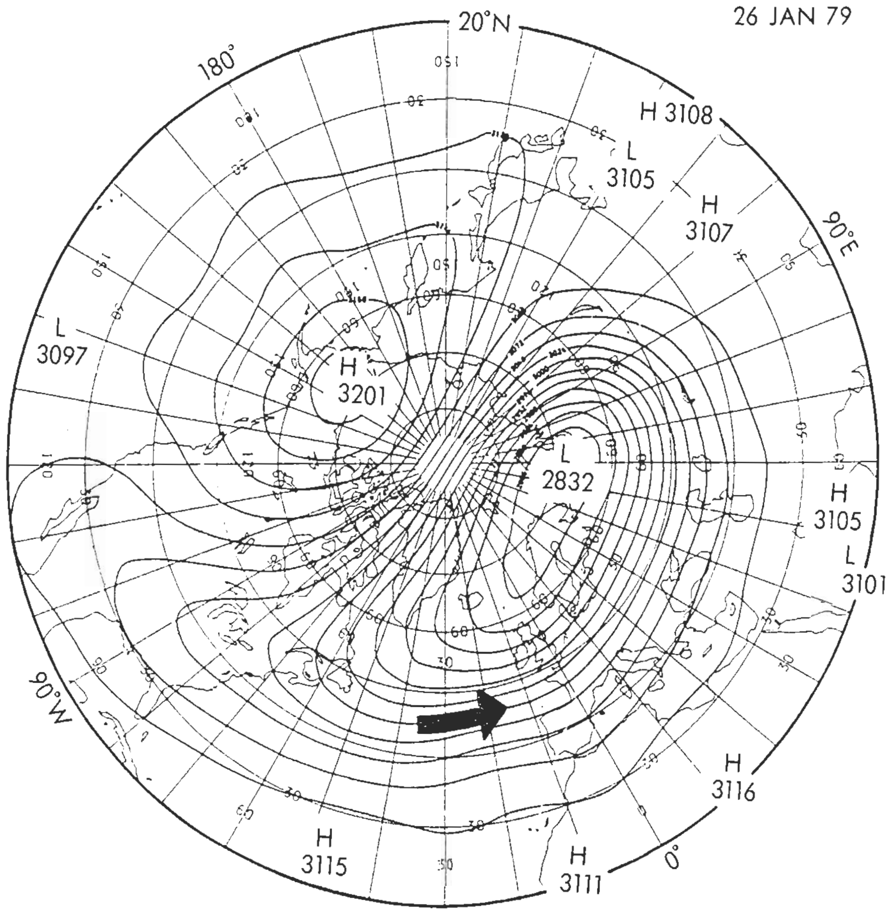


Fig. 9. NMC-based geopotential height of the 10 mbar isobaric surface, in dekametres, for the same time as the Q map shown in Fig. 7a (26 January 1979 at 00Z). Contour interval is 24 dekametres. Note the 'comma shape' of the outer contours surrounding the main vortex.

satellite were available (Clough *et al.*, 1984). In these cases it was concluded, purely from a consideration of the data, that at least some of the apparent blobs are likely to represent real features, since different satellites with different orbits showed the blobs in the same positions. If this is typical, it could be important for making estimates of the completeness or incompleteness of the mixing due to irreversible contour deformations like that suggested by Fig. 10. In cases where the Q distribution winds itself up into features large enough, in terms of the scale effect discussed in Section 2, to resist further deformation, the mixing could be much less complete than might be suggested by Fig. 3. *

5. Erosion versus diabatic effects

Figures 11a, b and c, the FGGE-based Q maps for 17 January, 16 February and 23 February 1979, summarize the development and eventual breakup of the main vortex, surf-zone structure during the second half of the winter. Note the reduction in the area of the main vortex, followed by the splitting of that vortex into two apparently separate pieces during the major warming of late February. (At higher resolution the two pieces would probably be seen to be joined by a narrow thread of high-Q air.) The overall picture agrees qualitatively with that suggested by the sequence of ozone maps for the same period presented by Leovy *et al.* (1984), who in addition present maps for the preceding autumn showing what seems to be the first appearance of surf-zone structure as planetary-wave amplitudes build up.

Inspection of the daily sequence of Q maps, and the ozone maps as well, indicates that the area $A(t)$ of the main vortex was a monotonically decreasing function of time t during the period spanned by Figs. 11a, b. This is consistent with the erosion hypothesis. In the case of January–February 1979, much of the net erosion appeared to be accountable for by the two large wave-breaking events seen in Figs. 7 and 8.

* Some further light on this is beginning to be shed by numerical experiments at very high resolution; a first report is in Juckes & McIntyre 1987, Nature 328, 590-6.

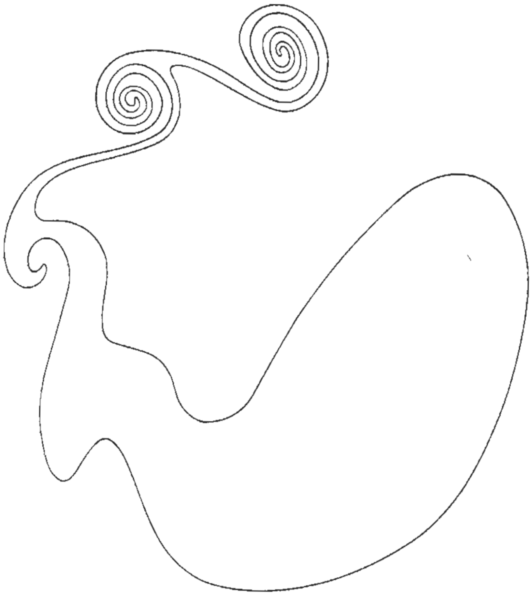


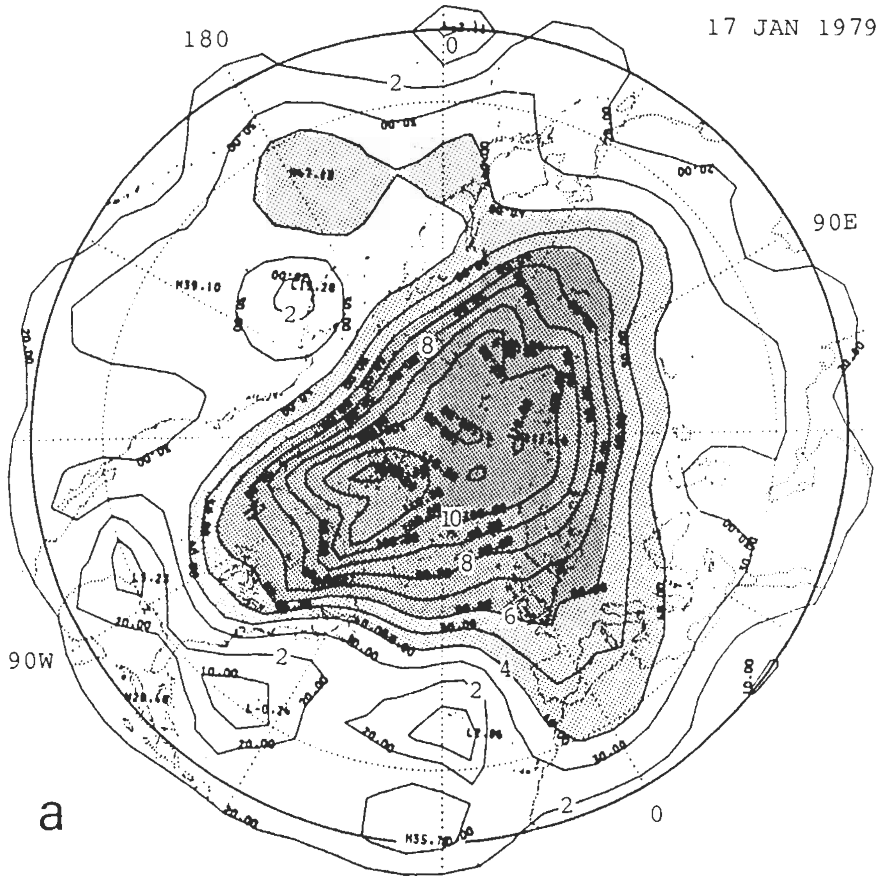
Fig. 10. A guess at the qualitative shape on 27 January 1979 of a material contour previously lying near the edge of the main vortex on or just before 23 January (see text, section 4, and Fig. 7b). The thin wisps of high-Q air joining the rolled-up vortices may themselves be unstable, and so on in the manner of L. F. Richardson: 'Big whorls have little whorls...'. The concept of 'vortex rollup' (e.g. Batchelor 1967, p. 590) may be relevant, as well as the concept of 'instability', especially for the furthermost vortex. If this interpretation is correct, then the vortices, once rolled up, will tend to be resistant to further mixing.

The relatively sharp edge of the main vortex, and the relatively steep gradients just inside it, are also consistent with the erosion hypothesis. Both are characteristic features of any situation where a conservable quantity Q, approximately satisfying an equation of the form (3), has an overall gradient in some region (in this case an isentropic surface spanning the northern hemisphere), but is being strongly mixed in part of that region. For a given overall gradient, and therefore a given total number of isopleths of Q, the result of such mixing will be to crowd most of the isopleths into the remaining space available. An attempt to suggest this process and the resulting gross structure pictorially was made in Fig. 5 of the review article by McIntyre (1982), here reproduced as Fig. 12 (see left-hand heavy curve). Parallel situations have been extensively studied on a small scale in the laboratory (Turner, 1973, §9.1.1), in order to help understand the formation of oceanic and atmospheric mixed layers and inversions. There, the relevant conservable quantity is potential temperature, density or chemical composition. (Thus the erosion hypothesis, in its simplest form, says that the stratospheric surf zone is like a mixed layer turned on its side, the steep isentropic Q gradients at the edge of the main vortex being analogous to the steep vertical potential-temperature gradients in the inversion which caps the mixed layer.) An essentially similar phenomenon is well known in magnetohydrodynamics under the heading 'magnetic flux expulsion' (e.g. Weiss, 1966; Moffatt and Kamkar, 1983; Rhines and Young, 1983), and another very striking example involving Q itself has been found by Dr W. R. Holland in recent experiments with a high-resolution numerical ocean model (Fig. 3 of Rhines and Young, 1982). Observational data from the real oceans seem to show the same signature (McDowell *et al.*, 1982; Sarmiento *et al.*, 1982; Holland *et al.*, 1984). In the concluding section of this paper we speculate that the tropopause may be yet another example of an erosion interface maintained, in part, by quasi-horizontal, isentropic erosion of Q.

p.846 b

For the middle stratosphere, it remains to consider whether processes other than erosion and mixing could have been responsible for the observed space–time structure in the ozone and Q distributions. We are grateful to S. B. Fels, D. L. Hartmann, C. B. Leovy and A. F. Tuck for some helpful conversations on this point. Since the diurnally-averaged flux S of absorbable solar radiation arriving in the middle stratosphere (mainly shortwave ultraviolet, $\lesssim 300$ nm) is a function of latitude ϕ and time t, it might be argued that the existence and time evolution of what we have called the surf zone, and its interface with the main vortex, could be the result of radiative–photochemical processes. Diabatic changes will certainly be important over time intervals of the order of a month, such as that covered by Figs. 11a, b. Indeed, they are responsible for the Q values in the main vortex being relatively high in the first place, and they are probably needed, moreover, to maintain the high Q values there in the second half of the winter, since the LIMS ozone results suggest that a certain amount of mixing of high-ozone (low-Q) air takes place across the interface (Leovy *et al.*, 1984). We think, however, that diabatic effects would be unlikely by themselves to give rise to a structure exactly like the observed one, particularly the interface itself. There are two reasons for this, neither of which depends on the fine details of radiative-transfer calculations.

First, the large-scale flow field in the middle stratosphere is sufficiently unsteady, and lacking in zonal symmetry, especially in highly disturbed winters like the winter of 1978–79, that no air parcel will actually be subject to the flux $S(\phi)$ corresponding to a single latitude ϕ. A zonally symmetric model of the stratosphere would be very different from the real stratosphere in this respect. In the presence of large-amplitude planetary-wave activity, most air parcels travel over a wide range of latitudes, often covering



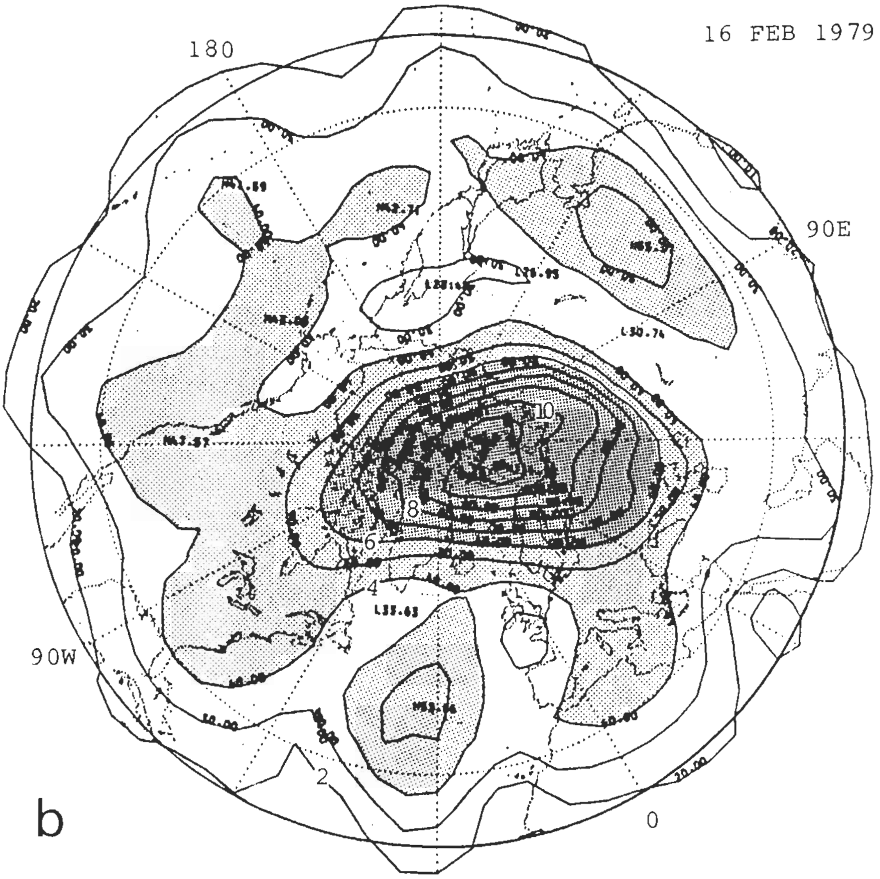


Fig. 11 a,b
(see over)

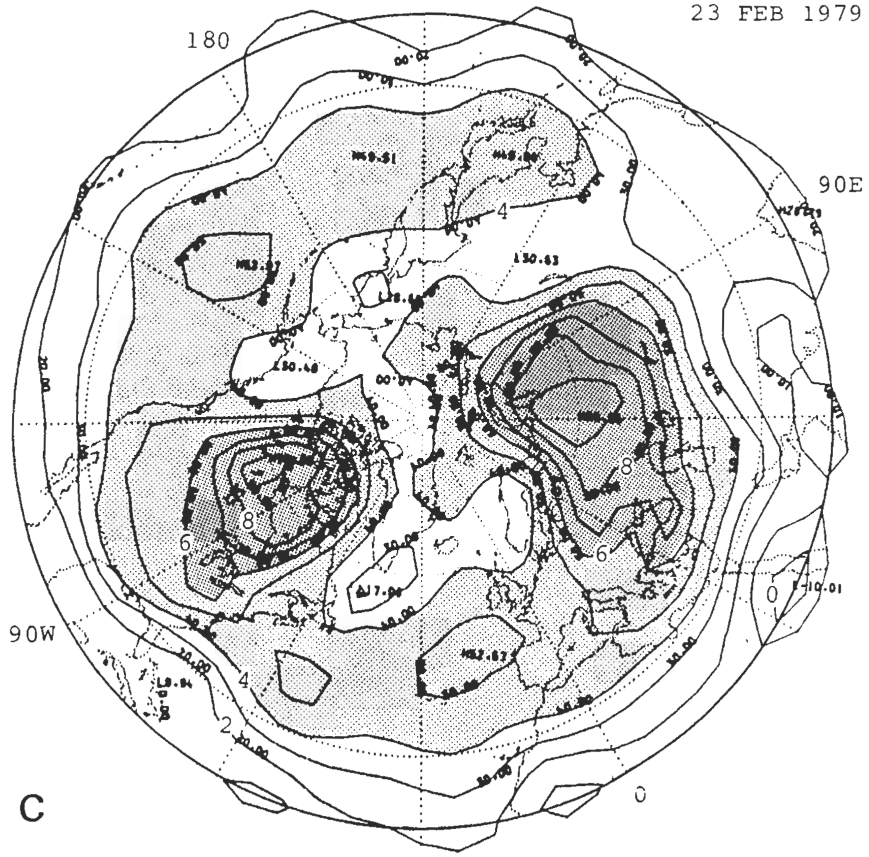


Fig. 11. FGGE-based potential vorticity as in Fig. 6, but (a) for 17 January, (b) for 16 February and (c) for 23 February 1979. The shading refers to contour values greater than 4 and 6 units, as before. Notice the decrease in the area of the main vortex between mid-January and mid-February (see text, section 5).

many tens of degrees latitude every few days. Such parcels must receive insolation corresponding to a very smeared-out version of the function $S(\phi)$ when averaged over a time scale of weeks. This smearing effect would be very likely to nullify the effect of any surf-zone-like structure which might be present in the latitude dependence of $S(\phi)$ itself.

Second, the existence of any such structure in $S(\phi)$ itself would seem unlikely in the first place, if not already present in the distributions of radiatively active constituents as a result of *non*-radiative processes, such as those we have postulated. In part this is because of the geometry of the situation. Near the edge of the polar night, which would appear to be the only place qualifying as a *prima facie* candidate for a sudden change in $S(\phi)$ as latitude ϕ varies, optical path lengths through the upper-stratospheric ozone shield, starting say at 50 km altitude, are extremely long for solar radiation arriving in the middle stratosphere at 30 km, even at mid-day. Because of these long path lengths, over 100 km within 10° latitude of the polar night and over 50 km within 20° of it, $S(\phi)$ will tend to be extremely small in comparison with its values at low latitudes: 50 km is considerably longer than typical *e*-folding distances for shortwave ultraviolet extinction in the ozone layer, and optical paths are still longer near sunset and sunrise. The short length of day reinforces the extinction effect still further. Thus it will not be until substantially lower latitudes ϕ are reached, where solar elevations are much higher and the day longer, that $S(\phi)$ (for shortwave ultraviolet) can be expected to rise to significant values. The foregoing expectations appear to be corroborated by the results of recent careful, line-by-line calculations of zonally symmetric radiative-equilibrium temperature fields $T_{\text{rad}}(\phi, p)$ (S. B. Fels, personal communication). In the middle stratosphere these are smooth, monotonic functions of latitude, showing no sudden changes near the edge of the polar night.

It is important to keep in mind that the diabatic evolution of Q is governed by values of the diabatic heating near the isentropic surface of interest, and not, at least not directly, by heating at lower levels of the atmosphere. More precisely, what is relevant is the

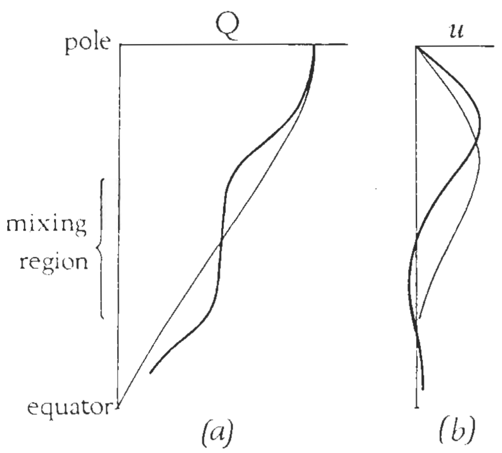


Fig. 12. (a) Schematic isentropic profiles of the gross distribution of Q in the middle stratosphere before and after quasi-horizontal mixing in the surf zone (light and heavy curves, respectively), on the assumption that mixing is fairly complete. Details such as those suggested by Figs. 3 and 10 are ignored. The profiles may be thought of as representing a zonally symmetric state to which the stratosphere might approximately revert if planetary wave amplitudes diminished to zero. (b) Zonal velocity profiles which might correspond qualitatively to (a) under parameter conditions typical of the middle stratosphere. (After McIntyre, 1982.)

Lagrangian history of the diabatic heating, and its gradient in the direction of the absolute vorticity vector, for each air parcel which arrives on the isentropic surface of interest. This follows from the generalization of equation (3) to take account of diabatic heating (e.g. Pedlosky, 1979, §2.5), which implies that

$$\mathrm{D}Q/\mathrm{D}t = \rho^{-1}(2\mathbf{\Omega}+\nabla\times\mathbf{u})\cdot\nabla H, \tag{4}$$

where H is the diabatic material rate of change of θ. It is particularly to be noted that 'radiative-equilibrium zonal wind profiles' derived from $T_{\mathrm{rad}}(\phi, p)$ and the thermal-wind relation by vertical integration from the ground upwards, assuming zero motion at the ground, have dubious relevance to the question of how the isentropic Q distribution might evolve purely diabatically as winter sets in. Such zonal-wind profiles tend, in any case, to be artificially sensitive to errors in T_{rad} in the air column extending between the middle stratosphere and the ground. The corresponding estimates of isentropic Q distributions are more sensitive still.

The formal simplicity of equation (4) and its partner

$$\mathrm{D}\theta/\mathrm{D}t = H \tag{5}$$

is to be contrasted with the relative complexity of the corresponding equations for the diabatic evolution of the height, wind and temperature fields, which are coupled vertically and horizontally by ageostrophic circulations and whose calculation again involves the inversion of Laplacian-like operators (e.g. Eliassen, 1951). This is why a direct consideration of the Q distribution is conceptually the simplest way of picturing what is happening, diabatically as well as adiabatically. Of course, a complete solution of the problem over a long time interval depends on solving for all the fields, including the meridional and vertical motions needed to evaluate D/Dt in (4) and (5) and to compute the feedback on H itself due to advectively-induced temperature changes. Such a calculation is needed before the matter can be further assessed.*

6. An objective definition of $A(t)$

It was suggested in McIntyre and Palmer (1983) that the area $A(t)$ of the main vortex at time t, as it appears on the daily sequence of Q maps for some suitable isentropic surface, should be valuable as a circulation index for the winter stratosphere. Its use should lead to an enhanced signal-to-noise ratio in statistical studies of such questions as the interannual variability of the winter stratosphere and the conditions under which major warmings are liable to occur. Arguing from considerations of planetary-wave 'focusing', we predicted that the conditions leading to major warmings will be found to be characterized by unusually small values of $A(t)$. The important property of such an index is that it avoids the loss of information incurred by Eulerian space and time averaging in the presence of fluctuations in the position and shape of the main vortex.

If the well-defined change from strong to weak isentropic Q gradients at the edge of the main vortex is a typical feature of the winter stratosphere, then the following procedure, or something close to it, should yield an objective definition of $A(t)$. We have not had the opportunity to try it in practice, and it is not yet clear how successfully it could be automated, since Q maps are not yet available for a sufficient number of other winters. It assumes that the isentropic distribution of Q is available as gridded values for each day t.

1. For a given day t, arrange all the grid points in descending order of Q value.
2. Let $A_{>}(Q_{*})$ be the total area associated with all the grid points having Q values of Q_{*} or greater. (Note that $A_{>}(Q_{*})$ is a monotonically decreasing function of Q_{*}, no matter how noisy the Q distribution.)
3. Starting from the maximum value of Q_{*}, follow the graph of $A_{>}(Q_{*})$ as Q_{*} decreases and find the first major *increase* in the magnitude of its slope $|\mathrm{d}A_{>}(Q_{*})/\mathrm{d}Q_{*}|$. Define $A(t)$ to be the value of $A_{>}(Q_{*})$ at the break-point where the slope changes in this sense.

* First attempts: Butchart & Remsberg JAS 43, 1319; Butchart, in Visconti & Garcia 1987: Transport Processes in the Middle Atmosphere pp. 121-136 Dordrecht, Reidel.

Perhaps the simplest robust algorithm for detecting the break-point would be to use least-squares local fitting of quadratic polynomials. That is, a quadratic polynomial would be fitted to the function $A_>(Q_*)$ over a small interval I centred on one value of Q_*, and the process repeated as Q_* decreases from grid-point to grid-point. The break could then be defined as the first prominent maximum in the leading (square-term) coefficient of the polynomial as Q_* decreases. The size of the interval I would need to be judiciously chosen so as to take in a sufficient number of grid points for the least-squares fit to be noise-resistant, but small enough to avoid smearing the break in the curve unnecessarily. If further noise reduction is required, the best procedure might be to extract the function $A_>(Q_*)$ for two or three successive days and then construct a composite $A_>(Q_*)$. This would preserve the advantage gained by the avoidance of Eulerian averaging, even if the main vortex was moving or changing shape significantly during the time interval considered.

7. CONCLUDING REMARKS

Even with imperfect data, and relatively crude data-processing and graphical techniques, the isentropic Q map viewpoint seems already to be leading to a flood of new insights in atmospheric dynamics. It is proving to be especially illuminating as regards the interplay between different locations and different scales of motion, and between different dynamical mechanisms. Similar developments have been taking place in dynamical oceanography (e.g. HOLLAND *et al.*, 1984), and we appear to be entering a period of renewed cross-fertilization between the two subjects. For the reasons mentioned in the introduction, Q maps provide information about the motion in a form which makes the dynamics highly visible and which is ideally suited to judging both the relevance and the limitations of theoretical constructs such as 'wave propagation', 'instability', 'turbulence', 'horizontal diffusion', 'critical layers', and so on, and to making meaningful comparisons with the results of numerical simulations and other kinds of model.

The implications of the present work for stratospheric pollutant and tracer-transport modelling were briefly indicated in MCINTYRE and PALMER (1983), where we outlined a possible procedure for improving two-dimensional modelling which takes advantage of the observed main vortex, surf zone structure. The calibration of the resulting two-dimensional models, which would depend on observational estimates of mixing and erosion rates, would be greatly helped if long time series of the area $A(t)$ of the stratospheric main vortex were to become available (Section 6). Moreover, should $A(t)$ prove to be well-defined and computable for long runs of winter stratospheric data, then it could be used, as already mentioned, as a general-purpose stratospheric circulation index which should lead to improved statistical significance in studies of the conditions leading to major warmings and of the interannual variability of the stratosphere and its possible connections with the quasi-biennial and southern oscillations. Progress in investigating other aspects of stratospheric behaviour, as hypothesized, for instance, in the resonance theory of stratospheric warmings (TUNG and LINDZEN, 1979a, b; PLUMB, 1981; MCINTYRE, 1982, section 5), will likewise depend on a clearer appreciation of the morphology of isentropic Q distributions and the mechanisms by which they evolve.

The same general considerations hold for the troposphere, and it is of interest to consider what tropospheric phenomena might turn out to be analogous to the phenomena we believe we are seeing on a larger scale in the stratosphere. It seems likely that although the troposphere and lower stratosphere appear to be more complicated in detail, there are, indeed, significant similarities, aspects of which have long been familiar to synoptic meteorologists, although not always seen clearly as part of the dynamical whole.

For example, we may reasonably expect that the high subtropical troposphere has the character of a surf zone, in which upward and equatorward propagating Rossby waves arriving from various sources in the lower troposphere break in much the same way as in the middle stratosphere. The relevant isentropic surfaces lie roughly in the range 330–350 K, depending on the time of year. The evidence from general circulation statistics and numerical modelling experiments seems broadly suggestive, both for stationary and for travelling waves, of the same kind of upward-equatorward pattern of propagation as we tend to see in the winter stratosphere. Some of the travelling waves (with positive zonal phase speeds) appear to be generated directly by mid-latitude cyclones as they occlude at low levels (EDMON *et al.*, 1980, section 4; HOSKINS, 1983, section 7.3.4; HOSKINS and MCINTYRE, 1984). The upward-equatorward propagation pattern, and the implied net transfer of (negative) angular momentum from the wave-generation region to the surf region, provide a physical picture which accounts entirely naturally for the so-called 'negative viscosity phenomenon' (STARR, 1968). EDMON *et al.* (1980) give a brief review of the relevant theoretical concepts in their simplest form (see also GILL, 1982, sections 13.9 and 13.10.2). Among earlier versions of the idea one may mention the study by DICKINSON (1969), which was based on linear Rossby wave critical-layer theory.

Rossby wave breaking may, in fact, occur near the tropopause at a variety of latitudes. In particular, the high-latitude convergence of the Eliassen–Palm flux seen in Figs. 1 and 4a, b of Edmon *et al.* (1980) may be partly due to the breaking of Rossby waves which appear in a zonally-averaged picture to be propagating straight up from below (and some of which may, for example, be actually propagating equatorward, but also downstream, relative to northeast-oriented storm tracks). Direct synoptic evidence for tongues of high-Q air in the high troposphere has often been seen in the form of 'shear lines' or 'upper-air fronts' near the ends of storm tracks, one of the places where one would expect to find the 'surf' produced by upward-propagating waves resulting from mid-latitude baroclinic instabilities. For a recent example, in which the local Q distribution was evaluated on the 330 K isentropic surface, see Holopainen and Rontu (1981). The Q contrast across the observed tongue was quite strong, and order-of-magnitude estimates suggests that it could have been locally unstable in the manner suggested in Fig. 10. Again, quantitative instability or vortex rollup calculations would be of interest here, especially since Holopainen and Rontu, referring to general synoptic experience with the phenomenon, comment that 'the southern part ... quite often forms a cut-off low'. This is reminiscent of the behaviour suggested in Fig. 10.

The same thing may happen on a larger scale whenever the stationary long-wave pattern amplifies in the troposphere. It is possible, for instance, that the associated erosion tongues are the essential cause of the upper-air cold vortices commonly seen in satellite infrared cloud pictures over the western tropical Atlantic and Pacific Oceans in summer (e.g. Frank, 1970; Shimamura, 1981; Sugi and Kanamitsu, 1982; Bengtsson *et al.*, 1982, section 2a). These vortices may, in other words, be essentially the same phenomenon as the middle-stratospheric phenomenon hypothesised in Fig. 10. The suggestion that such disturbances may have the nature of dynamical instabilities is not new, of course, but the wave-breaking picture provides a natural explanation of where the local isentropic potential-vorticity gradients required by the instability mechanism might come from. It also predicts that the upper-air disturbances in question should be most in evidence soon after the long-wave pattern amplifies. This prediction should be amenable to testing observationally.*

The wave-breaking *ansatz* leads one to expect the existence of erosion-sharpened interfaces on isentropic surfaces intersecting the lower stratosphere and upper troposphere. Once again, these would appear to be familiar, synoptically, as the 'dynamic tropopauses' of Reed and Sanders (1953), Reed (1955), Danielsen (1968, 1981), Shapiro (1976), and others, who locate them as sharp transitions from high-Q to low-Q air at major upper-air jet streams. This suggests immediately that the observed structure of the tropopause itself might be due partly to large-scale, quasi-horizontal erosion, at least in middle and low latitudes, in fundamentally the same way as we have postulated for the middle stratosphere, although the details are likely to be more complicated (references above). For some purposes the tropopause may, as the isentropic viewpoint itself suggests, be more like a 'wall' than a * 'ceiling'. Striking evidence indicating that upper-air jet streams approximately mark the positions of material contours, as would be expected from this picture, has been presented by Danielsen (1968), from measurements of nuclear test debris using instrumented aircraft. In one case, two quite different air masses, clearly distinguished from each other by different radioactive debris originating from different nuclear tests, flowed side by side in an upper-air jet over North America. Locally there was little small-scale mixing. Therefore, large-scale, *quasi-horizontal* mixing of the type suggested so vividly by the stratospheric Q maps, and also by Q maps near 300–310 K such as those presented by Reed (1955), Obukhov (1964) and Danielsen (1968), does seem to be a crucial part of the observed tracer transport and stratospheric–tropospheric exchange, as many investigators have emphasised. The present picture goes deeper, however, by suggesting the intimate connection between such tracer transports and dynamical phenomena like 'negative viscosity'.

Finally, we remark that daily sequences of Q maps on isentropic surfaces between 300 and 350 K, at present being constructed from operational meteorological data analyses (B. J. Hoskins and G. J. Shutts, personal communication), are promising not only to throw more light on the questions we have just been discussing, but also to help answer old questions about phenomena, like blocking, which have a very direct impact on weather forecasting. It seems no exaggeration to say that isentropic Q maps, from high-resolution numerical forecasting models as well as from real data, will soon greatly sharpen our understanding of these and a whole host of other large-scale dynamical processes.

Acknowledgements—We are again grateful to members of the Met.O.20 stratospheric analysis team at the U.K. Meteorological Office for supplying the data and for helping generously in many ways, particularly A. O'Neill, S. A. Clough, M. D. Brookes, N. S. Grahame and K. B. Robertson. We thank J. Austin for his valuable contributions to the isentropic trajectory calculations, C. B. Leovy for showing us the LIMS ozone results in advance of publication, S. B. Fels for showing us his stratospheric radiative-

* Liebmann & Hartmann '84, J Atmos. Sci <u>41</u>, 3333 !!

* Further discussion in Lindzen
1993, J. Atmos. Sci. 50, 1148
1994 " " " 51, 757.

equilibrium temperature fields, also unpublished, and P. D. KILLWORTH and J. VENN for kindly producing Fig. 2. Useful discussion and criticism were contributed by D. G. ANDREWS, J. AUSTIN, S. A. CLOUGH, S. B. FELS, A. E. GILL, D. L. HARTMANN, P. H. HAYNES, W. R. HOLLAND, E. O. HOLOPAINEN, J. R. HOLTON, B. J. HOSKINS, C.-P. F. HSU, M. KANAMITSU, R. J. KURZEJA, K. LABITZKE, C. B. LEOVY, R. S. LINDZEN, J. D. MAHLMAN, T. MATSUNO, R. MCINTYRE, TS. NITTA, A. O'NEILL, D. R. PICK, R. S. QUIROZ, P. B. RHINES, J. M. RUSSELL, M. L. SALBY, G. J. SHUTTS AND A. F. TUCK.

REFERENCES

AUSTIN J. and TUCK A. F. 1984 *Q. Jl R. met. Soc.* (submitted).

BATCHELOR G. K. 1967 *An Introduction to Fluid Dynamics*, p. 590., Cambridge University Press, Cambridge, U.K.

BATCHELOR G. K. 1969 *Phys. Fluids* Suppl. II, 233.

BENGTSSON L., KANAMITSU M., KÅLLBERG P and UPPALA S. 1982 *Bull. Am. met. Soc.* **63**, 277.

BENJAMIN T. B. and OLVER P. J. 1982 *J. Fluid Mech.* **125**, 137.

BRETHERTON F. P. 1966a *Q. Jl R. met. Soc.* **92**, 325.

BRETHERTON F. P. 1966b *Q. Jl R. met. Soc.* **92**, 335.

BRETHERTON F. P. and HAIDVOGEL D. B. 1976 *J. Fluid Mech.* **78**, 129.

CHARNEY J. G. and STERN M. E. 1962 *J. atmos. Sci.* **19**, 159.

CLOUGH S. A., GRAHAME N. S. and O'NEILL A. 1984 *Q. Jl R. met. Soc.* (to ~~be submitted)appear~~) 111, 335–358

Al-Ajmi et al " 359–389

DANIELSEN E. F. 1968 *J. atmos. Sci.* **25**, 502.

DANIELSEN E. F. 1981 *J. atmos. Sci.* **38**, 1319.

DICKINSON R. E. 1969 *J. atmos. Sci.* **26**, 73. ← Further discussion: Dritschel & McI J. Atmos Sci. 2007

EDMON H. J., HOSKINS B. J. and MCINTYRE M. E. 1980 *J. atmos. Sci.* **37**, 2600. (See also Corrigendum, *J. atmos. Sci.* **38**, 1115, especially second-last item.)

ELIASSEN A. 1951 *Astrophys. norv.* **5** (2), 19.

FJØRTOFT R. 1950 *Geofys. Publr* **17** (6), 1.

FRANK N. L. 1970 *Proc. Honolulu Symposium on Hurricanes and Tropical Meteorology* EIV-1–EIV-5. American Meteorological Society, Boston.

GILL A. E. 1982 *Atmosphere–Ocean Dynamics.* Academic, New York. Academic, New York.

HAYNES P. H. 1984 *J. Fluid Mech.* ~~(submitted)~~. 161, 493–511.

HERRING J. R., ORSZAG S. A., KRAICHNAN R. H. and FOX D. G. 1974 *J. Fluid Mech.* **66**, 417.

HODGES R. R. 1967 *J. geophys. Res.* **72**, 3455.

HOLOPAINEN E. O. and RONTU L. 1981 *Tellus* **33**, 351.

HOLLAND W. R., KEFFER T. and RHINES P. B. 1984 *Nature* **308**, 698.

HOLLOWAY G. and KRISTMANNSSON S. S. 1984 *J. Fluid Mech.* **141**, 27, figure 2.

HOSKINS B. J. 1983 *Large-Scale Dynamical Processes in the Atmosphere*, section 7.3.4, Ed. B. J. Hoskins and R. P. Pearce, Academic, New York.

HOSKINS B. J. and MCINTYRE M. E. 1984 In preparation.

HOSKINS B. J., MCINTYRE M. E. and ROBERTSON A. W. 1984 *Q. Jl R. met. Soc.* (to ~~be submitted). appear~~) 111, 8778 113, 402

HSU C.-P. F. 1981 *J. atmos. Sci.* **38**, 189.

KILLWORTH P. D. and MCINTYRE M. E. 1984 *J. Fluid Mech.* ~~(submitted)~~. 161, 449–492.

LEOVY C. B., SUN C.-R., HITCHMAN M. H., REMSBERG E. E., RUSSELL J. M., GORDLEY L. L., GILLE J. C. and LYJAK L. V. 1984 *J. atmos. Sci.* ~~(in press)~~. 42, 230–244.

LIGHTHILL M. J. 1963 *Laminar Boundary Layers*, Ed. L. Rosenhead, p. 93. Oxford University Press, Oxford, U.K.

LINDZEN R. S. 1967 *Q. Jl R. met. Soc.* **93**, 18.

LINDZEN R. S. 1981 *J. geophys Res* **86C**, 9707.

MATSUNO T. 1970 *J. atmos. Sci.* **27**, 871.

MCDOWELL S., RHINES P. and KEFFER T. 1982 *J. phys. Oceanogr.* **12**, 1417.

MCINTYRE M. E. 1982 *J. met. Soc. Japan.* **60**, 37.

MCINTYRE M. E. and PALMER T. N. 1983 *Nature* **305**, 593.

MOFFATT H. K. and KAMKAR H. 1983 *Stellar and Planetary Magnetism*, Ed. A. M. Soward, p. 91. Gordon and Breach, London.

OBUKHOV A. M. 1964 *Meteorologiya Gidrologiya* **1964** (2), 3.

PALMÉN E. and NEWTON C. W. 1969 *Atmospheric Circulation Systems*, Academic, New York.

PEDLOSKY J. 1979 *Geophysical Fluid Dynamics.* Springer New York.

PEIERLS R. 1979 *Surprises in Theoretical Physics*, p. 76. Princeton University Press, Princeton, U.S.A.

Phillips O. M. 1966 *The Dynamics of the Upper Ocean*, 1st edn., p. 186. Cambridge University Press, Cambridge, U.K.

Pick D. R. and Brownscombe J. L. 1981 *Adv. Space Res.* **1,** 247.

Plumb R. A. 1981 *J. atmos. Sci.* **38,** 2514.

Reed R. J. 1955 *J. Met.* **12,** 226.

Reed R. J. and Sanders F. 1953 *J. Met.* **10,** 338.

Rhines P. B. 1979 *A. Rev. Fluid Mech.* **11,** 401.

Rhines P. B. and Young W. R. 1982 *J. Fluid Mech.* **122,** 347.

Rhines P. B. and Young W. R. 1983 *J. Fluid Mech.* **133,** 133.

Salby M. L. 1982 *J. atmos. Sci.* **39,** 2577, 2601.

Sarmiento J. L., Rooth C. G. H. and Roether W. 1982 *J. geophys. Res.* **87C,** 8047.

Shapiro M. L. 1976 *Mon. Weath. Rev. Am. met. Soc.* **104,** 892.

Shimamura M. 1981 *Geophys. Mag.* **39,** 119.

Staley D. O. 1960 *J. Met.* **17,** 591.

Starr V. P. 1968 *Physics of Negative Viscosity Phenomena.* McGraw-Hill.

Stewartson K. 1978 *Geophys. Astrophys. Fluid Dyn.* **9,** 185.

Sugi M. and Kanamitsu M. 1982 *J. met. Soc. Japan* **60,** 932.

Tung K. K. and Lindzen R. S. 1979a *Mon. Weath. Rev. Am. met. Soc.* **107,** 714.

Tung K. K. and Lindzen R. S. 1979b *Mon. Weath. Rev. Am. met. Soc.* **107,** 735.

Turner J. S. 1973 *Buoyancy Effects in Fluids*, section 9.1.1. Cambridge.

Warn T. and Warn H. 1978 *Stud. appl. Math.* **59,** 371.

Weiss N. O. 1966 *Proc. R. Soc.* **A293,** 310.

Welander P. 1955 *Tellus* **7,** 141.–156

Woods J. D. 1968 *J. Fluid Mech.* **32,** 791.

Further relevant publications:
McIntyre & Palmer 1985: PAGEOPH 123, 964-975
Juckes & McIntyre 1987: Nature 328, 590-596.

APPENDIX A

The units in terms of which Q is contoured in the maps presented here, and in McIntyre and Palmer (1983), are defined in the following way. The quantity plotted is

$$\hat{Q} = g^{-1}H_0^{-1}p_0Q \qquad \text{(A1)}$$

expressed in units of $10^{-4}\,\mathrm{K\,m^{-1}\,s^{-1}}$, p_0 being a standard sea-level pressure taken as 1000 mbar and H_0 a standard pressure scale height taken as 6.5 km. A large-Richardson-number approximation to (1) is used, namely

$$Q = -g(f+\zeta)\partial\theta/\partial p, \qquad \text{(A2)}$$

where f and ζ are the vertical components of $2\mathbf{\Omega}$ and $\nabla \times \mathbf{u}$, ζ being evaluated isobarically for reasons of data-processing convenience. and where the hydrostatic relation has been used so that

$$\rho^{-1}\partial\theta/\partial z = -g\partial\theta/\partial p, \qquad \text{(A3)}$$

z being altitude and g the acceleration due to gravity.

In order to get a feel for the numerical values, one may note from these relations, and the conservation relation (3), that, for a given numerical value of $\hat{Q}$, the vertical absolute vorticity component Z_0 which would be realized if $\partial\theta/\partial p$ were brought to a standard value $\partial\theta_0/\partial p$, by means of an adiabatic, frictionless rearrangement of mass, is given by

$$Z_0 = \frac{-Q}{g\,\partial\theta_0/\partial p} = \frac{-H_0\hat{Q}}{p_0\,\partial\theta_0/\partial p}. \qquad \text{(A4)}$$

This is $0.224\hat{Q}$ in $\mathrm{s^{-1}}$ when $\partial\theta_0/\partial p$ is taken as $-29\,\mathrm{K\,mbar^{-1}}$, which is approximately the area average of $\partial\theta/\partial p$ on the 850 K isentropic surface north of 45°N on 26 January according to the satellite data analysis scheme. The contour marked 10 units then corresponds to $\hat{Q} = 10 \times 10^{-4}\,\mathrm{s^{-1}}$, i.e. to $Z_0 = 2.24 \times 10^{-4}\,\mathrm{s^{-1}}$. On 31 January the area average of $\partial\theta/\partial p$ north of 45°N was about $-27\,\mathrm{K\,mbar^{-1}}$, giving $Z_0 = 2.4 \times 10^{-4}\,\mathrm{s^{-1}}$ on the contour marked 10 units. On 17 January it was very close to $-32.5\,\mathrm{K\,mbar^{-1}}$, giving $Z_0 = 2.0 \times 10^{-4}\,\mathrm{s^{-1}}$ on the contour marked 10 units. These numbers may be compared with the maximum planetary vorticity $2\Omega = 1.458 \times 10^{-4}\,\mathrm{s^{-1}}$.

The data-processing procedures were outlined in McIntyre and Palmer (1983) and were based on a square grid on a polar stereographic projection, the grid size being about 580 km at 50°N. The procedures were essentially the same as those described in more detail in Clough *et al.* (1984), except that we used the gradient-wind approximation instead of the geostrophic approximation. One reason for doing this is that it gives more accurate Q values in regions of high wind speed and curvature, such as parts of the main vortex. The comparison between Figs. 1a and 1b gives a rough idea of the difference.

A minor omission in the description of data-processing presented in McIntyre and Palmer (1983) was to fail to note that, for the NMC-based Q maps, FGGE data at 50 mbar were used to help determine the cubic splines from which $\partial\theta/\partial p$ was calculated. This was because temperatures at the 50 mbar level are not well determined by the SSU data. The effect on the analysis near 10 mbar is believed to have been slight, like the effect of the missing data from the topmost SSU channel mentioned in McIntyre and Palmer (1983). As pointed out in McIntyre and Palmer (1983), all the vertical interpolations were carried out with respect to ln p, roughly equivalent to using the true altitude z. FGGE data were not used in any other way in the NMC-based maps, and the difference between Figs. 1b and 7b illustrates this. Fig. 7b used FGGE data not only for the 50 mbar contribution to the cubic splines, but also for the more crucial base analysis.

The deceleration correction, used in some of the isentropic trajectory calculations and referred to in connection with Figs. 4 and 5, is derived by taking the component of the momentum equation along the true trajectory and is given by the formula

$$\delta\eta = \frac{\delta|\mathbf{u}|}{f-\dot{\psi}_g}, \tag{A5}$$

where $\delta\eta$ is the lateral displacement, in a sense to be defined precisely below, of an air parcel to the side of the direction of the gradient wind during one time step (the direction of the gradient wind being assumed constant during the time step); $\delta|\mathbf{u}|$ is the change in $|\mathbf{u}|$ for the air parcel during the time step (the true change, not the change in the gradient wind, although in practice the latter is a good approximation), f is the vertical component of $2\mathbf{\Omega}$ as before, and $\dot{\psi}_g$ is the material rate of change of the direction of the *gradient* wind, usually fairly small in comparison with f. The direction of the gradient wind is used in the denominator, and not that of the true wind (this makes even less difference in practice), because the denominator times the magnitude $|\mathbf{u}_g|$ of the gradient wind represents the given isobaric geopotential height gradient or isentropic Montgomery-streamfunction gradient. The sign convention for $\delta\eta$ is such that the displacement is towards high geopotential if $\delta|\mathbf{u}|$ is negative (parcel decelerating). The sign of $\dot{\psi}_g$ is opposite to that of f in the denominator for trajectories which curve anticyclonically, as in Fig. 5.

The precise definition of $\delta\eta$ is $\delta\xi_g \sin\alpha$, where α is the angle between the corrected and uncorrected directions and where $\delta\xi_g = |\mathbf{u}_g|\,\delta t$, the (uncorrected) displacement along the direction of the gradient wind $\mathbf{u}_g$ during the time step δt. [The true displacement along the corrected trajectory is $\delta\xi_g \cos\alpha$. $(f-\dot{\psi}_g)/(f-\dot{\psi}_{\text{true}})$ in the present notation. The difference between this and $\delta\xi_g$ is of order Rossby number squared and is again unimportant, since α $(=\psi_{\text{true}}-\psi_g)$ is typically no more than a few degrees, as in Fig. 5, except when the wind field is locally very nonuniform, in which case other questions about the meaning of any kind of trajectory calculation are much more serious.]

One can get an immediate idea of the order of magnitude of the cumulative correction $\Delta\eta$ over a finite segment of the trajectory by integrating (A5) along the uncorrected, gradient-wind trajectory using a mean value D_m of the denominator (which can usually be roughly equated to a typical value of f), to give

$$\Delta\eta = \Delta|\mathbf{u}|/\mathrm{D}_m, \tag{A6}$$

where $\Delta|\mathbf{u}|$ is the total change in $|\mathbf{u}|$ along that segment. For the decelerating portion of the heavy dotted curve in Fig. 5 this is about 300 km, if we take $\Delta|\mathbf{u}| = 38\ \mathrm{m\ s^{-1}}$ and $\mathrm{D}_m = 2\Omega \sin 65^\circ = 1.3\times10^{-4}\ \mathrm{s^{-1}}$. The main uncertainty in making such a rough estimate *a priori* comes from the fact that the integrated lateral displacement $\Delta\eta$ may take the corrected trajectory into locations with very different values of $\mathbf{u}$, making the original gradient-wind trajectory irrelevant as a first approximation. For an accurate estimate, the correction must be applied at each time step, as was done for Figs. 4 and 5.

APPENDIX B

In any dynamical system described by time-symmetric equations, it is always *possible* to have an exactly reversed motion, for example a motion which would undo the pattern of contour deformations in Fig. 2, or patterns like those suggested by Figs. 3e and 10. The same applies to the problem of a gas let out of a bottle into a vacuum. The equations of motion for the colliding gas molecules are time-symmetric, and there is a dynamically possible motion in which the gas goes back into the bottle by itself. But in all such problems, the results of countless laboratory and computer experiments tell us that the reversed motion is very unlikely to be observed in practice.

The aim of statistical mechanics is to set up mathematical models which express this fact of experience. This is done, in the familiar way, by introducing an ensemble of dynamically possible motions chosen so as to be representative of those seen experimentally. The improbability of the reversed motions in examples like the foregoing is expressed, in the statistical-mechanical model, by the fact that very few members, if any, of this ensemble, consist of a realistic motion followed by its exact reverse within a time interval small enough to be of practical interest. Conversely, the probable, irreversible behaviour means the behaviour typical of most members of the chosen ensemble. This behaviour is constructed mathematically as the ensemble-averaged behaviour. It exhibits an arrow of time which comes not from the equations of motion, nor from the idea of considering an ensemble as such, but rather from the particular way in which the ensemble must be prescribed in order to be representative of real motions. The prescription is always found, in practice, to involve the imposition of a given ensemble of initial, rather than final, conditions, either tacitly or explicitly—an obvious time-asymmetry. For further discussion of this principle, as it is used in statistical mechanics, the reader is again recommended to consult, for example, PEIERLS (1979).

In the case of the fluid-dynamical irreversibility of present interest, another factor crucial to the construction of a suitable ensemble of dynamically possible motions is a correct representation of the advective nonlinearity in the equations of fluid motion, which is responsible for the potential-enstrophy 'cascade' and related phenomena. Particular examples like those of Figs. 2, 3 and 10 provide a sufficient illustration: any such fluid motion evidently involves the advective nonlinearity in an essential way. It is therefore clear that linearized dynamical equations, such as those commonly employed in linear models, or in zonally truncated 'wave-mean' models, of stratospheric Rossby waves, are incapable of correctly describing the complete wave-breaking process, even though the fields of motion predicted by linear theory (with invalidly large values of the wave amplitude substituted into the linear solution) may in some circumstances bear a superficial resemblance to the fields of motion occurring in dynamically possible wave-breaking scenarios. For instance it is known that such calculations can produce long, tongue-like features resembling those of Figs. 1 and 2, in some respects, especially when seen 'out of focus' (e.g. the height fields in HSU, 1981, figure 11, left-hand column); cf. end of Section 3 above. The real, nonlinear behaviour, including that of the Q field, and any associated potential-enstrophy cascades, cannot in general be correctly modelled in this way, if only because Q is not conserved by linear and 'wave-mean' models in the absence of diabatic and frictional effects. Notice how going to an approximation slightly closer to the nonlinear reality (right-hand column of Hsu's figure 11) completely changes the appearance of the tongue-like features in question.