Chapter 6

100 Years of Progress in Understanding the General Circulation of the Atmosphere

ISAAC M. HELD

NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

ABSTRACT

Some of the advances of the past century in our understanding of the general circulation of the atmosphere are described, starting with a brief summary of some of the key developments from the first half of the twentieth century, but with a primary focus on the period beginning with the midcentury breakthrough in baroclinic instability and quasigeostrophic dynamics. In addition to baroclinic instability, topics touched upon include the following: stationary wave theory, the role played by the two-layer model, scaling arguments for the eddy heat flux, the subtlety of large-scale eddy momentum fluxes, the Eliassen–Palm flux and the transformed Eulerian mean formulation, the structure of storm tracks, and the controls on the Hadley cell.

1. Introduction

In his monograph on the general circulation of the atmosphere, Lorenz (1967) includes an elegant chapter on the history of research on this topic from Hadley through Ferrel, Helmholtz, Vilhelm Bjerknes, Jacob Bjerknes, Jeffreys, Defant, and Rossby among many others. Lorenz mostly focuses on the period before the mid-twentieth-century advances in baroclinic instability and quasigeostrophic (QG) theory. It would be foolhardy to try to improve on Lorenz's coverage of the early twentieth century. After making contact with some key advances in this earlier period, this chapter will focus on the mid-twentieth-century turning point and the insights that emerged from these developments over the next half-century. Theories for large-scale transient eddy fluxes as well as the stationary eddies will be considered part of the problem of the general circulation. The focus is primarily on the extratropical troposphere, with some discussion of the Hadley cell and the intertropical convergence zone. Other essays in this collection, covering 100 years of advances in atmospheric science, focus on stratospheric dynamics, extratropical cyclones from a synoptic perspective, and numerical general circulation models of the climate, so these topics will be passed over or touched on very briefly here. But the subject is still vast enough that the emphases will necessarily be somewhat subjective, so apologies are in order to those with different perspectives on the important milestones in our field. The dominant storyline is less clear in recent years, with many active research paths being explored, and the limited references from this period are likely to be considered more idiosyncratic.

Needless to say, new observational systems and syntheses of observations paved the way for many of the key advances in our understanding of the general circulation. The ferment and progress in large-scale atmospheric dynamics around midcentury was undoubtedly catalyzed by the greater density of upper-air observation after WWII. Later breakthroughs of special importance to general circulation research included the advent of global satellite datasets and the creation of archives of the analyses used as initial conditions by weather prediction centers, which were followed by reanalyses obtained by passing the historical data stream through an up-to-date numerical atmospheric model (Kalnay et al. 1996; Gibson et al. 1997). The reanalyses, steadily improving over time, evolved into the prime repository of atmospheric circulation statistics. The important role played, before the advent of reanalyses, by the compilation of radiosonde-based circulation statistics in Oort and Rasmusson (1971), building on earlier work by V. Starr and colleagues, is also deserving of special mention.

Corresponding author: Isaac M. Held, isaac.held@noaa.gov

DOI: 10.1175/AMSMONOGRAPHS-D-18-0017.1

For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
2. Before the mid-twentieth-century breakthrough

Much of the early discussion of the general circulation of the atmosphere was focused on the explanation for the distribution of zonal mean surface winds and the three-cell tropospheric meridional overturning circulation. The explanation for the counterintuitive Ferrel cell served as a key goal in much of this work. It was generally understood that the angular momentum balance needed to be taken into account, in addition to the energy balance, to make progress on this problem.

An important constraint from angular momentum considerations, prominent in some of these discussions, is that the surface winds and the mean meridional circulation are tightly related because the dominant terms in the zonal momentum equation near the surface are the Coriolis force on the meridional flow, the pressure gradient, and drag associated with turbulent boundary layer stresses. With \( \mathcal{F} \) the turbulent stresses confined to the boundary layer, we have the following for the zonal momentum equation:

\[
\frac{\partial \rho u}{\partial t} = \rho f v - \frac{\partial p}{\partial x} - \frac{\partial \mathcal{F}}{\partial z}. \tag{6-1}
\]

Averaging in time, over longitude, and over the boundary layer, setting the surface stress equal to \( \mathcal{F}_s \), and using the fact that the zonal average of the pressure gradient vanishes,\(^1\) we have

\[
f \int_{\text{BL}} [\rho u] dz \approx -[\mathcal{F}_s], \tag{6-2}
\]

where square brackets refer to an average over time and longitude. As long as the stress exerted by the surface on the atmosphere is directed opposite to the mean surface winds, then in the region of surface westerlies the mass transport in the boundary layer will be poleward.

The growing appreciation that large-scale eddy fluxes of heat and momentum need to be at the forefront of any theory for the zonal mean circulation and thermal structure distinguished important early twentieth-century work. A key contribution was Jeffreys (1926), emphasizing the importance of meridional eddy momentum fluxes in the maintenance of the midlatitude surface westerlies, starting from the vertically integrated angular momentum balance in the presence of surface friction. The importance of eddies for this balance was later confirmed observationally (Starr and White 1951).

Perhaps most importantly, Jeffreys argued that the causal chain in explanations for the mean surface winds should start from the vertically integrated zonal angular momentum balance—not from the zeroth-order geostrophic balance between the Coriolis force due to the surface westerlies and the meridional pressure gradient. In Jeffreys’s argument, the pressure gradient between the subtropical highs and subpolar lows should be thought of as a consequence, not as a cause, of the surface westerlies.

When confronted with a diagnostic fact that A balances B, how does one justify a statement that A forces or controls B but that B does not force or control A? Historically, there have been disagreements about what the diagnostic balances between terms actually are in the atmosphere, but many of these have been settled by new observations. Often the more subtle questions and controversies have revolved around the causal implications of these balances. Theories for the time mean tropical flow (section 11) provide another example of this kind of question.

It was understood by midcentury that the observed large-scale angular momentum fluxes are primarily confined to the upper troposphere and do not contaminate the boundary layer linear frictional balance in Eq. (6-2). The momentum balance in the upper troposphere, with the zonal component of the Coriolis force on the upper branch of the Ferrel cell balancing the eddy momentum flux convergence, completes the simple low Rossby number picture of a Ferrel cell driven by the eddy momentum fluxes. Figure 6-1 portrays these balances schematically. The figure is from Eady (1950), an early depiction of this idea of an eddy momentum flux-driven Ferrel cell.

Unlike the subtle eddy momentum fluxes, the positive correlation between poleward flow and temperature is clear from surface observations. So the importance of midlatitude large-scale eddy heat transports was already evident to Dove in the nineteenth century, as discussed in Lorenz (1967), as well as to the Bergen School (Bjerknes 1919) early in the twentieth. An important milestone was provided by Defant (1921) in which turbulent diffusion was proposed as an appropriate model for poleward heat transport in the troposphere, and in which the diffusivity was estimated from observations. This concept of a macroturbulent diffusivity has had its proponents and detractors since its introduction by Defant. But despite its limitations it continues to provide a conceptual picture that motivates both theoretical research into geostrophic turbulence and the construction of simple energy balance models of the tropospheric climate.

Once one has accepted the importance of these large-scale eddies in both energy and angular momentum balances, one cannot go much further in the absence of a

\(^1\)Ignoring orographic drag for simplicity.
Theoretical picture for how the larger-scale environment determines the strength and shapes of eddies and how the eddies in turn shape the mean flow. Jeffreys did not have a cogent argument for why the eddy momentum fluxes have the latitudinal structure that they do, converging positive zonal angular momentum into mid-latitudes, thereby generating both the surface westerlies and the Ferrel cell. And Defant had no theory for his Austauch coefficient for the macroturbulent eddy diffusivity that controls the equator-to-pole temperature difference.

The synthesis by the Bergen School of surface observations of midlatitude cyclones and fronts into the polar front model has had lasting value for synoptic meteorology, while also providing ingredients for later theories of eddy dynamics. Building on earlier work, Bjerknes (1919) already had an appreciation that the potential energy stored in tilted isentropic surfaces was the energy source for cyclones, a starting point for thinking about eddy amplitudes. And the conclusion that frontal cyclones are not individual entities, but that a train of cyclones typically evolves from weaker wavelike perturbations (Bjerknes and Solberg 1922), was an important inspiration for instability theories with wavy underpinnings. Yet the polar front model did not in itself lead to a fully coherent theoretical picture.

The initial developers of classical hydrodynamics in the late nineteenth century had shown how circulation and vorticity, related by Stokes’s theorem, can provide remarkable insights into the evolution of fluids, especially homogeneous incompressible fluids. These ideas had been generalized so as to be more directly relevant to the atmosphere by V. Bjerknes2 by the start of the twentieth century and then by Rossby (1938) and Ertel.3 The result was a clear picture of the factors that can modify the circulation around material loops and the associated local conservation law for potential vorticity (PV). But QG theory struggled to emerge. For example, Rossby (1938) provided a classic study of the adjustment of an initially unbalanced flow to a geostrophically balanced state through the radiation of gravity waves, a calculation that critically involves the conservation of PV. Yet he did not push on to the next step of developing an equation for the slow evolution of a balanced flow based on PV conservation.

Rossby was, in fact, thinking about the slow evolution of large-scale waves, but focusing on the importance of the $\beta$ effect, the positive northward vorticity gradient produced by the solid-body rotation (Rossby 1939). He chose the simplest model that could capture this effect, the barotropic vorticity equation. Importantly, he also focused on a local dispersion relation using a Cartesian $\beta$-plane geometry, provided lasting insights into midlatitude vorticity dynamics. The famous Rossby stationary wavelength, following in a simple way from this.

---

2 See Thorpe et al. (2003) for an historical essay on Bjerknes’s circulation theorem.
3 See Schubert et al. (2004) for English translations of Ertel’s papers.
local dispersion relation, served as a point of departure for the developments to come in stationary wave theory.

That a change in sign of the vorticity gradient was a requirement for shear instabilities had been appreciated by Lord Rayleigh in the nineteenth century. This theory was extended to include the $\beta$ effect by Kuo (1949), a simple extension but an important one that clearly emphasized the crucial stabilizing effect of the monotonic vorticity distribution automatically provided by the rotating sphere. The ubiquity and importance of the $\beta$ effect would remain a central distinguishing feature of large-scale dynamical meteorology and oceanography.

3. Baroclinic instability and QG theory

The middle of the twentieth century was marked by the publication of the seminal papers on baroclinic instability theory by Charney (1947) and Eady (1949), both of whom utilized the QG approximation. The classic scaling arguments underlying the full nonlinear QG equations were simultaneously laid out by Charney (1948). The simplifications due to the QG approximation are profound, making many basic calculations analytically tractable. (It may be necessary to remind ourselves of the importance, in the precomputer era, of analytical tractability, or at least reduction to a calculation feasible on a mechanical calculator.) But more importantly in the long run, QG theory provides intuitive physical understanding not just of linear dynamics but also wave–mean flow interactions and fully nonlinear “‘geostrophic turbulence.’” It shows how the “vorticity thinking” familiar in studies of 2D flows can be generalized to “PV thinking” ([see Hoskins et al. (1985) for a review]. There obviously are important aspects of the general circulation for which QG theory is inadequate, such as the equatorial waveguide and the maintenance of the extratropical static stability. But QG theory continues to provide a core understanding of the circulation with which theories for these non-QG aspects must make contact.

There were many important antecedents to the work of Charney and Eady, that of Sutcliffe on cyclone development deserving special mention (Sutcliffe 1947). Classic QG theory can be thought of as consisting of approximations to vorticity and thermodynamic equations, the ageostrophic motion appearing only through 1) the effect on the vorticity evolution of stretching of planetary vorticity and 2) the effect on temperature evolution of the vertical advection of the base-state potential temperature. Eliminating the vertical motion one obtains a QG–PV equation involving only the geostrophic flow. One can also derive a diagnostic $\omega$ equation for the vertical motion. One can then equivalently solve for the evolution of the flow from the QG–PV equation in isolation (with the boundary condition provided by the surface temperature equation), or from the QG vorticity equation coupled with the $\omega$ equation, or, for that matter, from the QG thermodynamic equation coupled with the $\omega$ equation. Sutcliffe used the vorticity-$\omega$ formulation of QG. Hoskins (1999) describes Sutcliffe as deriving QG theory “‘on the back of a weather map’” (p. 26).

The remarkable similarity in the mathematics of baroclinic and barotropic instability became clearer through the papers of Charney and Stern (1962), Pedlosky (1964), and Bretherton (1966). It emerged unambiguously that baroclinic instability has essentially the same dynamics as the wavelike instabilities in barotropic shear flow, with instability associated with a change in sign of the gradient of PV. Importantly, Bretherton (1966) provided motivation for including the surface temperature field as fundamentally a part of the PV distribution.

The idea of baroclinic instability as a form of “‘slantwise convection’” analogous to gravitational instability, in which parcel trajectories lie within a “‘wedge of instability’” between horizontal and isentropic surfaces, is appealing from the point of view of energetics, and has had some traction historically. But the PV perspective highlights the fundamentally wavelike character of the instability that, unlike gravitational and inertial (symmetric) instability, cannot be easily understood from a parcel perspective in isolation. The PV perspective allows one to unify barotropic and baroclinic instability, to understand why internal jets in the stratosphere (for which surface temperature gradients play a negligible role) can be stable despite the presence of substantial mean available potential energy, and conversely to understand the fundamental role of the surface temperature gradient for tropospheric instabilities, providing the crux of our understanding of how tropospheric midlatitude eddies are generated.

Prior to the baroclinic instability breakthrough, the popularity of the polar front theory naturally encouraged a focus on the instabilities of fronts, for which small Richardson or large Rossby numbers are potentially relevant. The new QG theories focused instead on the instability of a smooth planetary-scale flow characterized by both large Richardson and small Rossby numbers. The implication is that the nonlinear evolution of instabilities of a smooth planetary-scale temperature gradient generates the frontal structures analyzed by the Bergen School. The definitive confirmation of this picture was provided later by Hoskins and Bretherton (1972) using the geostrophic momentum approximation, a
generalization of QG, in the idealized setting of the Eady instability, and then by Hoskins and West (1979) for an initial condition that generates a realistic configuration of cold and warm fronts.

4. Stationary waves

The analysis of models for stationary eddies in mid-latitudes followed rapidly after the initial Eady–Charney instability papers, illustrating the power of the new QG framework. The bold claim implicit in this work is that the basic structure of the large-scale deviations of the climate from zonal symmetry could be understood from the theory of linear stationary Rossby waves, in spite of the fact that the zonal mean flow clearly depends fundamentally on transient eddy fluxes.

In retrospect it is interesting how this picture—of linear stationary Rossby waves retaining their integrity while propagating (with their eastward group velocity) through a sea of nonlinear baroclinic disturbances—was rarely questioned. Wallace (1978) mentions the centrality of this assumption in a historical essay, arguing that the tendencies due to transient eddy fluxes are dominant on the relatively slow frictional and radiative relaxation time scales in the zonal mean balances, but they are less successful at competing on the shorter advective time scales of relevance to local budgets. In this way qualitative support can be provided for the idea of proceeding with the simplest linear stationary wave models. This history provides an interesting lesson as to the value of pushing ahead with simplified dynamical assumptions, often without any formal analysis of one’s approximations, and then, based on the results, assessing the appropriateness of this simplified picture as a starting point for further elaboration.

Charney and Eliassen (1949) and Bolin (1950) laid the foundation for the orographic problem using a linear QG shallow-water model. A quote from Bolin suggests the novelty of the idea that the climate of his native Scandinavia could be controlled in part by remotely forced waves:

... qualitatively we are justified in assuming that both the upper and surface high pressure ridge from the Azores to France in winter and the northerly position of the mean storm tracks over the eastern Atlantic resulting in relatively mild winters in northwestern Europe are not exclusively to be looked upon as the effect of the warm ocean but also as a result of the influence of the Rocky Mountains. (p. 195)

Smagorinsky (1953) used a continuously stratified QG model with no upper boundary to lay the foundation for the next generation of theories for thermally forced stationary waves. While the solution to this problem (as well as the topographic problem in this same semi-infinite setting) generally involves vertically propagating Rossby waves, these were hidden in Smagorinsky’s solutions in part because he assumed a linear shear with height, following Charney, a flow that produces evanescence rather than propagation at high enough altitudes. In retrospect, it would have been pedagogically useful to have first documented the solution to the QG thermally forced problem for the simpler case of a uniform mean flow with no vertical shear, in which vertically propagating waves are manifestly present. (Of course, there are many such cases where, in retrospect, one can imagine ways in which the progress of theory might have evolved more logically.) The concept of a vertically propagating Rossby, or “planetary,” wave was not fully appreciated until Charney and Drazin (1961). It is interesting to speculate how this concept, unknown to Rossby, could have been developed without the benefit of QG theory.

The observed stationary waves, in northern winter most clearly, have a large component with an equivalent barotropic structure, with little phase variation in the vertical but with maximum amplitude near the tropopause. This vertical structure is due to waves that are trapped within the troposphere. They are well described by linear theory which predicts this external mode vertical structure to be dominant far from the source and within the troposphere (Held et al. 1985), creating warm highs and cold lows that contrast with the warm lows and cold highs that are the dominant local response to extratropical heating.

QG is most naturally applied on a β plane rather than a full sphere, since the approximation breaks down in the tropics. As a result, QG models are often constructed in a midlatitude channel, bounded by vertical walls to the north and south. All of the original work by Charney, Eliassen, Bolin, and Smagorinsky on stationary waves, and the bulk of the work that followed in the next two decades, assumed this bounded channel geometry, typically assuming the simplest standing wave structure in latitude. This framework results in zonal propagation along latitude circles, with the propagation of the component trapped within the troposphere limited only by friction, resulting in exaggerated potential for resonance.

Hoskins and Karoly (1981), using a barotropic model and the full primitive equations on a sphere, convincingly demonstrated the importance of meridional propagation for stationary Rossby waves. With sufficiently smooth mean flow, the horizontal propagation of these waves tends to be oriented eastward along great circles and not latitude circles. Rather than the potential for resonance, focus turned to the fate of these waves as they propagate into the tropics (Dickinson 1968; Killworth and McIntyre 1985), where observations show
that they break and are at least partially absorbed (the observed poleward momentum flux in the stationary eddies being a signature of predominately equatorward propagation). Thermal forcing in the tropics generating waves that propagate into midlatitudes was the natural flip side to this new perspective. A tropical heating element was thereby added to the other factors (orography and midlatitude heating) maintaining extratropical stationary wave patterns. This work meshed nicely with the definitive isolation from observations, in this same time frame, of the wavelike equivalent barotropic pattern of the extratropical response to ENSO and of the teleconnections in low-frequency variability (Horel and Wallace 1981; Wallace and Gutzler 1981).

There followed a series of studies (Hoskins and Ambrizzi 1993; Branstator 2002) that focused attention back on the potential of sharp westerly jets to create waveguides that capture stationary Rossby waves and prevent meridional propagation. Stationary wave theory provides a good example of progress that starts with one idealization of the dynamics, and then considers distinct perspectives emerging from other idealizations, returning to a compromise position with a more complete understanding. Lorenz (2006), in a (tongue-in-cheek?) discussion of general circulation research as itself a chaotic dynamical system, refers to the damped oscillation as an underlying motif in the trajectory of this research.

Our understanding of large-scale stationary eddies, the deviation from zonal symmetry of the climate, has evolved from these linear theories in a number of ways, encompassing nonlinearity and interactions with transient eddies and diabatic processes. The importance of nonlinearity is illustrated by the problem of the extratropical response to ENSO variability. ENSO modifies the distribution of convective heating in the tropics, generating external Rossby waves that propagate into midlatitudes. After the initial linear studies with zonally generating external Rossby waves that propagate into midlatitudes, it became clear from linear studies with asymmetric backgrounds and from GCMs (Simmons et al. 1983; Geisler et al. 1985) that the boreal winter response to tropical heating could not be explained by linearizing about a zonally symmetric flow, since the largest extratropical response was found in the Pacific–North American sector, irrespective of the longitude of the tropical forcing. Simmons et al. (1983) related this behavior to the enhanced response when the tropically forced wave propagates through the strong jet exit region in the Pacific where the zonal flow weakens with increasing longitude. This dependence on the longitude of the heating can be thought of as the linear response given a zonally varying background state or, more fundamentally, as the nonlinear interaction between the tropically forced wave and the waves forced by midlatitude orography and heating that create the preexisting climatological asymmetries.

5. The two-level model

Immediately after the Charney–Eady baroclinic instability breakthrough, Phillips (1951) introduced the two-layer QG model, in part to help interpret the results of Charney’s analysis. The mathematics of Eady's model is relatively straightforward and intuitive, but with the inclusion of the β effect the details of Charney’s analysis is more opaque (at least to those of us not graced with intuition for hypergeometric functions). This two-layer model is directly relevant for the case of two ideal gas layers of differing potential temperature, but the troposphere’s continuous stratification bears no resemblance to the latter idealization. Rather, application to the atmosphere is typically defended by pointing to the observation that midlatitude eddies generally have minimal vertical structure in the streamfunction or geopotential, a structure describable to first approximation with two degrees of freedom in the vertical, one characterizing the lower-tropospheric flow and one characterizing the upper-tropospheric flow. Despite this rough empirical justification, in a seminal paper Phillips (1956) described the solution to a forced, dissipative fully nonlinear QG two-layer model, often called the first GCM, that has many realistic features, including baroclinic eddies of realistic amplitude producing plausible eddy heat and momentum fluxes and a three-cell mean meridional circulation. A model of interacting upper- and lower-tropospheric PV distributions in a β-plane channel produces a statistically steady state that, remarkably, captures much of the essence of the extratropical general circulation.

This QG model was quickly eclipsed by primitive equation simulations on the sphere that probed tropical as well and extratropical dynamics and explicitly simulated the static stability (in QG models, the background static stability $N^2$ is not simulated but is simply prescribed as part of the model formulation.) Because of this rapid development of more realistic GCMs, there was little time for the dynamics community to analyze the Phillips model and understand the parameter dependence of the resulting climate. Developing a fuller understanding of the turbulent, chaotic flows in the two-layer QG model remains, in fact, a challenging ongoing project (see Fig. 6-2) that aims to provide a natural foundation to our understanding of the general circulation in the simplest possible dynamical setting.

One particular feature of the two-layer model has played an important role in discussions of the maintenance
of the temperature gradients in the troposphere: the existence of a critical vertical shear required for baroclinic instability. This is the vertical shear required to change the sign of the lower-layer PV gradient, counterbalancing $\beta$. If the radiative forcing trying to increase the horizontal temperature gradient is relatively weak compared to the tendency of the eddy heat fluxes to reduce these gradients when the flow is even moderately unstable, the system would presumably hover close to the instability threshold.

A simple picture of baroclinic adjustment presents itself (Smagorinsky 1963; Stone 1978)—analogous to the familiar convective adjustment to neutrality for gravitational instability. This critical shear is proportional to $\beta$ and can be expressed as a critical isentropic slope required for instability. Impressively, this critical isentropic slope is close to that observed in the extratropical troposphere.\footnote{See, in particular, the appendix to Smagorinsky (1963), often overlooked among the other results in this famous work introducing a two-layer primitive equation model on the sphere.}

In Eady’s model, the growth rate of the most unstable mode is proportional to $f (dU/dz)N^2 \sim f/\sqrt{Ri}$, and so proportional to the vertical shear. This quantity is often referred to as the Eady growth rate and is frequently used as a guide to the location of the midlatitude storm tracks (e.g., Hoskins and Valdes 1990). There is no critical shear for instability in Eady’s model. If $\beta$ is set to zero in the two-layer model, the critical shear for instability also disappears, as expected. However, Charney’s model, despite the presence of nonzero $\beta$, does not have a critical shear for instability either, the maximum growth rate being essentially the same as in Eady’s model (the Eady growth rate could equally well be called the Charney growth rate). In the Charney problem the form of the instability depends on a parameter that can be expressed as the ratio of the isentropic slope to the critical isentropic slope in the two-layer model. When the isentropic slope drops below this critical value, the most unstable waves in the Charney model are shallow compared to the depth of the troposphere. The stability of the two-layer model for these subcritical isentropic slopes can be thought of as a consequence of the inability to resolve these shallow modes.

From the perspective of linear theory for a flow with uniform vertical shear and stratification, the critical shear for instability is an artifact. Therefore, one cannot justify a baroclinic adjustment perspective simply by appealing to the two-layer stability threshold. Justification requires an argument that supports the notion that this particular nondimensional measure of the isentropic slope has a strong effect on eddy dynamics and the resulting mean thermal structure. One possibility, giving the flavor for the kind of argument that might be possible, was offered in Held (1982), building on a related argument by Lindzen and Farrell (1980), pointing out that a troposphere supported by very shallow eddies is inconsistent if the height of the extratropical tropopause is constrained by the depth to which these eddies penetrate. Alternatively, if the extratropical tropopause is held up by other means (moist convection in particular) one can argue that the poleward heat flux drops rapidly as these eddies become shallower than the depth of the troposphere. Discussion of this topic has continued into recent years.

QG models with high vertical resolution have played a relatively small role in tropospheric general circulation modeling compared to two-layer QG models on the one hand and global primitive equation models on the other, despite their historically central role in studies of baroclinic instability. Continuously stratified QG theory requires that the isentropes that intersect the surface extend only a small distance into the interior; formally this distance is assumed to be infinitesimal. This assumption is fine.
for the analysis of linear instability but clearly is not a good approximation for the statistically steady state of our atmosphere. In the two-layer model, the surface-driven dynamics created by isentropes intersecting the surface is effectively spread through the lower of the two layers—the upper layer playing the role of the interior of the troposphere that is relatively uninterrupted by isentropes intersecting the surface. [One can think of the lower and upper layers as corresponding to the underworld and middlerow in the terminology of Hoskins (1991)]. There is no obvious advantage to using multilayer QG models in which interrupted isentropes are effectively assumed to be confined to the lowest, very shallow, model layer. One can move to more elaborate balanced models that avoid this unrealistic feature, but then the advantages of the conceptually and numerically simple QG framework over full primitive equation models tend to be lost. Adding in the fact that higher vertical resolution QG models do not help in simulating realistic frontal formation, static stability maintenance, or tropical dynamics, the dearth of papers studying the statistical steady state of such models becomes more understandable.

6. Scaling arguments for the poleward heat flux

The concluding paragraph of Eady (1949), addresses the difficulty of moving from a linear instability theory to a theory for the statistically steady state of the atmosphere, for which

... our techniques must resemble statistical mechanics. But we do not yet know enough about the “atoms” (the life histories of disturbances) nor are we concerned with “atoms” with a clear-cut individuality. (p. 52)

Fully satisfying statistical theories for midlatitude eddy structures and fluxes remain aspirational. But scaling theories for the poleward eddy heat flux have been proposed, based on ideas for how eddy equilibration is controlled, which may or may not involve turbulent cascades, as exemplified by the early work of Green (1970) and Stone (1972). One can think of these theories as providing expressions for a kinematic thermal diffusivity associated with the macroturbulence of the troposphere of the form $D \propto VL$, where $V$ and $L$ are characteristic eddy length and velocity scales. Even if no clear-cut scale separation between eddies and their larger-scale environment exists, the diffusivity can be thought of as a convenient way to express the dependence of the heat flux on eddy scales. In the large-$Ri$ limit of Stone’s theory, for example, the length scale is taken to be the radius of deformation $NH/f$, the zonal length scale of the most unstable mode in Eady’s model. The velocity scale $V$ chosen by Stone, $\partial U/\partial z$ where $U$ is the mean zonal wind, can be rationalized by assuming that eddy temperatures are determined by this mixing length and the mean temperature gradient and that eddy energy is roughly equipartitioned between available potential and kinetic energies.$^5$

Green’s approach was similar, but he assumed that the eddy mixing length was fixed by the width of the unstable region, independent of the radius of deformation. Both Green and Stone obtain diffusivities proportional to the zonal mean temperature gradient, and, therefore, heat fluxes proportional to the square of this gradient. But their dependencies on static stability and rotation rate are different, and neither theory has any explicit dependence on $\beta$. These scaling arguments provided intriguing starting points for rationalizing a kinematic diffusivity of the observed magnitude, $\approx 10^7$ m$^2$ s$^{-1}$, and, therefore, for the north–south temperature gradient when coupled to a model of radiative forcing of this gradient.

 Meanwhile, developments in the fluid dynamics of two-dimensional incompressible flows (Onsager 1949; Fjortoft 1953; Kraichnan 1967; Batchelor 1969) revealed the remarkable fact that, unlike the well-known 3D turbulent downscale energy cascade, the 2D energy cascade is upscale. Whether the energy cascade is upscale or downscale, the spectral shape is still expected to be proportional to $k^{-5/3}$, where $k$ is the horizontal wavenumber. But the implications of this spectral shape are very different in 2D and 3D: the kinetic energy is maximized at the forcing scale in 3D but at the scale at which the inverse cascade is halted in 2D. Salmon (1980) showed that an inverse cascade does occur in an idealized baroclinically unstable QG two-layer model, with the energy-containing eddies responsible for the bulk of the heat transport moving to scales much larger than the radius of deformation and losing their close connection to linear theory.

Rhines (1975) had earlier demonstrated how the presence of nonzero $\beta$ shapes the character of the final steady state of an inverse energy cascade. Because the phase speeds of Rossby waves increase with horizontal scale, as a given amount of energy cascades up the spectrum it reaches a scale (the Rhines scale) at which these phase speeds become comparable to characteristic eddy velocities. At this scale, the inverse cascade becomes very anisotropic, with energy flowing into zonal jets. In this regime the flow can be characterized as

\[ \frac{\partial^2}{\partial y^2} - (1/N^2) \frac{\partial^2}{\partial z^2} - \frac{\partial^2}{\partial y^2} (\partial B/\partial y) \sim (\partial U/\partial z)^2 H^2. \]

Using buoyancy $b$ rather than temperature to simplify the notation (with $B$ the mean buoyancy): $\vec{v}^2 - (1/N^2) \vec{B}^2 - (1/N^2)(NH/f)^2 (\partial B/\partial y)^2 \sim (\partial U/\partial z)^2 H^2$. 

$^5$ Using buoyancy $b$ rather than temperature to simplify the notation (with $B$ the mean buoyancy): $\vec{v}^2 - (1/N^2) \vec{B}^2 - (1/N^2)(NH/f)^2 (\partial B/\partial y)^2 \sim (\partial U/\partial z)^2 H^2$. 

Unauthenticated | Downloaded 03/11/24 02:53 PM UTC
consisting of nonlinear waves energizing jets, not a bad qualitative description of extratropical atmospheric flows. Scaling arguments based on the Rhines scale as a mixing length have been developed by Held and Larichev (1996), for example, and explain some aspects of idealized two-layer simulations in which a substantial inverse cascade occurs.

These arguments continue to be critically analyzed and their relevance for the atmosphere questioned. An important confounding factor was raised by the response to an increase in the radiatively forced meridional temperature gradient in some idealized GCMs, in which the static stability adjusts to prevent large changes in isentropic slope. The consequence is that the radius of deformation adjusts to the Rhines scale, effectively erasing evidence of an inverse cascade—behavior that cannot be studied in standard QG models (Schneider and Walker 2006).

Despite these complexities, results are accumulating from simulations with idealized and realistic GCMs that the Rhines scale plays the role of mixing length for poleward heat transport over a substantial range of parameters [see Barry et al. (2002) for a study with a realistically configured GCM]. One of the goals of the geostrophic turbulence perspective is to better understand this role of the Rhines scale as the eddy mixing length.

7. Momentum fluxes and life cycles

The initial baroclinic instability analyses, with zonal flow varying in the vertical but not in latitude, produce no momentum fluxes. Studies of more complex mean flows with latitudinal dependence followed eventually but had difficulty generating qualitative fits to the observed fluxes using exponentially growing normal modes. It seemed that momentum fluxes would be a more difficult nut to crack than the eddy heat fluxes, being tangential to the dominant energetics of the instability. This perspective is described by Eady (1950) in an essay on the general circulation that is remarkably modern in its outlook, in which he refers to the eddy momentum fluxes as a “secondary feature.”

The appropriate starting point in the context of general circulation theory was presented by Kuo (1951) with a barotropic model. The key was to start with a stable zonal flow, with a positive vorticity gradient dominated by $\beta$, and consider the evolution of a pre-existing disturbance, focusing on vorticity rather than momentum. If this disturbance is initially localized in midlatitudes, then as it spreads in latitude the result is convergence of angular momentum into the source region. A natural way of understanding this result is in terms of a conservation law for wave activity or pseudomomentum, as discussed briefly below, as was explicit in the much earlier work of Taylor (1915). Extending to a statistically steady picture, the argument is that plausible upper-tropospheric angular momentum fluxes are obtained if the effective source of these eddies (baroclinic instability) is located in midlatitudes, while the dissipation of the eddies is at least in part spread to lower or higher latitudes.

The idea that eddies created by stirring a rapidly rotating barotropic zonal flow on a sphere converge angular momentum into the stirred region is important, striking, and counterintuitive. But the history of the emergence of this insight is confusing. Taylor’s analysis was in the context of nonrotating vertically sheared flow in a nonstratified fluid and seems to have had no impact on Kuo’s or any other discussions of the general circulation until the 1970s. It is especially striking that Jeffreys, a prominent colleague of Taylor’s at Cambridge University, makes no reference in his classic discussion of eddy momentum fluxes to Taylor’s earlier insight into the value of focusing on the vorticity flux rather than momentum flux. Rhines (1977) cites Taylor in his encyclopedic article on ocean eddies, while Held (1975) cites Kuo as a starting point in a discussion of OG momentum fluxes. If the flow is sufficiently slowly varying to justify thinking in terms of a local dispersion relation, the same result can be obtained by working with the meridional Rossby wave group velocity, which is opposite in direction to the meridional flux of eastward momentum (Dickinson 1968; Thompson 1971). From this perspective the result appears to be dependent on the structure of a particular dispersion relation, whereas, in fact, it is best understood as a direct consequence of Kelvin’s circulation theorem that is not dependent on a slowly varying mean flow or, for that matter, on linearity (e.g., see Held 2000).

Kuo’s paper had a relatively modest impact on the general circulation literature in the following two decades, but its implications were clearly understood by Phillips, for example. After citing the Eady and Charney instability papers and Kuo’s study of barotropic decay on a stable flow, he suggested in Phillips (1954) that the kinematics of these theoretical motions are such that the disturbances one encounters on actual weather charts seem to be best described by a combination of the amplifying baroclinic wave and the damped barotropic wave. (p. 274)

---

One wonders how the history of general circulation theory would have been altered if Jeffreys and Taylor had actively collaborated on explaining the qualitative structure of large-scale eddy momentum fluxes.
The life cycle simulations of Simmons and Hoskins (1978) made this picture more explicit and helped it gain general acceptance. These authors followed the nonlinear development of baroclinically unstable modes in a primitive equation model on the sphere. The typical eddy life cycle could be described qualitatively as an asymmetric irreversible two-stage process: baroclinic growth followed by barotropic decay. Averaging over the life cycle, the eddy heat fluxes resemble those in the initial phase of the growing normal mode, while the horizontal momentum fluxes occur preferentially in the decaying phase, with less resemblance to the fluxes in the growth stage, consistent with Phillips’s description.

Studies of baroclinic eddy life cycles and eddy statistics in idealized GCMs have highlighted the sensitivity of the details of the upper-level wave breaking to factors such as the barotropic component of the zonal flow (Simmons and Hoskins 1980; Hartmann and Zuercher 1998) or the amplitude of the eddies (Orlanski 2003). These results presaged the difficulties that persist to this day in understanding the role of eddy momentum fluxes in climate variability and change.

An important case study is the tropospheric response to the ozone hole, which includes strengthening and poleward displacement of the surface westerlies in the Southern Hemisphere (Thompson and Solomon 2002; Gillett and Thompson 2003). The qualitative character of this response is robust and is captured in a hierarchy of turbulent climate models of different levels of complexity. But arguably there is as yet no consensus on the essence of the mechanism by which changes in stratospheric winds cause the response in eddy momentum fluxes required to support the modified surface winds. The same dynamical problem was introduced earlier in analyses of the wintertime response to volcanic aerosols (Kirchner et al. 1999) and, even earlier, in studies of the sensitivity of the tropospheric simulation in GCMs to biases in the stratospheric circulation (Boville 1984). Incorporating these stratospheric influences within a fully satisfactory understanding of the various mechanisms that exert control on upper-tropospheric wave breaking and the associated tropospheric eddy momentum fluxes is an important area of ongoing research.

Adding to this complexity, James and Gray (1986) and James (1987) describe a mechanism, referred to as the barotropic governor, by which momentum fluxes can feed back on eddy amplitudes through the surface winds. If momentum fluxes strengthen, for example, these will generate stronger surface winds, but these surface winds have a meridional structure, so if they strengthen then barotropic horizontal shears will strengthen throughout the troposphere even if we hold temperature gradients and thermal winds fixed. In addition to direct effects of these shears on upper-level wave breaking, they can interfere with baroclinic eddy production through the mechanism described in the papers cited above. The result is negative feedback that can potentially help control the equilibration of eddy amplitudes.

8. The E–P flux, TEM formulation, and isentropic coordinates

Andrews and McIntyre (1978) unified an array of earlier work on wave–mean flow interactions, especially that of Eliassen and Palm (1961), into a structure that has influenced a wide spectrum of atmospheric research. Much of the initial motivation came from stratospheric studies (Andrews and McIntyre 1976), but many aspects of this formalism are in wide use in analyses of the tropospheric general circulation as well. In the form in which it has primarily affected general circulation theory, it consists of two interrelated parts: a conservation law for “wave activity,” the flux of wave activity being the Eliassen–Palm (E–P) flux; and a reformulation of the zonal mean equations, referred to as the transformed Eulerian mean (TEM), in which the divergence of the Eliassen–Palm flux appears as a zonal force. A key additional feature is that for large-scale extratropical dynamics, the divergence of the E–P flux is well approximated by the northward eddy flux of QG potential vorticity.

The TEM formalism is closely related to the zonal mean equations in isentropic coordinates. The use of potential temperature as vertical coordinate has a long history in meteorology, with isentropic analyses of the zonal mean energy and mass balances emerging once sufficiently complete datasets became available, as described in Johnson (1989). When analyzing the circulation of mass in the meridional–vertical plane in isentropic coordinates, the interesting quantity is not the zonal mean of the meridional flow itself, \([\nu]\), but the meridional mass transport within each infinitesimal isentropic layer, a quantity that can be broken into mean and eddy contributions:

\[ [\nu H] = [\nu][H] + [\nu H']. \]  \hspace{1cm} (6-3)

Here \(H\) is the isentropic thickness, the mass per unit horizontal area within the layer in question. The time-averaged mass transport is observed to be poleward in upper-tropospheric layers and equatorward in the

---

1 For the zonal averages of interest here, wave activity is also referred to as “pseudo-momentum.”
2 In this section, we denote the zonal mean at fixed potential temperature by a square bracket and the deviation from this zonal mean by a prime.
colder layers near (or below) the mean surface temperature. It is the thickness weighting, needed to generate the meridional mass flux, that changes the midlatitude circulation from indirect to direct. In Eq. (6-3), \( [\nu H] \) is still equatorward while \( [\nu^* H^*] \) is poleward and larger in magnitude. It is in this sense that the Ferrel cell is said to disappear in isentropic coordinates as depicted in Fig. 6-3.

The zonal mean angular momentum balance also has a fundamentally different flavor when averaged over an isentropic layer. The key is that the upper and lower boundaries of this layer undulate so that the pressure varies on these surfaces. If this pressure is correlated with the east–west slope of one of these interfaces, a net zonal force—a form drag—is exerted by one layer on the other. The zonal momentum budget needs to account for the vertical derivative of this form drag—the difference between the form drag on the upper and lower surfaces of the layer in question. This difference also happens to be proportional to the Coriolis force associated with the eddy mass flux \( f[\nu^* H^*] \) due to the geostrophic flow within this layer. A growing baroclinic wave on a midlatitude zonal flow generates form drag that decelerates upper levels and accelerates lower levels, or, equivalently, creates eddy mass fluxes that redistribute mass within isentropic layers so as to reduce tropospheric isentropic slopes. The vertical transfer of momentum due to form drag is the essence of the mean flow modification due to baroclinic instability from this isentropic view. It provides an alternative perspective for how, in the time-averaged balance, angular momentum is transferred from the midlatitude region of upper-level convergence to the planetary boundary layer where it is lost to the drag on the surface westerlies (cf. Fig. 6-1).

Rhines and Holland (1979) provided a discussion of the importance of form drag in an oceanographic context in this same time frame, and it has remained at the forefront of discussions of ocean mesoscale eddy dynamics ever since. The fact that the ocean interior is nearly adiabatic has encouraged the prominence of this perspective in oceanography, more so than in meteorology where diabatic processes in the interior of the atmosphere are larger and have tended to discourage emphasis on an isentropic perspective at times.

The isentropic mass transport \([\nu H]\) is also referred to as the diabatic circulation, since it can only exist in the time mean if there are heat sources/sinks that create “vertical” motion in isentropic coordinates that balances the meridional convergence or divergence of mass within the layer.

The TEM reformulation of the momentum budget mimics this isentropic perspective without requiring an explicit transformation to isentropic coordinates. In the TEM equations, the E–P flux in the meridional–vertical plane encapsulates in a single quantity the large-scale eddy fluxes that force the zonal mean circulation. For large-scale dynamics, the vertical component of the E–P flux is proportional to the poleward eddy heat flux and the horizontal component to the eddy angular momentum flux. The divergence of the E–P flux appears as a zonal force, along with the Coriolis force acting on a residual meridional circulation analogous to the isentropic mass transport while the horizontal convergence of eddy heat flux disappears from the zonal mean.
thermodynamic equation, except at the surface. The dominant term in the interior thermodynamic equation, and the only term remaining in the QG approximation, is the vertical advection of potential temperature by the residual mean vertical motion, playing the role of the diabatic circulation in the isentropic framework.

The wave activity conservation law described by Andrews and McIntyre, with the E–P flux acting as the flux of wave activity, builds on the earlier work of Eliassen and Palm (1961), Charney and Stern (1962), and Dickinson (1969). Denoting the wave activity by $A$, the E–P flux, a vector in the meridional–vertical plane, by $F$, and the dissipation of wave activity by $D$, then this conservation law takes the following form:

$$\frac{\partial A}{\partial t} = -\nabla \cdot F - D. \quad (6-4)$$

The assumption here is that the eddies are produced by an internal instability rather than an external source. Wave activity is not created or destroyed on average by the growth of an instability because the wave activity is not positive definite when instability exists. The implication is that the E–P flux divergence appearing in the zonal mean momentum budget can be written in terms of the transient growth or decay and the dissipation of the eddies. In a statistically steady state, the E–P flux divergence can be expressed in terms of the dissipation alone. One can think of the dissipation as creating asymmetry between eddy growth and decay so that the decay phase does not undo the mean flow modification created during eddy growth. Within the troposphere, especially above the boundary layer, this dissipation is typically the end result of large-scale wave breaking.

Additionally, the direction of the E–P flux can be thought of as the direction of eddy propagation in the meridional–vertical plane, agreeing with simple group velocity concepts when the use of a local dispersion relation is justified. A poleward eddy heat flux (downward momentum transfer due to form drag) is a sign of upward propagation just as a poleward eddy momentum flux is a sign of equatorward propagation. As described in Edmon et al. (1980) the E–P fluxes and TEM momentum budget provide an appealing way of describing the Simmons–Hoskins life cycles (see Fig. 6-4) in terms of upward and equatorward meridional propagation resulting in an oppositely directed transfer of angular momentum from the subtropical upper troposphere to the midlatitude surface.

The fact that the divergence of the E–P flux in QG theory reduces to the meridional PV flux has additional interesting implications, since PV is itself conserved following the geostrophic flow in the OG limit. Considering a zonally symmetric climate, we expect a conserved tracer advected on a two-dimensional surface to produce a zonal mean flux down the mean meridional gradient on that surface. So a starting point for an eddy closure theory might be downgradient turbulent diffusion of PV. But one cannot diffuse PV arbitrarily since the meridional PV flux is itself the divergence of the E–P flux. Equivalently, one cannot diffuse PV without constraining the diffusivity so as to conserve angular momentum. In addition, the creation of mixing barriers by the strong PV gradients at the centers of westerly jets must be reflected in theories purporting to explain the distribution of upper-level mixing (McIntyre 1990; Ferrari and Nikurashin 2010). From these perspectives, it seems more natural for a closure theory to be based on a theory for the E–P flux itself (i.e., one based on a more quasi-linear rather than turbulent picture in which the propagation of waves is at the center of the theory). Green (1970) had a clear understanding of this dilemma of simultaneously conserving angular momentum and constraining the QG PV flux to be downgradient. Linear unstable modes know how to mix QG PV downgradient while conserving angular momentum, so Green used their structure as a guide to satisfying these constraints. But linear mode structures are biased, by definition, toward those that promote linear growth, which generally are not representative of a statistically steady state.

A possible alternative starting point for an understanding of baroclinic energy production is provided by “type B” cyclogenesis (Petterssen and Smeybe 1971). In this picture, waves in the upper troposphere induce flow near the surface that creates poleward heat flux and temporary growth when passing over surface temperature gradients. Temporary growth can be sufficient to create irreversible mixing and wave breaking, without the sustained phase locking between upper- and lower-layer disturbances that is a signature of type-A normal mode growth. Farrell and Ioannou (1993) provide an approach to making this kind of picture concrete, by perturbing a mean flow stochastically and adding sufficient damping to eliminate normal mode growth, while allowing the linear dynamics to shape the eddies and generate the E–P flux. These linear, stabilized, stochastic theories are enticing, but they are dependent on the availability of appropriate stochastic stirring and damping formulations.

In this spirit, a qualitative picture that reconciles diffusive theories focusing on the eddy heat flux and quasi-linear theories focusing on wave breaking in the upper troposphere might take the following form, thinking in terms of a two-layer troposphere. Starting with the eddy heat flux as given, being the vertical flux of wave activity
it serves as the source of upper-level waves. The theory for the upper-level PV flux can then be based on a quasi-linear picture for wave breaking and mixing of PV. The PV in the upper-level reservoir of wave activity could then induce low-level flows that drive the turbulent diffusion of the temperature field, closing the loop. There is a rough analogy here with the description by Salmon (1980) of how eddies equilibrate in the turbulent homogeneous two-layer model, with the barotropic mode playing the role of the reservoir of wave activity in that case.

9. Storm tracks

The zonal structure of eddy statistics in the atmosphere is difficult to study from surface data or radiosondes alone, even in the Northern Hemisphere extratropics, because of data deficiencies over the oceans. Therefore, research on storm-track structure was greatly accelerated when archives of analyses from the numerical prediction centers, the initial conditions used in forecasts, started to be utilized in general circulation research. These analyses provide dynamical consistency between different fields and extrapolate in a nontrivial way, through wave propagation and downstream development, from regions of high data coverage over land to regions of low data coverage over the oceans. In the following decades, storm-track research and other studies of regional circulations were invigorated by the development of the more homogeneous reanalysis products.

Fig. 6-4. (a)–(c) Snapshots of the E–P flux (arrows) and its divergence (contours) in the linear stage and then at two later times during the Simmons–Hoskins baroclinic life cycle. (d) The time average over the life cycle. Arrow lengths are normalized in the same way in (b), (c), and (d), while the contour interval is smaller in (d). [From Edmon et al. (1980).]
The first studies of this type (Blackmon et al. 1977; Lau 1978; Lau and Wallace 1979) made several choices that influenced future work. In particular, they departed from previous studies of storm counts and tracks using surface pressure maps (Peterssen 1956), focusing instead on Eulerian variances and covariances after subtracting a mean flow. Both approaches are important, but proportionally more research has focused on the variance/covariance rather than storm-tracking perspective since this pioneering work of Wallace, Blackmon, Lau, and others (although the tide may be turning once again toward storm-tracking analyses as more attention is focused on extreme events). Because the decomposition into a mean flow and eddies is the starting point for linear and quasi-linear theories, the Eulerian covariance perspective has proven to be more easily connected to theory—witness the difficulty in developing theories for tropical cyclone counts. A defense of the Eulerian variance perspective is provided by Wallace et al. (1988). Wernli and Schwierz (2006) provide an update of cyclone-track climatology from a modern reanalysis. A more seamless integration of these two perspectives remains an important goal.

Another significant choice made by Blackmon et al. (1977), and work that followed closely after, was to examine these variances and fluxes at different frequencies, focusing in particular on eddies with periods of 2–7 days typical of midlatitude disturbances. This time filtering is helpful for the sharp definition of the Atlantic and Pacific wintertime storm tracks, the lower-frequency, more barotropic variance having a different horizontal structure. The pattern of eddy fluxes—with the eddy momentum fluxes maximizing farther downstream than the eddy heat fluxes—also emerges more clearly from this time filtering. This zonal structure in the eddy fluxes recapitulates the barotropic growth/barotropic decay in the classic life cycles on zonally symmetric mean states, but with the evolution playing out in the zonal dimension as these eddies develop downstream with their eastward group velocities.

The importance of the fact that Rossby waves have a zonal group velocity that is eastward compared to their phase speed was discussed in a barotropic context by Rossby (1945) and documented clearly as ubiquitous in the atmosphere by Hovmöller (1949). Simmons and Hoskins (1979) placed this downstream development in the context of linear baroclinic instability theory, analyzing the evolution of disturbances initially localized in longitude, while Orlanski and Chang (1993) emphasized the profound difference in the energetics of cyclone growth in the presence of downstream development as compared to the baroclinic growth of a sinusoidal disturbance around a latitude circle.

A key theoretical concept is that of absolute as contrasted with convective instability.9 When an unstable flow is perturbed locally, the instability might grow locally or the instability might decay locally and only grow when viewed from a reference frame moving downstream. If there is local growth, the instability is referred to as absolute; if it only grows downstream it is referred to as convective. When the flow is convectively unstable in a reentrant geometry like that of the atmosphere, growth can only be maintained everywhere by propagating around the globe, a situation also referred to as global instability. The importance of this distinction, and the fact that baroclinic instability in the midlatitude storm tracks is almost exclusively global and not local, was emphasized by Pierrehumbert (1984). When global instability is dominant, a storm track cannot easily grow from local seeds but is dependent on disturbances entering from upstream. The implication is that upstream seeding is an important ingredient in theories of zonally asymmetric storm-track structure. For example, differences in the structure of the Atlantic and Pacific storm tracks can potentially be affected by the asymmetry of the seeding, with the Pacific storm track seeding the Atlantic storm track efficiently, but with the Atlantic returning the favor less vigorously.

Nakamura (1992) documented the striking midwinter suppression of the Pacific storm track in the Northern Hemisphere. Eddy amplitudes and fluxes are larger in November and March than in midwinter, despite the stronger meridional temperature gradient and Eady growth rates. This intriguing seasonal cycle is a warning against an overreliance on baroclinic growth rates in isolation as controlling storm-track structure. Among the multiple mechanisms suggested as potentially relevant to this behavior, one is that upstream seeding of the Pacific storm track is especially suppressed in midwinter. Chang et al. (2002) provides a review of this and other outstanding issues in our understanding of storm-track structure.

One can try to generalize the qualitative two-layer “type-II cyclogenesis picture” mentioned in section 8 to a zonally asymmetric climate. The quasi-linear upper layer can be thought of as having a capacity for holding wave activity that is a function of longitude and roughly proportional to the strength of the jet [see Swanson et al. (1997) for an analysis of this holding capacity in a barotropic model]. If the waves encounter a region of decreasing mean winds, they are more likely to slough off

---

9 This terminology was borrowed from plasma physics, and the term convective is unrelated to its use in describing gravitational instability.
some wave activity through wave breaking, creating momentum fluxes. The resulting zonal structure in upper-level eddy activity can be thought of as inducing lower-level flows that diffuse temperature down-gradient. The zonal structure in the upper-level activity and the lower-level temperature gradient then combine to create zonal structure in the eddy heat flux and, therefore, in the source of upper-level wave activity, closing the loop.

In this qualitative picture, the zonally asymmetric mean flow is taken as given. A variety of dynamical frameworks have been utilized to shed light on storm-track structure in this context of a prescribed time mean flow. These include linear normal model theory (Frederiksen 1983), stochastically forced linear models following the lead of Farrell and Ioannou (e.g., Zhang and Held 1999), and idealized fully nonlinear simulations in which the radiative forcing is manipulated to create a time mean flow with the observed structure (Chang 2006). These approaches simulate basic aspects of variances and fluxes, given the time mean flow, providing evidence that this way of describing the observations, removing the mean, and computing variances and covariances, has dynamical significance. In particular, several of these studies simulate aspects of the observed Pacific midwinter suppression.

But storm-track eddies also help sculpt the time mean zonal asymmetries. As emphasized by Shutts (1983), there is potential for strong wave–mean flow feedback in jet exit regions, with the momentum fluxes localized in these regions decelerating the jet and encouraging more wave breaking. Shutts uses a nondivergent barotropic framework; the interaction can be even more dramatic in a QG shallow-water model, with wave accumulation due to stagnation of eastward group velocities potentially playing a major role (Swanson 2000, 2002).

It is difficult to pursue the intricate issues underlying the interplay between localized storm tracks and the climatological asymmetries in the mean flow with strongly idealized models. Much recent and current research in this area utilizes GCMs, but forced with idealized boundary conditions, in which the mean flow and eddies are free to adjust to each other. Placing this work in the context of earlier research on both stationary waves and storm tracks remains challenging.

10. The Hadley cell

As implied in the schematic in Fig. 6-1, in his essay on the general circulation Eady (1950) did not single out the Hadley cell for separate treatment; the convergence of the large-scale angular momentum flux in the upper troposphere drives the Ferrel cell while, by a natural extension of this argument, implicit in Eady’s text, the associated divergence drives the Hadley and polar cells. The strength of the Hadley cell, in particular, is determined by the strength of the eddy momentum flux divergence due to eddies generated in midlatitudes. This low Rossby number momentum balance leaves no room for tropical heating to affect the strength of the Hadley cell except by altering these large-scale eddy stresses. But this is a very circuitous connection considering that these stresses are due to eddies generated in midlatitudes. In contrast, the founders of tropical meteorology generally argued for the tropical circulation as having an existence independent of eddies generated in midlatitudes. For example, Riehl (1969) concludes that Wallace (1978) describes a contentious exchange between Starr and Palmen on this issue.

A more intuitive description of the tropical upper-tropospheric momentum balance results when one breaks away from the low-Rossby number approximation by replacing

\[ f[v] \approx \frac{\partial [u' v']}{\partial y}, \]  

with

\[ (f + [\zeta]) [v] \approx \frac{\partial [u' v']}{\partial y}, \]  

where \( \zeta \) is the relative vorticity. In response to an increase in tropical heating, with fixed eddy stresses, the Hadley cell can increase in strength by decreasing the absolute vorticity or, equivalently, moving closer to an angular momentum conserving flow—a plausible outcome since the eddy stresses then have less time to act.

In contrast to the implications of the low Rossby number balance, the Hadley cell can survive in the limit of vanishing stresses if \( f + \zeta \) vanishes. Schneider (1977) and then Held and Hou (1980) (hereafter SHH) emphasized that in this limit, the Hadley cell can be constrained by the requirement of energy conservation and the need to maintain thermal wind balance. The resulting Hadley cell has a well-defined meridional extent which, for Earth-like parameters, places its terminus in the subtropics. In this limiting case, the circulation can also be thought of as existing so as to avoid inertial instability in the upper troposphere, a perspective emphasized by Plumb and Hou (1992) and Emanuel (1995) and one that provides a natural generalization to zonally asymmetric, especially monsoonal, circulations.

Direct overturning circulations spanning the domain are present in the well-known phase diagram of the
rotating annulus (Hide and Fowlis 1965); however, the domain and the role of boundary layers are such that these laboratory flows are not a close analog for Earth’s Hadley cell. Early work with GCMs added to this confusion. When these models are spun up from an isothermal state of rest, a Hadley cell that extends to the pole does initially form. This transient circulation exists so as to maintain thermal wind balance as radiation generates an equator-to-pole temperature gradient. As the spinup proceeds, this Hadley cell breaks up because of baroclinic instability. It is not uncommon to see this scenario misinterpreted as the instability of a steady axisymmetric flow that would exist in the absence of eddies. But if an axisymmetric model is run out to equilibrium, the equator-to-pole Hadley cell contracts close to its SHH limit if the upper troposphere is sufficiently inviscid.

Even if the Hadley cell has a natural horizontal extent in the angular-momentum conserving limit, this is unlikely to be the primary controlling factor in Earth’s atmosphere; the subtropical jet is then too strong compared to observations. Instead, we can think of the Hadley cell boundary as being determined by the onset of baroclinic instability. This has long been a large part of the conventional wisdom regarding the termination of the Hadley cell. For example, the comprehensive monograph by Palmén and Newton (1969) begins with the statement, the sense of which is attributed to Bjerknes et al. (1933), that

... a simple toroidal circulation with rising motion in low latitudes and sinking in high latitudes would, on the rotating earth, generate excessively large zonal wind speeds in the connecting meridional branches of the flow. The observed breakdown of the subtropical and higher latitude circulation into cyclonic and anticyclonic eddies was viewed as a condition necessary to avoid this difficulty. (p. 2)

The only thing missing here is that the Hadley cell can also be stopped by axisymmetric dynamics, as discussed above, so it is not self-evident but rather a quantitative question whether or not baroclinic instability limits the extent or not.

Once baroclinic instability theory became available, one could use the instability criterion for the Phillips model for example, or possibly the Eady growth rate compared to some frictional damping time, combined with the vertical shear implied by the angular-momentum conserving profile, to estimate this latitude of transition. Curiously, it is hard to find examples of this computation in the literature, perhaps because of doubts as to the validity of using linear instability theory in this way. If one barges ahead, using two-layer baroclinic adjustment as a guide, the schematic in Fig. 6-5a results for an Earth-like parameter setting.

These assumptions can be modified, for example by thinking in terms of a diffusivity that is a strong function of isentropic slope, or, as in Fig. 6-5b, considering the effects of eddy momentum fluxes in reducing the amplitude of the subtropical jet, using that angular momentum to spin up surface westerlies. An approach to the low-Rossby number regime in the tropics results if these stresses are large enough. The westerlies form where the eddy generation is strongest, poleward of the subtropical jet, the latter marking the onset rather than the center of the instability. The eddy-driven surface westerlies create an upper-level “eddy-driven jet,” even if the thermal wind is unchanged. Upper-level wave breaking, to the extent it avoids the jet center, can then
sharpen this eddy-driven jet (Robinson 2006). Whether the eddy driven and subtropical jets are distinctly separated or effectively merged can be of order-one importance when thinking of how the eddy storm track responds to warming (Lee and Kim 2003; Son and Lee 2005).

11. Moist dynamics of the general circulation

a. The tropics

A fundamental aspect of the tropical circulation was described in the classic Riehl and Malkus (1958) analysis of the tropical energy balance. Given observations of a relative minimum in the moist static energy in the midtroposphere, plus observed characteristics of tropical convection, they concluded that essentially all of the vertical motion in the Hadley cell occurs in the small fraction (they estimated about 0.1%) of the area covered by deep convective “hot towers.” In the simplest picture, there is weak subsidence between these actively convecting cores, the heating associated with this subsidence balanced by radiative cooling. In regions where the upward convective mass flux is larger than the subsidence, averaged over space and time scales of interest, there is “large-scale” upward motion. The skewness of the vertical motion makes the separation of the convection from its environment more natural than the separation between the spatial average over some region and the deviation from that average, the extreme skewness motivating “mass flux” convective parameterization schemes for GCMs, following Yanai et al. (1973) and Arakawa and Schubert (1974).

Based on a scaling argument for balanced tropical flows, Charney (1963) added another important element to this picture, that the different horizontal layers of the tropical atmosphere are very weakly coupled on low frequencies, except by the convection itself. The same scaling argument, based on the growth of the radius of deformation as one moves equatorward, implies that it is hard to maintain significant horizontal temperature gradients in the deep tropics, as long as geostrophic balance is relevant. The implications for tropical dynamics of these constraints were elaborated by Sobel and colleagues into a framework referred to as weak temperature gradient theory (Sobel and Bretherton 2000; Sobel et al. 2001).

Historically, many studies pointed to an understanding of the spatial distribution of deep convection as the central goal in theories of the tropical climatology, with the rest of the tropical state following from this distribution. In particular, the influential steady-state model of Gill (1980) demonstrated that realistic simulation of the zonally asymmetric component of the tropical flow, both its rotational and divergent components, could be obtained by simply linearizing the response to prescribed latent heating about a state of rest.

However, precipitation in the tropics is balanced to first approximation by low-level convergence of moisture rather than local evaporation. So moisture convergence, it seems, is controlled in large part by the same circulation that is driven (as in the Gill model) by the latent heating resulting from this moisture convergence. Should the resulting picture be one of strong positive feedback, or is it better to work with equations in which the zeroth-order balances are eliminated so as to more easily discern causality?

Along these lines, Neelin and Held (1987) suggested using the vertically averaged horizontal energy transport as a way to eliminate the dominant zeroth-order balances in the vertically averaged temperature and moisture equations. The ratio of the total energy transport to the mass transport in the divergent part of the flow, referred to as the “gross moist stability,” is a key ingredient in this picture on which, it is suggested, theory should be focused.

Emanuel et al. (1994) forcefully presented the more general perspective emerging from a large body of work, not just on the mean climatology but also on tropical cyclones and intraseasonal variability, which considers tropical convective heating as internally determined to enforce thermodynamic consistency and not in any useful sense as forcing the circulation. This approach has driven much recent research on tropical dynamics.

In the case of models with zonally symmetric climates, this complex of issues and approaches plays itself out in theories for the intertropical convergence zone (ITCZ). Research on the ITCZ has a long history that we do not try to summarize here, but see Sobel (2007). In a large part of this history the problem is phrased as developing a theory for tropical precipitation given the sea surface temperature (SST) distribution. But another branch of theory focuses on the energy balance of the atmosphere, concentrating, in particular, on the thermal equator at which the poleward energy transport in the Hadley cell changes sign. The latter perspective has been energized by studies of how Northern Hemisphere glaciation, or twentieth-century aerosol forcing, or low-frequency variations in the Atlantic meridional overturning circulation affect tropical precipitation (Chiang et al. 2003; Broccoli et al. 2006; Kang et al. 2009; Schneider et al. 2014). The SSTs in much of this work are thought of as internal to the system and at least in part responding to energy balance requirements. While one can often
relate the SSTs and precipitation in the final state, this may not be the best way to uncover the underlying causal relationships.

There is an interesting contrast between this energy balance perspective with the very different emphasis in the theory of Lindzen and Nigam (1987). This theory focuses on the horizontal structure of the SSTs, the associated boundary layer temperatures, and the winds driven by the pressure field in hydrostatic balance with these temperatures. It is hard to avoid a boundary layer contribution to low-level moisture convergence of this Lindzen–Nigam type, especially in regions of sharp SST gradients. Reconciling these different approaches is an active area of research. This may be another example of a research trajectory that has an oscillatory character that has not yet fully converged.

Recent research on the ITCZ provides a good example of systematic hierarchical modeling, in which one actively looks across models of differing complexity, rather than focusing on either the comprehensive or very idealized limit, to create a coherent dynamical picture—a promising approach to many issues in climate theory.

b. The extratropics

While moist processes may not exert as profound an effect on the extratropical as on the tropical general circulation, it is a role that is centrally important to an understanding of the effects of warming on the global climate. One way of stating the goal of this research is to create a dictionary that translates our understanding of dry theory to the moist precipitating case. Lorenz's extension of the concept of available potential energy to a moist atmosphere (Lorenz 1978) provides one valuable point of contact [see O’Gorman (2010) for an application of this concept to the analysis of storm-track responses to global warming]. Another is the inclusion of moist effects in an approximate way in baroclinic instability theory (e.g., Emanuel et al. 1987). Generalizations of Andrews–McIntyre wave–mean flow interaction theory to the case with precipitation, retaining as much as possible of its power to relate mean flow modification by waves to the sources and sinks of those waves, are potentially very important. But attempts in this direction, tracing back to Stone and Salustri (1984), remain a work in progress.

New diagnostic studies, such as analyses of mean mass transports within layers of different equivalent potential temperature, rather than potential temperature—see Fig. 6-6—are needed to guide theory. Interesting issues raised by this figure (see Pauluis et al. 2010) include the nearly uniform difference across all latitudes in the equivalent potential temperature between poleward and equatorward branches of the mass transport circulation, despite the differences in moisture content and meteorological regime between the tropics and extratropics—suggestive that there might be untapped underlying simplicity.

12. Exploring the space of planetary atmospheres

Much recent work on the general circulation can be described as trying to place the structure of Earth’s atmosphere in the context of other possible atmospheres, either idealized or realized in our solar system or, possibly, on extrasolar planets. An early work on classifying planetary atmospheres using dimensional analysis was Golitsyn (1970), focusing in particular on the key role of a thermal Rossby number. Recent discussions of the
effects of a wide set of parameter variations on the atmospheric circulation are provided by Schneider (2006) and Kaspi and Showman (2015).

Two remarkable qualitative features appear as one varies parameters starting from Earth-like settings. One is the development of multiple jets in each hemisphere, each with their own eddy activity, their own surface westerlies, and their own Ferrel cell. These multiple jet circulations are robust and can be created by simply increasing the rotation rate in a standard GCM (Williams and Holloway 1982) or in much more idealized QG systems (Panetta 1993), the key being the creation of a Rhines’s scale that is much smaller than the radius of the planet.

Another qualitatively different circulation that can be obtained by manipulating models, and that is common in planetary atmospheres, is equatorial superrotation, with strong westerlies at the equator in the upper troposphere. The routes to such a state are limited by the need for countergradient eddy angular momentum fluxes to maintain such a state (Hide 1969).

One route involves stirring of the tropics and the resulting generation of Rossby waves that escape from the tropics transporting angular momentum toward their source, the stirring typically being performed by convection (e.g., Liu and Schneider 2011). Another is the unstable interaction between extratropical Rossby waves and equatorially trapped Kelvin waves (Iga and Matsuda 2005). Superrotation is a generic characteristic of slowly rotating atmospheres (Mitchell and Vallis 2010) and the Kelvin–Rossby instability may help explain why this is so. Mechanisms accelerating the transport by the atmospheric heat engine. Nature, 415, 774–777, https://doi.org/10.1038/415774a.


Acknowledgments. The author acknowledges advice from Geoff Vallis on several issues covered in this historical review. Geoff’s collection of difficult-to-find classic papers, http://empslocal.ex.ac.uk/people/staff/gv219/classics.d/index.html, was also very useful. Reviews from Brian Hoskins, an anonymous reviewer, and other colleagues resulted in substantial improvements and, especially, better balance. Biases in the choice of which works to cite undoubtedly remain and are naturally the sole responsibility of the author.

REFERENCES


