Large scale inhomogeneities and mesoscale ocean waves: a single, stable wave field

by James C. McWilliams[1]

ABSTRACT

Horizontally propagating wave solution forms are assumed for mid-ocean mesoscale currents (i.e., those with spatial and temporal cycles of a few hundred kilometers and several months). The wave environment—defined to include the Coriolis parameter, bottom topography, and mean currents—is assumed to be inhomogeneous, but only on much larger scales. Mutual compatibility between these assumptions is derived. The wave numbers and frequency for the mesoscale motions evolve in a characteristic coordinate transformation defined by group velocity propagation; the wave amplitude is determined by a conservation of action density. Particular focus is placed upon the manner of rectified mesoscale forcing of large scale currents. The response consists of a combination of f/(depth) contour flow, nondispersive baroclinic wave propagation, and forced cross-contour currents. All three of these small amplitude laws remain valid at finite amplitude (i.e., a wave steepness of order unity or larger).

1. Modeling motivation

Mesoscale ocean currents are characterized by horizontal eddy diameters of a few hundred kilometers, periods of a few months, and vertical scales comparable to the ocean depth. In mid-ocean locations these currents typically are the ones with the greatest kinetic energy. In the theoretical description of ocean currents (as elsewhere), the concept of waves has had considerable utility, largely for its near uniqueness in yielding explicit solutions to the transient equations of motion. For example, there have been many studies of low frequency, quasigeostrophic waves (see the summaries of Platzman, 1968, and Lighthill, 1971). Furthermore, these wave theories have had empirical relevance. The presently completed mid-ocean, mesoscale experiments, Polygon and MODE, resolved little more than single cycles in space and time. On these scales, though, linearized waves were shown to be a successful description of much of the observed variability (McWilliams and Robinson, 1974; McWilliams and Flierl, 1975a). Clearly, however, this is a dynamically incomplete description: to address the generation, evolution, and decay of a mesoscale wave field requires a greater scope than a single cycle.

For the obvious advantage of mathematical simplicity, the wave environment

1. National Center for Atmospheric Research, P.O. Box 3000, Boulder, Colorado, 80303, U.S.A.

has usually been regarded as homogeneous. However, the ocean is an inhomogeneous medium. Currents in different locations are generally dissimilar due to variations of forcing, latitude, basin geometry, and, over a large enough separation, water mass characteristics. The present purpose is to extend the theory of mid-ocean waves to accommodate inhomogeneities of large scale in the wave environment– defined to include the Coriolis parameter, the bottom topography, and broad, slowly evolving currents (which will be called mean currents, implying an average over the shorter mesoscales). For the Coriolis parameter the modeling is accurate, since its spatial variability is only on a global scale. Bottom topography has variability at all scales; nevertheless, the greatest variance is associated with scales approaching those of the ocean basins (Lee and Kaula, 1967), which are generally much larger than mesoscale lengths. For the third environmental element, however, neither of these arguments applies: there are nonmesoscale motions on shorter as well as longer scales, and the velocity variance of the shorter is not generally the smaller of the two. The focus on the mean currents is thus somewhat arbitrary, but it is both mathematically compatible with the other environmental elements and permits a linking of mesoscale currents and the general circulation. Great effort has gone into observation of the latter, and surely it is of crucial importance in the climate regulating interaction of the ocean and atmosphere. It is with the expectation that the mesoscales do importantly contribute to the general circulation through eddy fluxes of heat and momentum that this model is formulated.

The ocean currents are assumed in the model to be incompressible, hydrostatic, and inviscid. The last assumption is probably appropriate for mesoscale currents over a few cycles, since no decay of synoptic features was observed during MODE or Polygon. It may also be qualitatively correct for much larger times: individual Gulf Stream current rings have been shown to exist for at least two years (Cheney and Richardson, 1974).

Observations indicate a simple vertical structure for mesoscale currents. Only two empirical eigenfunctions account for over 90% of the measured velocity and dynamic height variances during MODE (Davis, 1975), and these eigenfunctions are approximately combinations of the barotropic and first baroclinic, linearized, dynamical modes (McWilliams and Flierl, 1975b). The present model has a comparable number of vertical degrees of freedom: two immiscible layers of constant (but differing) densities. While this is adequate for a local description of mesoscale currents, its applicability for mean currents has weaker observational support. The MODE region in particular may have a mean velocity profile with much different structure (McWilliams, 1974). However, multiple-level numerical simulations of the general circulation conform approximately to a two mode description, whether prognostic (Bryan and Cox, 1967) or diagnostic (where observations fix the density structure: Holland and Hirschman, 1972).

The scale disparity between mesoscale currents and the subtropical gyres and

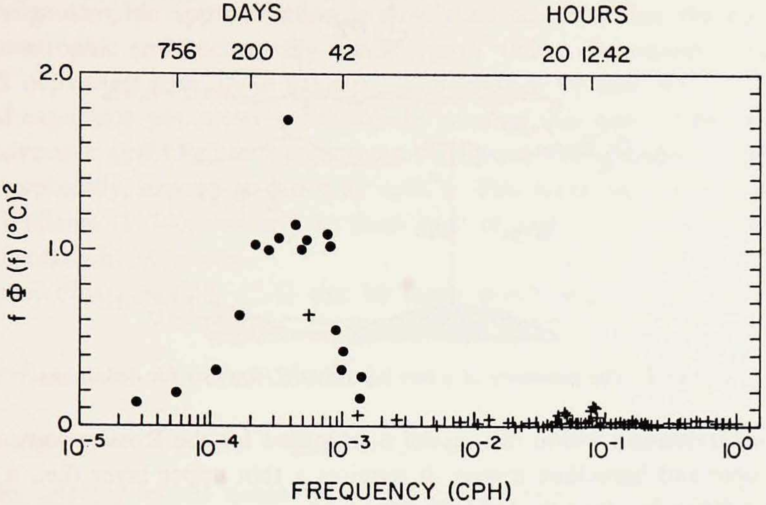


Figure 1. The power spectral density times frequency (such that area under the curve represents contribution to the variance) for temperature measurements taken near Bermuda (from Wunsch, 1972). There is a clear peak in the mesoscale regime (periods of 50-250 days).

ocean basins suggests using a method of multiple scale analysis, which depends upon a scale ratio ϵ being small ($\epsilon = l/L$, where l is characteristic of a mesoscale length and L of the environment). This partition, equivalent to the assumption of a scale gap between the mesoscale and the environment, is correct for the Coriolis parameter and asserted to be so for the topography and mean currents. Present observations cannot resolve the spectral distributions of the currents broader and slower than the mesoscales. While mesoscale eddies appear to be a scale-distinct phenomenon—they cause a peak in frequency spectra (Fig. 1) and probably also in velocity wave number spectra—it is not known whether interactions with currents slightly broader ($\epsilon \lesssim 1$) or much so ($\epsilon << 1$) are most important. The latter is assumed in what follows, but the consequences of relaxing this condition will also be discussed.

The structure of the paper is as follows. In §2 a two-parameter ordering of the equations of motion is derived; these parameters are the scale ratio ϵ defined above (which also serves as a Rossby number) and a measure, R_w, of the mesoscale current intensity. Before proceeding to derive the wave laws, we consider in §3 the equations for the mean currents, compatible with the other environmental elements, but without any rectified mesoscale forcing. In §4 the mesoscale solution form for a single, stable wave field in an inhomogeneous environment is introduced. §5 contains the infinitesimal amplitude wave laws (i.e., $R_w << \epsilon << 1$), and in §6 they are examined for finite amplitude ($R_w \gtrsim \epsilon$). In §7, for a simple environment, solutions are obtained for large scale currents driven by rectified mesoscale forcing. Finally,

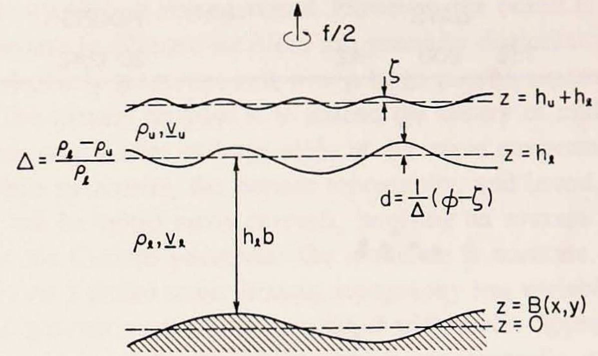


Figure 2. The geometry of a two layer model. See text for definitions.

in §8 a generalization within this model is proposed for the Rossby wave concepts of barotropic and baroclinic modes; it requires a thin upper layer (i.e., a thermocline depth much less than the total depth).

In recent years and other contexts, well dispersed waves in slowly varying inhomogeneous media have been much studied. Typical are the achievements of Benny (1971), Bretherton (1966), and Whitham (1965). Although explicit references will be sparse, this study is partly within the conceptual framework they have established, but the methodology is not wholly the same. The concern here will be with the particular physics of mesoscale waves rather than with wave generality.

2. A quasigeostrophic and nonlinear expansion

For depth independent, inviscid, hydrostatic, incompressible motions in two immiscible fluid layers with constant densities—as discussed above—the Navier Stokes equations in a rotating reference frame reduce to

$$\partial_t \mathbf{v}_u + (\mathbf{v}_u \cdot \nabla)\mathbf{v}_u + \hat{e}_z \times f\mathbf{v}_u + g\nabla\zeta = 0$$

$$\frac{1}{\Delta}\,\partial_t(\zeta(1+\Delta)-\phi) + \nabla \cdot [(h_u + \frac{1}{\Delta}\,(\zeta(1+\Delta)-\phi))\mathbf{v}_u] = 0$$

$$\partial_t \mathbf{v}_l + (\mathbf{v}_l \cdot \nabla)\mathbf{v}_l + \hat{e}_z \times f\mathbf{v}_l + g\nabla\phi = 0$$

$$\frac{1}{\Delta}\,\partial_t(\phi-\zeta) + \nabla \cdot [(h_l - B + \frac{1}{\Delta}\,(\phi-\zeta))\mathbf{v}_l] = 0\ . \tag{2.1}$$

The quantities $\mathbf{v}_u$ and $\mathbf{v}_l$ are the upper and lower layer velocities, f the Coriolis frequency (twice the rotation rate), g the gravitational acceleration, ζ and ϕ the upper and lower layer pressure heads, h_u and h_l the average layer thicknesses, $B(x,y)$ the level of the bottom, and Δ the positive relative density difference between the layers (see Fig. 2). The interface displacement d is defined by

$$d = \frac{\phi-\zeta}{\Delta}\ .$$

A quasigeostrophic approximation is simply an assertion that the currents are nearly geostrophic and horizontally nondivergent, with a consequent expansion in the small departures from these balances. This route is by now well-travelled, and the usual expansion parameter is the Rossby number, the ratio of the magnitudes of the advective and Coriolis forces (e.g., Phillips, 1963). One can alternatively, but isomorphically, expand in the scale ratio ϵ. This latter was done in Robinson and McWilliams (1974), and will be done here as well to emphasize the role of environmental inhomogeneity.

The preceding equations (2.1) can be made nondimensional by the following scales:

$$[f]\sim f_0 \qquad [x,y]\sim l \qquad [z]\sim h_l$$
$$[t]\sim(\epsilon f_0)^{-1} \qquad [\zeta,\phi]\sim\epsilon(f_0 l)^2/g \qquad [\mathbf{v}_u,\mathbf{v}_l]\sim\epsilon f_0 l \ . \tag{2.2}$$

The scales selected for the quantities on the lower line in (2.2) anticipate the geostrophic expansion in ϵ. The horizontal length l is characteristic of mesoscale motions; it has been observed to be comparable to the internal deformation radius

$$\frac{1}{f_0}\left(\frac{\Delta g h_u h_l}{h_u+h_l}\right)^{\frac{1}{2}} .$$

The time scale has been chosen consistent with waves whose restoring force is the environmental gradient (as, for example, in a Rossby wave whose acceleration is balanced by northward advection of the planetary vorticity f). The environment is assumed to have only broad scales of variation, characterized not by l but by $L = l/\epsilon$. Consequently, we define nondimensional environment coordinates, relative to mesoscale coordinates, by

$$[X,Y,T] = \epsilon[x,y,t] \ . \tag{2.3}$$

Thus, $f(Y)$, $B(X,Y)$, and the mean currents are functions only of X, Y, and T.

Nondimensional momentum and mass conservation equations from (2.1) become

$$\epsilon(\partial_t+\mathbf{v}_u\cdot\nabla)\mathbf{v}_u + \hat{e}_z\times f\,\mathbf{v}_u + \nabla\zeta = 0$$
$$\epsilon\gamma^2(\partial_t+\mathbf{v}_u\cdot\nabla)\,(\zeta(1+\Delta)-\phi) + (1+\epsilon\gamma^2(\zeta(1+\Delta)-\phi))\nabla\cdot\mathbf{v}_u = 0$$
$$\epsilon(\partial_t+\mathbf{v}_l\cdot\nabla)\mathbf{v}_l + \hat{e}_z\times f\,\mathbf{v}_l + \nabla\phi = 0$$
$$\epsilon\delta\gamma^2(\partial_t+\mathbf{v}_l\cdot\nabla)\,(\phi-\zeta) + \epsilon\mathbf{v}_l\cdot\nabla b + (b+\epsilon\delta\gamma^2(\phi-\zeta))\nabla\cdot\mathbf{v}_l = 0 \ . \tag{2.4}$$

The several nondimensional parameters and functions in (2.4) are defined by

$$\delta = h_u/h_l \ , \qquad \gamma = l f_0/(\Delta g h_u)^{\frac{1}{2}}$$
$$f(Y) = 1+\beta Y \ , \qquad \beta = L\,\mathrm{ctn}(\theta_0)/R$$
$$b = 1-B/h_l \ , \tag{2.5}$$

where θ_0 is the latitude and R the earth's radius. A convention has been introduced for the topographic term which will be followed throughout: all differential operators are O(1) for the functions to which they are applied. Thus, ∇b is interpreted as $\nabla_{(X,Y)} b(X,Y)$, while $\nabla\zeta$, treated as if the pressure varied only on the mesoscales, is $\nabla_{(x,y)}\zeta(x,y,t)$.

The equations (2.4) are to be systematically ordered in the small parameter ϵ. This is perhaps most simply done from potential vorticity equations, derived by eliminating the velocity divergence between the vorticity and mass conservation equations. For each layer the result is of the form

$$\epsilon h\left(\frac{D\chi}{Dt}+\beta v\right)-(f+\epsilon\chi)\frac{Dh}{Dt}=0\ , \tag{2.6}$$

where $\chi\ (= v_y - u_x)$ is the vorticity, $D/Dt(=\partial_t+\mathbf{v}\cdot\nabla)$ is the advective operator, and h is the instantaneous layer thickness,

$$\begin{aligned} h &= \delta + \epsilon\delta\gamma^2(\zeta(1+\Delta)-\phi) && \textit{upper} \\ h &= b + \epsilon\delta\gamma^2(\phi-\zeta)\ . && \textit{lower} \end{aligned} \tag{2.7}$$

Each of the preceding quantities is expanded in ϵ (e.g., $\mathbf{v}=\mathbf{v}_0+\epsilon\mathbf{v}_1+\epsilon^2\mathbf{v}_2+\ldots$). The momentum equations in (2.4) can be used to obtain $\mathbf{v}$ (hence χ and D/Dt) in terms of ζ and ϕ, h can be eliminated using (2.7), and the potential vorticity equations (2.6) then yield the governing balances for the pressures alone. Lengthy though the equations are, contributions through $O(\epsilon^2)$ will be required.

The pressures, however, must be partitioned into mean and mesoscale components:

$$\zeta\rightarrow\frac{R_M}{\epsilon}Z(X,Y,T)+R_w\zeta(x,y,t,X,Y,T),\qquad \phi\rightarrow\frac{R_M}{\epsilon}\Phi+R_w\phi\ . \tag{2.8}$$

The mean pressures depend upon only the environmental coordinates, while the mesoscale ones will in general depend upon both sets. The new parameters which appear are mean and mesoscale (wave) Strouhal numbers,

$$R_M=\frac{V_M}{f_0L\epsilon^2}\ ,\qquad R_w=\frac{V_w}{f_0L\epsilon^2}=\frac{V_w}{f_0l\epsilon}\ , \tag{2.9}$$

where V_M and V_w are particle speeds. Since the scale ratio ϵ measures the magnitude of particle accelerations relative to the geostrophic balance (i.e., it is a Rossby number), the Strouhal numbers are useful as measures of the ratio of advection to acceleration. The ϵ^{-1} factor for mean pressures in (2.8) allows $R_M=R_w$ to represent equal mean and mesoscale velocity magnitudes through geostrophy.

The insertion of (2.8) into the ϵ-ordered pressure equations from (2.6) yields a correct expansion of the differential operators acting on the various pressures. The pressures themselves must be expanded in ϵ (and, if appropriate, in R_M and R_w as well), but this will be deferred until §5. The resulting operators, explicit through $O(\epsilon^0)$, are the following:

$$\begin{aligned}
\textit{upper} \quad & R_M\left[\frac{H_u}{f^2}\beta Z_X-\gamma^2 D_u(Z(1+\Delta)-\phi)\right] \\
& + R_w\left[\frac{H_u}{f^2}D_u(\nabla^2\zeta)-\gamma^2 D_u(\zeta(1+\Delta)-\phi)-J\left(\zeta,\frac{H_u}{f}\right)\right] \\
& + R_w{}^2\left[\frac{H_u}{f^3}J(\zeta,\nabla^2\zeta)+\frac{\gamma^2}{f}J(\zeta,\phi)\right]=O(\epsilon) \\
\textit{lower} \quad & R_M\left[\frac{H_l}{f^2}\beta\phi_x-\delta\gamma^2 D_l(\Phi-Z)-\frac{1}{f}J(\Phi,b)\right] \\
& + R_w\left[\frac{H_l}{f^2}D_l(\nabla^2\phi)-\delta\gamma^2 D_l(\phi-\zeta)-J\left(\phi,\frac{H_l}{f}\right)\right] \\
& + R_w{}^2\left[\frac{H_l}{f^3}J(\phi,\nabla^2\phi)+\frac{\delta\gamma^2}{f}J(\phi,\zeta)\right]=O(\epsilon) \quad .
\end{aligned} \tag{2.10}$$

The large scale thicknesses of the layers are defined by

$$\delta H_u=\delta(1+\gamma^2 R_M(Z(1+\Delta)-\Phi)) \qquad H_l=b+\delta\gamma^2 R_M(\Phi-Z) \ ,$$

and the mean advective operators by[2]

$$D_u=\partial_T+\frac{R_M}{f}J(Z,\quad),\ D_l=\partial_T+\frac{R_M}{f}J(\Phi,\quad.) \tag{2.12}$$

In this paper, terms in the operator through $O(\epsilon^2)$ will be required; however, due to their complexity, they are not recorded here.

Because the purpose of this section has been to establish a useful ordering of the model, the appearance of Δ (of order 10^{-3} for the ocean) in the final equations should be commented upon. It represents the contribution to upper layer thickness changes due to upper surface displacements; setting Δ to zero where it appears explicitly in these equations is equivalent to imposing a rigid lid (with ζ becoming the pressure against the lid). The $O(\Delta)$ terms are almost never quantitatively important in what follows but are retained where they contribute to the formalism in a simple and understandable manner [e.g., the available potential energy of the free surface added to that of the interface in eq. (3.7)]. An exception occurs when both relative vorticity and interface displacements are small, such as for long (X,Y scale), barotropic ($\Phi=Z$) motions. There the $O(\Delta)$ terms imply a time scale of $f_0{}^{-1}L^2f_0{}^2/g(h_u+h_l)$, too fast to be characterized by the coordinate T (representing a scale $f_0{}^{-1}\epsilon^{-2}$). These motions, mixed gravity-Rossby waves with a scale near the external

2. Recall the convention described following (2.5): this operator should be re-interpreted as involving ∂_t, ∂_x, and ∂_y when applied to a function of (x,y,t).

deformation radius, are not consistent with the model. Whenever such behavior might arise in what follows, the O(Δ) terms are deleted.

3. Large scale currents

Prior to examining the mesoscale motions, it is of interest to examine one element of their environment in isolation: the large scale currents which are consistent with the preceding expansion in ϵ. We disregard for now any interaction with mesoscale waves and set $R_w = 0$.

The dominant balance comes from (2.10) at $O(\epsilon^0)$: strictly geostrophic potential vorticity is conserved following a particle path in each layer.

$$D_u\left(\frac{f}{H_u}\right) = 0, D_l\left(\frac{f}{H_l}\right) = 0 \ . \tag{3.1}$$

The balance is between planetary vorticity changes and vortex stretching due to layer thickness changes. Relative vorticity contributions would enter only at $O(\epsilon^2)$. This is a consequence of the large horizontal length scale L, and is sufficient to prevent the occurrence of classical barotropic instability (Kuo, 1949). Such instabilities as may arise on these broad currents must be fed from the vertical rather than the horizontal velocity shear (i.e., baroclinic instability).

The relations (3.1) are most often presented in their steady form. Then they require flow parallel to contours of f divided by the layer thickness, including the contributions from interfacial and free surface displacements. For weak mean currents [$R_M << 1$ in (2.11)], these displacements can be neglected, and the flow directions in the upper and lower layers are zonal and along f/b contours respectively. When is this steady flow limit appropriate? The ratio of the magnitudes of $\partial h/\partial T$ and the planetary vorticity advection βv (which should participate in a steady balance) is

$$\left(\frac{R\tan\theta_0}{f_0 L}\ \frac{1}{T_*}\right)\frac{(f_0 L)^2}{\Delta g h_u} \ ,$$

where T_* is a dimensional time scale. For L = 1000 km, an internal deformation radius of 50 km, and $\theta_0 = 30°$, this ratio can be small only for time scales much in excess of 230 days. For different L values, the critical T_* would vary in direct proportion. Thus, there is a large class of motions with T_* values greater than the mesoscale times of 10-40 days for which the limit of steady contour flow is incorrect.

The conservation laws (3.1) may be linearized when their Strouhal number is small. From the scaling of (2.2) and (2.3), $T_* = \epsilon^{-2} f_0^{-1}$, and (2.9) becomes

$$R_M = \frac{V_M T_*}{L} \ . \tag{3.2}$$

To avoid a trivial balance (i.e., arbitrary contour currents), only transient flows can be linearized. For a time scale T_* which permits the balance of $\partial h/\partial T$ and βv as above, $R_M \approx V_M/(5 \text{ cm sec}^{-1})$, which should be small for many mid-ocean circumstances. The linearized forms for (3.1) are

$$\gamma^2 \, \partial_T(Z-\Phi) - \beta/f^2 \partial_x Z = 0$$
$$\delta\gamma^2 \, \partial_T(\Phi - Z) + \frac{1}{f} J(\phi,b) - \beta b/f^2 \, \Phi_x = 0 \tag{3.3}$$

with error $O(R_M,\Delta,\epsilon^2)$. In each layer there is a time change in the vertical shear of pressure $(Z-\phi)$ due to flow across contours of either f or f/b. Solutions of (3.3) are propagating, baroclinic, planetary, topographic waves whose restoring forces are the cross-contour advections. The only complication in presenting explicit solutions arises from the nonconstant coefficients. For the special case of a flat bottom ($b = 1$), the general solution to (3.3) is

$$Z - G_1(Y,T) = \frac{1}{\delta} [G_1 + G_2(Y) - \Phi] = G_3 \left(X + \frac{\beta T}{f^2\gamma^2(1+\delta)} , Y \right), \tag{3.4}$$

where the G_i are the arbitrary functions of their arguments. Three phenomena are present in this solution: (1) barotropic, transient, zonal flow; (2) steady zonal currents, different in each layer; (3) westward propagation, with only meridional pattern distortion, of currents intensified above the thermocline (when $\delta<1$) and 180° out of phase between the layers. The phase speed is that of long, nondispersive, baroclinic Rossby waves. Barotropic Rossby waves have been suppressed by neglecting free surface displacements in the upper layer thickness (see §2).

The above solution illustrates the restrictions inherent in laws of the type (3.1). General boundary conditions cannot be imposed in X (along contours), but this is to be expected for first order differential equations. At the edges of the oceans, at least, other physical processes must be involved. Similarly, general initial conditions are not possible. From (3.1) we can derive the following synoptic constraint valid in particular at $T = 0$:

$$J\left(Z, \frac{\delta H_u}{f}\right) + J\left(\Phi, \frac{H_l}{f}\right) = J\left(Z, \frac{\delta}{f}\right) + J\left(\Phi, \frac{b}{f}\right) + O(R_M,\Delta) = 0 \ . \tag{3.5}$$

For steady solutions these terms vanish individually. These restrictions will be further explored in the context of mesoscale forcing of the mean flow. However, insofar as the preceding scale assumptions about the large scale, mid-ocean currents are correct, the restrictive laws of (3.1) must apply. But they cannot apply ubiquitously.

The first mean field correction to the conservation of geostrophic potential vorticity (3.1) occurs at $O(\epsilon^2)$. Consequently, we can expand the mean pressures in ϵ, without an $O(\epsilon)$ contribution,

$$Z = Z_0 + \epsilon^2 Z_2 + \ldots, \Phi = \Phi_0 + \epsilon^2 \Phi_2 + \ldots,$$

and, by continuing the operator expansion begun in (2.10), derive the following equations for Φ_2 and Z_2:

$$J\left(Z_2, \frac{H_u}{f}\right) + f\gamma^2 D_u\left(\frac{Z_2(1+\Delta)-\Phi_2}{f}\right) = \nabla \cdot \left(\frac{H_u}{f} D_u\left(\frac{\nabla Z_0}{f}\right)\right)$$

$$J\left(\Phi_2, \frac{H_l}{f}\right) + f\delta\gamma^2 D_l\left(\frac{\Phi_2 - Z_2}{f}\right) = \nabla \cdot \left(\frac{H_l}{f} D_l\left(\frac{\nabla \Phi_0}{f}\right)\right). \tag{3.6}$$

Thus, the expressions (2.11)-(2.12) and (3.1)-(3.5) should be interpreted as applying only to the leading order pressures, Z_0 and ϕ_0. The flux divergences in (3.6), which force the $O(\epsilon^2)$ pressures, represent a combination of depth-integrated relative vorticity changes and ageostrophic advections of geostrophic potential vorticity (f/H). Even for steady flows, the solutions can no longer simply be currents parallel to geostrophic potential vorticity contours; cross–contour flow is forced. Transient, linearized solutions for Z_2 and Φ_2 would satisfy a wave equation of the type (3.3) forced by the flux divergence in (3.6).

In summary, the large scale balances can be described as an expansion through $O(\epsilon^2)$ of the conservation of potential vorticity. Alternatively, a depth integrated mean energy balance to this order can be obtained by multiplying the layer potential vorticity equations by $\delta(Z_0 + \epsilon^2 Z_2)$ and $(\Phi_0 + \epsilon^2 \Phi_2)$ respectively. The result is

$$\frac{\partial}{\partial T}(E_{PI} + E_{PS} + \epsilon^2 E_{Ku} + \epsilon^2 E_{Kl}) = -\nabla \cdot [\mathbf{V}_u(P_u + \epsilon^2 E_{Ku}) + \mathbf{V}_l(P_l + \epsilon^2 E_{Kl})] \tag{3.7}$$

The energy density can change locally due to the divergence of the particle fluxes of pressure and energy. The available potential energies of the interface and free surface are

$$E_{PI} = \frac{1}{2}\delta\gamma^2 [(Z_0 - \Phi_0)^2 + 2\epsilon^2(Z_0 - \Phi_0)(Z_2 - \Phi_2)] ,$$

$$E_{PS} = \frac{1}{2}\delta\Delta\gamma^2 [Z_0^2 + 2\epsilon^2 Z_0 Z_2] ,$$

and, with a relative contribution in (3.7) only at $O(\epsilon^2)$, the depth integrated kinetic energies in the layers are

$$E_{Ku} = \frac{1}{2}\delta H_u |(\nabla Z_0)/f|^2,$$

$$E_{Kl} = \frac{1}{2} H_l |(\nabla \Phi_0)/f|^2$$

The particle velocities include an ageostrophic component at $O(\epsilon^2)$; for example,

$$f\mathbf{V}_u = \hat{e}_z \times \nabla(Z_0 + \epsilon^2 Z_2) - \epsilon^2 D_u[(\nabla Z_0)/f] \ .$$

The formulation (3.7) is similar to that for general fluid motions given by Landau and Lifshitz (1959). Notice, however, that in (3.7) there is no flux of available potential energy; that is because of our use of depth integrated pressures,

$$P_u = \delta H_u(Z_0 + \epsilon^2 Z_2) + \epsilon^2 \, \delta\gamma^2 R_M Z_0(Z_2(1 + \Delta) - \Phi_2)$$

$$P_l = H_l(\Phi_0 + \epsilon^2 \Phi_2) + \epsilon^2 \delta\gamma^2 R_M \Phi_0(\Phi_2 - Z_2) \ .$$

An area integral of the energy balance law (3.7) shows that the total energy of a region can change only by fluxes through its boundaries. But it is incorrect to identify them with geographical boundaries since the equations (3.1) and (3.6) cannot be expected to apply for an entire closed basin.

This completes a survey of free, large scale currents; they will be altered when mesoscale motions of sufficient intensity are present (§§ 5-7).

4. A single, stable wave field

The mesoscale solutions in an ϵ and R_w expansion must have multiple scales to accommodate the inhomogeneous environment, and, by our assertion, they must also exhibit wave-like behavior on the faster scales (x,y,t). A solution form with these characteristics is the WKB form,

$$\zeta \sim C(X,Y,T) e^{i\theta(X,Y,T)/\epsilon}. \tag{4.1}$$

C and θ are functions only of the slow variables; the fast coordinates do not explicitly appear. We call C the amplitude function and θ the phase function. To demonstrate the mesoscale variability, we expand the phase function in a Taylor series.

$$\zeta \sim C_0 e^{i\theta_0} \exp\{i[\nabla\theta_0 \cdot \mathbf{X}/\epsilon + \theta_{T_0} T/\epsilon + \ldots]\} \propto e^{i[\mathbf{k}.\mathbf{x} - \omega t]} + \ldots \tag{4.2}$$

The phase constant θ_0 can be incorporated into the amplitude, nonlinear phase terms are irrelevant over an O(1) mesoscale distance, and the linear terms make ζ resemble a mesoscale wave when $\nabla\theta$ and $-\theta_T$ are identified with the local wave number vector $\mathbf{k}$ and frequency ω.

The structure of the mesoscale solution in ϵ and R_w depends upon the local character of the waves, and a categorization will be adopted based upon this. Waves are usually defined as stable whenever ω is real for real $\mathbf{k}$. Since in the WKB form both are derivable from the phase function, the word stable in this context implies real θ. Furthermore, only the situation where a single wave is present, rather than a discrete spectrum, will be considered. Since both unstable and multiple wave fields can be solutions to the eqs. (2.10), however, these restrictions must be imposed.

This will be done by limiting the inquiry to initial conditions (Z, Φ, C, and θ at $T = 0$) consistent with the assumptions, and then following solutions only while they remain so.

5. Small amplitude waves

The solution form for the mesoscale wave field includes the terms

$$\zeta = \frac{1}{2} C e^{i\theta/\epsilon} + \ldots + (-i)$$
$$\phi = \frac{1}{2} C(A + \epsilon A_{10}) e^{i\theta/\epsilon} + \ldots + (-i) \ , \quad (5.1)$$

where $(-i)$ indicates the complex conjugate of the preceding. This solution is to be inserted in the operator expansion of §2. For small ϵ and R_w, the terms in (5.1) are sufficient to fully determine the dominant wave field (including the amplitude, the phase, and the ratio A of the lower and upper layer mesoscale velocities) and its forcing of the large scale currents. Although there exists higher order structure to the solution—and even contributions comparable in magnitude to the $O(\epsilon)$ term in (5.1)—its consideration can be postponed until the ϵ and R_w limits are relaxed.

a. Phase. When (5.1) is inserted into (2.10), the only contributions at $O(\epsilon^0 R_w^1)$ are from derivatives of the phase function. With a factor of $(-i/2)Ce^{i\theta/\epsilon}$, the layer equations which determine θ and A are

$$\left[\frac{H_u}{f^2} \nabla\theta \cdot \nabla\theta + \gamma^2(1+\Delta-A)\right] \quad D_u(\theta) + J\left(\theta, \frac{H_u}{f}\right) = 0 \quad (5.2a)$$

$$\left[A \frac{H_l}{f^2} \nabla\theta \cdot \nabla\theta + \delta\gamma^2(A-1)\right] \quad D_l(\theta) + AJ\left(\theta, \frac{H_l}{f}\right) = 0 \ . \quad (5.2b)$$

Only first spatial and time derivatives of θ are present; hence, with the identifications following (4.2), (5.2) can be interpreted as algebraic relations between frequency, wave number vector, and amplitude ratio—or, with the elimination of A between the pair of equations, as a dispersion relation. A more specific description is that, within each layer, the mean potential vorticity gradient $\nabla(H/f)$ balances the product of a Doppler frequency ($D(\theta)$) and an energy coefficient (see (5.11)). Except for minor notational differences and $\Delta \neq 0$, these laws are equivalent in their local (algebraic) interpretation to eqs. (2.8) of Robinson and McWilliams (1974). There, however, the focus was upon unstable solutions.

The range of behavior described by these equations is considerable and includes well studied examples. Rossby waves between a rigid lid and a flat bottom ($\Delta = Z = \Phi = 0$, $b = 1$) are described by

$$\textit{barotropic}: \quad \theta_T = \beta\theta_X/K^2; \; A = 1$$
$$\textit{baroclinic}: \quad \theta_T = \beta\theta_X/[f^2\gamma^2(1+\delta)+K^2]; \; A = -\delta \ , \quad (5.3)$$

where $K^2=\mathbf{k}^2=\nabla\theta\cdot\nabla\theta$. Another example is for constant Coriolis parameter and a sloping bottom. Then the solutions would be either steady flow parallel to topographic contours and with arbitrary vertical structure or a topographic wave (e.g., Rhines, 1970), concentrated in the bottom layer and described by

$$\theta_T = -J(\theta,b)\,\frac{K^2+\gamma^2}{K^2[K^2+\gamma^2(b+\delta)]}\,;\; A = 1 + \frac{K^2}{\gamma^2}\;. \tag{5.4}$$

However, an algebraic interpretation of these laws is incomplete: with $\mathbf{k}$ a parameter, there is an infinity of acceptable (ω,A) pairs in any particular environment. But a treatment of these as differential equations fully determines the evolution of θ and A once $\theta(X,Y,0)$ is initially prescribed. The above applies only for nontrivial, initial value problems; the external forcing of either the mesoscale or the mean currents from an idealized state of rest is not considered.

An economical manner of solving the phase equations (5.2) is in terms of characteristic coordinates. For any first order partial differential equation which we can write in the functional form

$$F(\theta_T,\nabla\theta,T,\mathbf{X}) = 0\;, \tag{5.5}$$

the solution is also a solution of the equivalent characteristic system (Courant and Hilbert, 1962) defined by

$$\begin{aligned}\frac{\partial\theta_T}{\partial s} &= -F_T/F_{\theta_T} & \frac{\partial\mathbf{X}}{\partial s} &= F_{\nabla\theta}/F_{\theta_T}\\ \frac{\partial\nabla\theta}{\partial s} &= -F_X/F_{\theta_T} & \frac{\partial\theta}{\partial s} &= \nabla\theta\cdot\frac{\partial}{\partial s}\mathbf{X} + \theta_T\;,\end{aligned} \tag{5.6}$$

where a transformation of the independent variables has been made from (X,Y,T) to (ξ,η,s). We define $s = T$, and ξ and η by the initial condition.

$$X(\xi,\eta,\, s=0) = \xi, \qquad Y(\xi,\eta,\, s=0) = \eta\;. \tag{5.7}$$

A further initial condition required for (5.6) is $\theta(\xi,\eta,0)$ [plus $\theta_T(\xi,\eta,0)$ and $A(\xi,\eta,0)$ such that (5.2) is satisfied]. A grid in ξ and η, initially uniform in X and Y, translates and distorts in time with the velocity $\partial\mathbf{X}/\partial s$. But since (5.5) is the dispersion relation, the pertinent velocity is

$$\frac{\partial\mathbf{X}}{\partial s} = -\frac{F_{\nabla\theta}}{F_{\theta_T}} = \frac{F_{\mathbf{k}}|_{X,T}}{F_{\omega}|_{X,T}} = \left.\frac{\partial\omega}{\partial\mathbf{k}}\right|_{X,T}\;, \tag{5.8}$$

which is the group velocity $\mathbf{c}_g$. The derivatives along characteristics can be written

$$\left.\frac{\partial}{\partial s}\right|_{\xi,\eta} = \left.\frac{\partial}{\partial T}\right|_{X,Y} + \left.\mathbf{c}_g\cdot\nabla\right|_{T}\;. \tag{5.9}$$

Thus the group velocity defines the moving reference frame within which changes in the frequency and wave number vector can be most simply expressed. The evolution of the phase function here conforms to the general pattern described by Lighthill (1965).

Any combination of the eqs. (5.2) is an acceptable choice for F. In general, then, F will depend on A as well. To preserve the simplicity of the characteristic system (5.6), we might make a choice which eliminated A between the layer equations. The same advantage, however, comes from a choice such that $\partial F/\partial A = 0$; for example, $\delta/D_u(\theta)$ times (5.2a) plus $A/D_l(\theta)$ times (5.2b) yields[3]

$$F \equiv \delta e_u + Ae_l + J\left(\theta, \frac{\delta H_u}{f}\right) / D_u(\theta) + A^2 J\left(\theta, \frac{H_l}{f}\right) / D_l(\theta) \,, \tag{5.10}$$

where "energy coefficients" for the two layers are defined by

$$e_u = \gamma^2(1-A) + \Delta\gamma^2 + \frac{H_u}{f^2}\nabla\theta\cdot\nabla\theta,\ e_l = \delta\gamma^2(A-1) + A\frac{H_l}{f^2}\nabla\theta\cdot\nabla\theta \tag{5.11}$$

To evaluate F and its derivatives, the amplitude ratio A can, for example, be algebraically calculated from

$$A = 1 + \Delta + \frac{H_u}{\gamma^2 f^2}\nabla\theta\cdot\nabla\theta + \frac{J\left(\theta, \frac{H_u}{f}\right)}{\gamma^2 D_u(\theta)} \,. \tag{5.12}$$

b. Amplitude. The wave amplitude function C is a passive factor at $O(\epsilon^0 R_w^1)$. For linearized waves ($R_w << \epsilon << 1$), it can be determined at $O(\epsilon^1 R_w^1)$ with contributions from both the expanded operator at that order plus first derivatives of C from the $O(\epsilon^0 R_w^1)$ operator (2.10). After collecting these terms and factoring $\frac{1}{2}e^{i\theta/\epsilon}$, the layer equations can be written

$$D_u(Ce_u) + J(C,H_u/f) + 2\frac{H_u}{f^2}D_u(\theta)\nabla\theta\cdot\nabla C + CD_u(\theta)\nabla\cdot\left(\frac{H_u}{f^2}\nabla\theta\right) \tag{5.13a}$$

$$+ Cfe_u D_u(1/f) - iC\,A_{10}\gamma^2 D_u(\theta) = 0$$

$$D(Ce_l) + J(CA,H_l/f) + 2\frac{H_l}{f^2}D_l(\theta)\nabla\theta\cdot\nabla CA + CAD_l(\theta)\nabla\cdot\left(\frac{H_l}{f^2}\nabla\theta\right) \tag{5.13b}$$

$$+ Cfe_l D_l(1/f) + iCA_{10}\left[\left(\delta\gamma^2 + \frac{H_l}{f^2}K^2\right)D_l(\theta) + J\left(\theta, \frac{H_l}{f}\right)\right] = 0 \,.$$

Both C and A_{10} must be determined here. The latter can be eliminated by forming the same linear combination of layer equations as in (5.10). When this is done,

3. This particular choice is such that $\frac{1}{2}C^2F$ could serve as the averaged Lagrangian in the Whitham (1965) formalism in order to derive the phase and amplitude laws of this section.

then by use of (3.1) and (5.2) the following law governing the wave amplitude can be derived:

$$\frac{\partial}{\partial t}\hat{A} + \nabla\cdot(\hat{A}\mathbf{c}_g) = 0 \ . \tag{5.14}$$

This implies the conservation of action density $\hat{A}$ in a volume whose boundaries translate with the group velocity. We define $\hat{A}$ by

$$\hat{A} \equiv \tfrac{1}{2}\, C F_{\theta_T} = \tfrac{1}{2}\, C^2 \left(\frac{\delta e_u}{D_u(\theta)} + \frac{A e_l}{D_l(\theta)}\right) . \tag{5.15}$$

Within the frame of reference defined by $\mathbf{c}_g$, $\hat{A}$ can change only because $\mathbf{c}_g$ diverges (n.b., $\mathbf{c}_g$ must be spatially variable). For the characteristic representation (5.6)-(5.9), however, this divergence can be related to the time change of the Jacobian of the coordinate transformation, $J\left(\equiv \dfrac{\partial(X,Y)}{\partial(\xi,\eta)}\right)$; namely,

$$\left.\frac{\partial \ln \hat{J}}{\partial s}\right|_{\xi,\eta} = \nabla\cdot\mathbf{c}_g|_T \tag{5.16}$$

Consequently, the law (5.14) describing the evolution of $\hat{A}$ in physical coordinates should be written in (ξ,η,s)–space as $\dfrac{\partial}{\partial s}(\hat{A}\hat{J}) = 0$.

This equation has the integral

$$\hat{A}(\xi,\eta,s) = \frac{\hat{A}(\xi,\eta,0)}{\hat{J}(\xi,\eta,s)} , \tag{5.17}$$

since $\hat{J}(s{=}0){=}1$ from (5.7). The action density can change along a characteristic only due to the stretching or contraction of the transformation. In addition to the initial condition for the phase equation, $C(X,Y,0)$ is also required to evaluate (5.17). If $C(X,Y,0)$ is zero outside of a finite region, then C will always be zero outside of the transformed image of the region. Thus, there is always a period of time within which lateral boundary effects can be ignored.

The action laws (5.14) and (5.17) are formally identical to previous presentations (e.g., Bretherton and Garrett, 1968). However, the definition of $\hat{A}$ in (5.15) is not quite the usual form. Commonly, the action is defined as the ratio of an averaged (over wave cycles) wave energy density and a frequency that is Doppler-shifted by the local mean current. As such, action is spatially a point-value quantity. However, $\hat{A}$ from (5.15) represents a vertically integrated property of a three-dimensional ocean, and has only horizontal spatial variability. Each of the two terms in $\hat{A}$ might individually be interpreted as energy-frequency ratios, if we identify $\tfrac{1}{2}\, C^2\delta e_u$ and $\tfrac{1}{2}\, C^2 A e_l$ with layer energy densities and $D_u(\theta)$ and $D_l(\theta)$ with shifted frequencies.

However, it is only the sum (vertical integral) of the individual layer ratios, not each separately, that is conserved following $\mathbf{c}_g$. Furthermore, the individual layer energy densities need not always be positive. If we form their sum, we obtain a satisfactory total energy density,

$$\tfrac{1}{2} C^2 (\delta e_u + A e_l) = \tfrac{1}{2} C^2 \left[\delta\gamma^2(1-A)^2 + \delta\Delta\gamma^2 + (\delta H_u + H_l A^2) \frac{K^2}{f^2} \right] .$$

It consists of available potential energies of the interface and free surface as well as depth-integrated kinetic energy. The separate energy coefficients in (5.11), though, do not have positive definite interfacial energy terms. Their sum does, as above, but in $\hat{A}$ this summation is prevented by any differences between the mean layer velocities (via the Doppler frequencies).

Nevertheless, physical interpretations of the wave field evolution can more familiarly be made in terms of amplitudes and energies. In terms of (5.15) and (5.17), it is perhaps helpful to categorize heuristically certain behavior in $\hat{A}$. The vanishing of $\hat{J}$ corresponds to concentrating the wave field into a vanishingly small region, with usually a singularity of the local energy density and wave amplitude (see the example in §7). An exception may occur at a critical layer [i.e., where either $D_u(\theta)$ or $D_l(\theta)$ vanishes]. There wave amplitude may remain finite while $\hat{A}$ becomes infinite; alternatively, vanishing C at a critical layer may allow finite $\hat{A}$ and non-zero $\hat{J}$. If the separate terms in (5.15) cancel each other—which would not be possible for an action proportional to a positive definite energy—then C, hence total wave energy, may become infinite unless the region of influence has greatly expanded (i.e., $\hat{J} \rightarrow \infty$). Whether any of these possibilities would occur depends upon the integration in time of the eqs. (5.6).

c. Rectification. Since the nonlinear terms of the original layer equations (2.1) are quadratic, so are those of the expanded eqs. (2.10) to leading order in ϵ and R_w. A quadratic product of oscillatory real functions should in general have a nonoscillatory component; in the present context, this would be a term with no mesoscale variability. This rectification represents a forcing of the large scale currents by the mesoscale wave-field, and may be of considerable importance in modeling the mean circulation.

In (2.10) a quadratic product first occurs at $O(\epsilon^0 R_w^2)$, i.e., second order in the wave amplitude. Rectification might therefore be expected at this order. However, all $O(\epsilon^0 R_w^2)$ operators include a Jacobian which, to leading order, operates only on the phase function in (5.1). The result is proportional to $J(\theta,\theta)$ and thus vanishes. Even at the next order in ϵ, the $O(\epsilon R_w^2)$ operators are either Jacobians or have a combination of even and odd derivatives of ζ and ϕ. These also do not rectify. Mesoscale rectification is postponed until second order in the scale ratio, $O(\epsilon^2 R_w^2)$. Contributions come from all three of the $O(\epsilon^n R_w^2)$ operators, where $n = 0, 1, 2$. Consequently, the result is exceedingly cumbersome. We define the rectification functions, B_u and B_l, so that they represent mean pressure forcing in the form

$$R_M\left[\frac{H^2}{f}\,D\left(\frac{f}{H}\right)+\mathrm{O}(\epsilon^2)\right]=\epsilon^2 R_w{}^2 B\ .$$

Then, with opportunistic condensation of the diverse contributions described above, we can derive the following formulas:

$$B_u=\tfrac{1}{2}\left\{J\left(\theta,\frac{|C|}{D_u(\theta)}\left[\frac{1}{f}\,J\left(|C|,\frac{H_u}{f}\right)+D_u\left(\frac{|C|e_u}{f}\right)\right]\right)+J\left(|C|,\frac{|C|e_u}{f}\right)+J\left(\frac{H_u}{f}\,,\,\frac{|C|^2|\nabla\theta|^2}{2f^2}\right)\right\} \tag{5.18a}$$

$$B_l=\tfrac{1}{2}\left\{J\left(\theta,\frac{|C|A}{D_l(\theta)}\left[\frac{1}{f}\,J\left(|C|A,\frac{H_l}{f}\right)+D_l\left(\frac{|C|e_l}{f}\right)\right]\right)+J\left(|C|A,\frac{|C|e_l}{f}\right)+J\left(\frac{H_l}{f}\,,\,\frac{|C|^2A^2|\nabla\theta|^2}{2f^2}\right)\right\} \tag{5.18b}$$

This is certainly a central result of this paper. Neither the order of magnitude nor the functional form of the rectification have any precedents to the author's knowledge.

The presence of mesoscale rectification alters the large scale energy eq. (3.7) to the following one:

$$\frac{\partial}{\partial T}(E_{PI}+E_{PS}+\epsilon^2E_{Ku}+\epsilon^2E_{Kl})+\nabla\boldsymbol{\cdot}[\mathbf{V}_u(P_u+\epsilon^2E_{Ku})+\mathbf{V}_l(P_l+\epsilon^2E_{Ku})] = -\frac{\epsilon^2R_w{}^2}{R_M}(\delta Z_0B_u+\Phi_0B_l)\ . \tag{5.19}$$

If the mean and mesoscale velocities are equal ($R_M = R_w$), then the rectification only forces a slight alteration to the mean currents: Z_2 and Φ_2, rather than Z_0 and Φ_0, absorb the wave forcing. On the other hand, in order that the rectified forcing contribute to the mean at the same order as mean planetary vorticity advection, we require

$$\frac{\epsilon^2R_w{}^2}{\beta R_M}=1 \tag{5.20a}$$

or, from (2.9),

$$\frac{V_M}{V_w}=\frac{R}{L\ \mathrm{ctn}\theta_0}\ \frac{V_w}{f_0L}\ . \tag{5.20b}$$

This latter is a small number; the mean flows are expected to be weaker than the waves which drive them. For values of $L = 500$ km, $V_w = 15\mathrm{cm}\ s^{-1}$ and $\theta_0 = 30°$, V_M would be .5 cm/sec. This value for the velocity magnitude is similar to those below the thermocline in the MODE region (Schmitz, Luyten, and Sturges, 1976). What is observationally unknown, however, is the mean length scale L (the above ratio varies with L^2).

The rectification functions in (5.18) are real and quadratic in the wave amplitude. The way in which terms are grouped does not permit a clear distinction between heat and vorticity flux divergences (i.e., $\nabla\cdot\overline{h'\mathbf{v}'}$ and $\nabla\cdot\overline{\chi'\mathbf{v}'}$); these transports are combined, for example, through the energy coefficients (5.11), which contain both thickness changes and relative vorticity terms. One can specify a *local interpretation*, incorporating many of the concepts of a homogeneous environment and uniform plane waves, by treating C, A, $\nabla\theta$, θ_T, β, ∇b, ∇Z, $\nabla\Phi$ as constants (with vanishing higher derivatives) and simultaneously setting $f=b=1$ and $Z=\Phi=0$. This procedure is the one which permits an interpretation of the phase equations (5.2) as a dispersion relation. In this sense, then, we note from (5.18) that most contributions to rectification are nonlocal. The wave forcing is predominantly an expression of the inhomogeneity of both the environment and the mesoscale wave field initial conditions. There is no inherent sign constraint on the rectification terms in either (5.18) or (5.19). They can act as either a source or sink of energy for the mean flow. Interpretation of their character is best done by specific examples, several of which are described below.

This completes the derivation of the important small amplitude results. To judge their oceanic relevance requires considering processes which might alter or negate them. Partly this involves the relaxation of the $R_w \to 0$ assumption, as in the next section.

6. Nonlinearity

We can estimate the wave Strouhal number R_w from observations. The value $R_w = .25$ is obtained for typical scales of $V_w = 10$ cm/sec, $l = 60$ km, $\theta_0 = 30°$, and $\epsilon = .1$. This is small, but not so small that wave nonlinearities can be generally neglected—particularly for mesoscale currents with relatively large speeds, short lengths, and long time scales.

A "single wave" solution structure, generally valid for small ϵ and finite R_w is the following one:

$$\zeta = \tfrac{1}{2}\, C e^{i\theta/\epsilon} + \epsilon/2\, \{[C_{10} + R_w{}^2 C_{12}] e^{i\theta/\epsilon} + R_w D e^{i2\theta/\epsilon}\}$$
$$+ \epsilon^2/2\, \{C_2 e^{i\theta/\epsilon} + R_w[D_{10} + R_w{}^2 D_{12}] e^{i2\theta/\epsilon} + R_w{}^2 E e^{i3\theta/\epsilon}\} + (-i) \qquad (6.1a)$$

$$\phi = \tfrac{1}{2}\, C A e^{i\theta/\epsilon} + \epsilon/2\, \{[C_{10}A + CA_{10} + R_w{}^2 C_{12} A] e^{i\theta/\epsilon} + R_w D A_D e^{i2\theta/\epsilon}\}$$
$$+ \epsilon^2/2\, \{[C_2 A + C_{10} A_{10} + C A_{20} + R_w{}^2 C_{12} A_{12}] e^{i\theta/\epsilon}\} \qquad (6.1b)$$
$$+ R_w[D_{10} A_{D10} + R_w{}^2 D_{12} A_{D12}] e^{i2\theta/\epsilon} + R_w{}^2 E A_E e^{i3\theta/\epsilon}\} + (-i)\ .$$

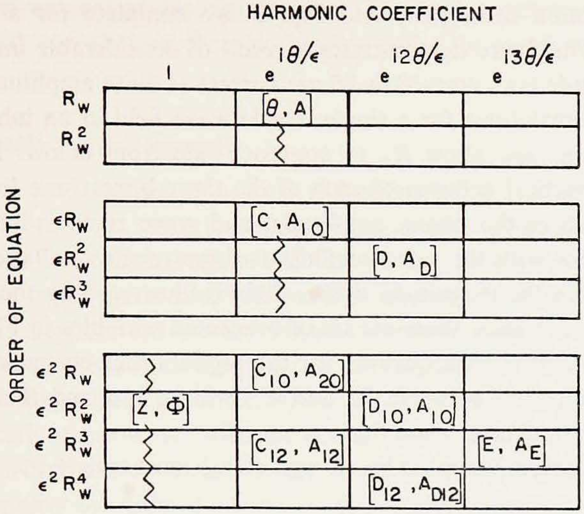


Figure 3. The structure of nontrivial dynamical balances in an ϵ and R_w expansion for a single stable wave field. The boxes are identified by their order of magnitude and coefficient of mesoscale oscillation.

This expansion form is the completion of (5.1). Amplitude functions for the fundamental, second, and third harmonics are labeled by C, D, and E; the corresponding amplitude ratios are A, A_D, and A_E. All quantities in (6.1), except C_2 (see below), can be determined by the expanded equations through $O(\epsilon^2)$. The amplitudes and phase functions are all functions of only the environmental scale coordinates. The method of analysis is to insert (6.1) into the operators (2.10) and their continuation through $O(\epsilon^2)$, and then set to zero each coefficient of distinct powers of ϵ and R_w and $e^{i\theta/\epsilon}$.

A matrix of the resulting equations is shown in Fig. 3. Bracketed entries indicate which of the quantities defined in (6.1) are determined by the equations at a particular order in ϵ and R_w with a particular coefficient of mesoscale oscillation. Also, for the rectification at $O(\epsilon^2 R_w^2)$, we have indicated that the wave-driven mean currents are determined (see 5.18).

With reference to (6.1), we see that to leading order in ϵ only a single wave is present. Systematic corrections to C and A occur at higher orders in ϵ, as do higher harmonics of the fundamental wave: the second harmonic arises with nondimensional pressure magnitude $O(\epsilon R_w^2)$ and the third with magnitude $O(\epsilon^2 R_w^3)$. There are no higher harmonics through $O(\epsilon^2)$. Thus, there is a postponement of the nonlinear higher harmonics to higher orders in ϵ in a manner analogous to that for rectification. Both postponements are due to the structure of quasigeostrophic advection (a Jacobian operator to leading order in ϵ) and the single wave field assumption.

Both the solution form (6.1) and Fig. 3 are complete for all powers of R_w through $O(\epsilon^2)$. The figure demonstrates a result of considerable importance. All of the small amplitude laws derived in §5 are correct at finite amplitude as well. There is no nonlinear breakdown for a single, stable wave field in an inhomogeneous environment. To see this, allow R_w to approach one from below. In Fig. 3 this is equivalent to a vertical collapse of each of the three boxes (one for each power of ϵ). Then for each of the phase, amplitude, and wave rectification laws there are no other equations with the same coefficient of mesoscale oscillation and power of ϵ—no matter what the magnitude of R_w. This is illustrated by the zagging vertical lines not intersecting more than one set of bracketed variables in Fig. 3. There will, of course, be nonlinear consequences for the higher order corrections in (6.1). For example, the distinction between C_{10} and C_{12} will be lost and their governing balance will become nonlinear when R_w is not small.

The second harmonic amplitudes (D and DA_D) must satisfy the following pair of layer equations:

$$D\left[\frac{3H_u|\nabla\theta|^2}{\gamma^2 f^2} + A\right] - DA_D = -\frac{C^2}{4\gamma^2}\left[\frac{J(\theta,e_u/f)}{D_u(\theta)} + \frac{3e_u|\nabla\theta|^2}{f^2}\right]$$

$$-\delta D + DA_D\left[\frac{3H_l|\nabla\theta|^2}{\gamma^2 f^2} + \frac{\delta}{A}\right] = -\frac{C^2}{4\gamma^2}\left[\frac{J(\theta,e_l/Af)}{D_l(\theta)}A + \frac{3Ae_l|\nabla\theta|^2}{f^2}\right] \tag{6.2}$$

The left-hand sides are related to the phase laws (5.2). If $(2\theta, A)$ as well as (θ,A) were a solution to those equations, then (6.2) would be an indeterminate system for (D, A_D). This could be described as a self-resonance. It is unlikely to occur in general since (5.2) yields a nonlinear dispersion relation, but, even if it did, it could have no important consequences in distances or times of the order of the environmental scales whenever R_w is small. The forcing of the second harmonic amplitudes is simply proportional to C^2—no amplitude derivatives enter. Most of the contributions are local (in the sense of §5), but there are nonlocal ones as well. These equations are algebraic rather than differential.

Any amplitude function in (6.1) which is a coefficient of $e^{i\theta/\epsilon}$ and $Ae^{i\theta/\epsilon}$ in the upper and lower layer pressures respectively cannot be determined except at an order higher in ϵ than it first appears. We saw this in §5, where the $O(R_w)$ amplitude C was solved for at $O(\epsilon R_w)$. Similarly, C_{10} and C_{12} cannot be obtained until $O(\epsilon^2 R_w)$ and $O(\epsilon^2 R_w{}^3)$ respectively. Also C_2, which includes terms of several orders in R_w, is indeterminate through $O(\epsilon^2)$. To complete the description of the solution elements in (6.1), we remark that, just as the second harmonic amplitude D is forced by the quadratic interactions of the dominant wave C, so also are D_{10} and D_{12} forced by the interaction of C with C_{10} and C_{12}.

The fundamental feature of this fully nonlinear solution is its equivalence to that for a linearized wave in its important characteristics. The smallness of the ratio of mesoscale to environment scales assures this. The only requirement on wave intensity is the mild one that its Rossby number (ϵR_w) be small enough to permit quasigeostrophy.

7. An (almost) uniform environment

The interpretation in terms of a local dispersion relation defined in §5 is an analog of perfectly periodic plane waves in a homogeneous environment. This does not imply a trivial environment ($\beta = Z = \Phi = B = 0$), for then θ_T would vanish (see 5.2), and no wave propagation could occur. Non-zero gradients of f, $B(X,Y)$, Z, or Φ are required to support the waves; however, periodic plane wave solutions are consistent only when these functions for the environment are replaced by average values and when derivatives of order higher than first are ignored. As has been shown, these locally plane-wave solutions are modified on the larger scales by the true variability of both the environment and the initial wave field [e.g., nonuniform $C(\xi,\eta,0)$]. To simplify the laws of §5, we make the approximation that the shortest scale nonuniformities in the evolution of the wave field are caused by initial conditions of the wave field rather than the environment. In other words, we posit yet another scale L', characterizing the scale of wave amplitude variations, and such that $L >> L' >> l$.

To derive the laws analogous to the more general preceding ones, we define intermediate scale coordinates

$$(X',Y',T') = \epsilon'(x,y,t);\ \epsilon << \epsilon' << 1\ . \tag{7.1}$$

The total pressure solution forms—replacing (2.8) and (5.1)—are given by

$$\begin{aligned}\zeta &\rightarrow \frac{R_M}{\epsilon} Z\,(X,Y,T) + \frac{R_M'}{\epsilon'} Z'\,(X',Y',T') + \tfrac{1}{2} R_w[C(X',Y',T')e^{i\theta(X',Y',T')/\epsilon} + (-i)] \\ \phi &\rightarrow \frac{R_M}{\epsilon} \phi\,(X,Y,T) + \frac{R_M'}{\epsilon'} \phi' + \tfrac{1}{2} R_w\,[C(A + \epsilon A_{10})e^{i\theta/\epsilon} + (-i)]\ ,\end{aligned} \tag{7.2}$$

with errors $O(\epsilon R_M,\ \epsilon' R_M',\ \epsilon' R_w,\ \epsilon R_w)$. This solution represents (1) a very broad flow, (Z, Φ), which exists independent of the mesoscale [i.e., R_M is larger than the value specified by 5.20b]; (2) a weaker, shorter scale mean flow, (Z', Φ'), driven by wave rectification at $O(\epsilon'^2 R_w^2)$; and (3) a single, stable wave field with nonuniformities[4] of scale L'. The magnitude of R_M' in order to balance wave forcing is as in Eq. (5.20b) with $V_M \rightarrow V_M'$ and $L \rightarrow L'$.

When the solution (7.2) is inserted into equations expanded as in §2, a crucial simplification results. The parameter ϵ remains the relevant quasigeostrophic expansion parameter (n.b., initial wave field inhomogeneities cannot support propaga-

4. Inevitably the environment imposes nonuniformities of scale L. However, if $C(\xi',\eta',0)$ vanishes outside a region $O(1)$ in (ξ',η'), then the largest scale inhomogeneities can be ignored for a time $O(1)$ in T'.

tion). Thus, for $\epsilon << \epsilon'$, only the operator (2.10) is relevant through $O(\epsilon'^2)$; no operators from $O(\epsilon)$ or higher can contribute. The action law at $O(\epsilon' R_w)$ and rectification at $O(\epsilon'^2 R_w^2)$ no longer incorporate higher order expressions of environmental variability.

a. Phase. The resulting equations at $O(R_w)$ are the same as (5.2). However, now θ and A are functions of (X',Y',T') while H_u, H_l, Z, Φ, b, f, and their derivatives vary only with (X,Y,T). Over any $O(1)$ increment in (X',Y',T'), the latter therefore are effectively constant. Within the equivalent characteristic system (5.6), both F_T' and F_X' are $O(\epsilon/\epsilon')$ and negligible. Thus $\theta_{T'}$, and $\nabla'\theta$ are independent of s'; that is, $\mathbf{k}$ and ω are constants along a characteristic. Similarly the group velocities are straight lines in $\mathbf{X}'$-T' space, but not necessarily parallel ones. If, however, θ is initially linear in ξ' and η', then $\mathbf{k}$, ω, and $\mathbf{c}_g$ are constants uniformly in (ξ',η',T').

b. Amplitude. At $O(\epsilon R_w)$ one can again derive action laws of the forms (5.14) and (5.17). However, the derivatives are in X', Y', T' or s' and thus apply to only θ, A, and C as they appear in $\mathbf{c}_g$, $\hat{A}$, and $\hat{J}$. Because the group velocity cannot change along a characteristic, we can explicitly solve (5.6) for the coordinate transformation.

$$X' = \xi' + c_g^{(x)} s', \qquad Y' = \eta' + c_g^{(y)} s'$$
$$\hat{J} = 1 + s'\nabla\cdot\mathbf{c}_g + s'^2 J(c_g^{(x)}, c_g^{(y)}) \ . \tag{7.3}$$

From (5.17), the initial tendency for action changes is governed by the group velocity divergence, but the long time behavior, if relevant, is due to the Jacobian product of the group velocity components. Either $\hat{J}$ has at least one zero crossing or, if not, it must eventually diverge to positive infinity. Along a characteristic, therefore, the action and the wave amplitude must either become singular or else asymptotically vanish. While the latter spreading of energy might be stopped at large times of $O(\epsilon'/\epsilon)$ by environmental focusing, the singular case implies a failure of the present model in a time T' of order one. This failure may be interpreted as amplitude variability on increasingly shorter scales, which is in violation of the original multiple scale assumption. This situation is analogous to a turning point in traditional WKB problems (e.g., Ferry and Mount, 1972), and extension through the singularity can be made; this will not be done here. With a linear initial phase function, $\hat{J}$ is a constant and the action law (5.14) reduces to

$$\left(\frac{\partial}{\partial T'} + \mathbf{c}_g\cdot\nabla'\right) C = 0 \ . \tag{7.4}$$

This implies simple propagation, without distortion, of the initial wave pulse in the direction of the group velocity.

c. Rectification. The simplification of an almost uniform environment is perhaps greatest for the wave field rectification. From the $O(R_w^2)$ operators in (2.10), the $O(\epsilon'^2 R_w^2)$ terms become

$$B_u = \frac{1}{2f}\left\{ J\left(\theta, \frac{|C|}{D_u(\theta)} D_u(|C|e_u)\right) + J\left(\frac{|C|^2}{2}, e_u\right)\right\}$$
$$B_l = \frac{1}{2f}\left\{ J\left(\theta, \frac{|C|A}{D_l(\theta)} D_l(|C|e_l)\right) + J\left(\frac{|C|^2A^2}{2}, \frac{e_l}{A}\right)\right\} \tag{7.5}$$

analogous to those of (5.18). Again, only derivatives of the wave functions contribute. For a linear initial phase function, we keep only derivatives of the amplitude magnitude:

$$B_u = \frac{e_u}{2fD_u(\theta)} J\left(\theta, D_u\left(\frac{|C|^2}{2}\right)\right),\ B_l = \frac{Ae_l}{2fD_l(\theta)} J\left(\theta, D_l\left(\frac{|C|^2}{2}\right)\right) . \tag{7.6}$$

This final form can be rewritten in terms of more familiar wave quantities by the $(\omega,\mathbf{k})$ identifications of (4.2), a local interpretation of the environment, and the simple amplitude law (7.4). Then

$$B_u = \tfrac{1}{2} \frac{e_u K^2}{\omega(\omega - \mathbf{V}_u\cdot\mathbf{k})} [\hat{e}_z \times \mathbf{c}_p]_j [\mathbf{c}_g - \mathbf{V}_u]_k \frac{\partial}{\partial X_j} \frac{\partial}{\partial X_k} \frac{|C|^2}{2}$$
$$B_l = \tfrac{1}{2} \frac{Ae_l K^2}{\omega(\omega - \mathbf{V}_l\cdot\mathbf{k})} [\hat{e}_z \times \mathbf{c}_p]_j [\mathbf{c}_g - V_l]_k \frac{\partial}{\partial X_j} \frac{\partial}{\partial X_k} \frac{|C|^2}{2} , \tag{7.7}$$

where $\mathbf{V}_u$ and $\mathbf{V}_l$ are mean upper and lower layer currents and $\mathbf{c}_p$ is the phase velocity,

$$\mathbf{V}_u = 1/f\hat{e}_z \times \nabla' Z',\ \mathbf{V}_l = 1/f\ \hat{e}_z \times \nabla'\Phi',\ \mathbf{c}_p = -\frac{\theta_{T'}}{|\nabla'\theta|^2} \nabla'\theta . \tag{7.8}$$

In (7.7) a summation convention is used (n.b., the indices j and k are 1 or 2). Thus, the rectified wave forcing is proportional to the energy coefficients in each layer, the square of the scalar wave number, and the inverse of both the intrinsic and local Doppler frequencies. Its shape is determined by the spatial curvature in the wave energy projected along directions perpendicular to the phase velocity and parallel to the difference between $\mathbf{c}_g$ and the mean velocity of the layer.

d. Large scale currents. The mean currents driven by the wave forcing in (7.5) (and with $R'_M = \epsilon'^2 R_w^2$) satisfy linearized equations which are inhomogeneous versions of (3.3):

$$\gamma^2(Z' - \phi')_{T'} + J\left(Z', \frac{1}{f}\right) = -B_u,\ \delta\gamma^2(\Phi' - Z')_{T'} + J\left(\Phi', \frac{b}{f}\right) = -B_l . \tag{7.9}$$

The synoptic constraint, analogous to (3.5), becomes

$$J\left(Z',\frac{\delta}{f}\right)+J\left(\Phi',\frac{b}{f}\right)=-\left(\delta B_u+B_l\right)\ . \tag{7.10}$$

Except for fortuitous cancellation ($\delta B_u = - B_l$), it is inconsistent to consider an initial state with a non-zero wave field and zero mean currents. An adjustment process between waves and mean flow must be presumed to have occurred prior to the period where (7.10) applies. The large scale barotropic motions described at the end of §2, but neglected here, must participate in this adjustment.

A particular example of flow above a flat bottom ($b \equiv 1$) is described below. There the free, mean currents would be as in (3.4); they may be arbitrarily superimposed upon the forced response. The synoptic constraint (7.10) can be integrated zonally to show that the large scale currents cannot vanish everywhere at large distances from even a bounded region of wave forcing. Both Z' and Φ' cannot asymptotically approach constants unless, for all Y',

$$\int_{-\infty}^{\infty} dX'(\delta B_u + B_l) = 0\ , \tag{7.11}$$

which is unlikely in general. The forced meridional currents will be confined to the region of nontrivial forcing, but the zonal currents will not. For this reason, as well as the general Y-dependent arbitrariness of (3.4), the present mid-ocean dynamics must be judged a more fundamental description of flow across, rather than along, contours of f/(thickness). For a closed basin, additional dynamical processes must be included near boundaries and narrow mean currents in order to fully determine the contour flows.

e. Illustration. Consider the special case of a baroclinic Rossby wave field (i.e., $\beta \neq 0$, $b = 1$, $Z_0 = \Phi_0 = 0$). The dispersion relation and amplitude ratio are as in (5.3). Assume further that the initial spatial envelope of the oscillations—the amplitude function—is Gaussian (see Fig. 4) and $\theta(\xi,\eta,0)$ is linear. Then a single frequency and wave number vector characterize the wave field and are preserved with propagation; amplitude is conserved as in (7.4). The forced large scale currents satisfy (7.9) and (7.10). For their initial conditions, we assume no lower layer flow ($\Phi' = 0$) and vanishing upper layer flow to the west ($Z' \to 0$ as $X' \to -\infty$).

Eliminating Φ' from (7.9) yields a forced wave equation for the upper layer meridional velocity:

$$\left\{\gamma^2(1+\delta)\frac{\partial}{\partial T'} - \frac{\beta}{f^2}\frac{\partial}{\partial X'}\right\}\frac{\partial Z'}{\partial X'} = \delta\frac{f^2\gamma^2(1+\delta)}{\beta}\frac{\partial B_u}{\partial T'} - \frac{\partial B_u}{\partial X'}\ . \tag{7.12}$$

For baroclinic waves, $B_l = B_u$. The forcing is thus more intensified in the upper layer than is the wave structure (by an extra factor of δ^{-1}).

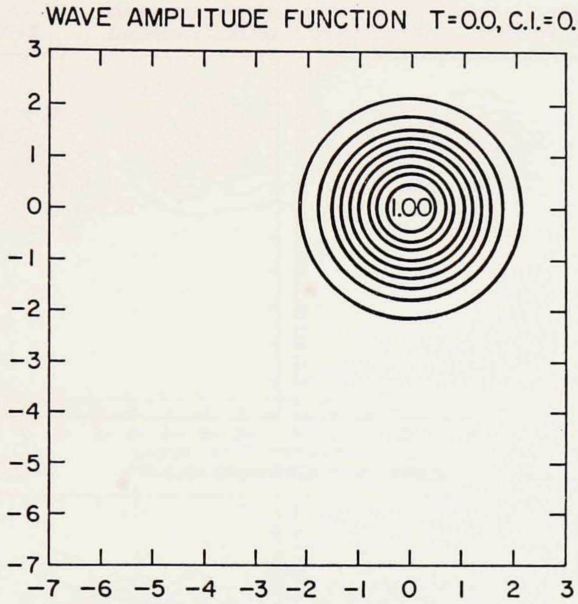


Figure 4. The initial structure of the amplitude function $C(X', Y', 0)$. This is the envelope of mesoscale oscillations which force the large scale flows shown in Figs. 5 and 6. The number of wave crests within this envelope would be approximately $1/\epsilon'$. C.I. is an abbreviation for "Contour Interval".

Because the wave field evolution is so simple under these circumstances, so also is the rectified forcing. The initial shapes for B_u and the right-hand side of (7.12) will simply propagate, without distortion, at the group velocity. The coincidence of $\mathbf{c}_g$ with the large scale baroclinic phase speed $-\beta\hat{e}_x/\gamma^2 f^2(1+\delta)$ should, therefore, lead to a resonantly effective forcing. This condition ($\mathbf{c}_g$ for small scales matching $\mathbf{c}_p$ for large) is reminiscent of Landahl's (1973) for shear flow instability.

Illustrations of two wave-driven mean flows are given in Figs. 5 and 6. The two have identical initial conditions for wave amplitude (i.e., Fig. 4) and wave number magnitude ($\gamma^2 = \mathbf{k}^2 = 1$) but the wave vectors differ in direction. One has phase propagation towards NNW (Fig. 5) and the other towards WNW (Fig. 6). The resulting mean currents are quite different.

(*i*) *Wave properties.* The wave vectors chosen were

$$\mathbf{k} \equiv (m,n) = \begin{cases} (-.31, .95) & \text{(NNW)} \\ (-.95, .31) & \text{(WNW)} \end{cases} \tag{7.13}$$

Since frequency is proportional to zonal wave number m [see (5.3)], the NNW wave field has a small frequency relative to the WNW wave (by a factor of 1/3) with $\mathbf{c}_g$ towards WSW at slightly more than half the large scale wave speed (a "near

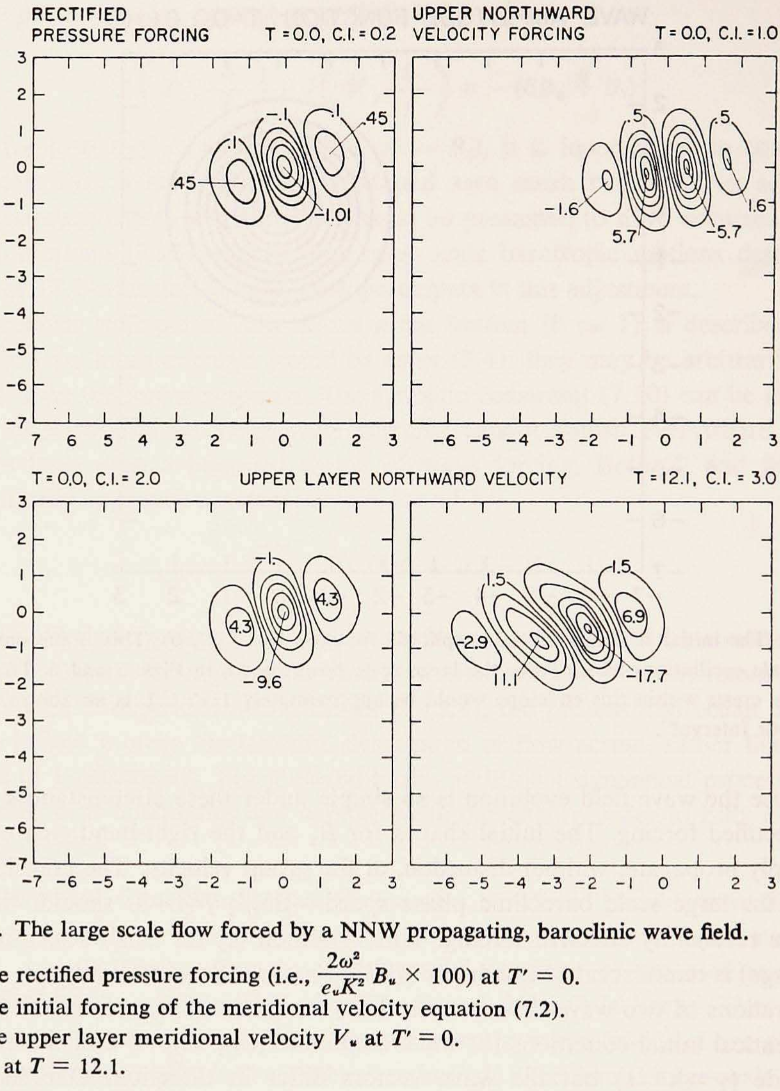


Figure 5. The large scale flow forced by a NNW propagating, baroclinic wave field.

(a) The rectified pressure forcing (i.e., $\frac{2\omega^2}{e_u K^2} B_u \times 100$) at $T' = 0$.

(b) The initial forcing of the meridional velocity equation (7.2).

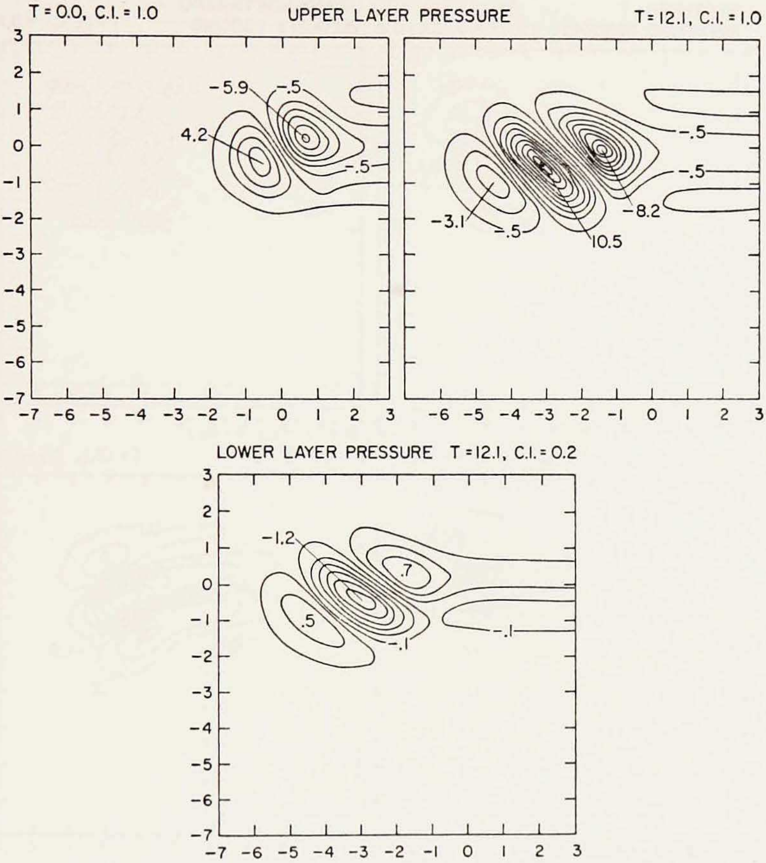
(c) The upper layer meridional velocity V_u at $T' = 0$.

(d) V_u at $T = 12.1$.

resonance"). The more rapid WNW wave has a slower group velocity, mostly to the south.

(*ii*) *Rectified pressure forcing*. Figures 5a and 6a show $2\omega^2 B_u / e_u \mathbf{k}^2$ at $T' = 0$ as calculated from (7.7). The two patterns are very similar but for a rotation; the pattern strengths do not much differ, in spite of differing $\mathbf{c}_p$ and $\mathbf{c}_g$ magnitudes, because the variations between the two cases are compensating in B_u.

(*iii*) *Meridional velocity forcing*. Figures 5b and 6b show the quantity on the right-hand side of (7.12) at $T' = 0$. With time it simply translates with $\mathbf{c}_g$. The



(e) The upper layer pressure Z' at $T' = 0$.

(f) Z' at $T' = 12.1$.

(g) The lower layer pressure Φ' at $T' = 12.1$.

NNW wave forcing is more intense by an order of magnitude. To understand this, rewrite the (7.12) forcing using (5.3) and the corresponding $\mathbf{c}_g$ to eliminate all but wave number factors. Thus,

$$\frac{\delta\gamma^2(1+\delta)}{\beta}\ \frac{\partial B_u}{\partial T'} - \frac{\partial B_u}{\partial X'} = \delta\Gamma\left[\frac{m^2-n^2-\Gamma}{K^2+\Gamma} + \frac{1}{\delta\Gamma}\ ,\ \frac{2mn}{K^2+\Gamma}\right]_j \tag{7.14}$$

$$\left[\ -n,m\ \right]_k \left[\frac{m^2-n^2-\Gamma}{m}\ ,2n\right]_l \frac{\partial}{\partial X_j}\ \frac{\partial}{\partial X_k}\ \frac{\partial}{\partial X_l}\ \frac{|C|^2}{2}\ ,$$

where $\Gamma = \gamma^2(1+\delta)$ and $K^2 = m^2+n^2$. The notation $[A,B]_j$ denotes the jth component of a vector whose x and y components are A and B. As m gets small with n of

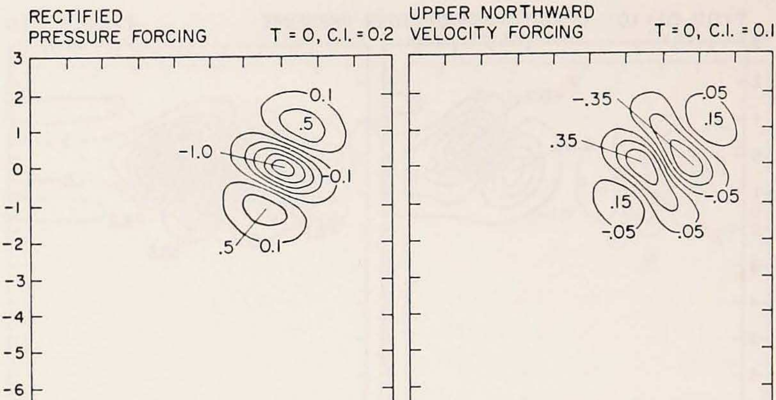
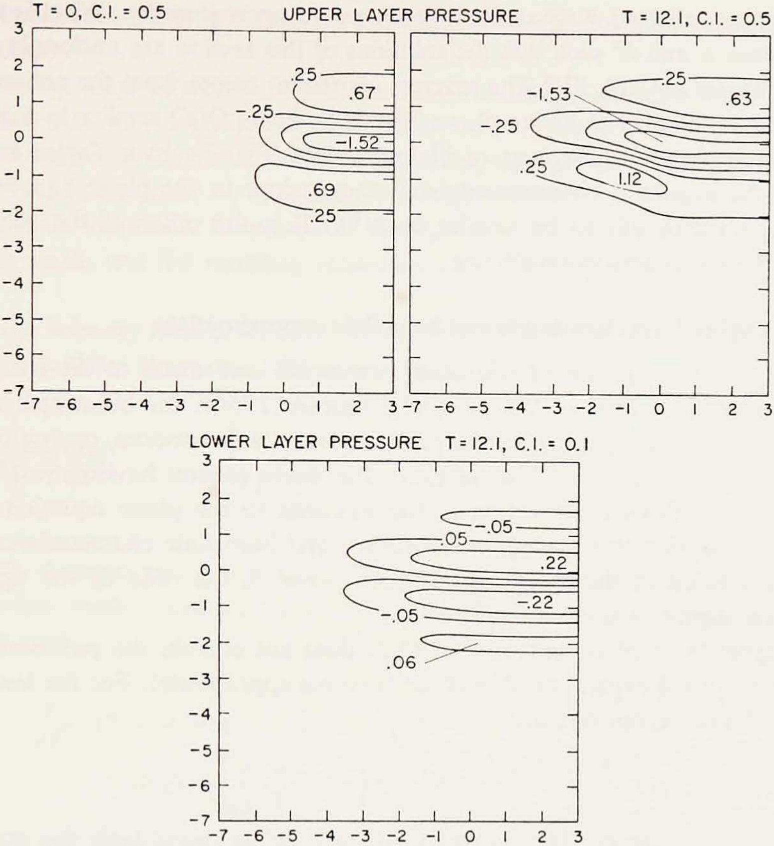


Figure 6. The large scale flow forced by a WNW propagating, baroclinic wave field. The sequence is as in Fig. 5.

order one (the NNW) wave), the third vector gets large, whereas when m is $O(\gamma^2)$ and n is small (the WNW wave) it becomes small.

(iv) Pressure patterns. Figures 5e, 5f, 6e, and 6f show Z' initially and for a much later time ($T' \approx 12$). The mean flow forced by the NNW wave starts out more intense and forms a nearly closed pair of gyres. The weaker WNW case has banded zonal flow. Both patterns develop westward. The NNW case continues to intensify as the forcing and large scale, baroclinic response move (approximately) together. There is little intensification in the more skew WNW situation, but a clear mean field wake develops along f contours. The lower layer pressure is initially zero, but grows weakly and generally out of phase with the upper layer flow (Figs. 5g and 6g) in order to satisfy (7.10).



(*v*) *Upper layer meridional velocities.* Since the governing eq. (7.12) is most directly a forcing of V_u', the large scale response is perhaps most clearly seen in V_u'. These patterns, initially and at $T' \approx 12$, are shown in Figs. 5c, 5d, 6c, and 6d. In both cases the loci of translation with mesoscale $\mathbf{c}_g$ and large scale $\mathbf{c}_p$ are clearly exhibited at longer times.

(*vi*) *Dimensional interpretation.* These solutions correspond to $\beta = .4$, which implies $L = 1500$ km and $\epsilon = .03$. With $\epsilon' = .2$ and $V_w = 10$ cm/sec, spatial units in Figs. 4-6 correspond to $L' = 250$ km, the large scale $\mathbf{c}_p$ is 3 cm/sec to the west, the mesoscale $\mathbf{c}_p$ and $\mathbf{c}_g$ are O(1 cm/sec), pressure units $(R_M'\epsilon(f_0 l)^2/\epsilon' g)$ in Figs. 5e, f, g and 6e, f, g, are .6 cms of head, velocity units $(V_w{}^2\epsilon'^2/f_0 L\epsilon^2)$ in Figs. 5c, d and 6c, d are .4 cm/sec^{-1}, and the integration time $(T' = 12)$ is in excess of a year. For a deformation radius of 45 km, the mesoscale wavelength $(2\pi l)$ is 280 km. Notice that in the case of near resonance (Fig. 5d), large scale, upper layer, meridional velocities greater than 5 cm/sec have been forced after only a year.

f. Limit of validity for the uniform environment approximation. What is the relation between L and L' such that the solutions of this section are uniformly valid in the coordinates (X', Y', T')? The severest constraint comes from the action law. It is presumed independent of the phase equations, being $O(\epsilon')$ smaller in magnitude, even though sharing a common oscillatory factor (Fig. 3). In a Taylor series expansion, the neglected environmental inhomogeneities in the phase equation arise at $O(R_w\epsilon/\epsilon')$. For this to be smaller than $O(\epsilon' R_w)$, the order of the action law, requires the severe restriction $\epsilon'^2 >> \epsilon$.

8. A thin upper layer: barotropic and baroclinic approximations

Theoretical descriptions of mid-ocean transience owe much to the two vertical normal modes described by Veronis and Stommel (1956): the barotropic and first baroclinic modes. They were derived for the case of a flat-bottom, resting ocean on the β-plane, and are reproduced in (5.3). For more general environmental conditions, however, this categorization of the solutions to the phase equation (5.2) is not meaningful due to a mixing of barotropic and baroclinic characteristics. Yet a natural extension of these concepts emerges when δ, the ratio of the upper and lower layer depths, is small.

The upper layer phase equation of (5.2) does not contain the parameter δ and is unaltered in a δ expansion (though $\Delta=0$ seems appropriate). For the lower layer equation, however, one obtains

$$A\left[\frac{b\nabla\theta\cdot\nabla\theta}{f^2}\,D_l(\theta) + J\left(\theta\,,\frac{b}{f}\right)\right] = O(\delta)\ . \tag{8.1}$$

The vanishing of the two factors on the left side define respectively the generalized baroclinic and barotropic solutions. They are explicitly

$$\textit{barotropic}\qquad D_l(\theta) = -\frac{f^2J(\theta,b/f)}{b\nabla\theta\cdot\nabla\theta}\,,\qquad A = 1 + \frac{H_u\nabla\theta\cdot\nabla\theta}{\gamma^2 f^2} + \frac{J\left(\theta,\dfrac{H_u}{f}\right)}{\gamma^2 D_u(\theta)}$$

$$\textit{baroclinic}\qquad D_u(\theta) = -\frac{J\left(\theta,\dfrac{H_u}{f}\right)}{\gamma^2 + H_u\nabla\theta\cdot\nabla\theta/f^2}\,,\ \mathrm{A}=0 \tag{8.2}$$

with an error of $O(\delta)$, except in the case where the two solutions are coincident [the error is then $O(\delta^{\frac{1}{2}})$]. The barotropic wave has a combined planetary and topographic restoring force and extends throughout the water column. In contrast the baroclinic wave responds to β and the vertical shear of the large scale velocity, but is confined to the upper layer. The frequencies are Doppler shifted by the mean velocities in the respective layers. Surface currents will more strongly influence baroclinic propagation, but it is effectively the depth averaged currents which influence the barotropic.

The consequences of a shallow thermocline ($\delta << 1$) can be extended throughout the solution structure described in §§5 and 6. In particular, if $\delta = 0(\epsilon^n)$ for n any positive integer, the solution form of (6.1) is appropriate here as well; the phase function is fully specified by (8.2) since its higher order corrections in δ can be absorbed in the amplitude functions. At $O(\epsilon R_w)$ the layer equations (5.13) decouple when δ is small, and the resulting action laws involve only quantities for a single layer.

They are formally analogous to the relations of §5b if we replace the functional in (5.10) with either barotropic (F^T) or baroclinic (F^C) forms:

$$F^T \equiv A^2 \left[\frac{b|\nabla\theta|^2}{f^2} + \frac{J\left(\theta, \frac{b}{f}\right)}{D_l(\theta)} \right], \quad F^C \equiv \gamma^2 + \frac{H_u \nabla\theta \cdot \nabla\theta}{f^2} + \frac{J(\theta, H_u/f)}{D_u(\theta)} \qquad (8.3a) \quad (8.3b)$$

The action densities and group velocities then follow from (5.15) and (5.8). For the barotropic mode, there is strict conservation of A^T as in (5.14), whereas for the baroclinic mode,

$$\frac{\partial}{\partial T} A^C + \nabla(A^C \mathbf{c}_g{}^C) = i\left(\frac{\delta}{\epsilon}\right) A^C \; \frac{2\gamma^2}{\hat{F}^T} \frac{D_u(\theta)\, D_l(\theta)}{\left[\gamma^2 + \frac{h_u|\nabla\theta|^2}{f^2}\right]}, \qquad (8.4)$$

where $\hat{F}^T = F^T \cdot D_l(\theta)/A^2$. When $\delta/\epsilon \sim 1$, A^C is not conserved; note, however, that the forcing is purely imaginary and causes changes along characteristics only in the phase, not the magnitude, of C. Other relations for the single wave field solution, such as the rectification and the nonlinearly forced second harmonic can be similarly simplified.

These results are valid only away from the coalescence of the two modal frequencies in (8.2). Coalescence is also a necessary condition for wave instability. The unexpanded dispersion relation (5.2) is quadratic; hence the two modal frequencies must be either distinct and real or conjugates. Since the $O(\delta^0)$ frequencies are always real, complex values at higher orders require them to be equal. The analysis of this situation is beyond the scope of this paper. However, the following heuristic argument shows that, if a wave field is initially not at coalescence everywhere and δ is small, it cannot reach coalescence in any O(1) increment in the environmental coordinates.

Inserting A from (5.12) into (8.1) yields the dispersion equation

$$\hat{F}^T \hat{F}^C = O(\delta) \;, \qquad (8.5)$$

where $\hat{F}^C = \hat{F}^C D_u(\theta)$. This functional is an acceptable choice for the characteristic system (5.6), which defines the evolution in $s(=T)$, of θ_T, $\nabla\theta$, T, and $\mathbf{X}$, hence also of $\hat{F}^T$ and $\hat{F}^C$. If initially there is no coalescence in a baroclinic[5] wave field, then $\hat{F}^C = O(\delta)/\hat{F}^T$, and it is approximately zero. From the characteristic equations (5.6), it follows that

$$\left.\frac{\partial \hat{F}^C}{\partial s}\right|_{\xi,\eta} = O(\delta)/\hat{F}^T \left.\frac{\partial F^T}{\partial s}\right|_{\xi,\eta} = O(1)\hat{F}^T + O(\delta)/\hat{F}^T \ .$$

The waves remain approximately baroclinic along characteristics, while F^T can make O(1) changes. If coalescence is somewhere approached (say $F^T \sim \delta^\alpha$ where α increases from zero), then $\partial F^T/\partial s$ also becomes small and $O(\delta, \delta^{1-\alpha})$, thus retarding the approach. With uniform validity in (X,Y,T), coalescence can never be reached. Similarly, one can show that an initially coalesced wave field can never leave coalescence in an O(1) environmental scale interval.

9. Comments and conclusions

Even in a broadly variable environment, mesoscale, wave-like solutions can exist and their environmental scale evolution can be predicted. For mesoscale distances and times, over which the environment is effectively uniform, their structure is that of plane, propagating waves. However, due to their accommodation of the broader scale variability, the solutions have features which are trivial for plane waves: the propagation and distortion of the wave packet, the transformation of the oscillation frequency and wavelength, and the forcing of the mean circulation by the rectified eddy fluxes. The character of each of these laws is complicated analytically, yet fairly easily solved for in full detail by numerical computation.

If there is a true spectral gap between mesoscale and environment, then important postponements of nonlinear behavior must occur. Nonlinear waves ($R_w \gtrsim 1$) will still have propagation characteristics described by linear wave dispersion relations: the simulations of Rhines (1975) suggest that once a spectral evolution has settled on a dominant scale (i.e., approached a single wave field situation) its phase propagation is uniform and as predicted by linear theory. Furthermore, nonlinearity should be ineffective in either altering the distribution of wave energy (the amplitude function) or generating higher harmonics of the fundamental wave.[6] Rectified forcing by the mesoscale wave field will be weak. Since observations indicate a generally weak mid-ocean mean flow, however, wave forcing may often be its cause. Note that for a single, stable wave field the mean field interaction is uni-directional. For comparable mean and wave particle speeds, the mean currents are part of the wave environment, but are not driven by rectified forcing to leading order. On the other hand, wave-driven mean currents are too weak to influence the wave field evolution. Within this spectral gap model, the interaction can become mutual only for baroclinically unstable waves—but this is the subject of a future study.

5. The argument could be applied to barotropic as well.

6. The preceding characteristics must be reconsidered when more than a single wave is present.

Since there is such a disparity between the internal radius of deformation (~50 km) and ocean basin dimensions (~3000 km), numerical simulations resolving both mesoscale eddies and the general circulation are difficult. Of considerable interest would be a successful parameterization of mesoscale currents. The solution structure (4.1) has a considerable advantage in this issue—it only requires solving equations in the large scale, not mesoscale, coordinates. If there were a clear mesoscale spectral peak, then, the preceding formulas might be useful, eq. (5.18) in particular. However, several crucial questions remain unanswered: for example, (1) which wave field(s) should be present initially? (2) What complications are introduced by multiple wave fields (a discrete spectrum)? (3) What alternative relations are required in the special locations—near boundaries or narrow, intense mean currents—where this model does not apply?

These questions can be partly answered in a spectral gap model. For small amplitude waves ($R_w << 1$), boundary conditions could be satisfied by a superposition of waves of the form (4.1), but such a representation would be cumbersome. The character of a multiple wave field can be explored in the present model—this also is a subject for the future—with only a weak restriction on R_w. Many of the single wave laws persist, including the rectification structure (5.18). Baroclinic instability in a spectral gap model may partially indicate which waves should be assumed present. Nevertheless, the questions are formidable, and only as components of a more general parameterization scheme could the present formulas be applied.

It is crucial to the formulation of the model that the scale ratio ϵ be small. If it is not, then one loses the concept of waves altogether (phase and amplitude changes merge). Yet as ϵ becomes less small, all wave processes speed up: the frequencies are higher, and the amplitude and wave number evolutions are more rapid and occur over shorter distances. Similarly the rectified forcing would be more intense, even to the point of forcing wave-comparable mean flows ($R_M \sim R_w$). All of these tendencies are characteristic of synoptic scale atmospheric waves, for example, where the wave and environmental scales are more similar.

Acknowledgments. This work had its beginnings in discussions with Professor A. R. Robinson, and its presentation has been much improved through the suggestions of Professor G. Veronis. Support was provided by the Office of Naval Research under contract N000 14-67-A-0298-0011 to Harvard University and by the National Science Foundation through the National Center for Atmospheric Research. This is contribution No. 35 of the Mid-Ocean Dynamics Experiment.

REFERENCES

Benny, D. J. 1971. Non-linear waves. Mathematical Problems in the Geophysical Sciences, *1:* 103–121.

Bretherton, F. P., 1966. The propagation of groups of internal gravity waves in a shear flow. Quart. J. Roy. Met. Soc., *92* (394): 466–480.

Bretherton, F. P., and C. J. R. Garrett, 1968. Wavetrains in inhomogeneous moving media. Proc. Royal Soc. London, *A:* 302, 529–554.

Bryan, K., and M. D. Cox, 1967. A numerical investigation of the oceanic general circulation. Tellus, *19:* 54–80.

Courant, R., and D. Hilbert, 1962. Methods of Mathematical Physics, II, New York, Interscience, 830 pp.

Cheney, R. E., and P. L. Richardson, 1974. The observed decay of a cyclonic Gulf Stream ring. U. of Rhode Island Tech. Rpt. No. 74–2.

Davis, R. E., 1975. Statistical methods, *in* Dynamics and the Analysis of the MODE Data, A. R. Robinson, ed., Cambridge, Massachusetts Institute of Technology 54–1417, 250 pp.

Ferry, M. V., and K. E. Mount, 1972. Semiclassical approximations in wave mechanics. Reports on Progress in Physics, *35:* 315–397.

Holland, W. R., and A. D. Hirschman, 1972. A numerical calculation of the circulation in the North Atlantic Ocean. J. Phys. Oceano., *2:* 336–354.

Kuo, H. L., 1949. Dynamic instability of two-dimensional non-divergent flow in a barotropic atmosphere. J. Meteor., *6:* 105–122.

Landahl, M. T., 1973. Wave mechanics of breakdown. J. Fluid Mech., *65:* 775–802.

Landau, L. D., and E. M. Lifshitz, 1959. Fluid Mechanics, London, Pergamon Press, 536 pp.

Lee, W. H. K., and W. M. Kaula, 1967. A spherical analysis of the earth's topography. J. Geophys. Res., *72:* 753–758.

Lighthill, M. J., 1965. Group velocity. J. of the Inst. of Math. Appl., *1:* 1–28.

—— 1971: Time-varying currents. Phil. Trans. Roy. Soc. Land., *A:* 270, 371–390.

McWilliams, J. C., 1974. The MODE mean currents and eddy surface intensification. MODE Hot Line News, No. *57*, unpublished document.

McWilliams, J. C., and A. R. Robinson, 1974. A wave analysis of the Polygon array in the tropical Atlantic. Deep-Sea Res., *2:* 359–368.

McWilliams, J. C., and G. R. Flierl, 1975a. Optimal quasigeostrophic wave analysis of MODE array data. Deep-Sea Res., *23:* 285–300.

—— 1975b. Quasigeostrophic wave analyses, *in* Dynamics and the Analysis of the MODE Data, A. R. Robinson, ed., Cambridge, Massachusetts Institute of Technology 54–1417, 250 pp.

Phillips, N. A., 1963. Geostrophic Motion. Rev. Geophys. *1:* 123–176.

Platzman, G. W., 1968. The Rossby wave. Quart. J. Roy. Met. Soc., *94:* 225–248.

Rhines, P., 1970. Edge-, bottom-, and Rossby waves in a rotating, stratified fluid. Geophys. Fluid Dyn., *1:* 273–302.

—— 1975. Waves and turbulence on a β-plane. J. Fluid Mech., in press.

Robinson, A. R., and J. C. McWilliams, 1974. The baroclinic instability of the open ocean. J. Phys. Oceano., *4:* 281–294.

Schmitz, W., J. Luyten, and W. Sturges, 1976. The role of eddies in the general circulation of the deep western North Atlantic. Deep-Sea Res., *69:* 417–443.

Veronis, G., and H. Stommel, 1956. The action of a variable wind stress on a stratified ocean. J. Mar. Res., *15:* 43–75.

Whitham, G., 1965. A general approach to linear and non-linear dispersive waves using a Lagrangian. J. Fluid Mech., *233:* 273–282.

Wunsch, C., 1972. The spectrum of temperature near Bermuda from two minutes to two years. Deep-Sea Res., *19:* 577–593.

Received: *14 March, 1975;* revised: *2 March, 1976.*