J. Fluid Mech. (1981), *vol.* 102, *pp.* 169–209
Printed in Great Britain

The nonlinear dynamics of baroclinic wave ensembles

By JOSEPH PEDLOSKY

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543

(Received 18 December 1979 and in revised form 2 April 1980)

A theory is developed to describe the weakly nonlinear dynamics which applies in the simultaneous presence of several, long, baroclinic waves. The geometry is flat (i.e. $\beta = 0$) and dissipation is modelled by Ekman friction in the context of the quasi-geostrophic two-layer model. Three main problems are discussed.

(1) For free, unstable waves it is shown that the wave which is realized in finite amplitude is *not* the linearly most unstable wave. Rather a longer wave, capable of achieving the single largest steady amplitude, is favoured in the competition for the potential energy of the basic state. This result is shown necessary if the end state is steady and numerous numerical calculations indicate the pre-eminence of the same wave if the final state is vacillatory. The notion of conjugate waves, capable of identical final amplitude, is also discussed.

(2) If the free waves are subject to time-varying supercriticality so that intervals of stability ensue, the response is asymmetric over the period of the forcing. Sufficiently rapid 'seasonal' forcing leads to long-term aperiodic response.

(3) If each wave in the spectrum is directly forced a wave hysteresis phenomenon occurs. Sudden jumps in the wave amplitude at critical values of the forcing are intrinsic to the wave response. Again, sufficiently rapid wave forcing produces an aperiodic response.

The forced wave problem exhibits multiple equilibria. Each solution branch corresponds to a different dominant wave. The determination of the realized branch depends on the relative stability criteria developed for the *free* waves.

1. Introduction

The theory for the finite-amplitude behaviour of unstable baroclinic waves has usually focused attention on the dynamics of a single wave. Weakly nonlinear theories (e.g. Pedlosky 1970, 1971, 1972; Drazin 1970, 1972) allow such a restriction since within those theories the dynamics for a single-wave component becomes closed in a consistent way if the *initial* wave spectrum is limited to a single wave. This *a priori* restriction to a single wave in the initial conditions would be a more serious difficulty were it not for the frequent observation in the well-known annulus experiments of wave states containing essentially a single wave.

Nevertheless the restriction to a single wave, i.e. a disturbance of a single wavelength in the downstream direction, raises certain important questions that the single-wave, finite-amplitude theories cannot alone address. Perhaps the most obvious is the question of which wave, at a given parameter setting, is the most appropriate to examine in finite amplitude. Of course, the wave should be linearly unstable on the basic current but does the realized wave in finite amplitude have the wavelength of the

0022-1120/81/4596-2770 $02.00

most unstable wave according to linear theory? Although that wave will probably dominate the growing spectrum of waves initially, there is no reason to believe ultimately, as the zonal flow is altered by the waves, that it will remain the most successful predator of the potential energy available in the zonal flow. If linear theory is not a good predictor is there an alternative predictor that does not require explicitly testing the joint behaviour of an ensemble of waves? One of the purposes of this paper is to present a study of multiple-wave baroclinic dynamics to begin to answer these questions.

At the same time certain other fundamental questions seem to require the formulation of a model capable of handling the simultaneous presence of several waves. In a fascinating experimental study, Buzyna, Pfeffer & Kung (1978) showed how cyclic variations of the imposed cross-stream temperature gradient produced both regularly periodic and also aperiodic changes of observed wavenumber. To some extent my study was motivated by the report of these experiments and I felt it was desirable to formulate a theory that could allow, in the initial data, several wavelengths and thus equilibrium states at some parameter values which contain at least two co-existing wavelengths.

Finally, the problem of directly *forced* waves introduces dynamical issues that require a multiple wave theory. The finite-amplitude response of a set of waves to forcing at each wavelength in the spectrum depends on the collective response to the forcing. The amplitude each forced wave attains depends on the vigour of the forcing, the degree of instability at that wavelength and the nonlinear interaction of the wave with the zonal flow. The last effect in turn couples each forced wave with its fellows in the spectrum. This produces certain profound alterations in the dynamics. In particular, the simultaneous presence of several waves yields *multiple* equilibrium states. Which of these (if any) are stable turns out to be very closely related to the first question posed in this section, namely which *free* wave is realized in finite amplitude.

The next section discusses the model to be used to approach these questions. Section 3 discusses the nonlinear dynamics of free waves. Section 4 re-examines free waves in cases where the cross-stream temperature gradient or supercriticality is a slowly varying, cyclic function of time. Section 5 deals with the problem of forced baroclinic waves. Some final remarks and speculations are given in §6. While the detailed discussion of the results is contained in each appropriate section, I simply remark here that a central result of this paper is that although the single-wave theory may describe well the finite-amplitude dynamics of a baroclinic wave it is generally *not* the linearly most unstable wave that dominates the finite-amplitude problem. Thus single-wave theories should not, in general, focus on the most unstable wave according to linear theory. The aim of the following discussion is to describe why this is so and what the alternative finite-amplitude selection principle is for the determination of the realized wavelength.

2. The model

The physical and mathematical model is identical to that used in Pedlosky (1971). That is, a quasi-geostrophic, two-layer model, in a flat ($\beta = 0$) geometry is considered wherein the only dissipation is due to Ekman layers at the two rigid, horizontal,

bounding surfaces. The reader is referred to the earlier paper for a discussion and derivation of the basic quasi-geostrophic equations.

If U_1 and U_2 are the uniform, non-dimensional zonal velocities present in the upper and lower layers in the wave-free state, then the finite-amplitude problem for the disturbance stream function, ϕ_n, is

$$\left(\frac{\partial}{\partial t}+U_n\frac{\partial}{\partial x}\right)(\nabla^2\phi_n+(-1)^nF(\phi_1-\phi_2))-(-1)^nF(U_1-U_2)\frac{\partial\phi_n}{\partial x}=-r\nabla^2\phi_n$$
$$-J(\phi_n,\nabla^2\phi_n+(-1)^nF(\phi_1-\phi_2))+Q_n,\quad n=1,2. \tag{2.1}$$

In the above, the subscript n has values 1 and 2 for the upper and lower layers respectively. The parameter

$$F=4\Omega^2L^2\Big/g\frac{\Delta\rho}{\rho}H \tag{2.2a}$$

is the internal rotational Froude number. Ω is the rotation rate, L is the width of the channel containing the flow, while H is the undisturbed depth of each layer. The constant density of the lower layer exceeds the density of the upper layer by the small amount $\Delta\rho$. The effect of friction is measured by the parameter

$$r=E^{\frac{1}{2}}/\epsilon, \tag{2.2b}$$

where E is the Ekman number and ϵ the Rossby number. The nonlinearity is produced by the self-advection of potential vorticity in each layer,

$$J(\phi_n,\nabla^2\phi_n+(-1)^nF(\phi_1-\phi_2))\equiv\frac{\partial\phi_n}{\partial x}\frac{\partial}{\partial y}\{\nabla^2\phi_n+(-1)^nF(\phi_1-\phi_2)\}$$
$$-\frac{\partial\phi_n}{\partial y}\frac{\partial}{\partial x}\{\nabla^2\phi_n+(-1)^nF(\phi_1-\phi_2)\},\quad n=1,2. \tag{2.2c}$$

The single new term in (2.1) is a source term for potential vorticity in each layer, Q_n. The source term may be due to heating, external surface stress or topographic forcing by the mean flow U_n. The detailed specification of the agent producing the Q_n is not required. It will suffice that the Q_n can be partitioned into a portion independent of the downstream co-ordinate, x, and another portion which is periodic in x. Both components are functions of time and cross-stream co-ordinate y. Boundary conditions for (2.1) are

$$\frac{\partial\phi_n}{\partial x}=0\quad\text{on}\quad y=0,1, \tag{2.3}$$

while the x-independent portion of the stream field satisfies

$$\frac{\partial^2\phi_n}{\partial y\,\partial t}=0. \tag{2.4}$$

Linear theory (Pedlosky 1971) shows that a wave disturbance of the form

$$\phi_n=A\,e^{ik(x-ct)}\sin\pi y \tag{2.5}$$

will be unstable if F exceeds the critical value

$$F_c=\frac{a^2}{2}+\frac{a^2r^2}{2U_s^2k^2}, \tag{2.6a}$$

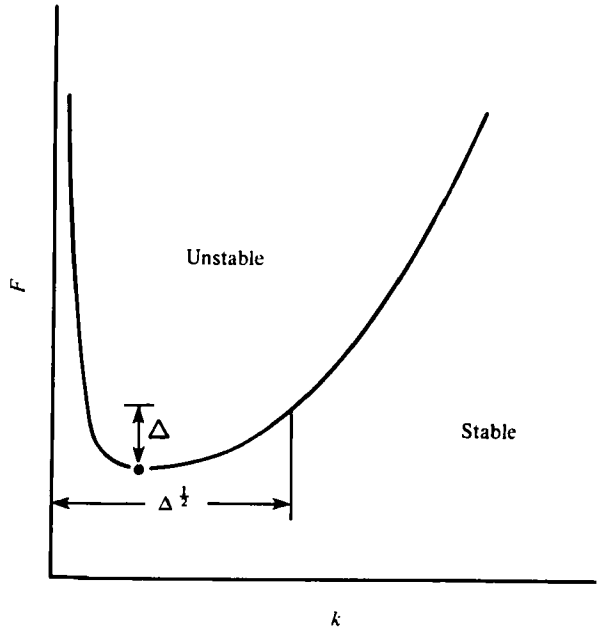

FIGURE 1. The schematic linear stability diagram for the internal, rotational Froude number $F = 4\Omega^2 L^2/g(\Delta\rho/\rho)\,H$ *versus* x wavenumber k. The stability cut-off for *small* k is due to Ekman friction. The large k cut-off is the Eady short-wave cut-off. For small friction the curve's minimum approaches $\frac{1}{2}\pi^2$ near $k = 0$. For small supercriticality, Δ, above the minimum, the range of unstable wavenumber is $O(\Delta^{\frac{1}{2}})$.

where

$$a^2 = k^2 + \pi^2, \qquad U_s = \tfrac{1}{2}(U_1 - U_2). \tag{2.6b, c}$$

The critical wave of F_c *versus* k is shown schematically in figure 1. For small r the minimum value of F_c occurs for long waves near $k = 0$. Friction stabilizes the *very* longest waves while the short waves suffer the familiar Eady (1949) cut-off. At values of F below F_c the waves are damped by friction. Now for small r the minimum critical value of F is very nearly $\frac{1}{2}\pi^2$. If F exceeds this value by an amount Δ a *range* of unstable wavenumbers in an $O(\Delta^{\frac{1}{2}})$ neighbourhood of $k = 0$ will be unstable if $r \leqslant O(\Delta)$. The theory to be developed pivots on this restriction. We will consider values of F such that

$$F = \frac{\pi^2}{2} + \Delta, \quad \Delta \ll 1, \tag{2.7}$$

for values of r which satisfy

$$r/|\Delta| = O(1). \tag{2.8}$$

The waves which are unstable are long waves, i.e. each of their wavelengths is $O(\Delta^{-\frac{1}{2}})$ times the cross-stream width. It may be helpful to the reader to think of an annulus with a narrow gap of width L and a mean radius $\mathscr{R}$. Think about the inviscid condition for instability. It is simply

$$F_c(m) = \frac{\pi^2}{2} + m^2 \frac{L^2}{\mathscr{R}^2}, \tag{2.9}$$

where m is the integral azimuthal *angular* wavenumber of the disturbance. The critical value for angular wavenumber n is given by (2.9) with n replacing m. Hence

$$F_c(n) = F_c(m) + \frac{n^2 - m^2}{2}\frac{L^2}{\mathscr{R}^2}. \tag{2.10}$$

Thus $F_c(n)$ will differ from $F_c(m)$ by $O(\Delta)$ if $L/\mathscr{R} \leqslant O(\Delta^{\frac{1}{2}})$. If wave m is supercritical, i.e. if F differs from $F_c(m)$ by $O(\Delta)$, wave n will also be supercritical as long as $n^2 - m^2 = O(1)$ and $L^2/\mathscr{R}^2 = O(\Delta)$. Hence a small supercriticality in F near $k = 0$ can embrace a *set* of long unstable waves each slightly unstable (and each possessing a different linear growth rate). To exploit this fact further let us note that according to linear theory the growth rate for each of these waves will be $kc_i(k)$. Since k is $O(\Delta^{\frac{1}{2}})$ and c_i is also $O(\Delta^{\frac{1}{2}})$ (Pedlosky 1971), the evolution time for the unstable waves is $O(\Delta^{-1})$ times the advection time. We therefore introduce the *slow* time

$$T = |\Delta| t \tag{2.11}$$

and for obvious reasons rescale x as

$$X = |\Delta|^{\frac{1}{2}} x. \tag{2.12}$$

To keep the theory as simple as possible, I restrict attention in this study to the case where

$$U_m = \frac{U_1 + U_2}{2} = 0 \tag{2.13}$$

so that initially there is no mean flow. This implies that the linearly unstable waves are stationary. For the free wave problem ($Q_n = 0$), this involves no loss of generality. In the final section I will discuss the effect of (2.13) on the forced problem.

Since the waves turn out to be stationary, even in finite amplitude when $U_m = 0$, the Q_n may be assumed to be functions of X, y, and T only.

If

$$\phi_T = \frac{\phi_1 - \phi_2}{2}, \quad Q_T = \frac{Q_1 - Q_2}{2}, \quad \phi_B = \frac{\phi_1 + \phi_2}{2}, \quad Q_m = \frac{Q_1 + Q_2}{2}, \tag{2.14}$$

then (2.1) may be rewritten, with the aid of (2.11), (2.12), (2.13) and (2.14), as

$$|\Delta|^{\frac{1}{2}} \frac{\partial}{\partial T}[\tilde{\nabla}^2 \phi_T - 2(F_c + \Delta)\phi_T] + U_s \frac{\partial}{\partial X}[\tilde{\nabla}^2 \phi_B + 2(F_c + \Delta)\phi_B]$$

$$= -\frac{r}{|\Delta|}|\Delta|^{\frac{1}{2}} \tilde{\nabla}^2 \phi_T + Q_T/|\Delta|^{\frac{1}{2}} - J(\phi_B, \tilde{\nabla}^2 \phi_T - 2(F_c + \Delta)\phi_T) - J(\phi_T, \tilde{\nabla}^2 \phi_B) \tag{2.15a}$$

and

$$|\Delta|^{\frac{1}{2}} \frac{\partial}{\partial T} \tilde{\nabla} \phi_B + U_s \frac{\partial}{\partial X} \tilde{\nabla}^2 \phi_T = -\frac{r}{|\Delta|}|\Delta|^{\frac{1}{2}} \tilde{\nabla}^2 \phi_B + Q_m/|\Delta|^{\frac{1}{2}} - J(\phi_B, \tilde{\nabla}^2 \phi_B) - J(\phi_T, \tilde{\nabla}^2 \phi_T), \tag{2.15b}$$

where

$$\tilde{\nabla}^2 \equiv \frac{\partial^2}{\partial y^2} + |\Delta| \frac{\partial^2}{\partial X^2}, \quad J(A, B) \equiv \frac{\partial A}{\partial X}\frac{\partial B}{\partial y} - \frac{\partial A}{\partial y}\frac{\partial B}{\partial X},$$

I further restrict attention to cases where Q_m and Q_T are $\leqslant O[(\Delta)^2]$.

Finally ϕ_T and ϕ_B are expanded in the asymptotic series

$$\left.\begin{aligned} \phi_T &= \alpha|\Delta|^{\frac{1}{2}}\{\phi_T^{(0)} + |\Delta|^{\frac{1}{2}}\phi_T^{(1)} + |\Delta|\phi_T^{(2)} + \ldots\}, \\ \phi_B &= \alpha|\Delta|^{\frac{1}{2}}\{\phi_B^{(0)} + |\Delta|^{\frac{1}{2}}\phi_B^{(1)} + |\Delta|\phi_B^{(2)} + \ldots\}, \end{aligned}\right\} \tag{2.16}$$

where α is an $O(1)$ scaling constant. If (2.16) is inserted into (2.15a,b) a sequence of *linear* problems are obtained by equating like orders in $|\Delta|^{\frac{1}{2}}$. The problems for $\phi_B^{(0)}$ and $\phi_T^{(0)}$ are (from the $O(\Delta^{\frac{1}{2}})$ problem)

$$\frac{\partial}{\partial X}\left\{\frac{\partial^2\phi_B^{(0)}}{\partial y^2}+2F_c\phi_B^{(0)}\right\}=0, \tag{2.17a}$$

$$\frac{\partial}{\partial X}\left\{\frac{\partial^2\phi_T^{(0)}}{\partial y^2}\right\}=0, \tag{2.17b}$$

whose solutions, satisfying (2.3), are

$$\phi_T^{(0)}=0, \quad \phi_B^{(0)}=A(X,T)\sin\pi y, \quad F_c=\tfrac{1}{2}\pi^2. \tag{2.18a,b,c}$$

The basic disturbance is primarily barotropic. Its cross-stream structure is taken to be the gravest cross-stream mode since higher cross-stream modes would be linearly stable and decaying due to friction. The temporal and downstream structure, as represented by the amplitude A, is yet an undetermined function of X and T. Since r is $O(\Delta)$ the critical value of F_c which occurs in (2.18c) is simply the inviscid criterion for long waves, as given by (2.9).

The $O(\Delta)$ portion of (2.15b) implies that

$$\frac{\partial}{\partial T}\frac{\partial^2\phi_T^{(0)}}{\partial y^2}+U_s\frac{\partial}{\partial X}\frac{\partial^2\phi_T^{(1)}}{\partial y^2}=-\frac{r}{|\Delta|}\frac{\partial^2\phi_B^{(0)}}{\partial y^2} \tag{2.19}$$

or

$$\phi_T^{(1)}=B(X,T)\sin\pi y+\Phi_T^{(1)}(y,T).$$

The function B must, by (2.19), satisfy

$$\frac{\partial B}{\partial X}=-\frac{1}{U_s}\left\{\frac{\partial A}{\partial T}+\frac{r}{|\Delta|}A\right\}, \tag{2.20}$$

which determines the X-shifted baroclinic component of the wave in terms of both the time rate of change of the barotropic field and the effect of dissipation on that barotropic field. The function $\phi_T^{(1)}$ represents an order-Δ alteration of the x-independent stream function and is, to this point, undetermined. However, it is clear that we may always choose $\Phi_T^{(1)}$ so that, by definition, B has a zero average on the infinite X interval.

Since $\phi_T^{(0)}$ vanishes, the $O(\Delta)$ portion of (2.15a) merely implies that $\phi_B^{(1)}$ is a multiple of $\phi_B^{(0)}$. Renormalization in that case allows us to satisfy (2.15a) to that order by simply letting $\phi_B^{(1)}$ be *zero*.

The problem for $\phi_B^{(2)}$, as determined by (2.15a), is then

$$\begin{aligned}U_s\frac{\partial}{\partial X}\left\{\frac{\partial^2\phi_B^{(2)}}{\partial y^2}+2F_c\phi_B^{(2)}\right\}=&-U_s\frac{\partial}{\partial X}\left\{\frac{\partial^2}{\partial X^2}\phi_B^{(0)}+2\frac{\Delta}{|\Delta|}\phi_B^{(0)}\right\}-\frac{\partial}{\partial T}\left(\frac{\partial^2}{\partial y^2}-2F_c\right)\phi_T^{(1)}\\&-\frac{r}{|\Delta|}\frac{\partial^2}{\partial y^2}\phi_T^{(1)}+\frac{Q_T}{|\Delta|^2}-\alpha\mathcal{J}\left(\phi_B^{(0)},\frac{\partial^2}{\partial y^2}\phi_T^{(1)}-2F_c\phi_T^{(1)}\right)\\&-\alpha\mathcal{J}\left(\phi_T^{(1)},\frac{\partial^2\phi_B^{(0)}}{\partial y^2}\right).\end{aligned} \tag{2.21}$$

If (2.18*b*), (2.18*c*) and (2.20) are used to evaluate the right-hand side of (2.21), we obtain

$$U_s\frac{\partial}{\partial X}\left\{\frac{\partial^2\phi_B^{(2)}}{\partial y^2}+2F_c\phi_B^{(2)}\right\}=-U_s\left[\frac{\partial^3 A}{\partial X^3}+2\frac{\Delta}{|\Delta|}\frac{\partial A}{\partial X}\right]\sin\pi y+4F_c\frac{\partial B}{\partial T}\sin\pi y$$
$$+\frac{r}{|\Delta|}2F_cB\sin\pi y+Q_T/|\Delta|^2-\alpha\frac{\partial A}{\partial X}\frac{\partial^3\phi_T^{(1)}}{\partial y^3}\sin\pi y$$
$$-\frac{\partial}{\partial T}\left(\frac{\partial^2}{\partial y^2}-2F_c\right)\phi_T^{(1)}-\frac{r}{|\Delta|}\frac{\partial^2\phi_T^{(1)}}{\partial y^2}$$
$$+\alpha\frac{\pi^3}{2}\sin 2\pi y\left[\frac{\partial}{\partial X}(AB)+\frac{1}{U_s}\left(\frac{\partial A^2}{\partial T}+2\frac{r}{|\Delta|}A^2\right)\right]. \tag{2.22}$$

The x average of (2.22) yields an equation for $\phi_T^{(1)}$, the correction to the x-independent stream function, i.e.

$$\frac{\partial}{\partial T}\left\{\frac{\partial^2}{\partial y^2}-\pi^2\right\}\phi_T^{(1)}+\frac{r}{|\Delta|}\frac{\partial^2\Phi_T^{(2)}}{\partial y^2}=\alpha\frac{\pi F_c}{U_s}\sin 2\pi y\left\{\frac{\partial\overline{A^2}}{\partial T}+\frac{2r}{|\Delta|}\overline{A^2}\right\}+\bar{Q}_T/|\Delta|^2, \tag{2.23}$$

where an overbar represents an average over the entire x interval. Equation (2.23) is the potential vorticity equation for the correction to the x-independent baroclinic field. The corrections are produced by both friction and the time dependence of $\overline{A^2}$ as well as by the slow change imposed by the external source of potential vorticity.

If (2.22) is now multiplied by $\sin\pi y$ and integrated in y over the interval (0, 1) we obtain

$$U_s\frac{\partial}{\partial X}\left[\frac{\partial^2 A}{\partial X^2}+2\frac{\Delta}{|\Delta|}A\right]-4F_c\frac{\partial B}{\partial T}-\frac{r}{|\Delta|}2BF_c$$
$$-2\int_0^1\frac{\{Q_T-\bar{Q}_T\}}{|\Delta|^2}\sin\pi y-\alpha\frac{\partial A}{\partial X}2\int_0^1\sin^2\pi y\frac{\partial^3\Phi_T^{(1)}}{\partial y^3}dy=0. \tag{2.24}$$

An X derivative of (2.24) in conjunction with (2.20) yields the amplitude equation for A,

$$\frac{\partial^2 A}{\partial T^2}+\frac{3}{2}\frac{r}{|\Delta|}\frac{\partial A}{\partial T}+\frac{r^2}{2|\Delta|^2}A+\frac{U_s^2}{4F_c}\left\{2\frac{\Delta}{|\Delta|}\frac{\partial^2 A}{\partial X^2}+\frac{\partial^4 A}{\partial X^4}\right\}$$
$$-\alpha\frac{U_s}{F_c}\pi\frac{\partial^2 A}{\partial X^2}\left\{2\int_0^1\sin 2\pi y\frac{\partial^2\Phi_T^{(1)}}{\partial y^2}dy\right\}=\frac{2}{|\Delta|^2}\frac{U_s}{2\pi^2}\int_0^1\frac{\partial Q_T}{\partial X}\sin\pi y\,dy. \tag{2.25}$$

Equations (2.23) and (2.25) comprise the governing equations for the disturbance field $A(X,T)$ and the alteration of the mean field $\Phi_T^{(1)}$.

Define

$$\sigma=U_s/(2F_c)^{\frac{1}{2}}=\frac{U_s}{\pi},\quad t=\sigma T,\quad \Psi=(U_s/\alpha\pi F_c)\,\Phi_T^{(1)} \tag{2.26}$$

and choose

$$\alpha=U_s/F_c. \tag{2.27}$$

Then the governing equations (2.23) and (2.25) become the neater set

$$\frac{\partial^2 A}{\partial t^2}+\tfrac{3}{2}\eta\frac{\partial A}{\partial t}+\frac{\eta^2}{2}A+\frac{\partial^2 A}{\partial X^2}\left\{\frac{\Delta}{|\Delta|}-2\int_0^1\frac{\partial^2\Psi}{\partial y^2}\sin 2\pi y\,dy\right\}+\frac{1}{2}\frac{\partial^4 A}{\partial X^4}=\frac{\partial H_1}{\partial X}, \quad (2.28a)$$

$$\frac{\partial}{\partial t}\left\{\frac{\partial^2}{\partial y^2}-\pi^2\right\}\Psi+\eta\frac{\partial^2\Psi}{\partial y^2}=\sin 2\pi y\left\{\frac{\partial\overline{A^2}}{\partial T}+2\eta\overline{A^2}\right\}+\mathscr{H}_0(y,t), \quad (2.28b)$$

where

$$\eta=r/\sigma|\Delta|=\frac{r\pi}{U_s|\Delta|}, \quad H_1(X,t)=\int_0^1\frac{Q_T}{U_s|\Delta|^2}\sin 2\pi y\,dy, \quad H_0(y,t)=\frac{\bar{Q}_T}{U_s|\Delta|^2}. \quad (2.29)$$

In (2.28a) and (2.28b) the physics of the long-wave problem is very simply exposed. The evolution of the disturbance amplitude A is governed by a *partial* differential equation since the spatial scale of A, i.e. the wave scale, contributes to the determination of the growth. The disturbance contains all scales in X embraced within the valley of the critical curve in figure 1 and the growth rate is different for each wavenumber within that valley and so a partial differential equation is the natural form for the relation between the time and space structure of the wave. The nonlinear effect in finite amplitude is solely the interaction of the wave-like disturbance with the mean field and the mean field in turn is determined only by the x average of A^2. This is vital for what follows. In spite of the fact that (2.28a, b) is a nonlinear problem the nature of the nonlinearity just described allows a rigorous Fourier decomposition of (2.28a) *without* truncation. That is, suppose $H_1(X,t)$ is represented by

$$H_1(X,t)=\frac{1}{2}\sum_k kY(k)e^{ikX}+*, \quad (2.30)$$

where $*$ denotes the complex conjugate of the preceding term. The factor k is introduced in (2.30) for subsequent convenience. $kY(k)$ then yields the amplitude of the potential vorticity source at each wavenumber. We may similarly write

$$A(X,t)=-i\sum_k A_k(t)e^{ikX}+*. \quad (2.31)$$

The sums in (2.31) are over all k in the spectrum required to represent A and H_1 and the sum may be finite or infinite depending on the particular problem.

Since

$$\overline{A^2}=\frac{1}{2}\sum_k|A_k|^2\equiv\frac{E(t)}{2}, \quad (2.32)$$

the system (2.28a, 2.28b) may be rewritten in its final form

$$\frac{d^2A_k}{dt^2}+\tfrac{3}{2}\eta\frac{dA_k}{dt}-k^2A_k\left\{q(k)-2\int_0^1\frac{\partial^2\Psi}{\partial y^2}\sin 2\pi y\,dy\right\}=-k^2Y(k,t), \quad (2.33a)$$

$$\frac{\partial}{\partial t}\left\{\frac{\partial^2}{\partial y^2}-\pi^2\right\}\Psi+\eta\frac{\partial^2\Psi}{\partial y^2}=\frac{\sin 2\pi y}{2}\left[\frac{\partial E}{\partial t}+2\eta E\right]+\mathscr{H}_0(y,t), \quad (2.33b)$$

where

$$E=\sum_k|A_k|^2, \quad q(k)=\frac{\Delta}{|\Delta|}-\frac{k^2}{2}-\eta^2/2k^2. \quad (2.34a,\ b)$$

For each wave in the spectrum an ordinary differential equation links its dynamics to the dynamics of the mean field. The mean field, in turn, is affected by the total

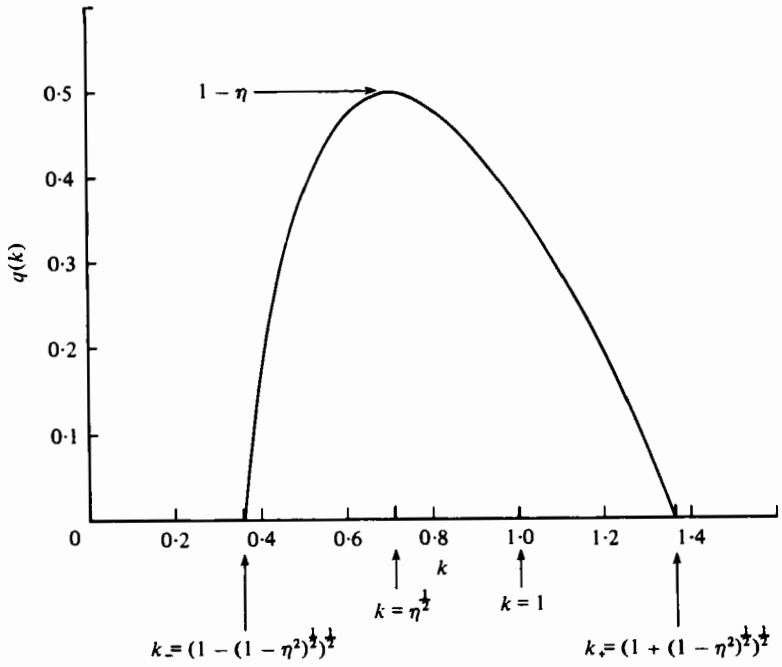

FIGURE 2. The function $q(k) = 1-\frac{1}{2}k^2-\eta^2/2k^2$. q is positive in the range

$$\{1-(1-\eta^2)^{\frac{1}{2}}\}^{\frac{1}{2}} < k < \{1+(1-\eta^2)^{\frac{1}{2}}\}^{\frac{1}{2}}$$

and achieves its maximum at $k = \eta^{\frac{1}{2}}$. The wave of maximum linear growth rate occurs at $k = 1$. The maximum value of q is $1-\eta$. This particular figure is drawn for the case $\eta = 0{\cdot}5$.

variance, E, of the wave field. Thus the waves interact with one another solely through the mean field. Each feeds on the mean and depletes its available potential energy and it is clear that the most voracious feeder will tend to crowd out less successful waves. However a precise characterization of a successful feeder in finite amplitude requires the further analysis of §3. Before moving to the next point some preliminary remarks about $q(k)$ are useful.

Consider the linearly unstable case, $\Delta/|\Delta| = 1$. Then $q(k)$, as shown in figure 2, is a positive function in the k interval

$$\{1-(1-\eta^2)^{\frac{1}{2}}\}^{\frac{1}{2}} \leqslant k \leqslant \{1+(1-\eta^2)^{\frac{1}{2}}\}^{\frac{1}{2}} \tag{2.35}$$

so that $q(k) > 0$ requires $\eta = r\pi/U_s|\Delta| < 1$. The function $q(k)$ has its maximum at

$$k_m = \eta^{\frac{1}{2}}, \tag{2.36}$$

where

$$q(k_m) = 1-\eta. \tag{2.37}$$

The condition that $q(k) > 0$ may also be written

$$\eta^{-1} > \eta_c^{-1} = \frac{1}{k(2-k^2)^{\frac{1}{2}}}, \tag{2.38}$$

which in our new units is simply (2.6a) rewritten with the shear (i.e. η^{-1}) as the critical parameter. Equation (2.38) reproduces, locally, the geometry of the valley of figure 1

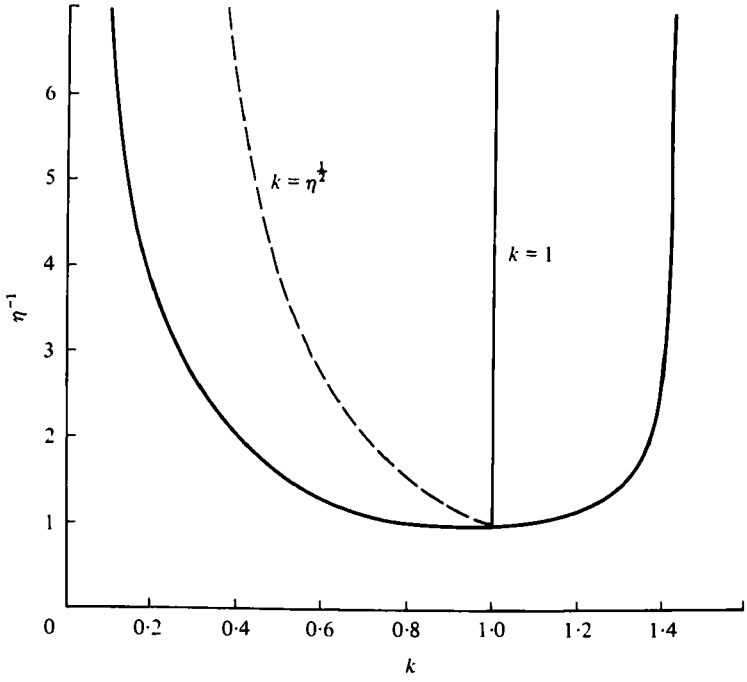

FIGURE 3. The local stability diagram $\eta_c^{-1} = \{k(2-k^2)^{\frac{1}{2}}\}^{-1}$. Note that the minimum occurs at $k = 1$, $\eta = 1$ so that at that point alone the $\eta^{\frac{1}{2}}$ wave shown with the dashed line coincides with the wave of maximum growth rate $k = 1$.

in the vicinity of $k = 0$. Figure 3 shows $\eta_c^{-1}(k)$. The dashed line is $\eta^{\frac{1}{2}}$, i.e. the locus of the k which maximizes $q(k)$. The thin line is $k = 1$. The significance of these two distinguished wavenumbers becomes clear when we consider the dynamics of free waves.

3. Free waves

In this section we begin our investigation of (2.33a, b) in the free case where $\mathscr{H}_0 = Y(k,t) = 0$. To begin with, let us first examine the *linear* problem where A_k is small enough so that E is very small and Ψ, the correction to the mean field, is small with respect to $q(k)$. Then each wave in the developing spectrum satisfies the *linear* equation

$$\frac{d^2A_k}{dt^2} + \tfrac{3}{2}\eta\frac{dA_k}{dt} - k^2q(k)A_k = 0. \tag{3.1}$$

For growth at wavenumber k, $q(k)$ must be > 0, i.e. $\Delta > 0$. The growth rate for each wavenumber k is

$$\omega(k) = -\tfrac{3}{4}\eta + \tfrac{1}{2}[\tfrac{9}{4}\eta^2 + 4k^2q(k)]^{\frac{1}{2}}, \tag{3.2a}$$

$$= -\tfrac{3}{4}\eta + \frac{1}{2}\left[\frac{\eta^2}{4} + 2k^2(2-k^2)\right]^{\frac{1}{2}}. \tag{3.2b}$$

It is clear from (3.2*a*) that the linear growth rate shares its maximum with the maximum of $k^2q(k)$ and *not* $q(k)$. This maximum occurs at $k = 1$ for all η, at which

$$\omega(1) = -\tfrac{3}{4}\eta + \tfrac{1}{2}(2+\eta^2/4)^{\frac{1}{2}}. \tag{3.3}$$

For comparison, at $k = \eta^{\frac{1}{2}}$ where $q(k)$ is maximized

$$\omega(\eta^{\frac{1}{2}}) = -\tfrac{3}{4}\eta + \tfrac{1}{2}(4\eta - \tfrac{7}{4}\eta^2)^{\frac{1}{2}}, \tag{3.4}$$

from which it follows that $\omega(1) > \omega(\eta^{\frac{1}{2}})$ for all $\eta < 1$. The two growth rates coincide only as $\eta \to 1$ where $\omega \to 0$, i.e. at the minimum critical value of η^{-1}.

The significance of the maximum of $q(k)$ is revealed most clearly by the consideration of the *steady* finite-amplitude problem for free waves. For all $0 < \eta < 1$, a steady solution exists. Whether it will be achieved as a result of arbitrary initial data remains to be seen but it is helpful to first examine the nature of the steady, free, finite-amplitude solutions of (2.33*a*, *b*).

It follows from (2.33*b*) that in a free, steady state

$$\frac{\partial^2 \Psi}{\partial y^2} = E \sin 2\pi y \tag{3.5}$$

so that in the steady state (2.33*a*) becomes

$$A_k\{q(k) - E\} = 0. \tag{3.6}$$

Although E is the total wave variance it is, after all, simply a number independent of k. Hence either $E = q(k)$ for some k or else, for that k, A_k must vanish. Since there are at most two values of k, k_1 and k_2, for a given $q(k)$, at most two waves can coexist as steady free finite-amplitude solutions. These waves have the same value of q, say q_*, and their amplitudes satisfy (with A_k real)

$$A_{k_1}^2 + A_{k_2}^2 = q_*. \tag{3.7}$$

All other A_k will be zero. Now the variance of the wave amplitude will be a maximum if E corresponds to the maximum of $q(k)$. This occurs at the *single* wavenumber $k = \eta^{\frac{1}{2}}$ for which

$$A_k = (1-\eta)^{\frac{1}{2}}; \quad k = \eta^{\frac{1}{2}}. \tag{3.8}$$

Thus the steady state with the *maximum* wave amplitude occurs as a state containing a single wave at the wavenumber $\eta^{\frac{1}{2}}$ which maximizes $q(k)$. This does *not* correspond to the wavenumber of maximum linear growth rate, $k = 1$. The amplitude which is *possible* for the steady state containing only the linearly most unstable wave ($k = 1$) is

$$A_{k=1} = q^{\frac{1}{2}}(1) = (\tfrac{1}{2}(1-\eta^2))^{\frac{1}{2}} = (\tfrac{1}{2}(1+\eta))^{\frac{1}{2}}(1-\eta)^{\frac{1}{2}} \tag{3.9}$$

or

$$A_{k=1} = (\tfrac{1}{2}(1+\eta))^{\frac{1}{2}} A_{k=\eta^{\frac{1}{2}}}, \tag{3.10}$$

so that, for all $\eta < 1$, the steady-state amplitude attainable by the most unstable wave (that is, the wave with the largest growth rate) is always *less* than the amplitude attainable by the more slowly growing wave at $k = \eta^{\frac{1}{2}}$. Note that, since $\eta < 1$, this latter wave, which maximizes the steady-state variance, is always *longer* than the wave of maximum linear growth rate.

Thus we have two distinguished waves. One, at $k = 1$, has the largest growth rate but a smaller steady-state amplitude than the wave at $k = \eta^{\frac{1}{2}}$, which has the largest possible amplitude in the steady state. Which of these two, if either, will be realized in finite amplitude? By (3.6) they cannot coexist.

We can approach this question by examining the stability of each of the possible steady states to a small disturbance consisting of background noise of the other waves. Only a wave which is stable to small disturbances can be realized as the asymptotic steady state.

Suppose our putative steady-state wave has wavenumber k_s. Then

$$A_{k_s}^2 = A_s^2(k_s) = E_s(k_s) = q(k_s). \tag{3.11}$$

When the steady state is perturbed the amplitude at k_s will be

$$A_{k_s} = A_s(k_s) + a_{k_s}(t) \tag{3.12}$$

while, for all $k \neq k_s$,

$$A_k = a_k(t), \tag{3.13}$$

where the $a_k(t)$ represent the small-disturbance amplitudes.

Now in the disturbed state

$$\Psi(y,t) = \Psi_s(y) + \Psi'(y,t), \tag{3.14}$$

where

$$\int_0^1 dy \sin 2\pi y \frac{\partial^2 \Psi_s}{\partial y^2} = E_s(k_s) = q(k_s). \tag{3.15}$$

For all $k \neq k_s$, the linearized version of (2.33a) (with $Y(k) = 0$) becomes

$$\frac{d^2 a_k}{dk^2} + \tfrac{3}{2}\eta \frac{da_k}{dt} - k^2\{q(k) - q(k_s)\} a_k = 0 \tag{3.16}$$

if (3.11) and (3.13) are used. If $q(k) > q(k_s)$ the amplitude of the disturbance at wavenumbers $k \neq k_s$ will grow. In that case the wave at wavenumber k_s will not be stable. *The only steady, linearly stable wave is the wave of maximum* $q(k)$. Thus, if a *steady state* is attained the realized wave will not be the wave with the maximum linear growth rate. Rather the realized wave will be the wave of maximum amplitude, i.e. at the wavenumber of maximum $q(k)$.

Of course it is possible that in these circumstances the final state may not be steady so that the considerations just described are not applicable. That is, the wave of maximum amplitude may be stable to perturbations only in some small neighbourhood of the steady solution but unstable to initial conditions which start a considerable distance from it. Or, for small enough η, the final states may be vacillatory or aperiodic, in which case the selection principle for wavenumber based on the presumption of final steadiness is irrelevant. In addition, only a small subset of k in the unstable interval might be allowed. If we think again of waves in an annulus the x wavenumber will be quantized and the attainable values of k may 'miss' both $k = 1$ and $k = \eta^{\frac{1}{2}}$. In this case we would suspect from (3.16) that the allowable wavenumber with the largest $q(k)$ will be obtained even though $k = \eta^{\frac{1}{2}}$ may not be allowed. Is this the case?

To examine the general time-dependent problem numerical integrations of (2.33a, b) were performed as follows.

First define

$$u(y,t) = -\frac{\partial \Psi}{\partial y}, \tag{3.17}$$

where $u(y,t)$ is the correction to the mean thermal wind. It must vanish at $y = 0, 1$ by (2.4), which is accomplished by expanding $u(y,t)$ in the Fourier sine series

$$u = \sum_{n=1}^{N} u_n(t) \sin(2n-1)\pi y. \tag{3.18}$$

Only the sines of arguments with odd integral multiples of πy are required by (2.33b). The further convenient partition

$$u_n(t) = \frac{4(2n-1)}{\pi^2((2n-1)^2-4)[(2n-1)^2+1]}\{E + V_n(t)\} \tag{3.19}$$

allows the unforced version of (2.33a, b) to be written

$$\frac{dA_k}{dt} = B_k - \eta A_k, \tag{3.20a}$$

$$\frac{dB_k}{dt} = -\frac{\eta}{2}B_k + \frac{\eta^2}{2}A_k + k^2 A_k\left\{q(k) - \frac{2}{\pi^2}\sum_n \frac{(n-0{\cdot}5)^2[E+V_n]}{[n(n-1)+0{\cdot}5][(n-0{\cdot}5)^2-1]^2}\right\}, \tag{3.20b}$$

$$\frac{dV_n}{dt} = \eta\left[\frac{E\{(2n-1)^2+2\} - (2n-1)^2 V_n}{(2n-1)^2+1}\right], \tag{3.20c}$$

where

$$E = \sum_k A_k^2.$$

The sum in (3.18) should, in principle, be an infinite sum. Considering the rapid convergence of the sum in (3.20b), I found that taking twelve terms in (3.18) sufficed (i.e. up to the $\sin 23\pi y$ term).

At $t = 0$, $A_k(0)$ and $dA_k(0)/dt$ are specified, and, since $u(y,0) = 0$,

$$V_n(0) = -E(0). \tag{3.21}$$

If both A_k and dA_k/dt are zero at $t = 0$ for some k, then that A_k will remain zero. Hence the number of waves to be carried in the calculation depends entirely on the initial data. In the calculations reported here either one, two, or three waves are considered.

Recent results by Pedlosky & Frenzen (1979) show that the single wave form of (3.20a, b, c) exhibits a rich range of behaviour as a function of η. The dependence on η, especially when η is small, is quite complex and sensitive to tiny changes in η. In this paper only a relatively few values of η are examined and so it must be borne in mind that there are undoubtedly some aspects of the transient dynamics left unexplored. Since for these long waves we have chosen r to be quite small ($O(\Delta)$) most of our attention will be focused on the situation where η is $O(1)$ (but less than unity).

Figure 4 shows the result of the equilibration of a *single* wave with $k = 0{\cdot}5$ at $\eta = 0{\cdot}5$. The equilibration occurs relatively swiftly, at $t \sim 40$, and the final value, $A_s = 0{\cdot}6123$, is precisely $\sqrt{q(0{\cdot}5)}$. Figure 5 shows the evolution of a three-wave spectrum. Each amplitude initially is set at $A_k = 0{\cdot}1$ with $dA_k/dt = 0$. The three wavenumbers (k_1, k_2, k_3) are $(1, 0{\cdot}75, 0{\cdot}5)$. We see from the figure that the most unstable wave A_{k_1}

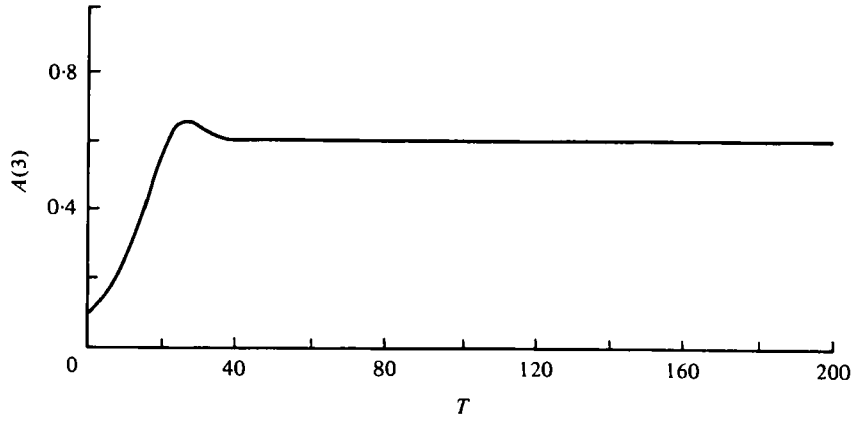

FIGURE 4. The equilibration of a single finite-amplitude wave. $\eta = 0{\cdot}5$, $k = 0{\cdot}5$. The final amplitude is 0·612, i.e. $(q(k))^{\frac{1}{2}}$.

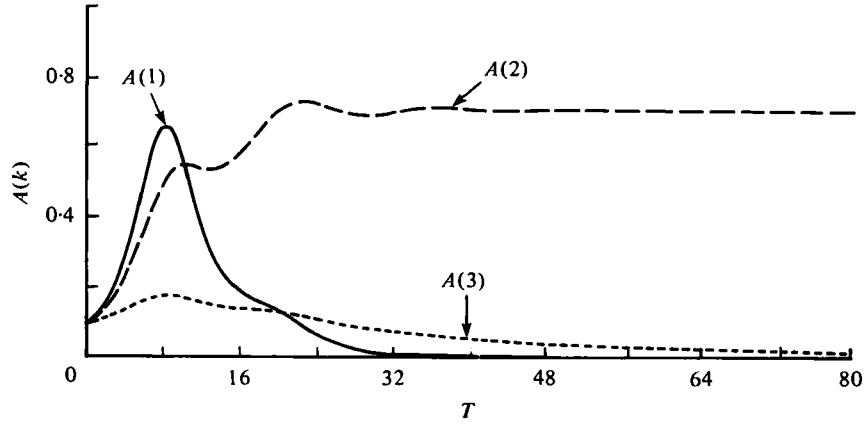

FIGURE 5. The amplitude history of the three-wave spectrum $(k_1, k_2, k_3) = (1, 0{\cdot}75, 0{\cdot}5)$ corresponding to $q_k = (0{\cdot}375, 0{\cdot}496, 0{\cdot}375)$. The $k = 1$ wave has the largest linear growth rate and initially attains the maximum amplitude. The suviving wave is the second wave, i.e. the wave of highest q, and the other two waves eventually vanish.

initially dominates the spectrum. It grows most rapidly and until $t \sim 10$ (about 10 linear e-folding times) maintains the largest amplitude. However, beyond $t \sim 10$ the second wave, A_{k_2}, dominates the amplitude response and the amplitude of the most unstable wave plunges to zero and effectively vanishes after $t \sim 30$. The third wave, with $k = 0{\cdot}5$, which was the equilibrated wave of figure 4, rises slowly above its initial amplitude and then slowly declines toward zero. The $k = 1$ wave is the most unstable wave according to linear theory, i.e. has the largest growth rate. Its value of q, $q(1)$ is 0·375 for this η. The wavenumber of maximum q, $\eta^{\frac{1}{2}}$, is 0·7071. The second wave, with $k = 0{\cdot}75$, is the closest to this peak and has a q of 0·4965 which exceeds that of the first wave.

The third wave is the *conjugate* wave to the first. That is, it is the wave with the *same* value of q as the first wave. It is a simple matter to show from (2.34*b*) that the wavenumbers of two conjugate waves, say k_1 and k_3, must satisfy

$$k_1 k_3 = \eta. \tag{3.22}$$

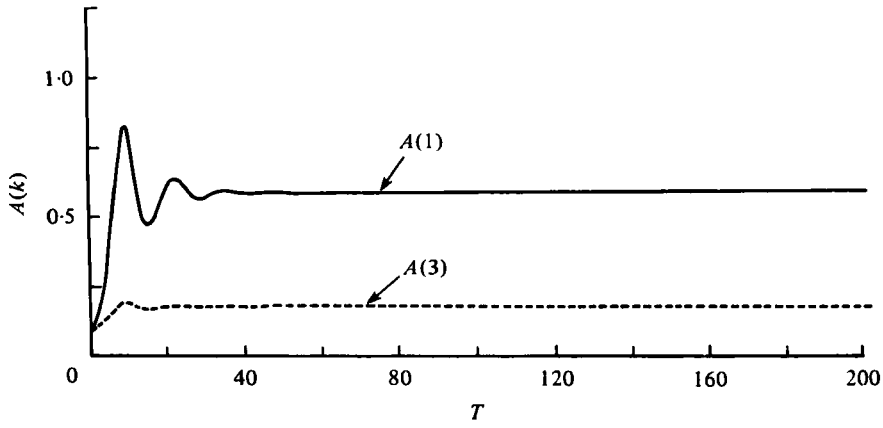

FIGURE 6. The amplitude history of a two-wave spectrum in which the two waves are conjugates, i.e. they have the same q. The end state is a mixed wave state whose relative proportions depend on the initial data. In the case shown, $\eta = 0{\cdot}5$, $k_1 = 1$, $k_2 = 0{\cdot}5$, $A_1(0) = A_2(0) = 0{\cdot}1$, while $dA_1(0)/dt = dA_2(0)/dt = 0$.

The conjugate wave at k_3 has the same q as the wave at k_1, but its growth rate is less (since the growth rate depends on $k^2q(k)$). Hence it never achieves even the transient prominence that A_{k_1} does. Note, however, that, during the decay phase of A_{k_1} and A_{k_3}, the third wave survives longer. Nevertheless, in the end the sole survivor of the competition for the available potential energy of the mean flow is A_{k_2}, which is the wave in the spectrum with the largest $q(k)$. Note that it is not necessary that k_2 should precisely equal $\eta^{\frac{1}{2}}$, only that its $q(k_2)$ should exceed all other available $q(k)$. The results of this calculation are typical of many numerical integrations with different initial conditions and different η as long as η is $O(1)$. In each case, when equilibrium is attained the sole surviving wave is the wave of maximum q and *not* the most unstable wave.

Consider now the case where, perhaps because of quantization of k, the wavenumber of maximum q is not attainable and only two waves at k_1 and k_3 are allowed to the unstable spectrum. The dominant wave in equilibrium will again be the wave of maximum q and whether it is the wave with k_1 and k_3 will of course depend on η. Suppose η is altered from case to case until finally η takes on the value given by (3.7) so that k_1 and k_3 are conjugate waves. Then we have no prediction from (3.16) as to which wave will survive, while the result of the steady-state theory (3.7) predicts only the total variance of the two-wave field and not the partition of amplitude between them. Figure 6 shows the result of a numerical calculation with two waves for which, at $t = 0$, $A_{k_1} = A_{k_3} = 0{\cdot}1$ for $k_1 = 1$, $k_3 = 0{\cdot}5$, and $\eta = 0{\cdot}5$ so that the two waves are conjugates. The end state is now a *mixed* wave state with both k_1 and k_3 present. The larger final amplitude is $A_{k_1} \sim 0{\cdot}57$ while $A_{k_3} \sim 0{\cdot}22$, which satisfies (3.7) with $q = 0{\cdot}375$. Hence in this case the dominant final wave is the most unstable wave since we have restricted attention to a spectrum where the $k = \eta^{\frac{1}{2}}$ is disallowed and the conjugate waves have the highest q. Extensive numerical calculations show for the conjugate wave case that the partition of the final variance between the two conjugate waves is a function of initial conditions, e.g. of the ratio $A_{k_1}(0)/A_{k_3}(0)$. It is easy to imagine a situation where, for a range of η and a quantization of k, only k_1 and k_3 are allowed. Then if η does not satisfy (3.22) either k_1 and k_3 will be the realized wave.

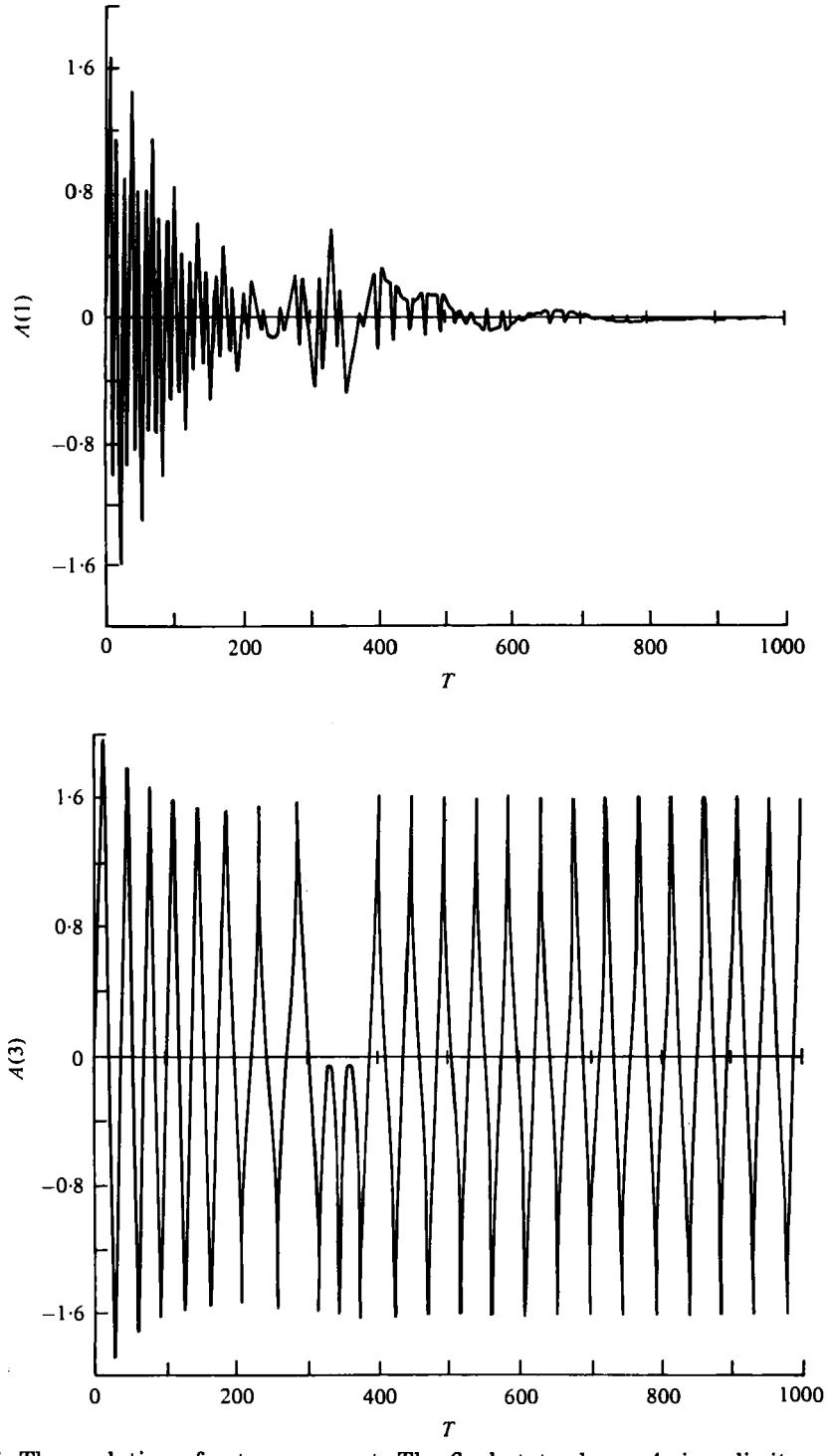

FIGURE 7. The evolution of a two-wave set. The final state shows A_3 in a limit cycle while $A_1 \to 0$. In this case $\eta = 0{\cdot}01$, $k_1 = 1$ (corresponding to the most unstable wave, $q_1 = 0{\cdot}49995$) and $k_3 = 0{\cdot}5$ (corresponding to a $q_3 = 0{\cdot}8748$). Although a steady state is not attained the asymptotic state contains only the wave of higher q. (*a*) $A_1(t)$; (*b*) $A_3(t)$.

Which of the two is realized depends on $q(k_1)$ and $q(k_3)$. Suppose $k_1 > k_3$. Since

$$q(k_3) - q(k_1) = \tfrac{1}{2}(k_1^2 - k_3^2)\left[1 - \left(\frac{\eta}{k_1 k_3}\right)^2\right], \tag{3.23}$$

$q(k_3)$ will exceed $q(k_1)$ if $\eta < k_1 k_3$ while $q(k_1)$ will exceed $q(k_2)$ if $\eta > k_1 k_3$. As η approaches the value $k_1 k_3$ the waves become conjugate but the final amplitudes will depend on the amplitudes present in the initial state and this will depend on whether η has been previously larger or smaller than $k_1 k_3$. Thus even in the free wave case the presence of conjugate waves implies a hysteresis effect in the observed wave field depending on the direction of change of η.

When η is decreased to much smaller values the adjustment time required to achieve equilibrium increases. I have not here sought to precisely define the value of η required to achieve complete, perpetual unsteadiness. Preliminary calculations indicate that this must, to some as yet undetermined extent, also depend on the spectrum of wavenumbers allowed and the initial data. However, figure 7 shows a numerical integration of a two-wave case with $\eta = 0{\cdot}01$ and $k_1 = 1$ and $k_3 = 0{\cdot}5$. Again the k_1 wave is the most unstable according to linear theory. Since $\eta < k_1 k_3$, it follows from (3.23) that $q(k_3) > q(k_1)$. The calculations were continued to $t = 1000$. A steady state was not attained. Rather, an apparent limit cycle behaviour for A_{k_3} was realized while A_{k_1}, after considerable oscillation, *vanished*. Thus *even in the absence of final steady states the surviving wave is the wave of maximum* $q(k)$, *a result that could not be predicted on the basis of the theory leading to* (3.16). So far, these results are typical of all the numerical calculations conducted.

This inescapably suggests that although a single wave theory may be relevant for the study of the finite-amplitude dynamics of baroclinic waves the appropriate single wave to consider is *not* the most linearly unstable wave. *Rather the wave to consider is the wave which, were it alone, could achieve the maximum steady-state amplitude.* This criterion is unfortunately a nonlinear one. However it is important to recognize that it is a criterion that can be applied to the result of *single* wave calculations. In short, consider the finite-amplitude dynamics of a wave of arbitrary wavelength. Find its steady amplitude. Maximize it with respect to wavelength. The heuristic theory developed here indicates that that wave will be the realized wave in finite amplitude even if the asymptotic state is unsteady.

Thus a linear instability calculation is not, by itself, a good predictor of the finite-amplitude observed wave either as to its wavelength or vertical structure. Any deficiency of the predictions of linear theory based on a selection principle of maximum growth rate does not imply that the wave field does not spring from a baroclinic instability. It only means that the *selection* principle for the wave is fundamentally nonlinear. The linear selection principle will be valid only *initially*, or, of course, in the limit where the supercriticality goes to zero so that only *one* wave is unstable. In that limiting case and in only that case the wave of maximum q and the wave of maximum growth rate coincide.

4. Free waves with time-dependent zonal heating

The adjustment time to mark equilibrium, especially for $\eta \ll 1$, is quite long; even at $\eta = 0{\cdot}5$ in the preceding examples it requires about 40 linear e-folding times to reach

steady state. With an atmospheric application in mind it is natural to ask how long-term changes in the store of available potential energy affect the picture developed in the preceding section. How will changes in the x-independent vertical shear produced by, say, seasonal changes affect the finite-amplitude wave state?

To study this problem it is necessary to return to (2.33a, b) and examine the nature of its solutions in the case when $\mathscr{H}_0 \neq 0$. Now a constant value of $\mathscr{H}_0$ can be absorbed directly in the supercriticality. This is a simple fact left for the reader to derive. Hence only time-dependent $\mathscr{H}_0$ with zero time average need be considered. It is somewhat less obvious, but also true, that the effect of time-dependent $\mathscr{H}_0$ is equivalent to the consideration of an *effective* $q(k)$ which is time dependent. Although this can be shown in general it is sufficient for the purpose of this paper to consider forcing fields such that

$$\mathscr{H}_0 = Z_0 \sin B_0 t \cos \pi y. \tag{4.1}$$

Then the solution for Ψ may be written

$$\Psi = \frac{-Z_0}{2\pi^2(B_0+\eta^2/4)}\left[\tfrac{1}{2}\eta \sin B_0 t - B_0 \cos B_0 t + B_0 e^{-\eta/2t}\right]\cos \pi y + \Psi_h(y,t), \tag{4.2}$$

so that (2.33a, b) may, in turn, be rewritten

$$\frac{d^2A_k}{dt^2} + \tfrac{3}{2}\eta \frac{dA_k}{dt} - k^2 A_k \left\{\tilde{q}(k,t) - 2\int_0^1 \frac{\partial^2 \Psi_h}{\partial y^2} \sin 2\pi y\right\} = -k^2 Y(k,t), \tag{4.3a}$$

$$\frac{\partial}{\partial t}\left\{\frac{\partial^2}{\partial y^2} - \pi^2\right\}\Psi_h + \eta \frac{\partial^2 \Psi_h}{\partial y^2} = \frac{\sin 2\pi y}{2}\left[\frac{\partial E}{dt} + 2\eta E\right], \tag{4.3b}$$

where

$$\tilde{q}(k,t) = \frac{\Delta}{|\Delta|} - \frac{k^2}{2} - \frac{\eta^2}{k^2} - \frac{4}{3\pi}\frac{Z_0}{(B_0^2+\eta^2/4)}\left\{\frac{\eta}{2}\sin B_0 t - B_0 \cos B_0 t + B_0 e^{-\eta/2t}\right\}. \tag{4.4}$$

Thus the problem with variable zonal heating can conceptually be directly related to the problem with no zonal forcing but simply a time-varying supercriticality $\tilde{\Delta}(t)$

$$\tilde{\Delta}(t) = \Delta - |\Delta| \frac{4}{3\pi}\frac{Z_0}{(B_0^2+\eta^2/4)}\left\{\frac{\eta}{2}\sin B_0 t - B_0 \cos B_0 t + B_0 e^{-\eta/2t}\right\}. \tag{4.5}$$

The virtue of considering the problem in this light is as follows. The wavenumber which maximizes $\tilde{q}(k,t)$ is *again* $k = \eta^{\frac{1}{2}}$. This feature is unchanged by the time dependence, i.e.

$$\tilde{q}(k,t) = \frac{\tilde{\Delta}(t)}{|\Delta|} - \frac{k^2}{2} - \eta^2/2k^2, \tag{4.6}$$

whose dependence on k is the same as that of $q(k)$. However, the wavenumber of quasi-steady growth rate will depend on $\tilde{\Delta}(t)$. That is, the maximum of $k^2\tilde{q}(k,t)$ will occur at

$$k_\omega(t) = \left\{\frac{\tilde{\Delta}}{|\Delta|}\right\}^{\frac{1}{2}}. \tag{4.7}$$

Of course when $\tilde{q}$ is time dependent the simple notions of linear exponential growth are not valid, but if $\tilde{\Delta}(t)$ changes slowly enough we can anticipate that linear theory would suggest $k_\omega(t)$ as a quasi-steady scale predictor. It should be noted that $k_\omega(t)$ can only coincide with $\eta^{\frac{1}{2}}$ if, as before, the supercriticality at k_ω tends to zero, i.e. only when $\tilde{q}(k_\omega, t) \to 0$, and then all the waves are stable.

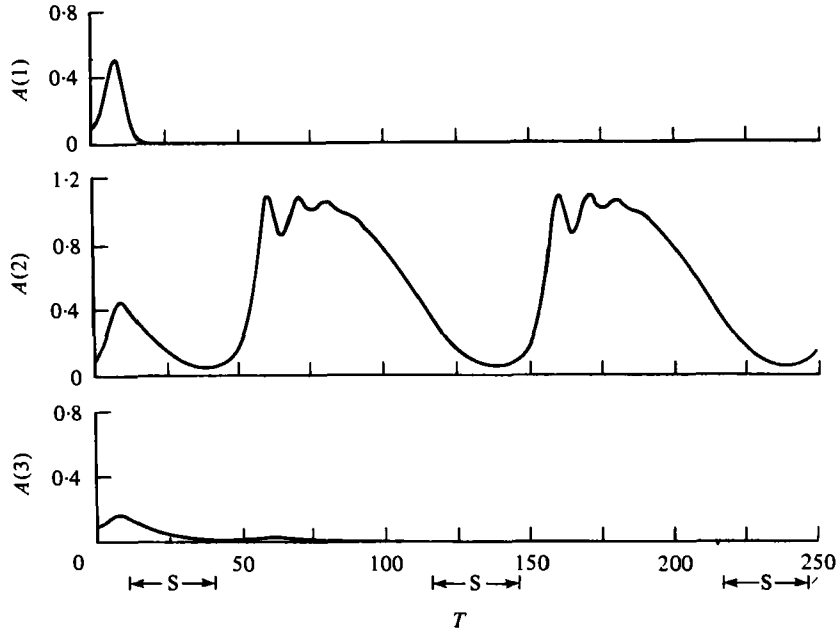

FIGURE 8. The evolution of the three waves with wavenumbers (1, 0·75, 0·5) at $\eta = 0{\cdot}5$ subject to periodic supercriticality. The asymptotic state again contains only the highest q wave. The intervals of quasi-steady stability are indicated by the symbol S. Note the steep rise after the onset of instability and the following slow decline. Upper panel, $A_1(t)$; middle panel, $A_2(t)$; lower panel, $A_3(t)$.

To examine the nature of the response of free waves to variable zonal heating numerical integrations of (2.33a, b) were carried out with $Y(k,t) = 0$ and $\mathscr{H}_0$ given by (4.1). This is equivalent, naturally, to integrating (4.3a, b). Again, the same spectral formulation was applied to (4.3b), whose spectral form becomes

$$\frac{dV_n}{dt} = \eta\left[\frac{E\{(2n-1)^2+2\}-(2n-1)^2V_n}{(2n-1)^2+1}\right] + \delta_{n1}(2{\cdot}35619)\,Z_0 \sin B_0 t, \tag{4.8}$$

while A_k is calculated from (3.20a, b). In (4.8)

$$\begin{aligned} \delta_{n1} &= 1, \quad n = 1, \\ &= 0, \quad n > 1. \end{aligned} \tag{4.9}$$

Figure 8 shows the response of a three-wave spectrum, where

$$(k_1, k_2, k_3) = (1, 0{\cdot}75, 0{\cdot}5),$$

for the case $\eta = 0{\cdot}5$, $B_0 = 2\pi/100$, and $Z_0 = 0{\cdot}35$. The period of the oscillation of $\mathscr{H}_0$ is thus about 100 linear e-folding times. As figure 8 demonstrates the sole surviving wave in finite amplitude is the wave of largest $\tilde{q}$, which again occurs for $k = 0{\cdot}75$. The other two waves, including the wave of maximum linear growth rate, appear only fleetingly and then forever disappear. With the value of Z_0 chosen, all three waves have intervals during which the supercriticality is negative. That is, wave 2, the survivor, has periods during which, according to a quasi-steady theory, it would be stable,

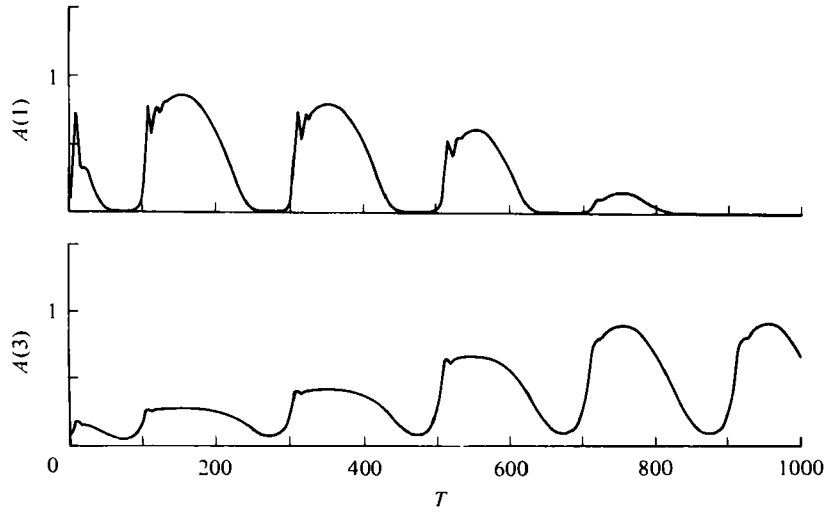

FIGURE 9. The evolution of two conjugate waves $k_1 = 1$, $k_3 = 0{\cdot}5$, $\eta = 0{\cdot}5$ subject to periodic supercriticality. The wave with the larger linear growth rate asymptotically vanishes. Upper panel, $A_1(t)$; lower panel, $A_3(t)$.

i.e. intervals over which $\tilde{q}(k_2, t) < 0$. For the dominant k_2 wave this interval exists as an epoch centred around the peaks of $\mathscr{H}_0$, i.e. at

$$\frac{t}{100} = (j + \tfrac{1}{4}), \quad j = 0, 1, 2, \ldots, \tag{4.10}$$

while the interval around the peak of quasi-stability is

$$\frac{\Delta t}{100} \sim 0{\cdot}24. \tag{4.11}$$

Thus, the basic state is 'stable' roughly one quarter of the overall cycle. The stable intervals are indicated in figure 8 in the graph for A_{k_3}. After an initial transient period a forced, nonlinear oscillation in the dominant amplitude occurs, whose period is identical to the forcing period. The response is very strongly skewed in time. Directly after a period of quasi-steady stability gives way to instability there is a sharp rise in the amplitude. It swiftly reaches its equilibrium amplitude whose maximum is well predicted by the maximum value of $\{\tilde{q}(k_2, t)\}^{\frac{1}{2}}$ during the cycle. As $\tilde{q}$ declines slowly, A_{k_2} similarly slowly declines and follows the quasi-steady equilibrium value quite closely. It reaches its minimum during the stable period and then responds swiftly to the recurrence of instability to renew the 'seasonal' cycle. It is once again true that a single-wave theory is apt for the asymptotic wave dynamics but it is also true again that the selection principle picks out the wave of largest $q(k)$ and not the wave of greatest growth rate.

A particularly interesting phenomenon occurs when two conjugate waves experience a changing supercriticality. The reader will recall that when $\mathscr{H}_0$ is zero the end result of the calculation, as shown in figure 6, is a mixed wave state. When the initial amplitudes of the two waves are equal the final mixed state is dominated by the more linearly unstable wave. Figure 9 shows the result of the same parameter setting with the addition of variable zonal heating. In this case $Z_0 = 0{\cdot}2644$, $B_0 = 2\pi/200 = 0{\cdot}0314$,

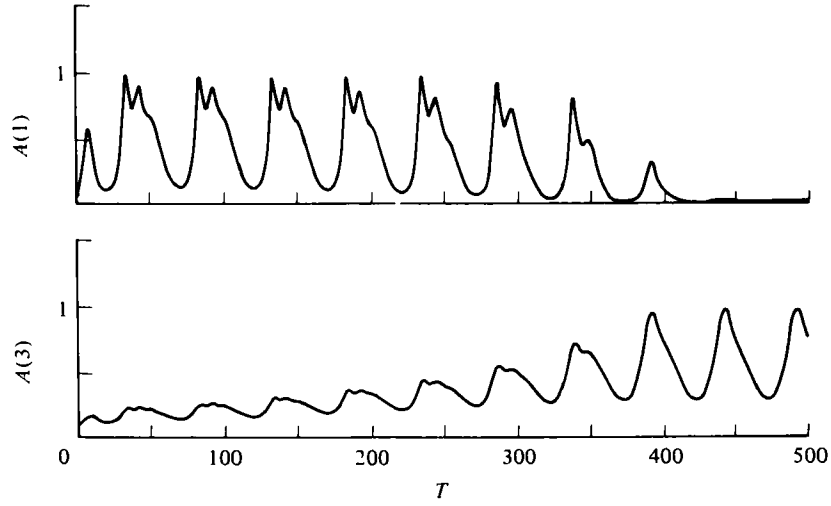

FIGURE 10. As in figure 9 but the interval of subcriticality is briefer and the 'seasonal' period is shorter. Upper panel, $A_1(t)$; lower panel, $A_3(t)$.

$\eta = 0{\cdot}5$ and the conjugate waves have wavenumbers $(k_1, k_3) = (1, 0{\cdot}5)$. With these values of Z_0 and B_0 the criticality is negative during roughly 18% of the cycle (whose period is 200). We observe that slowly, over several cycles, the most unstable wave disappears and the final state consists of the quasi-steady forced oscillation of the conjugate wave with the *lower* growth rate. This result is typical of the calculations over a considerable range of Z_0 and B_0 for $\eta = 0{\cdot}5$ as long as there exists at least a brief period of subcriticality (i.e. $\tilde{\Delta} < 0$). Figure 10 shows another example for which Z_0 is less (0·223) and B_0 greater (0·1256) so that the interval of stability is about 3 to 4 times briefer. The same qualitative result reappears.

The explanation for this rather remarkable fragility of the more unstable of the conjugate wave to the occurrence of subcriticality is more clearly displayed in figure 10. In periods of instability there is a rapid rise in amplitude and the time scale is determined by the intrinsic instability dynamics. After obtaining quasi-equilibrium the waves begin their decline during the stable phase at the slower rate determined by the time scale of $\mathscr{H}_0$. It is apparent in figure 10 that the wave with the larger $k^2q(k)$ is more sensitive to the decay phase. At each subcritical period its amplitude plunges to smaller and smaller values and never fully recovers during the growth period. The more sluggish wave declines more slowly and is better able to rebound during the growth phase of the cycle. We can make this argument more quantitative by exploiting the relative slowness of the time changes during the declining portion of the cycle. For the two conjugate waves whose q's are equal, (4.3a) may be written during the decline period approximately as (with $Y_k = 0$),

$$\tfrac{3}{2}\eta \frac{dA_{k_1}}{dt} + k_1^2 A_{k_1}\left\{-\tilde{q}(k_1) - 2\int_0^1 dy \frac{\partial^2 \Psi_h}{\partial y^2}\sin 2\pi y\right\} = 0, \tag{4.12a}$$

$$\tfrac{3}{2}\eta \frac{dA_{k_2}}{dt} + k_2^2 A_{k_2}\left\{-\tilde{q}(k_2) - 2\int_0^1 dy \frac{\partial^2 \Psi_h}{\partial y^2}\sin 2\pi y\right\} = 0. \tag{4.12b}$$

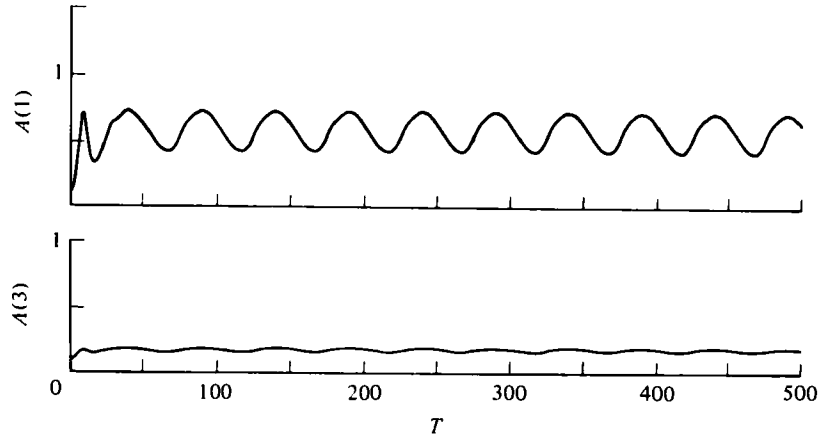

FIGURE 11. As in figures 9 and 10 but in this case the supercriticality is always positive over the cycle and both conjugate waves are present in the asymptotic state. Upper panel, $A_1(t)$; lower panel, $A_3(t)$.

Since $\tilde{q}(k_1) = \tilde{q}(k_2)$ the ratio of the two equations yields

$$\frac{1}{k_1^2}\frac{d}{dt}\ln A_{k_1} = \frac{1}{k_2^2}\frac{d}{dt}\ln A_{k_2} \tag{4.13}$$

or

$$\frac{A_{k_1}(t)}{A_{k_1}(t_m)} = \left\{\frac{A_{k_2}(t)}{A_{k_2}(t_m)}\right\}^{(k_1/k_2)^2}, \tag{4.14}$$

where $A_k(t_m)$ is the amplitude of the k wave at its maximum. Recall that (4.12a, b) and (4.14) are presumed valid *only* during the slow declining phase of the amplitude, in which case, for each wave, $A_k/A_k(t_m)$ is a number less than one. Since $k_1 > k_2$, (4.14) shows that the wave with the larger k will decline more strongly with respect to its maximum. That is in each decay cycle $A_{k_1}/A_{k_1}(t_m) \ll A_{k_2}/A_{k_2}(t_m)$ at the minimum. As noted above this will continue until the larger k wave is too small to recover in the growth cycle. This rationalization is perhaps made more convincing by the calculation shown in figure 11. In this case $Z_0 = 0{\cdot}1$, $B_0 = 2\pi/50$ and the criticality is always > 0. The end state is a periodic, *mixed* state wherein both conjugates can now exist.

When η is reduced the dynamical memory of the system increases and residual perturbations from a previous epoch can persist to grow up again. Figure 12 shows a two-wave case for *constant* supercriticality for $\eta = 0{\cdot}077$. Note that about 100 e-folding times are required to expunge A_k ($k = 1$) while A_k ($k = 0{\cdot}5$) equilibrates slowly after many oscillations whose 'natural' period is about 30 t units.

Figure 13 shows the very striking result attained with the same two, *non-conjugate*, waves in the presence of variable supercriticality. In this example $Z_0 = 0{\cdot}223$ and $B_0 = 0{\cdot}052$ so that the period of the 'seasonal' cycle is about four times the natural period of A_k ($k = 0{\cdot}5$) deduced from figure 12. For the length of the calculation (500 t units) the results indicate an *aperiodic* mixed wave response. The wave with the larger q, $k = 0{\cdot}5$, is dominant and as it grows during the unstable epochs the wave with the larger growth rate declines. The latter does not completely disappear since η is so small and when the period of growth recurs it briefly springs up. It never reaches the amplitude of the higher q wave but it also weakly persists. The advent of the unstable period occurs regularly but the amplitudes and phases of the A_k's still remember the

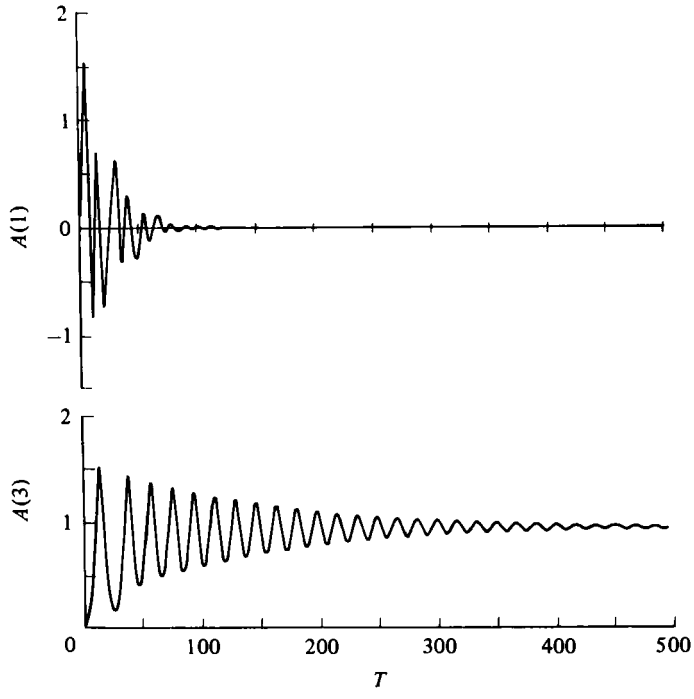

FIGURE 12. The evolution of two waves. $\eta = 0{\cdot}077$, $k_1 = 1$ (the most unstable) and $k_3 = 0{\cdot}5$. The supercriticality is constant (i.e. $Z_0 = 0$). The final state contains only $A(k_3)$. Upper panel, $A_1(t)$; lower panel, $A_3(t)$.

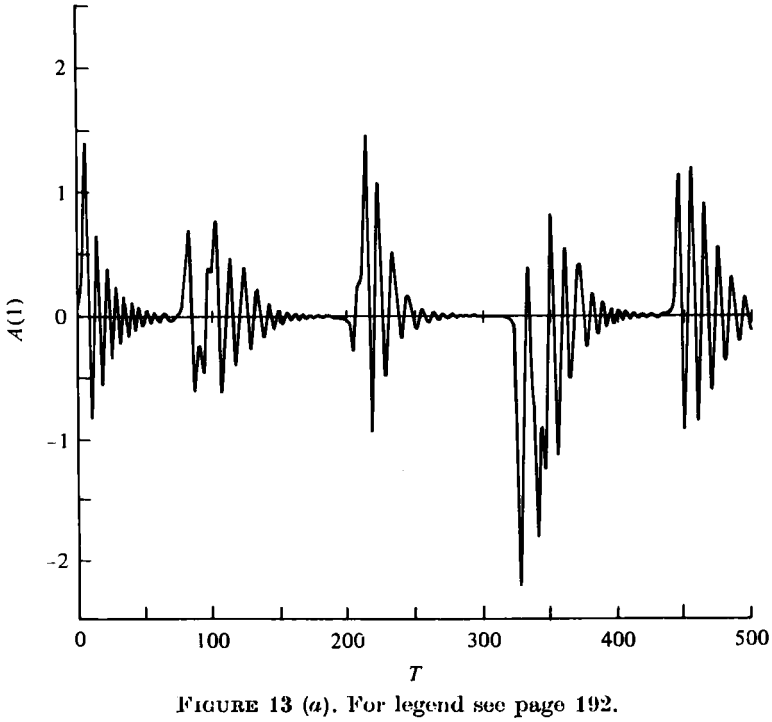

FIGURE 13 (*a*). For legend see page 192.

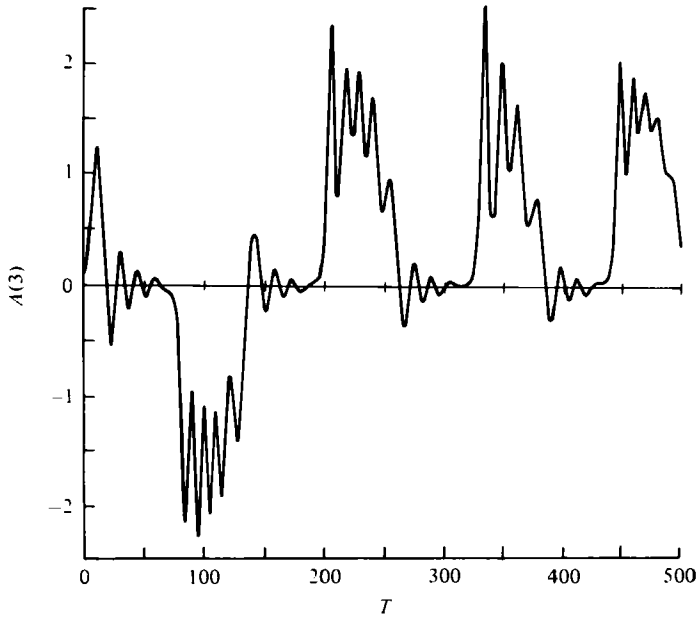

FIGURE 13. The *aperiodic* response of the two waves A_1, A_3 with wavenumbers $k_1 = 1$, $k_3 = 0{\cdot}5$. $Z_0 = 0{\cdot}223$ and the forcing period is 120 e-folding times. $\eta = 0{\cdot}077$. Contrast this aperiodicity under slow periodic forcing with the completely predictable asymptotic state in figure 12. (*a*) $A_1(t)$, (*b*) $A_3(t)$.

previous cycle, and are each time slightly altered. This gives effectively different initial conditions for each growth period and hence the consequent aperiodicity. It is important to stress the consistency of these results and the $\mathscr{H}_0 = 0$ calculations for which the wave of smaller $q(k)$ vanished. The same process is at work here but the combination of small dissipation and variable supercriticality prevents the attainment of complete asymptotic equilibration. A calculation for the identical parameters but with B_0 *increasing* so that the seasonal period is reduced by a factor of 4 to about 30 t units leads to the rapid disappearance of the $k = 1$ wave. In that case the frequency dependence of $\tilde{\Delta}$ (equation (4.5)) reduces the amplitude of the time-variable portion of $\tilde{\Delta}$ and the qualitative results of the steady problem re-emerge. It would be of interest to examine the parameter dependence of the small η cases more completely to determine the regions of mixed, aperiodic wave states but that parameter study is not yet completed. It is also possible that after a very long time the single-wave state may re-emerge.

5. Forced waves

In the previous sections it became clear that there was a natural tendency for the wave spectrum of free waves to sharpen around the wave with maximum $q(k)$. Indeed, in most cases all the waves except the wave (or conjugate waves) with the largest accessible $q(k)$ eventually vanished. This leads naturally to single- or at most double-wave end states with a rather simple dynamical structure.

However, when wave forcing exists, i.e. when each wave in the spectrum is forced to be non-zero, the dynamics becomes inherently multi-waved. The behaviour of the

multi-wave response to the forcing is rather complex because the amplitude response depends on the total wave variance E and this nonlinearity introduces the possibility of several states of response for the same forcing.

This can simply be seen by first examining the *steady* forced problem. With Y_k independent of time and $\mathscr{H}_0$ equal to zero steady solutions are possible. For these steady solutions (3.5) will still apply, but now the steady version of (2.33a) yields, instead of (3.6),

$$A_k(q(k)-E) = Y_k \tag{5.1}$$

for each wavenumber k. Since $E = \sum_k A_k^2$ equation (3.9) is a coupled set of algebraic equations for all the waves. Momentarily let us assume that $Y_k/(q(k)-E)$ is always bounded (as it turns out to be). Then solving (5.1) for A_k yields

$$A_k = Y_k/(q(k)-E), \tag{5.2}$$

so that

$$E = \sum_k \frac{Y_k^2}{(q(k)-E)^2}, \tag{5.3}$$

which is a single equation for E. Once E is determined each A_k is found from (5.2). Note that, if $Y_k = 0$, A_k can differ from zero *only* if $E = q(k)$.

It is helpful to write

$$Y_k = R_k Y \tag{5.4}$$

so that R_k gives the projection of the forcing on each wavenumber. It is illuminating to first consider (5.3) as an equation for Y in terms of E. The basic structure is already revealed in the three-wave case, i.e. when only R_{k_1}, R_{k_2} and R_{k_3} are different from zero. In that case

$$Y^2 = \frac{E(q_1-E)^2(q_2-E)^2(q_3-E)^2}{R_1^2(q_2-E)^2(q_3-E)^2+R_2^2(q_1-E)^2(q_3-E)^2+R_3^2(q_1-E)^2(q_2-E)^2}. \tag{5.5}$$

In (5.5) the short-hand notation

$$q_j = q(k_j), \quad R_j = R(k_j)$$

has been used. The generalization of (5.5) to N waves should be obvious.

A graph of Y *versus* E will show a zero at $E = 0$ and a zero at each $q_n > 0$. This is simply a restatement of the fact that free, ($Y = 0$), steady waves can exist if E equals one of the $q(k_j)$. For large E, $Y \sim E^{\frac{3}{2}}/(R_1^2+R_2^2+\ldots+R_N^2)$. A graph of Y versus E for the two-wave case ($R_1 = R_2 = 1, R_3 = 0), \eta = 0{\cdot}5, k_1 = 1, k_2 = 0{\cdot}75$ is shown in figure 14. For large Y and hence large E the Y, E relationship is single-valued. However, for $Y \leqslant 0{\cdot}104$ three solutions for E exist, while for $Y \leqslant 0{\cdot}04$ five E solutions are possible. In general for N waves, M of which have their q's > 0, there is a $2M+1$ multiplicity of solutions. That is, there are $2M+1$ separate branches, for the function $E(Y)$. In figure 14 they are numbered sequentially. The branches are separated by either zeroes of $Y(E)$ or zeroes of the derivative dY/dE. Each branch of E corresponds to a distinct branch of the steady solution, $A_k(Y)$. In figure 15(a, b) the steady solutions for A_{k_1} and A_{k_2} are shown.† The E-branch of each solution is labelled. There are several important facts to keep in mind. In the steady state all the amplitude solutions must

† The convenient notation $A_j = A(k_j)$ is used here.

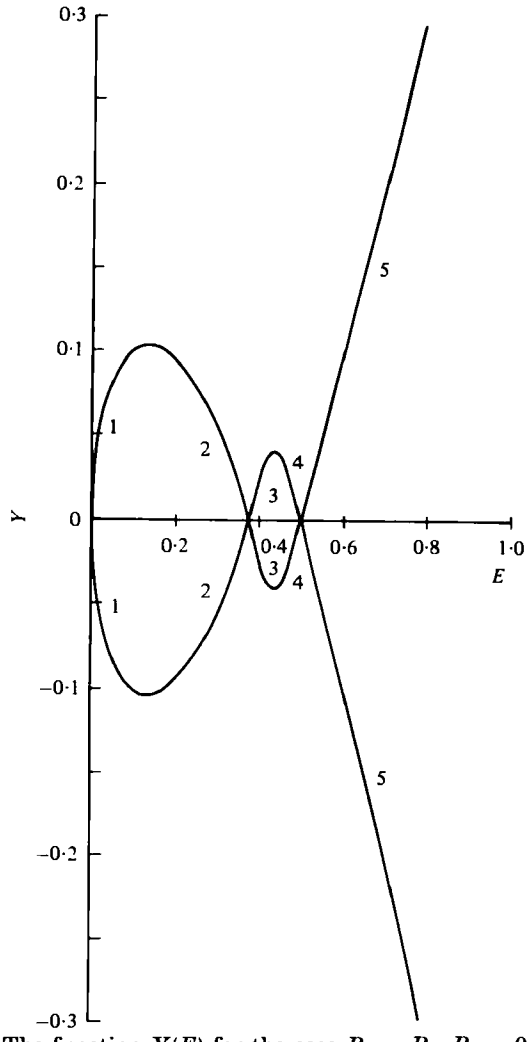

FIGURE 14. The function $Y(E)$ for the case $R_1 = R_2$, $R_3 = 0$. $\eta = 0{\cdot}5$, $k_1 = 1$, $k_2 = 0{\cdot}75$. Note that $E(Y)$ has five branches to its solution.

simultaneously be on the same E branch. That is, A_1 cannot be on branch 5 if A_2 is on branch 3. So that, for example in figure 15, it is apparent that if A_2 is on branch 5 near $Y = 0$, where its amplitude is large, then A_1 must be nearly zero. The same, but reversed, situation occurs if the solution is on the third E-branch. In addition consider a *subcritical* wave, $q(k_s) < 0$. Although the existence of such a wave will not add to the multiplicity of the steady solutions, the subcritical wave will share the $2M + 1$ multiplicity of possible solutions with the supercritical waves.

Let us examine figure 15 (a, b) in more detail. From (5.5) and (5.2) it follows that for a *given* branch

$$A_k(-Y) = -A_k(Y), \tag{5.6}$$

which allows figure 15 to be continued to negative Y as shown schematically in figure 16 (a, b). For the wave with the larger q (and for more than two waves this holds

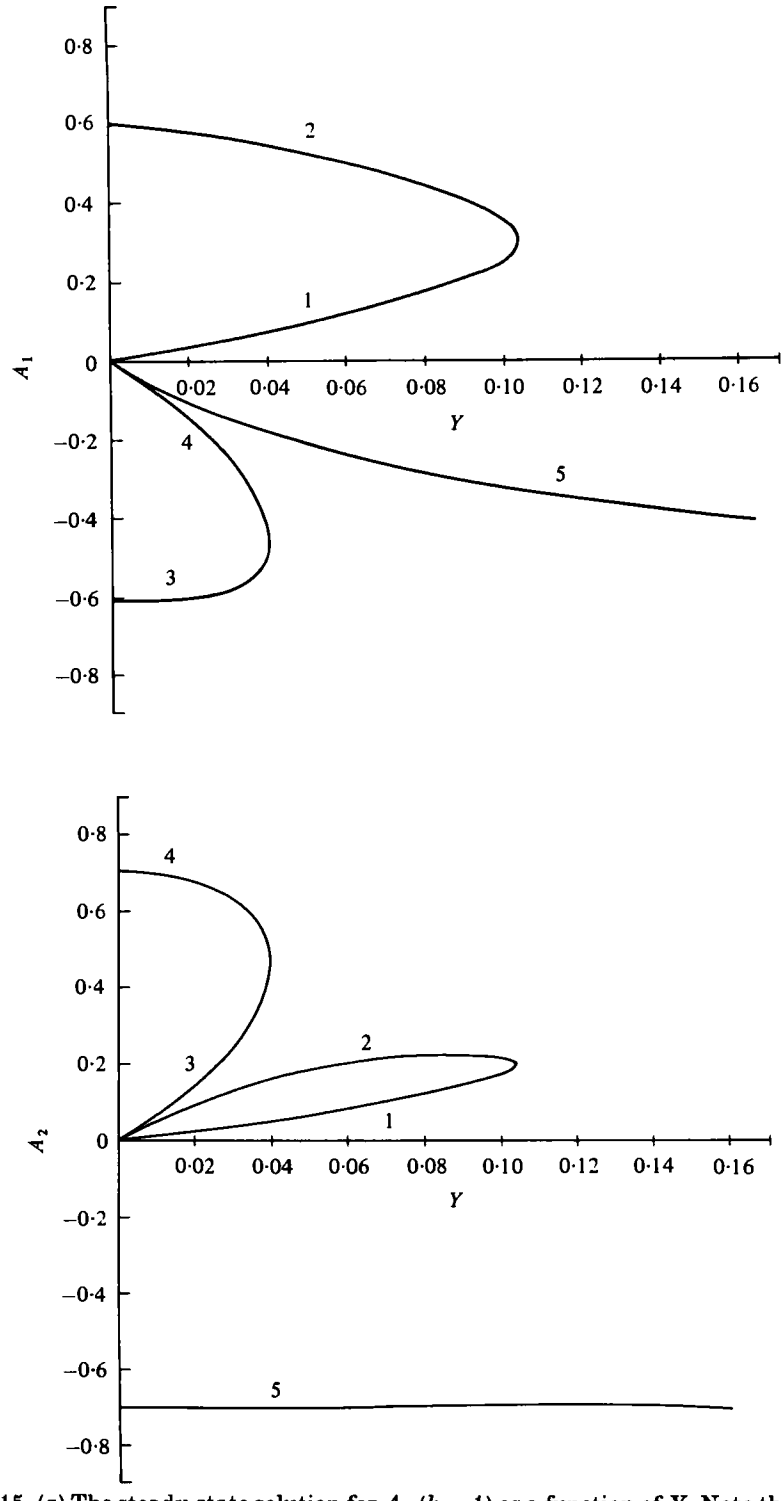

FIGURE 15. (*a*) The steady-state solution for A_k ($k = 1$) as a function of Y. Note the five branches of the solution corresponding to the five branches of $E(Y)$ in figure 14. (*b*) As in (*a*) except for A_k ($k = 0{\cdot}75$) which is the wave of higher q.

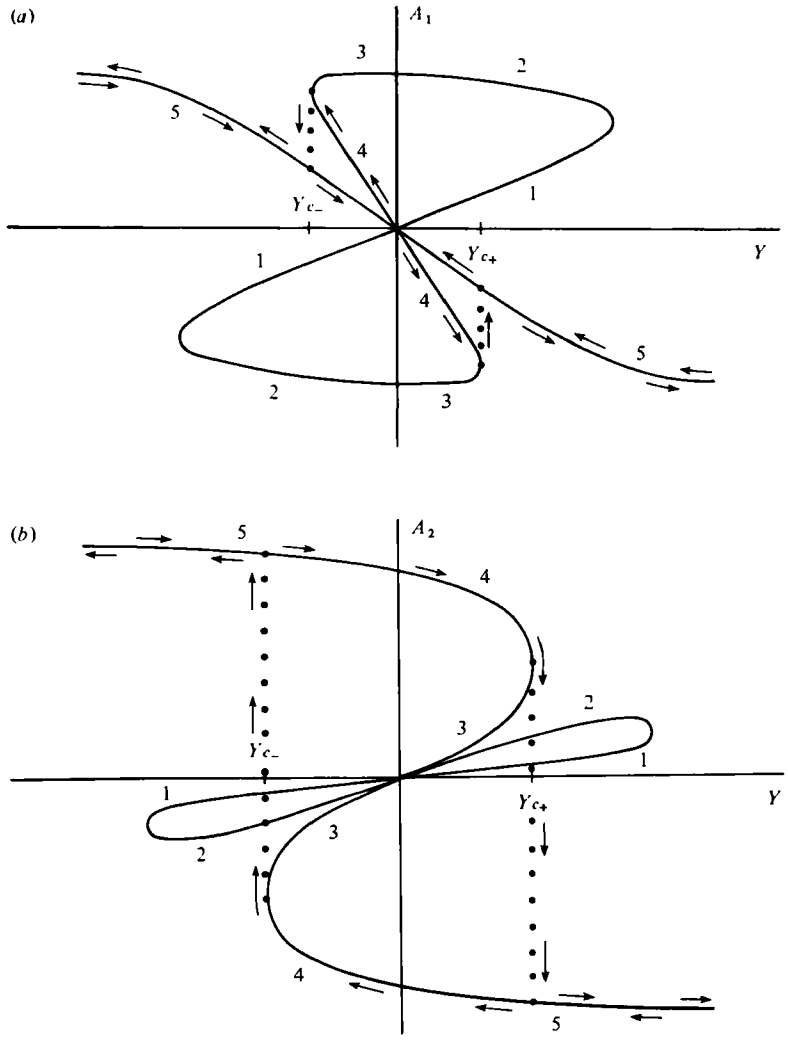

FIGURE 16. The hysteresis diagram for the wave amplitudes as a function of the slowly varying forcing. Arrows show the solution trajectories as Y changes and the dots indicate amplitude jumps experienced at the critical forcing points Y_{c_+} and Y_{c_-}. Please see text for a more complete description.

true for the wave of *largest* q), the continuation of branch 5 from large Y crosses the axis at the value $A_{k_2} = -(q_2)^{\frac{1}{2}}$ while $A_{k_1} \to 0$.

It will be necessary to examine the behaviour of the transient problem to determine which, if any, of the steady branches are stable but on the basis of the results of the investigation of free waves the essential result can be anticipated. For the free waves, $Y = 0$, the solution tended to the final state in which $A_2 \to \pm (q_2)^{\frac{1}{2}}$ while $A_1 \to 0$. In the present context that would imply that for small Y, where there is the greatest multiplicity of solution, branch 5, and its continuation branch 4, would be the stable branch. This turns out to be the case, as will be seen, but first let us examine the consequences of this result.

Consider a value of Y sufficiently large and positive so that the solutions for E, A_1 and A_2 must be on branch 5. Now, as shown by the arrows in figure 16, let Y decrease. Examine figure 16(*b*) first. The solution will travel leftward on the lower number 5 branch until $Y = 0$. At this point $A_2 = -(q_2)^{\frac{1}{2}}$. It then smoothly continues on branch 4 with A_2 decreasing in magnitude until the critical forcing Y_{c_-} is obtained. For larger *negative* values of Y the solution for A_2 cannot continue on branch 4. Since Y is decreasing branch 3 is unavailable. We can guess therefore that a jump in A_2 from branch 4 will occur upon further decrease of Y. But to which branch will it jump? Branches 1, 2 and 5 (upper) are available to it. If we guess (correctly as it turns out) that branches 1 and 2 are unstable branches, the solution will jump to the upper branch 5 and continue on that track with decreasing Y. The jump is indicated by the dots and steady theory clearly cannot describe that interval. If Y then *increases* the solution returns on the upper branches 5 and 4, as shown by the arrows, until $Y = Y_{c_+}$ occurs. At this point a similar quandary faces the steady solution and again we can imagine the jump at Y_{c_+} precipitating a sudden amplitude change from branch 4 to the lower branch 5. The value of the jump in A_2 is, at Y_{c_+}, from about $A_2 = 0{\cdot}5$ on the upper branch 4 to $A_2 = -0{\cdot}7$. This implies a sudden increase, of 40%, of the wave amplitude and an equally sudden phase shift in A_2 of 180°. Note that the shift in sign of A_2, which is sudden, *lags* the shift in sign of the forcing, i.e. Y must reach Y_{c_+} before A_2 again changes sign. Again, provisionally accepting this as the correct picture, let us examine now the response of A_1, the wave with the lower q. In a richer spectrum all the waves with q's less than the maximum q can be shown to share the qualitative behaviour to be described now. The solution for A_1, as A_2, starts at positive Y on branch 5. As $Y \to 0$, A_1 follows branch 5 and vanishes, consistent with the free wave result. The inaccessibility of branches 2 and 3 means that A_1 can never obtain a fully blossomed state, which is formally possible but unstable. At $Y = 0$ and decreasing, A_1 has, apparently, two continuous choices available for quasi-steady solutions. It could continue symmetrically on branch 5 or smoothly change tracks to branch 4. Since all solutions must be on the same E branch it must clearly do the latter. The significance of this is twofold. First the slope of the $A_1(Y)$ curve is greater on branch 4 than branch 5 so that as Y goes through zero A_1 will increase more rapidly with Y than it decreased toward zero. Referring to figure 15(*a*) shows that this effect is, however, numerically small. Second, of course, is the fact that after continuing on branch 4 a jump *down* to branch 5 will occur at Y_{c_-}. Whereas the jump in A_1 was, algebraically, $\sim 1{\cdot}2$, the jump in A_2 is $\sim 0{\cdot}3$. The jump in the lower q A's will always be smaller as long as the jumps occur between the $(2M+1)$th and $(2M)$th branch since by (5.2) $(q-E)$ becomes a progressively weaker function of E the greater $(q-E)$ becomes. It is left to the reader to retrace the A_1 *versus* Y solution path from negative to positive Y through the jump at Y_{c_+}.

It remains to be seen whether the behaviour described above will actually occur. Figure 17 shows the numerically calculated response of a two-wave field to the forcing

$$Y_1 = Y_2 = 0{\cdot}15 - 0{\cdot}3 \sin 2\pi t/T. \tag{5.7}$$

The two wavenumbers are $k_1 = 1$, $k_2 = 0{\cdot}75$ with $q_1 = 0{\cdot}375$, $q_2 = 0{\cdot}4965$. At $t = 0$, $A_1 = A_2 = 0{\cdot}1$ and $\eta = 0{\cdot}5$. The period of sweep along the Y axis was chosen to be very long ($T = 4 \times 10^3$) in order to render the quasi-steady theory valid. Figure 17 shows that as expected the solution for A_2 (and A_1) quickly converges to branch 5. As t increases,

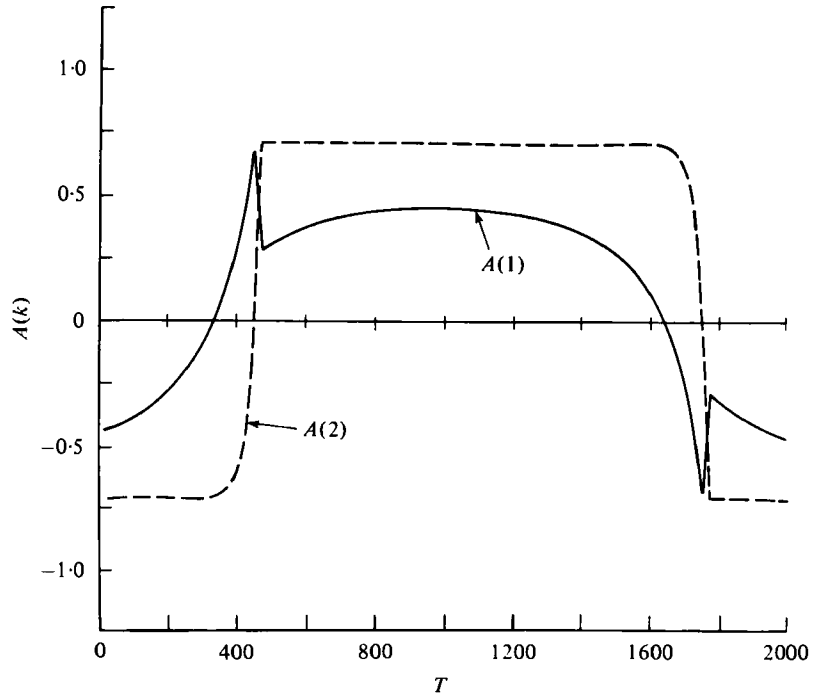

FIGURE 17. The time history of two waves A_1 and A_2 with $k_1 = 1$, $k_2 = 0{\cdot}75$, $\eta = 0{\cdot}5$. The wave forcing is $Y_1 = Y_2 = 0{\cdot}15 - 0{\cdot}3 \sin(2\pi t/T)$, $T = 4 \times 10^3$. The figure demonstrates the sudden amplitude changes experienced at $Y_{c+} = -Y_{c-}$. Note that the dominant wave changes sign and increases in amplitude. The most unstable wave maintains its phase and sharply decreases in magnitude at Y_c.

Y_k slowly decreases until, at the moment *predicted* by the quasi-steady theory, sudden jumps occur in both A_1 and A_2. In the case of A_1 the jump is a diminishment, as expected, from the peak it attained on branch 4 while A_2 after a smooth but fairly rapid decline on branch 4 suddenly jumps up to branch 5 as predicted. A second jump is also shown at $t \sim 1700$ as predicted again by the steady theory. On the other hand the maximum amplitude obtained by A_1 is somewhat larger than predicted by the steady theory. Even for $T = 4000$ there is a dynamical overshoot on branch 4 before and during the jump. This is quite likely due to the fact that $A_2 \to 0$ during the jump which allows A_1 a fleeting moment to increase. The minimum reached in the jump in A_1 is somewhat larger than the steady value would predict because during the finite period occupied by the jump Y_k has increased to a point where the steady value of A_1 is larger than its value is directly at Y_c. The amplitude solutions, except in the jump periods, are thus reasonably well predicted by the quasi-steady theory. Numerous numerical calculations with different starting values of Y_k were performed. In all cases the solutions rapidly converged to branches 4 or 5 and the qualitative behaviour just described was reproduced. It therefore seems an intrinsic characteristic of periodic changes of the wave forcing that if it sweeps past $|Y_{c+}|$ the wave response to such a smoothly varying forcing will nevertheless exhibit sudden nearly discontinuous changes in the magnitude of each of the wave amplitudes. Furthermore, during these jump epochs the dominant high q wave passes through zero so that in the case depicted

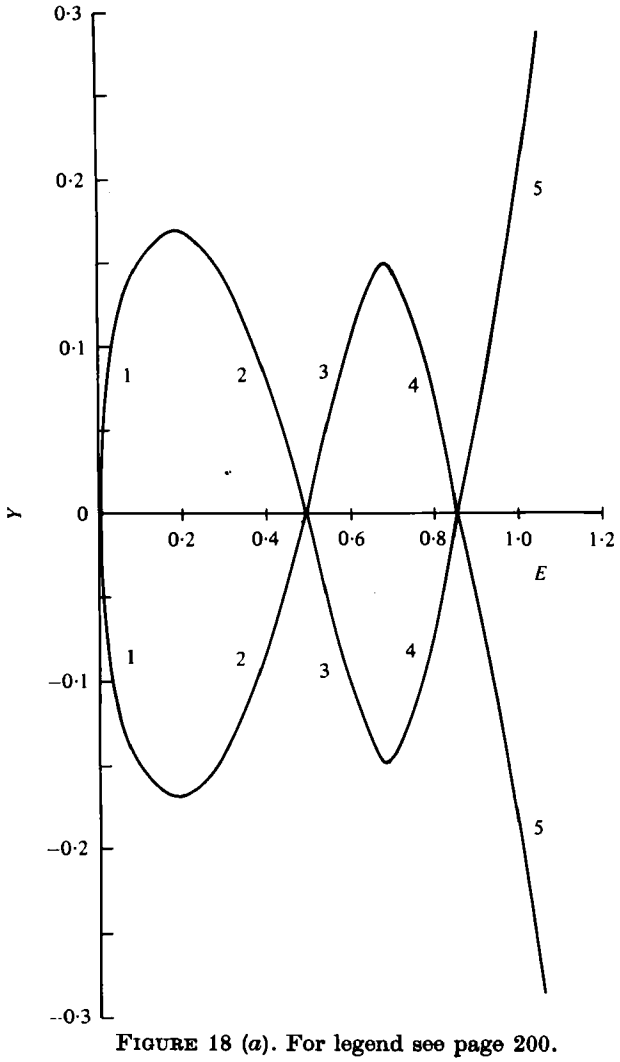

FIGURE 18 (*a*). For legend see page 200.

in figure 17 there is an interval of about twenty, linear e-folding times where, fleetingly, the wave of smaller q becomes the larger. After the jump event is concluded the high q wave quickly establishes its dominance. This, as we shall see, is to some extent sensitive to the equality of the wave forcing assumed and will be discussed more fully below.

As η is decreased the values of $q(k, \eta)$ obtainable increase since the maximum q is $1-\eta$. Figure 18(a, b, c) shows the $E(Y)$, $A_1(Y)$, $A_2(Y)$ steady solutions for the case $\eta = 0{\cdot}1$, $k_1 = 1$, $k_2 = 0{\cdot}5$, $R_1 = R_2 = 1$, where $(q_1, q_2) = (0{\cdot}495, 0{\cdot}855)$. The separation of the q's on the E-axis is greater but the general nature of the steady amplitude response curves is quite similar to the more dissipative case of figure 15(a, b). The dynamical response to slowly varying forcing is, however, somewhat different. At this lower value of η there is a tendency for the free waves to oscillate for some time before reaching their steady values. That is, we can expect considerably more dynamical overshoot from this more lightly damped system. Figure 19(a, b) shows the response of

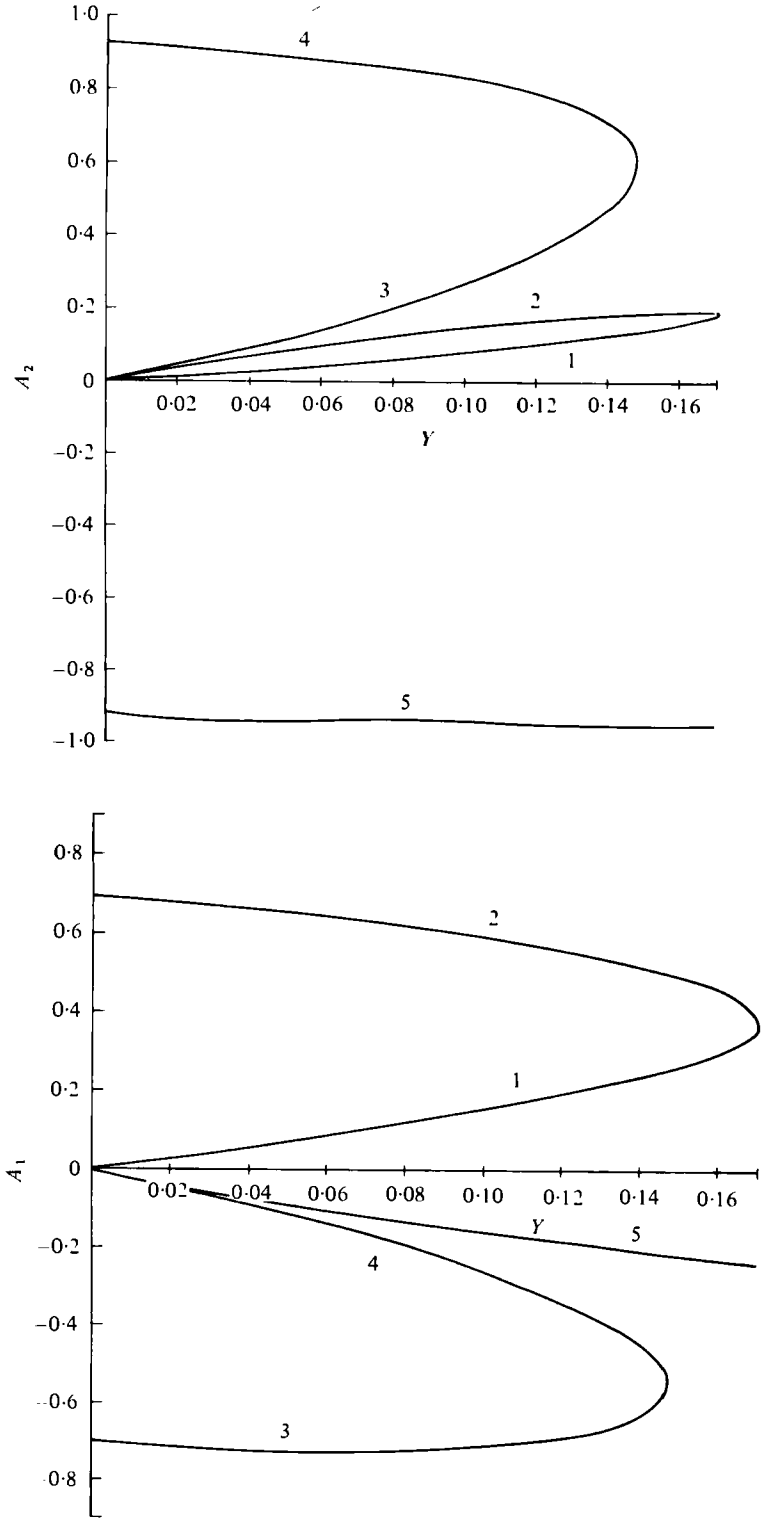

FIGURE 18. Wave hysteresis diagram for $\eta = 0{\cdot}1$, $k_1 = 1$, $k_2 = 0{\cdot}5$, $R_1 = R_2$. (*a*) $Y(E)$, (*b*) $A_{k_1}(Y)$, (*c*) $A_{k_2}(Y)$.

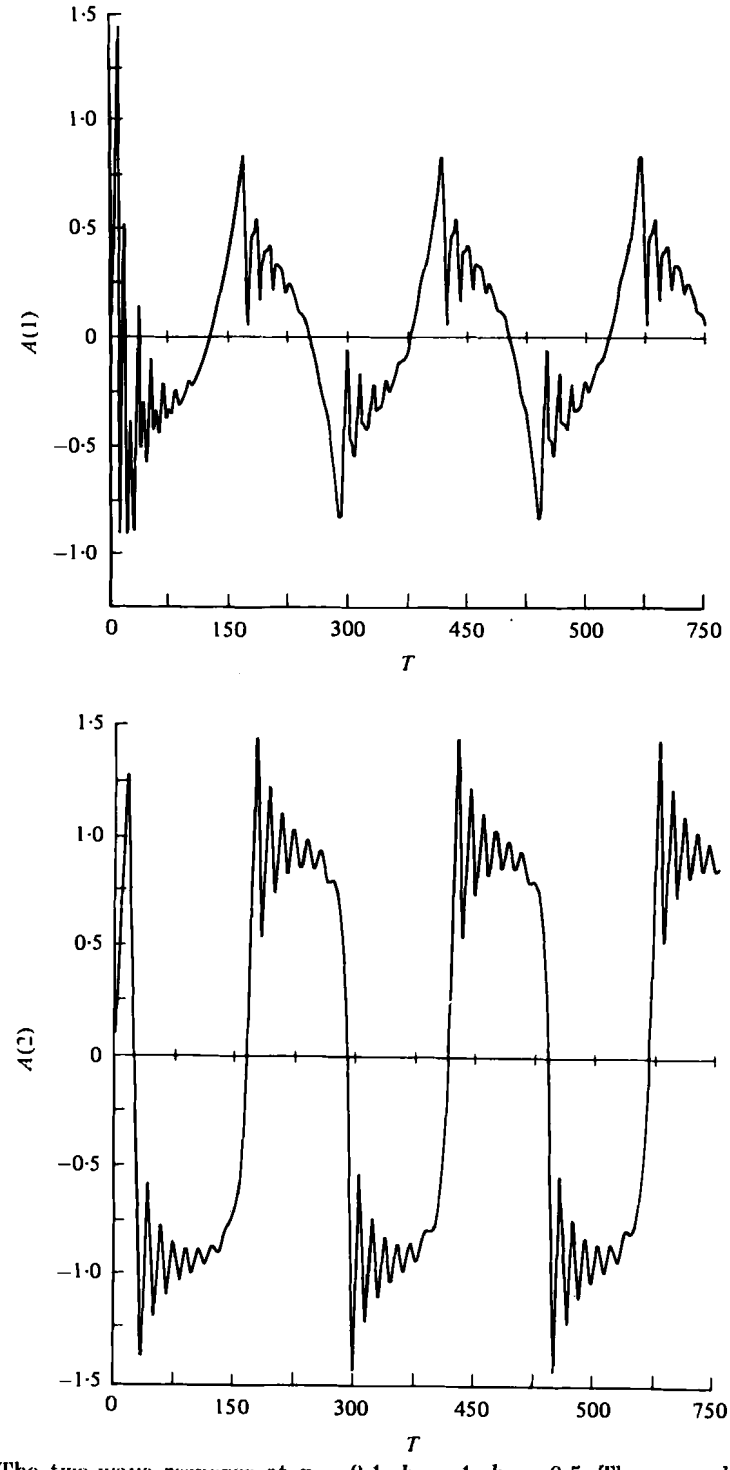

FIGURE 19. The two-wave response at $\eta = 0{\cdot}1$, $k_1 = 1$, $k_2 = 0{\cdot}5$. The general nature of the response is as in the higher η case of figure 17 but considerably greater dynamic overshoot exists and a longer time is required to achieve equilibration. The solution is periodic with the period of the forcing. (*a*) $A(k_1)$, (*b*) $A(k_2)$. The forcing period is 250.

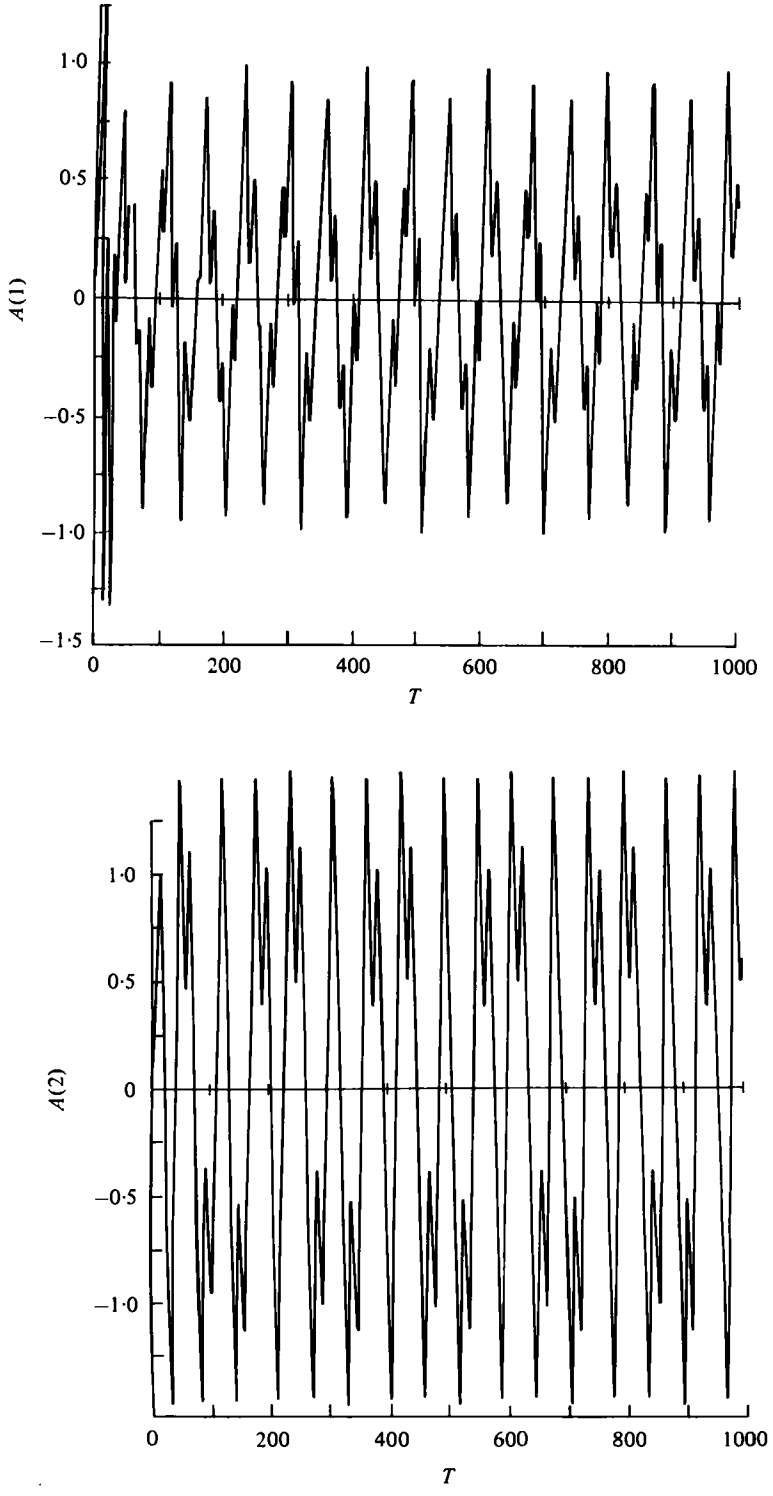

FIGURE 20. As in figure 19 except that the forcing period is 62. The response is once again periodic but *three* forcing periods are required before the solution repeats. (*a*) $A(k_1)$, (*b*) $A(k_2)$.

the system to the forcing

$$Y_{k_1} = Y_{k_2} = 0{\cdot}18 \sin(2\pi t/250). \tag{5.8}$$

The general nature of the response is as in figure 17 only now there is considerably greater range in the amplitude jumps due to the aforementioned dynamical overshoot. This is particularly evident in the response of A_1. During its jump from branch 4 back to branch 5 the overshoot is sufficiently great to bring the amplitude of A_1 momentarily near zero. A_2 on the other hand springs up to a value of about 1·4, considerably larger than its steady, past-jump value of $\sim 0{\cdot}92$. Once again the hysteresis response curves are valuable in predicting at what point dramatically large changes of amplitude will occur, but at low η the steady theory consistently underestimates the magnitude of the amplitude jumps experienced by the waves.

Examining figure 19 cannot help but suggest the thought of what might occur if the jump points $Y_{c_\pm}$ were obtained before the solution could at all settle down. If the occasion for a jump then finds the wave field in an unsettled state the starting values for different jump events might differ, leading to a continuously non-repetitive, or aperiodic, response.

What I did therefore was to hold the forcing amplitude in (5.8) fixed and consider the response to forcing of ever higher frequency. At a forcing period of 125, half that of (5.8) the response was qualitatively as in figure 19, i.e. the response, though complex, was periodic with the period of the forcing. At a forcing period of 62·5, i.e. halved again, the solution is still periodic but experiences four separate subcycles (figure 20) before repetition and the period is about three times the forcing period. Decreasing the forcing period to 41 t units reveals a hint of aperiodicity. At a forcing period of 25·13 the solution to periodic forcing is strongly chaotic as shown in figure 21. It is important to point out that this parameter range of η leads to eventually steady waves in the absence of forcing and the occurrence here of aperiodicity is caused only by the presence of *periodic* forcing. Note too, that at these relatively high frequency forcings the response shows little evidence of the jump phenomenon. For that to occur the solutions must *slowly* track on the steady solution branches. Since this study is meant to be a preliminary one, no exhaustive search has yet been made of the detailed dependence of the amplitude response on the forcing frequency. However, the results displayed here show that the dependence is likely to be significant even while the forcing periods are much longer than the linear e-folding times for instability.

It is clear from the analysis given above that the nature of the wave amplitude hysteresis is determined primarily by the presence of the highest q wave. In particular the amplitude of the jump is determined by the ΔE which exists between branch 4 and branch 5 at the value of Y where $dY/dE = 0$ on the summit of branch 4. The larger this cleavage is in the $Y(E)$ curve at that point the greater the jump will be. If the forcing amplitude, R_n, of the waves are not equal then the $Y(E)$ curve will vary accordingly. Consider the two-wave case. Then

$$Y^2 = \frac{E(q_1 - E)^2(q_2 - E)^2}{R_1^2(q_2 - E)^2 + R_2^2(q_1 - E)^2}. \tag{5.9}$$

Consider what occurs as $R_2/R_1 \to 0$, i.e. when there is very little forcing at the dominant

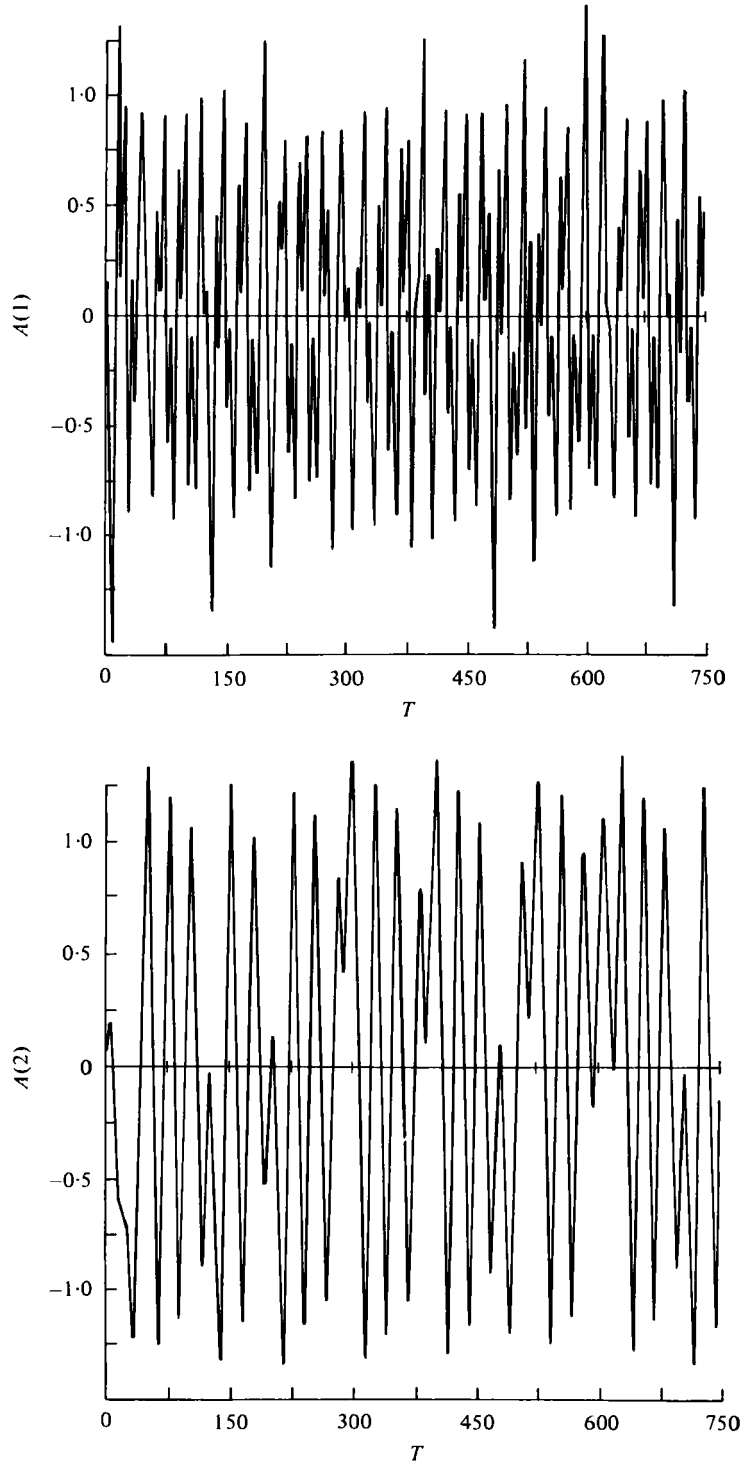

FIGURE 21. As in figures 19 and 20. However, the forcing period is now 25 t units. The response appears completely aperiodic. Although the forcing is simply periodic the wave response is non-repetitive and chaotic.

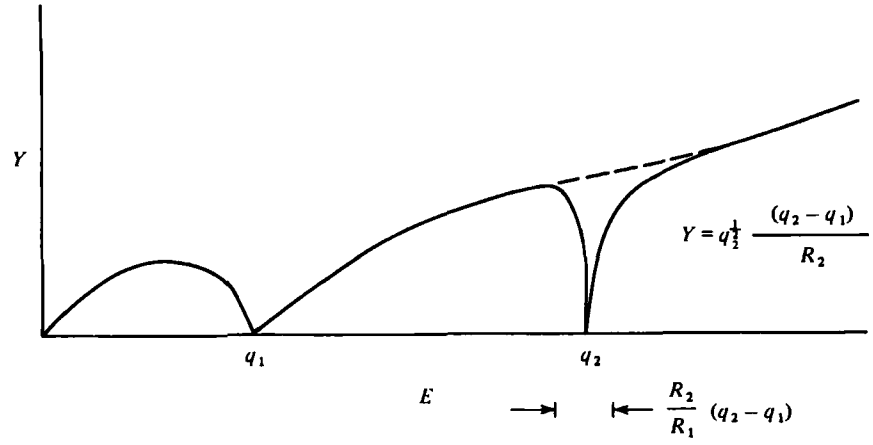

FIGURE 22. The $Y(E)$ curve in the case $R_2/R_1 = \epsilon \ll 1$. Except in the immediate vicinity of $E = q_2$ the curve is approximated by the single-wave response curve. However, in a narrow region around $E = q_2$, $Y \to 0$ and the free, maximum q wave emerges. The dotted line is the continuation of the *single*-wave curve past $E = q_2$.

wave. For $R_2/R_1 = \epsilon \ll 1$ an approximation to (5.9) which is valid for all $E \neq q_2$ is

$$Y = \frac{E^{\frac{1}{2}}(q_1 - E)}{R_1}, \tag{5.10}$$

which is the single-wave response curve, and it is shown schematically in figure 22. On the other hand when

$$E = q_2 + e, \quad e \ll q_2, \tag{5.11}$$

$$Y = \frac{q_2^{\frac{1}{2}}(q_2 - q_1)}{R_1} \frac{e}{\{e^2 + \epsilon^2(q_1 - q_2)^2\}^{\frac{1}{2}}}. \tag{5.12}$$

Thus, consider ϵ arbitrarily small but fixed. Then, as $E \to q_2$, $e \to 0$ and $Y \to 0$, dropping precipitously from its value $q^{\frac{1}{2}}(q_2 - q_1)/r_1$ at $e \gg \epsilon(q_2 - q_1)$. Outside this narrow region in E of order $\epsilon(q_2 - q_1)$ the single-wave response curve should apply and therefore

$$A_1 \sim q^{\frac{1}{2}}, \quad 1 \gg e \gg \epsilon(q_2 - q_1), \quad A_2 \sim \tfrac{1}{2}\epsilon(q_2 - q_1)q_2^{\frac{1}{2}}. \tag{5.13}$$

Outside the $E = q_2$ region the dominant *forced* wave, when $R_1 \gg R_2$, will be the wave with the *smaller* q but the *larger* forcing. However, as e goes to zero, in particular as

$$e \ll \epsilon(q_2 - q_1), \tag{5.14}$$

then

$$A_1 \sim q_2^{\frac{1}{2}}\left(\frac{e}{(q_2 - q_1)}\right), \quad A_2 \sim q_2^{\frac{1}{2}}. \tag{5.15a, b}$$

Thus when $e \to 0$, i.e. when the forcing goes to zero, the role of the dominant and subordinate waves are interchanged and $A_2 \gg A_1$. A little thought shows that this is as it must be. For zero overall forcing the high q wave must become predominant. If the overall forcing is primarily at wavelengths other than that of the maximum q those other waves will have $O(1)$ amplitudes while the maximum q wave will approximately satisfy

$$A_{\max}(q_{\max} - E) \sim 0.$$

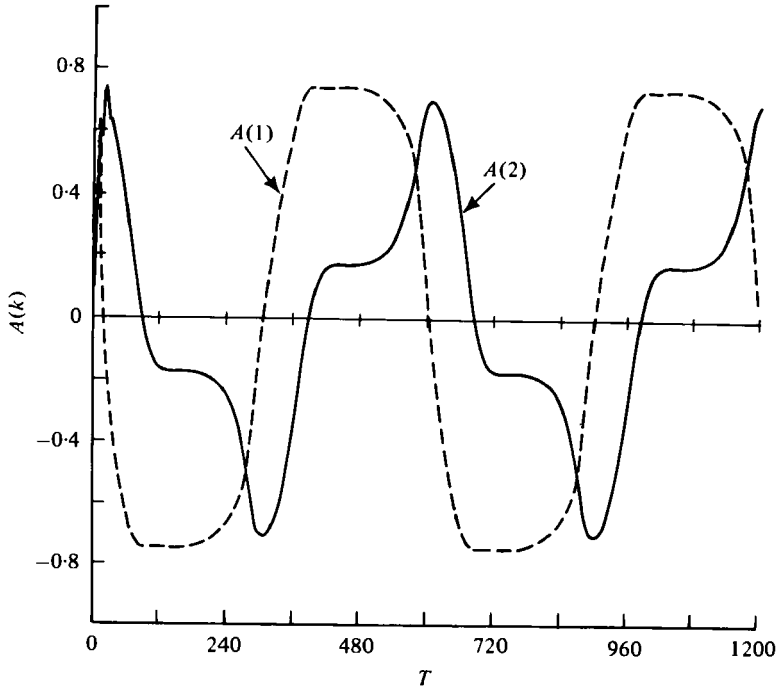

FIGURE 23. The two-wave response, $k_1 = 1$, $k_2 = 0{\cdot}75$ at $\eta = 0{\cdot}5$ to the forcing $Y_{k_1} = 0{\cdot}16 \sin(2\pi t/600)$: $Y_{k_2} = 0{\cdot}1 Y_{k_1}$. Note, as described in the text, the periodic interchange of dominance of the two waves.

Since E will be forced to be other than $q_{\max}$, $A_{\max}$ will essentially vanish unless all the $Y_k \to 0$.

It is important to note that the ΔE, in the small R_2/R_1 case, between branches 4 and 5 will be exceedingly small. Hence the A_1 wave will show little or no jump in amplitude for $\epsilon \ll 1$. Figure 23 shows the response of the two-wave system $k_1 = 1$, $k_2 = 0{\cdot}75$, $\eta = 0{\cdot}5$ for the case where

$$Y_{k_1} = 0{\cdot}16 \sin(2\pi t/600), \quad Y_{k_2} = 0{\cdot}016 \sin(2\pi t/600) \qquad (5.16a, b)$$

so that $R_2/R_1 = 0{\cdot}1$. Figure 23 shows that A_1 and A_2 *interchange* intervals of dominance. When $Y_k \to 0$ A_2 rises to its free wave maximum $q_2^{\frac{1}{2}}$ and A_1 vanishes. When Y_k is $O(1)$ it is A_1 that is forced to be dominant. It must be stressed that this interchange of dominance (and hence the periodic alteration of the wavenumber of the finite-amplitude state) is here accentuated by the smallness of R_2/R_1. As R_2/R_1 increases towards unity the previous cases are obtained where, over most of the cycle, the high q wave dominates.

6. Final remarks

Two important qualitative notions are suggested by the results of the present paper. First, the instability, through interaction with the mean flow, of all the long waves with respect to the wave of maximum q, will tend to concentrate the long wave spectrum in a single wave, or on occasion in two conjugate waves. To some extent this

result is model dependent, for the dynamics of shorter baroclinic waves will undoubtedly allow energy transfers by wave–wave interaction. Hence some leakage will occur to other waves. To the extent that the *primary* baroclinic interaction is between each wave and the mean flow the tendency for a sharpening of the long-wave spectrum should be valid. Thus finite-amplitude single-wave theories are not as artificial as might be imagined.

The most important idea this study suggests, however, is that the wave of maximum linear growth rate is only fleetingly the dominant wave in the finite-amplitude free wave response. It rapidly gives way to the wave of maximum $q(k)$ in the spectrum. This latter wave may have a much lower growth rate but its ability to consume more of the total available potential energy gives it a long-term advantage over the wave of maximum growth rate. One virtue of the present simple model is that the simplicity of the mathematics allows explicit representations of both the growth rate dependence on k and of the function $q(k)$. However, that very simplicity makes the generalization of the wavenumber selection principle necessarily heuristic. In the present case the growth rate maximum occurs at the maximum of $k^2q(k)$, a number obtainable from linear theory. The maximum of $q(k)$ is thus just dependent upon the ratio $k^2q(k)/k^2$. This particular result depends on the simplicity of the model and cannot be generally used. If it could be, linear theory alone would be sufficient to determine the nonlinear selection principle and that would be absurd. The proper generalization of the above criterion is most likely as follows. Consider the general amplitude equation for a *single* baroclinic wave. Its form will be

$$\frac{d^2A}{dt^2}+\gamma\eta\frac{dA}{dt}+A\{k^2c_0^2(k)-M(A,k)\}=0, \tag{6.1}$$

where M is some operator quadratic in A and γ is an $O(1)$ constant. Experience has indicated that, at least in the *steady* state,

$$M(A,k)=A^2k^2N(k), \tag{6.2}$$

where $N(k)$ is the so-called Landau constant. I believe that the function

$$A_s^2=\frac{c_0^2(k)}{N(k)}=Q(k) \tag{6.3}$$

is the proper generalization of the $q(k)$ of this paper. In the present study, because of the relatively weak dependence of the wave field on k the function N has turned out to be independent of k. This is not generally the case, as can be seen from earlier studies (Pedlosky 1970, 1971). I therefore suggest that, rather than fixing on the wave of maximum $kc_0(k)$, the fundamental wave is the wave of maximum $c_0/N^{\frac{1}{2}}$. To be sure, this is a nonlinear selection principle and much less easy to apply than the linear principle. On the other hand it requires the consideration of only single-wave baroclinic dynamics.

Whether this particular, heuristic suggestion is correct or not, I should like to stress that the present study at least serves as a strong counter-example to the more traditional view which regards the wave of largest growth rate as the dominant wave. This *will be true only initially*, so that the idea of its dominance, while initially correct, is

not uniformly valid in time. Thus nonlinear calculations of baroclinic flows which display wave dominance at wavenumbers other than that of maximum linear growth are not inconsistent with the theory of baroclinic instability. Such results are simply in accord with the divergence in finite amplitude between the wavelength of maximum growth and the wavelength which maximizes $c_0/N^{\frac{1}{2}}$.

The results of §§4 and 5 show that the response of the waves to periodic, 'seasonal' changes in either the basic temperature gradient or the wave-forcing is highly complex. Although the forcing is periodic the response may be aperiodic, i.e. no two 'seasons' need be alike if the relaxation time for nonlinear adjustment is long enough compared with the seasonal period. The precise frequency criterion for the advent of aperiodicity to periodic forcing has not yet been determined. It is significant, though, that this aperiodicity depends only on the simplest wave–mean flow interaction dynamics and does not require exotic, 'climatic' mechanisms. It is also significant that the aperiodicity occurs for those waves which intrinsically, i.e. in their free states, have asymptotically steady amplitudes.

Finally, the forced wave problem itself exhibits a wide range of fundamentally interesting phenomena. The quasi-steady problem inherently responds to smooth changes in the forcing smoothly except at critical values of the forcing. At these points, sudden changes, either of increase or decrease, in the wave amplitude occur. Sudden changes in phase of the dominant wave also are manifested whereby the position of a wave crest and trough are suddenly interchanged. One deficiency in the model introduced for the sake of simplicity is the stationarity of the free waves, i.e. the mean flow is purely baroclinic so that the unstable waves have zero real phase speed. This makes them particularly sensitive to essentially stationary (but slowly varying) wave forcing. The relaxation of this condition will require a more complex model. I present here some preliminary calculations to indicate the general nature of the changes to be expected.

If a barotropic mean flow of order $\Delta^{\frac{1}{2}}$ is included in the theory so that mean advection of perturbation potential vorticity is as important as its temporal development, it can be shown that the equation for the *steady* response to wave forcing will now be, instead of (5.2), rather

$$A_k = \frac{Y_k}{\{q(k) + \omega_a^2 - E - i\epsilon_k\}}. \tag{6.4}$$

The new parameters in (6.4) are ω_a, which is the advection velocity scaled by the imaginary part of the linear phase speed, and $\epsilon_k = \frac{3}{2}\eta(\omega_a/k)$. The relationship (6.4) is now *complex* so that $E \equiv \sum_k |A_k|^2$. The fact that (6.4) is complex reflects the fact that there is no *steady* solution (aside from $A_k = 0$) as $Y_k \to 0$. Instead, there the free waves will be slowly propagating. Nevertheless the steps leading to (5.3) and (5.5) can be obviously generalized to

$$E = Y^2 \sum_k \frac{|R_k|^2}{\{(q_k + \omega_a^2 - E)^2 + \epsilon_k^2\}}. \tag{6.5}$$

A plot of Y *versus* E for the two-wave case $k_1 = 0{\cdot}989$ ($q_1 = 0{\cdot}5$) and $k_2 = 0{\cdot}373$ ($q_2 = 0{\cdot}8$), $\omega_a = 0{\cdot}2$ is shown in figure 24. There is again a multiplicity of solutions and the multiplicity is the same as before, i.e. 5 in the two-wave case. However, now the range of multiple solutions has a lower as well as an upper critical value of Y. It seems unlikely that, starting on branch 5, the energy should jump for low Y to the small

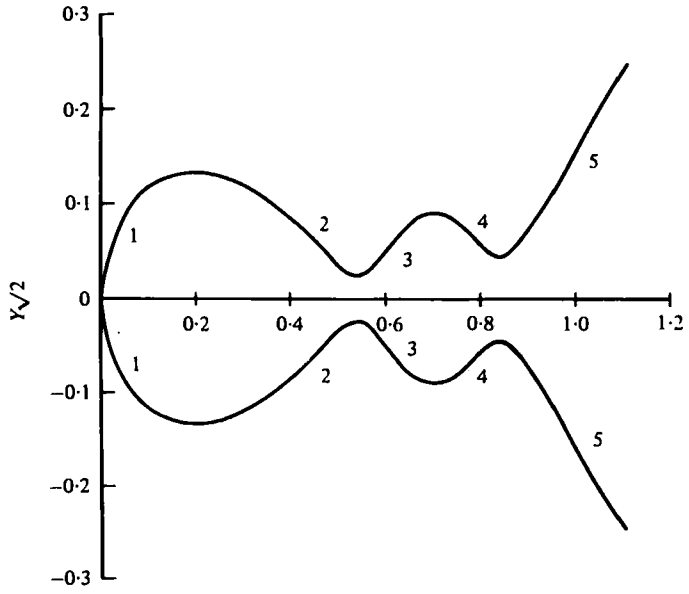

FIGURE 24. The $Y(E)$ curve in the case where the *free* waves are non-stationary. Multiple steady solutions exist but now a maximum as well as a minimum forcing is required. For small forcing only weak steady solutions are possible and these are likely to be supplanted by travelling unstable waves.

values of branch 1. It seems more likely that the low Y response for $\omega_a \neq 0$ may be dominated by the free waves (and probably the high q wave). This would be consistent with the results at $Y = 0$ of the study of §5. However, this will require further study.

This research was supported, in part, by a grant from the National Science Foundation's Office of Atmospheric Science.

REFERENCES

BUZYNA, G., PFEFFER, R. L. & KUNG, R. 1978 Cyclic variations of the imposed temperature contrast in a thermally driven rotating annulus of fluid. *J. Atmos. Sci.* **35**, 859–881.

DRAZIN, P. G. 1970 Non-linear baroclinic instability of a continuous zonal flow. *Quart. J. Roy. Met. Soc.* **96**, 667–676.

DRAZIN, P. G. 1972 Non-linear baroclinic instability of a continuous zonal flow of a viscous fluid. *J. Fluid Mech.* **55**, 577–588.

EADY, E. T. 1949 Long waves and cyclone waves. *Tellus* **1**, 33–52.

PEDLOSKY, J. 1970 Finite amplitude baroclinic waves. *J. Atmos. Sci.* **27**, 15–30.

PEDLOSKY, J. 1971 Finite amplitude baroclinic waves with small dissipation. *J. Atmos. Sci.* **28**, 587–597.

PEDLOSKY, J. 1972 Limit cycles and unstable baroclinic waves. *J. Atmos. Sci.* **29**, 53–63.

PEDLOSKY, J. & FRENZEN, C. 1979 Chaotic and periodic behaviour of baroclinic waves. *J. Atmos. Sci.* **37**, 1177–1196.