U. S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE

OFFICE NOTE 94

Analysis and Experiment in Nonlinear Computational Stability

Frederick G. Shuman
National Meteorological Center

FEBRUARY 1974

Presented at the Symposium on Difference and Spectral Methods for Atmosphere and Ocean Dynamics Problems, Novosibirsk, USSR, September 17-22, 1973. Published in the Proceedings of the Symposium, Part I, pp. 51-81, Novosibirsk, 1974.

Analysis and Experiment in Nonlinear Computational Stability

Frederick G. Shuman
National Meteorological Center
National Weather Service, NOAA

ABSTRACT

New extensions of Phillips' (1959) computational stability analysis to several finite-difference forms yield some interesting explanations of what is generally observed when nonlinear computational instability sets in. The one-dimensional shock equation has been adopted for the analyses, and in spite of its apparent simplicity it is capable of describing many features of both stable and unstable calculations with the full set of atmospheric equations.

The analyses, among other things, are used to investigate the validity of some ideas which have been proposed in the past as root causes of nonlinear instability. Among the analyses are cases in which cascade of energy is permitted in the spectrum but energy is not trapped in the high wave numbers, unstable cases in which aliasing is not permitted, and cases in which the computational mode can be a stabilizing factor. The so-called "energy-conservation" form is partially analyzed, and shown to have a stability criterion similar to the others.

Introduction

What this paper purports to do is to make a few extensions to a small body of knowledge to be found in the literature about the analysis of stability of non-linear partial finite-difference equations commonly used in meteorological research and operations. The classical paper in stability analysis is Courant et al. (1928), but Phillips (1959) provided a departure into the analysis of non-linear equations. Phillips' work was extended by Richtmyer (1963), Lilly(1965), Robert (1969), and Robert et al. (1970). From the standpoint of the effort in hand, Lilly (1965) does not fit directly in the sequence because he implicitly assumed that the time-difference-ratio converges to the derivative as the time step becomes vanishingly small. His results are borne out experimentally in stages of calculations where convergence is at least approximated, but experiments show that convergence does not universally obtain.

Use of smoothing devices in time, or forward differences which damp are now the rule, in both operations and research, which may quantitatively improve convergence. It is our purpose however to investigate systems centered in space and time, because in the linear case such systems are perfectly stable, i. e. , neither amplification nor damping arises from truncation error.

Notation

A convenient finite difference notation will be used, which is described in Robert et al. (1970) and elsewhere. Briefly, an independent variable used as a subscript (e. g. , x in u_x) will denote a partial finite difference ratio in that variable, involving values at immediately adjacent grid points. An independent variable used as a superscript following a superposed bar (e. g. , x in $\bar{u}^x$) will denote an algebraic mean of the variable at two immediately adjacent grid points. The prefix "2" to an operating independent variable (e. g., 2 in u_{2x} and $\bar{u}^{2x}$) will extend the action to two grid increments, with the central value not involved. Attachment of two operating variables denotes action twice.

Extension of Phillips' analysis

Phillips (1959) dealt with the so-called barotropic equation,

$$\frac{\partial \zeta}{\partial t} + \frac{\partial(\psi, \zeta)}{\partial(x, y)} = 0.$$

where

$$\zeta = \nabla^2 \psi .$$

The finite-difference analog he analyzed was

$$\zeta_{2t} + \psi_{2x}\, \zeta_{2y} - \psi_{2y}\, \zeta_{2x} = 0. \qquad (1)$$

where

$$\zeta = \psi_{xx} + \psi_{yy}$$

The general solution for this equation is not known. Phillips, however, found the special solution

$$\psi = (C \cos \tfrac{1}{2} \pi j + S \sin \tfrac{1}{2} \pi j + U \cos \pi j) \sin \tfrac{2}{3} \pi k$$

where

$$j = x/\Delta$$

$$k = y/\Delta$$

and Δ is the distance between adjacent points in a square array in the two space dimensions, x and y. The coefficients C, S, and U are all functions of n only, where

$$n = t/\Delta t$$

and Δt is the time step.

Anticipating the analyses of Richtmyer (1963), Robert (1969), and Robert et al. (1970), we will alter Phillips' solution in a deceptively simple way by adding a "V", thus:

$$\psi = (C \cos \tfrac{1}{2} \pi j + S \sin \tfrac{1}{2} \pi j + U \cos \pi j + V) \sin \tfrac{2}{3} \pi k. \quad (2)$$

The variable, V, here, like C, S, and U, is dependent on n only. Then, otherwise following Phillips, we substitute from (2) into (1), equate coefficients of like components, and obtain

$$C_{2t} = \frac{\sqrt{3}}{10\,\Delta^2}\, S(U - V) \quad (3a)$$

$$S_{2t} = \frac{\sqrt{3}}{10\,\Delta^2}\, C(U + V) \quad (3b)$$

$$U_{2t} = V_{2t} = 0 \quad (3c)$$

The general solutions of (3c) may immediately be expressed as

$$\begin{aligned} U &= U_1 + U_3 \cos \pi n \\ V &= U_0 + U_2 \cos \pi n \end{aligned} \quad (4)$$

where U_0, U_1, U_2, and U_3 are constants depending on the values of U and V at the two necessary initial time steps. We next difference (3a) with the operator $(\;)_{2t}$, using (4), and substitute for S_{2t} from (3b), obtaining the second-order difference equation

$$C_{2t2t} + \frac{3}{100\,\Delta^4} \left[(U_0 \pm U_3)^2 - (U_1 \pm U_2)^2 \right] C = 0 \quad (5)$$

The upper signs hold if the central n is even, the lower if odd.

The criterion for stability of (5) is

$$0 \le \frac{3}{100} \frac{(\Delta t)^2}{\Delta^4} \left[(U_0 \pm U_3)^2 - (U_1 \pm U_2)^2 \right] < 1 \qquad (6)$$

Whether the right hand inequality is satisfied or not, depends on the magnitude of Δt, and corresponds to the criterion analyzed by Courant et al. (1928). Whether the left hand inequality is satisfied, however, depends only on the balance between the squared quantities within the square brackets, and not at all on the magnitude of Δt, except for the trivial case of $\Delta t = 0$.

A few interesting points may be made from (6). First, the computational mode is often blamed for instability. The computational mode is represented by the coefficients U_2 and U_3 of the high frequency component, $\cos \pi n$, in (4). The constants U_0 and U_1, on the other hand, represent the low frequency physical mode. Note that in Phillips' original analysis, $V = 0$, and therefore $U_0 = U_2 = 0$. His stability criterion, therefore, reduces to

$$0 \le \frac{3}{100} \frac{(\Delta t)^2}{\Delta^4} (U_3^2 - U_1^2) < 1$$

Thus, the physical mode in his analysis is the source of any non-linear instability; the computational mode is a stabilizing factor.

In the more general criterion (6), however, we find that the coefficient U_0, representing the physical mode, is a stabilizing factor and U_2, representing a high frequency, and therefore a computational mode, is a destabilizing factor. Since U, and therefore U_1 and U_3, are coefficients of the high wave number component, $\cos \pi j$, in the solution (2), U_1 and U_3 represent high wave numbers in the solution, and likewise U_0 and U_2 represent low wave numbers.

The true situation is that high frequencies are either stabilizing or destabilizing, depending on their wave number. Similarly, high wave numbers in the solution are either stabilizing or destabilizing, depending on their frequency. The true situation is summarized in Table I.

It is also important to point out that neither the high frequencies nor the high wave numbers are in themselves unstable. Indeed, as shown by (4), they are perfectly stable, i. e., they are neither amplified nor damped. The same is true of low frequencies and low wave numbers. As Table I shows, however, the analysis indicates that a component with either high frequency or high wave number (but not both) causes instability in the middle wave numbers.

In this connection, note that the coefficient C in (2) multiplies $\cos \frac{1}{2} \pi j$, which cycles in four grid increments. The four-grid-increment wave is exactly in the middle of the spectrum, in the sense that there are just as many longer components as there are shorter.

The role of aliasing

The next question which will be considered here is the role aliasing plays. The shock equation,

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0$$

will be used for this investigation. Richtmyer (1963) showed that the analysis of the shock equation gives results in all important respects the same as Phillips' analysis of the barotropic equation. Analyses of the shock equation have often been interpreted in terms of the effects of the advective terms on stability of calculations with the more general hydrodynamical equations.

Table I

Space Time	Low wave number	High wave number
Low frequency (Physical mode)	U_0 Ri Stabilizing	U_1 Ph-Ri Destabilizing
High frequency (Computational mode)	U_2 Ro Destabilizing	U_3 Ph-Ri Stabilizing

Table I, showing schematically the effects on stability of the spectra extremes. The symbols, Ri, Ro, Ph, stand for Richtmyer (1963), Robert (1969), and Phillips (1959), respectively, and show the parts of the spectra included in their analyses.

Furthermore, the equations for a traveling gravity wave may be reduced to a single shock equation. For illustration, consider the system of equations which describe the dynamics of a gravity wave in a homogeneous incompressible fluid with a free surface and slab symmetry:

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + g \frac{\partial h}{\partial x} = 0$$

$$\frac{\partial h}{\partial t} + u \frac{\partial h}{\partial x} + h \frac{\partial u}{\partial x} = 0$$

where x is horizontal distance, t time, u velocity, g gravitational acceleration, and h the height of the free surface. Define a new variable, c, by

$$c^2 = g h$$

Then

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + 2 c \frac{\partial c}{\partial x} = 0$$

$$2 \frac{\partial c}{\partial t} + 2 u \frac{\partial c}{\partial x} + c \frac{\partial u}{\partial x} = 0$$

Addition and subtraction of these two equations yields

$$\frac{\partial (u+2c)}{\partial t} + (u+c) \frac{\partial (u+2c)}{\partial x} = 0$$

$$\frac{\partial (u-2c)}{\partial t} + (u-c) \frac{\partial (u-2c)}{\partial x} = 0$$

The last two equations each correspond to one of the two roots of the linearized gravity wave equation, and show that influences travel with the two speeds, u + c and u - c. Now let $c' = c - \bar{c}$, where $\bar{c}$ is the space-time average of c. If at an instant $u - 2c' = 0$ everywhere, the latter of the pair of equations shows that the condition holds for all time.

Letting, then,. $u = 2c'$ and eliminating u from the first of the pair, we find

$$\frac{\partial c'}{\partial t} + (3c' + \bar{c}) \frac{\partial c'}{\partial x} = 0 .$$

Introduction of the variable

$$U = 3c' + \bar{c}$$

leads to

$$\frac{\partial U}{\partial t} + U \frac{\partial U}{\partial x} = 0 .$$

The simple shock equation, therefore, represents a whole variety of phenomena described by the more general meteorological equations.

We will use the finite-difference form which Richtmyer analyzed,

$$u_{2t} = -\tfrac{1}{2}(u^2)_{2x}$$

The right-hand member may also be written in two precisely equivalent other ways:

$$u_{2t} = -\tfrac{1}{2}(u^2)_{2x} = -\overline{(\bar{u}^x \; u_x)}^{\,x} = -\bar{u}^{2x} \; u_{2x} \qquad (7)$$

The third member corresponds to the so-called semi-momentum form, which has been used at the National Meteorological Center in research and later in operations since 1959.

We will try as a special solution of (7)

$$u = C \cos \tfrac{1}{2} \pi j + S \sin \tfrac{1}{2} \pi j + U \cos \pi j + V \qquad (8)$$

where C, S, U, and V are functions of time (n) only. This trial solution differs from Richtmyer's in that he regarded V as a constant, which is a special case of (8). Now at grid points,

$$\begin{aligned} \bar{u}^{2x} &= - U \cos \pi j + V \\ u_{2x} &= \frac{1}{\Delta} \left(- C \sin \tfrac{1}{2} \pi j + S \cos \tfrac{1}{2} \pi j \right) \end{aligned} \qquad (9)$$

Therefore,

$$\bar{u}^{2x} \; u_{2x} = \frac{1}{\Delta} \left[\begin{aligned} &- C (- U \cos \pi j + V) \sin \tfrac{1}{2} \pi j \\ &+ S (- U \cos \pi j + V) \cos \tfrac{1}{2} \pi j \end{aligned} \right] (10)$$

But

$$\begin{aligned} \cos \pi j \cdot \cos \tfrac{1}{2} \pi j &= \tfrac{1}{2} \cos \tfrac{1}{2} \pi j + \tfrac{1}{2} \cos (3/2 \; \pi j) \\ \cos \pi j \cdot \sin \tfrac{1}{2} \pi j &= - \tfrac{1}{2} \sin \tfrac{1}{2} \pi j + \tfrac{1}{2} \sin (3/2 \; \pi j) \end{aligned} \qquad (11)$$

These relations are valid whether or not j is an integer.

Substitution from (11) into (10) and from (10) and (8) into (7) gives

$$\begin{aligned} &C_{2t} \cos \tfrac{1}{2} \pi j + S_{2t} \sin \tfrac{1}{2} \pi j + U_{2t} \cos \pi j + V_{2t} \\ &= \frac{1}{\Delta} \left[\begin{aligned} &S (\tfrac{1}{2} U - V) \cos \tfrac{1}{2} \pi j + \tfrac{1}{2} S U \cos \; 3/2 \; \pi j \\ &+ C (\tfrac{1}{2} U + V) \sin \tfrac{1}{2} \pi j - \tfrac{1}{2} C U \sin \; 3/2 \; \pi j \end{aligned} \right] (12) \end{aligned}$$

This will lead to a non-trival solution only if the components $\cos \; 3/2 \; \pi j$ and $\sin \; 3/2 \; \pi j$ can be related to the other components which appear in

(12), and indeed, at the grid-points, where j is an integer,

$$\begin{aligned} \cos(3/2\ \pi j) &= \cos(3/2 - 2)\pi j = \cos\tfrac{1}{2}\pi j \\ \sin(3/2\ \pi j) &= \sin(3/2 - 2)\pi j = -\sin\tfrac{1}{2}\pi j \end{aligned} \qquad (13)$$

In making a direct substitution from (13) into (12), then, we will have aliased $\cos(3/2\ \pi j)$ and $\sin(3/2\ \pi j)$ (components $4/3\ \Delta$ long) into $\cos\frac{1}{2}\pi j$ and $\sin\frac{1}{2}\pi j$ (components 4Δ long). We will, however, have done more. The components $4/3\ \Delta$ long interacting with the 2Δ components will generate $4/7\ \Delta$ components according to:

$$\cos\pi j \cdot \cos(3/2\ \pi j) = \tfrac{1}{2}\left[\cos\tfrac{1}{2}\pi j + \cos(7/2\ \pi j)\right]$$

The $4/7\ \Delta$ components will generate higher wave numbers, which in turn will generate even higher wave numbers, and so on. Since the $4/3\ \Delta$ component cannot be distinguished from the 4Δ component with the information at grid points, it is evident that the higher wave numbers also cannot. The upshot is that making a substitution from (13) is valid, but in doing so we will have aliased a multitude of high wave numbers into the 4Δ components. At any rate, making the substitution, we get,

$$C_{2t} = \frac{1}{\Delta} S(U - V)$$

$$S_{2t} = \frac{1}{\Delta} C(U + V)$$

$$U_{2t} = V_{2t} = 0$$

On the other hand, we can in principle achieve a non-trivial solution of (12) without aliasing, by modifying the numerical integration procedure. In particular, we are interested in the solution for a procedure in which all components shorter than 2Δ are removed immediately as they are generated.

This leads to

$$C_{2t} = \frac{1}{\Delta} S(\tfrac{1}{2}U - V)$$

$$S_{2t} = \frac{1}{\Delta} C(\tfrac{1}{2}U + V)$$

$$U_{2t} = V_{2t} = 0$$

Both numerical procedures may be indicated in a single set:

$$\begin{aligned} C_{2t} &= \frac{1}{\Delta} S[\tfrac{1}{2}(1+\delta)U - V] \\ S_{2t} &= \frac{1}{\Delta} C[\tfrac{1}{2}(1+\delta)U + V] \\ U_{2t} &= V_{2t} = 0 \end{aligned} \tag{14}$$

where if $\delta = 1$, the set describes the former procedure in which we have aliased. If $\delta = 0$, the set (14) describes the latter procedure in which we have removed all components shorter than $2\,\Delta$. In the latter case, not only have we aliased no components into others, we have not even included any components which can be aliased into the ones with which we started. Comparison of the stability of the two procedures, therefore, will shed considerable light on the role of aliasing.

We immediately write the solutions for U and V,

$$\begin{aligned} U &= U_1 + U_3 \cos \pi n \\ V &= U_0 + U_2 \cos \pi n \end{aligned} \tag{15}$$

Then proceeding as with the derivation of (6), we find the stability criterion to be

$$0 \leq \frac{(\Delta t)^2}{\Delta^2} \{ [U_0 \pm \tfrac{1}{2}(1+\delta)U_3]^2 - [\tfrac{1}{2}(1+\delta)\, U_1 \pm U_2]^2 \} < 1 \tag{16}$$

Again, the upper signs hold if the central n is even, the lower if odd.

The nature of the criterion is thus not changed whether $\delta = 0$ (no aliasing) or $\delta = 1$ (aliasing), although the balances among the various terms are somewhat different. Furthermore, the difference in the balance among the terms, depending on whether $\delta = 0$ or $\delta = 1$, will favor either stability or instability, depending on the values of U_0, U_1, U_2 and U_3 relative to each other.

Orszag (1971) has pointed out that, if components 3Δ ($\cos 2/3\ \pi j$ and $\sin 2/3\ \pi j$) and shorter are continually removed from the calculation, there will be no components to alias. Application of his idea to (7) and (8), leads to a numerical procedure in which U vanishes, and the stability condition becomes

$$0 \le \left(\frac{\Delta t}{\Delta}\right)^2 \left(U_0^2 - U_2^2\right) < 1 .$$

It should be remarked that (16) with $\delta = 1$, is identical to the stability criterion found by Robert et al. (1970) for the linear equation

$$f_{2t} + U f_{2x} = 0$$

$$U = U_0 + U_1 \cos \pi j + U_2 \cos \pi n + U_3 \cos \pi j \cdot \cos \pi n$$

The identity of the criteria follows from (9) and (15).

The nature of "aliasing"

Let us examine carefully the nature of "aliasing." First, consider a function, $f(x)$, given in the region $0 < x < P$. This may be expressed with a Fourier series:

$$f(x) = \sum_{k=0}^{\infty} \sigma_k \left(a_k \cos \frac{2\pi k x}{P} + b_k \sin \frac{2\pi k x}{P} \right)$$

where $\sigma_0 = \frac{1}{2}$

and $\sigma_k = 1$ if $k \neq 0$.

The discrete function, σ_k, has been introduced for convenience in the following derivations.

We now let $f(0) = f(P)$, and thereby let the region for which the series is valid be extended to $x = 0$ and $x = P$, i.e., $0 \le x \le P$. Since the corresponding component vanishes, we arbitrarily set $b_0 = 0$.

Now let the region $0 \le x \le P$ be divided into J equal intervals, Δ , and consider the situation where we as analysts are given $f(x)$ only at the points bounding the intervals. Let $j = x/\Delta$, a set of integers. Note that $J = P/\Delta$. We will denote the given discrete set of $f(x)$ by f_j, $1 \le j \le J$. Then by substitution,

$$f_j = \sum_{k=0}^{\infty} \sigma_k \left(a_k \cos \frac{2\pi k j}{J} + b_k \sin \frac{2\pi k j}{J} \right) \quad (17)$$

We immediately note that the subset of coefficients $b_{\frac{1}{2}NJ}$, where N is any integer, do not contribute to the sum at the given points, since the corresponding components vanish there. In other words, for $k = \frac{1}{2} N J$, $\sin (2\pi k j/J)$ is indistinguishable from its values when $k=0$. We may therefore "alias" the former into the latter. We may do so, however, only so long as our analysis includes no information about $f(x)$ except at the given points. The components have not themselves changed merely because certain information about $f(x)$ is being withheld from us. Other analysts who may have more information about $f(x)$ may not "alias" such components.

Now, we are given J values of f_j , but there are an infinite number of independent coefficients, a_k and b_k , in (17). Because of the disparity of information contained in f_j compared with a_k and b_k, we cannot determine the set of a_k and b_k from the set of f_j . Through the technique of "aliasing," however, sums of certain subsets of a_k and b_k can be determined. Note again, that it is we, the analysts, who are doing the "aliasing," nothing about the coefficients or their components is changed.

We proceed with the "aliasing" technique as follows: For any k in (17) a pair of integers, m and n, may always be found such that

$$k = m + n J$$

and $-\frac{1}{2} J \leq m \leq + \frac{1}{2} J.$

Furthermore, for a given k , m and n are unique, except for $m = \pm \frac{1}{2} J$. In that case, k may be expressed as

either

$$k = \frac{1}{2} J + n J$$

or

$$k = -\frac{1}{2} J + (n + 1) J$$

Equation (17) may now be expressed in terms of m and n in place of k :

$$f_j = \sum_{m=-\frac{1}{2}J}^{+\frac{1}{2}J} \sigma'_m \sum_{n=0}^{\infty} \sigma_n \left[\begin{array}{l} a_{m,n} \cos \dfrac{2\pi(m+nJ)j}{J} \\ +b_{m,n} \sin \dfrac{2\pi(m+nJ)j}{J} \end{array} \right] \tag{18}$$

where

$$\sigma'_m = \tfrac{1}{2} \text{ if } m = \pm \tfrac{1}{2} J$$

$$\sigma'_m = 1 \text{ if } m \neq \pm \tfrac{1}{2} J$$

The ambiguity about n when $m = \pm \frac{1}{2} J$ has here been resolved by including $\frac{1}{2}$ of each of the two possible pairs of m, n . As with σ, σ' has been introduced for convenience and economy in writing. The meaning of the limits on the first Σ in (18) is that the summation includes all terms at and inside, but no terms outside, the limits

$$-\tfrac{1}{2} J \leq m \leq +\tfrac{1}{2} J .$$

Thus for odd J, the first and last terms in the first sum of (18) are for $m = -\frac{1}{2}(J - 1)$ and $m = +\frac{1}{2}(J - 1)$, respectively. We also note that some of the terms in the summation for $n = 0$ imply $k < 0$, and therefore imply a subset of a_k and b_k not included in (17). Such a_k and b_k have been introduced with the condition

$$a_{-k} = a_k$$

$$b_{-k} = -b_k$$

Now, because n, J, and j are integers, and trigonometric functions are unchanged by a change in angle by any multiple of 2π, we write,

$$f_j = \sum_{m=-\frac{1}{2}J}^{\frac{1}{2}J} \sigma'_m \sum_{n=0}^{\infty} \sigma_n \left[\begin{array}{l} a_{m,n} \cos \dfrac{2\pi m j}{J} \\ + b_{m,n} \sin \dfrac{2\pi m j}{J} \end{array} \right] \qquad (19)$$

and have thus "aliased" all wave numbers higher than $\frac{1}{2}$ J into wave numbers lower than $\frac{1}{2}$ J. Equation (19) is not an identity, however, but only an equality, and its validity is restricted to the set of integer j .

Because the sign of the sine changes with a change in sign of the angle, and the sign of the cosine does not, (19) may be written:

$$f_j = \sum_{m=0}^{\frac{1}{2}J} \sigma_m \sigma_m' \sum_{n=0}^{\infty} \sigma_n \left[(a_{m,n} + a_{-m,n}) \cos \frac{2\pi mj}{J} + (b_{m,n} - b_{-m,n}) \sin \frac{2\pi mj}{J} \right]$$

or,

$$f_j = \sum_{m=0}^{\frac{1}{2}J} \sigma_m \sigma_m' \left(A_m \cos \frac{2\pi mj}{J} + B_m \sin \frac{2\pi mj}{J} \right) \quad (20)$$

where

$$\begin{aligned} A_m &= \sum_{n=0}^{\infty} \sigma_n \,(a_{m,n} + a_{-m,n}) \\ B_m &= \sum_{n=0}^{\infty} \sigma_n \,(b_{m,n} - b_{-m,n}) \end{aligned} \quad (21)$$

Through the technique of analysis called "aliasing," we have thus achieved an equation without disparate information content in the set of A_m and B_m compared with the set of f_j. Just as there are J independent values of f_j given for (20), there are J coefficients, A_m and B_m. If J is even, there are $(\frac{1}{2}J+1)$ of A_m and $(\frac{1}{2}J+1)$ of B_m, but two of B_m, B_o and $B_{\frac{1}{2}J}$, do not contribute to the sum because their corresponding components vanish for integer j . If J is odd, there are $\frac{1}{2}(J+1)$ each of A_m and B_m , but one of B_m, B_o, does not contribute to the sum. B_o, and $B_{\frac{1}{2}J}$ when it exists, may therefore arbitrarily be set to zero.

We analysts commonly determine only the set of A_m and B_m , because it is determinable from the set of f_j , whereas the set of a_k and b_k is not.

In the march forward in time with a non-linear finite difference equation such as (7), wave numbers higher than $\frac{1}{2}J$ are generated. Indeed their amplitudes can in principle be determined through the use of the predictive equation (7) and relations like (11), although in practice the number of components doubles each time step and it would therefore not be feasible to keep an accounting of them for very many time steps even with today's largest computers.

To relate this discussion to our earlier analysis, (8) should be interpreted as

$$u = \sum_{k=0}^{\infty} \sigma_k \left(a_k \cos \frac{2\pi kj}{J} + b_k \sin \frac{2\pi kj}{J} \right)$$

with k restricted to multiples of $\frac{1}{4}J$. If k is so restricted in the initial conditions, it will remain so restricted throughout the integration, for wave numbers of components generated by multiplication are sums and differences of k's, and therefore will themselves be multiples of $\frac{1}{4}J$.

Thus, with

$$k = m + 4n$$

and $-2 \leq m \leq +2$,

then

$$u = \sum_{m=0}^{2} \sigma_m \sigma_m' \left(A_m \cos \frac{2\pi mj}{4} + B_m \sin \frac{2\pi mj}{4} \right)$$

where

$$\sigma_0 \ \sigma_0' = \tfrac{1}{2}$$

$$\sigma_1 \ \sigma_1' = 1$$

$$\sigma_2 \ \sigma_2' = \tfrac{1}{2}$$

and

$$A_m = \sum_{n=0}^{\infty} \sigma_n \, (a_{m,n} + a_{-m,n})$$

$$B_m = \sum_{n=0}^{\infty} \sigma_n \, (b_{m,n} - b_{-m,n})$$

In (8), then,

$$V = \tfrac{1}{2} A_0 = \sum_{n=0}^{\infty} \sigma_n \, a_{0,n}$$

$$C = A_1 = \sum_{n=0}^{\infty} \sigma_n \, (a_{1,n} + a_{-1,n})$$

$$S = B_1 = \sum_{n=0}^{\infty} \sigma_n \, (b_{1,n} - b_{-1,n})$$

$$U = \tfrac{1}{2} A_2 = \tfrac{1}{2} \sum_{n=0}^{\infty} \sigma_n \, (a_{2,n} + a_{-2,n})$$

In determining stability in terms of whether V, C, S, and U are bounded or not, therefore, we are not looking at amplitudes a_k and b_k of individual components, but rather the sums, A_m and B_m, of amplitudes of selected subsets of those components. Aliasing is not anything happening in the calculation, but rather something which we analysts do when we group amplitudes into a finite number, J to be precise, of subsets. Semantically it is nonsense to say that one component aliases into another, although it may be acceptable as jargon if the meaning is precisely understood. More correct semantically is "one component is aliased (by the analyst) into another." In its nature, "aliasing" is not something done by the components in a calculation, but something which is done by we analysts in our analysis of the calculation. It is likewise incorrect to say that "aliasing" causes instability or any other characteristic of a calculation.

Trapping of energy and instability

An older idea, which is not very current, is that instability arises from interactions among the various parts of the spectrum such that there is a cascade of energy from low to high wave numbers, with energy trapped in the high wave numbers. Part of the idea, related to aliasing, is the supposition that the grid cannot express wave numbers higher than $\frac{1}{2}J$ (wave lengths shorter than 2Δ), and there is therefore a barrier in the spectrum at $\frac{1}{2}J$ near which energy accumulates. We have already dealt with the question of aliasing, and pointed out that a non-linear equation generates wave numbers higher than $\frac{1}{2}J$. In principle their amplitudes may be calculated as the integration proceeds. A practical problem would arise, though, with accounting for all wave numbers, since the numbers of components double with each time step. Another aspect of the problem is that if the wave number accounting is not kept as the integration proceeds, we analysts are faced with the fact that we cannot distinguish among the a_k ($a_{m,n}$) and b_k ($b_{m,n}$) which make up A_m and B_m in (21).

There is no barrier at $k = \frac{1}{2}J$, but merely insufficient information in the grid values at a single time step to distinguish all a_k and b_k from one another. To demonstrate more precisely, finite difference systems may be invented which allow cascade, but in which trapping does not occur. Consider, for illustration,

$$u_{2t} + \bar{u}^{xx} u_{2x} = 0$$

We again take (8) as the solution. Then

$$\bar{u}^{xx} = \tfrac{1}{2} C \cos \tfrac{1}{2} \pi j + \tfrac{1}{2} S \sin \tfrac{1}{2} \pi j + V$$

$$u_{2x} = \frac{1}{\Delta} (- C \sin \tfrac{1}{2} \pi j + S \cos \tfrac{1}{2} \pi j)$$

Substitution as before then leads to

$$
\begin{aligned}
C_{2t} &= -\frac{1}{\Delta}\, S\, V \\
S_{2t} &= \frac{1}{\Delta}\, C\, V \\
U_{2t} &= -\tfrac{1}{2}\, \frac{1}{\Delta}\, C\, S \qquad\qquad (22) \\
V_{2t} &= 0
\end{aligned}
$$

Following our earlier derivations, we find the criterion for stability to be

$$0 \le \frac{(\Delta t)^2}{\Delta^2}\,(U_0^2 - U_2^2) < 1$$

This criterion differs from the previous one (16) only in that high wave numbers (U_1 and U_3) play no role. The high wave numbers are active, however, as shown by (22). Interactions in the middle of the spectrum generate high wave numbers, but energy trapping does not occur. Instead the two grid increment wave oscillates, if the criterion is satisfied, and is therefore perfectly stable.

We grant that this derivation does not prove that trapping does not occur in more general cases, but it is a counter example. We have shown that the idea in the first place has fatal flaws in logic.

Energy conservation and stability

A principle used widely in designing finite difference forms for atmospheric modeling is due to Arakawa (1966). Under his basic principle the finite difference forms conserve a square, such as energy, when the functional dependence on time is not discretized. An example of Arakawa's principle applied to the shock equation is

$$\frac{\partial u}{\partial t} + \tfrac{1}{3}\,(u_{j+1} + u_j + u_{j-1})\; u_{2x} = 0$$

which may also be written

$$\frac{\partial u}{\partial t} + \tfrac{1}{3}\ (2\ \bar{u}^{2x} + u\ )\ u_{2x} = 0$$

Multiplying this equation by $2u$, we get, with a little manipulation,

$$\frac{\partial u^2}{\partial t} + \tfrac{2}{3}\ [u_{j+\frac{1}{2}}\ u_{j-\frac{1}{2}}\ \bar{u}^{x}\ ]_x = 0$$

If this is summed over j, from 1 to J,

$$\frac{\partial}{\partial t} \sum_{j=1}^{J} u_j^2 + \frac{1}{3\,\Delta} [\ u_{J+1}\ u_J\ (u_{J+1} + u_J\ )$$

$$-u_1\ u_0\ (u_1 + u_0)\ ]$$

and, if $u_0 = u_J = 0$, or if

$$u_j = u_{j+J},\ \sum_{j=1}^{J} u_j^2 \text{ is conserved.}$$

Of course in the larger atmospheric calculations, of which the shock equation merely represents some features, the functional dependence on time must be discretized, and therefore the conservation theorems are not satisfied. Nevertheless, it is typical of such calculations that the discrete approximations to the time derivative are highly accurate during a considerable period of the integration in time. Experience has shown that during such period, difference systems based on Arakawa's ideas do indeed conserve energy (or whatever other square is designed for conservation) to a correspondingly high degree of approximation. This can be a great advantage, especially in certain basic researches such as general circulation experiments, where it is often more important for the numerical model to copy the atmosphere in conservation of its energy and other fundamental quantities, than in ordinary accuracy of its evolution from one state to another.

The conservation argument cannot, however, as has sometimes been supposed, be used as an argument for stability. Using the shock equation as an example, if u^2 were indeed conserved, it is evident that u would be bounded, and therefore the calculation would be stable. When, however, the functional dependence on time is discretized, u^2 is not conserved. Rather than u^2, what is conserved is $u_{n+\frac{1}{2}}\ u_{n-\frac{1}{2}}$, with centered time-differences:

$$\left(\sum_{j=1}^{J} u_{j,n+\frac{1}{2}}\ u_{j,n-\frac{1}{2}} \right)_t$$

$$+ \frac{1}{3\Delta} \Big[u_{J+1,n}\ u_{J,n}\ (u_{J+1,n} + u_{J,n}) \\ - u_{1,n}\ u_{o,n} (u_{1,n} + u_{o,n}) \Big]$$

Let us seek stability conditions for

$$u_{2t} + \tfrac{1}{3}\,(2\ \bar{u}^{2x} + u)\ u_{2x} = 0$$

in the case of the restricted spectrum (8). Following the earlier procedure of substituting and equating coefficients, we find

$$\begin{aligned} C_{2t} &= \frac{1}{\Delta}\, S\,(\tfrac{1}{3}U - V) \\ S_{2t} &= \frac{1}{\Delta}\, C\,(\tfrac{1}{3}U + V) \\ U_{2t} &= -\frac{1}{\Delta}\, \tfrac{1}{3}\, C\, S \\ V_{2t} &= 0 \end{aligned} \qquad (23)$$

Again, we may immediately write the solution for V,

$$V = U_0 + U_2 \ \cos \pi n \,.$$

Otherwise, however, the set can apparently not be reduced to a linear form in the general case.

Special solutions, however may be found. For example,

$$C = \tfrac{1}{2} (1 + \cos \pi n) C'$$

$$S = \tfrac{1}{2} (1 - \cos \pi n) S'$$

where C' and S' are not restricted. This is the solution for any set of initial conditions with the restriction,

$$S_0 = C_1 = 0 .$$

Note that S is then zero for all even n , and C for all odd n . With this restriction,

$$U_{2t} = 0 , \text{ and}$$

$$U = U_1 + U_3 \cos \pi n$$

and the stability condition is

$$0 \leq \frac{\Delta t^2}{\Delta^2} \left[\left(U_0 \pm \tfrac{1}{3} U_3 \right)^2 - \left(U_2 \pm \tfrac{1}{3} U_1 \right)^2 \right] < 1$$

again with the upper signs holding for even n, the lower for odd n . This condition is only a small modification of (16), depending again on the same kind of balance among U_0, U_1, U_2, and U_3 .

In principle, the general stability conditions for the set (23) may be found through numerical experiment, if we modify our notion of stability somewhat. Instead of solutions being bounded for all n , we will accept as stable solutions those in which the solution does not exceed in magnitude a pre-selected large value, A^2, during a numerical integration to a pre-selected large value, N, of n·, i.e.,

$$|C|, \; |S|, \; |U| \leq A^2$$

for

$$0 \leq n < N .$$

In practice the calculation, and therefore the stability condition, depend on eight parameters, viz., the two initial values each of C, S, U, and V. A thorough investigation not therefore being practical, we conducted a limited one.

We let N = 8640 (corresponding to 60 days of 10 minute time steps) and called a run unstable if

$$E^2 = C^2 + S^2 + 2 U^2 > 1000(\Delta/\Delta t)^2$$

before n reached N . It may easily be demonstrated that E^2 is conserved if the difference ratios on the left hand sides of (23) are replaced by their corresponding derivatives. Equations (23), on the other hand, conserve as they stand the quantity

$$(C_{n+1} \; C_n \; + S_{n+1} \; S_n + 2 \; U_{n+1} \; U_n)$$

which was calculated at the end of each run and compared with its initial value as a check on the machine program. For one case we let N = 52560 (365 days of 10 minute steps) as a check on the appropriate magnitude for N .

Figure 1 shows a typical printout. It is for one experiment, consisting of 21 x 21 = 441 runs with (23). The printout is in the form of a table, with arguments U_0 and U_2 . The vertical argument is U_0, at intervals of $0.1 \times (\Delta/\Delta t)$, from zero at the top to two at the bottom. The horizontal argument is U_2 , at intervals of $0.1 \times (\Delta/\Delta t)$, from zero at the left to two at the right. In the experiment shown, the initial values of C, S, and U were all $0.1 \times (\Delta/\Delta t)$, and their values at the second time step were calculated with a forward time step.

The region between the two lines is the stable region. The lines were drawn connecting points representing runs which reached 8640 time steps without E^2 exceeding $1000 \times (\Delta/\Delta t)^2$; and at the same time enclosing all other such points. The figure shows

$$0 \leq \left(\frac{\Delta t}{\Delta}\right)^2 \left(U_0^2 - U_2^2\right) < 1 \qquad (24)$$

to be an approximation to the stability criterion. The upper straight diagonal line is $0.1 \times (\Delta/\Delta t)$ away from the left hand critical condition,

$$0 = U_0 - U_2$$

and the lower line is $0.1 \times (\Delta/\Delta t)$ away from the right hand critical condition,

$$\left(\frac{\Delta t}{\Delta}\right)^2 \left(U_0^2 - U_2^2\right) = 1$$

Thirty-two other experiments of 441 runs each were conducted, and sixteen, in which the initial values of C, S, and U were $0.1 \times (\Delta/\Delta t)$ or smaller, exhibited (24) as an approximation to the stability criterion. For the other sixteen, in which the initial C, S, and U were $0.5 \times (\Delta/\Delta t)$ or $1.0 \times (\Delta/\Delta t)$ the stable regions on the printouts disappeared or became very small.

Concluding remarks

Analyses such as performed here cannot describe all problems of stability in numerical atmospheric models. However, in a sense they set necessary conditions for stability. Because it is so trustworthy in practice, we overlook sometimes that the stability analysis of Courant et al. (1928) is also severely limited, and sets only necessary conditions for stability. Sufficient conditions for stability of the nonlinear equations have as yet not been defined.

These stability analyses, when applied to large calculations with a full spectrum, should not be interpreted only in terms of the initial conditions, which are usually smooth in space and time, and which therefore usually satisfy the conditions for stability derived here. The analyses are much less helpful than they would be if they were to set the initial conditions. Because of the severely restricted spectrum used, the analyses only tell us that certain conditions must not be violated during the integration if the variables are to remain bounded.

In large calculations, high-wave numbers and high frequencies usually have small amplitudes near the beginning of the calculation, but through nonlinear interactions they grow as the integration proceeds. We know little about the growth rates, other than what we observe in particular integrations with particular sets of difference equations and particular sets of initial data. The analyses do tell us, however, that feedback of high-wave numbers and high frequencies must be suppressed. This is commonly done, correctly, with either dissipative terms or with smoothing devices on variables in the equations, or in the case of high frequencies by adoption of forward time differences, or time-smoothing devices. The analyses tell us the nature of what we must do, but not knowing a priori the growth rates of high-wave numbers and high frequencies, we are unfortunately left to cut-and-try on a case-by-case basis in establishing quantitatively what to do.

Acknowledgments. The helpful criticism by Prof. Norman A. Phillips of an early draft is appreciated. Special acknowledgment is given to Dr. Joseph Sela of the National Meteorological Center, who criticised the rigor of my earlier derivations. I hasten to add that any faults of rigor which remain are my own. Finally, I publicly express my gratitude to Ruth N. Thornberg, who patiently and beautifully typed the manuscript replete with fussy equations, and under pressure of an overdue deadline.

REFERENCES

Arakawa, A., 1966: Computational design for long-term numerical integration of the equations of fluid motion: Two-dimensional incompressible flow. Part I. Journ. Computational Physics, 1, 119-143.

Courant, R., K. Friedrichs, and H. Lewy, 1928: Uber die partiellen Differenzengleichungen der mathematischen Physik. Mathematische Annalen, 100, 32-74.

Lilly, D. K., 1965: On the computational stability of numerical solutions of time-dependent non-linear geophysical fluid dynamics problems. Monthly Weather Review, 93, 11-26.

Orszag, S. A., 1971: On the elimination of aliasing in finite-difference schemes by filtering high-wave number components. J. Atmos. Sci., 28, 1074.

Phillips, N. A., 1959: An example of non-linear computational instability. The Atmosphere and the Sea in Motion, Rockefeller Institute Press, New York, 501-504.

Richtmyer, R. D., 1963: A survey of difference methods for non-steady fluid dynamics. NCAR Technical Notes 63-2, National Center for Atmospheric Research, Boulder, Colo., 25 pp.

Robert, A., 1969: An unstable solution to the problem of advection by the finite-difference Eulerian method. Office Note 30, National Meteorological Center, Weather Bureau, ESSA, Washington, D.C., 2 pp. (from an unpublished intraoffice series).

Robert, A., F. G. Shuman, and J. P. Gerrity, 1970: On partial difference equations in mathematical physics. Monthly Weather Review, 98, 1-6.

Figure 1

One of the computer outputs from a program designed to experimentally analyze the so-called energy conservation finite difference form of the shock equation.

	0.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80	1.90	2.00
0.00	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.30	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.40	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.50	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.60	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.70	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.80	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.90	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.00	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.30	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.40	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.50	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.60	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.70	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.80	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1.90	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2.00	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]