INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS, VOL. 4, 853–897 (1984)

THE STABILITY OF EXPLICIT EULER TIME-INTEGRATION FOR CERTAIN FINITE DIFFERENCE APPROXIMATIONS OF THE MULTI-DIMENSIONAL ADVECTION–DIFFUSION EQUATION

A. C. HINDMARSH AND P. M. GRESHO

Lawrence Livermore National Laboratory, Livermore, CA 94550, U.S.A.

AND

D. F. GRIFFITHS

University of Dundee, Dundee, Scotland

SUMMARY

A comprehensive study is presented regarding the numerical stability of the simple and common forward Euler explicit integration technique combined with some common finite difference spatial discretizations applied to the advection–diffusion equation. One-dimensional results are obtained using both the matrix method (for several boundary conditions) and the classical von Neumann method of stability analysis and arguments presented showing that the latter is generally to be preferred, regardless of the type of boundary conditions. The less-well-known Godunov–Ryabenkii theory is also applied for a particular (Robin) boundary condition. After verifying portions of the one-dimensional theory with some numerical results, the stabilities of the two- and three-dimensional equations are addressed using the von Neumann method and results presented in the form of a new stability theorem. Extension of a useful scheme from one dimension, where the pure advection limit is known variously as Leith's method or a Lax–Wendroff method, to many dimensions via finite elements is also addressed and some stability results presented.

KEY WORDS Stability Advection–diffusion von Neumann method Matrix method Explicit Euler

1. INTRODUCTION

In spite of its many concomitant problems, the simple explicit (forward) Euler method has been, and still is, widely used in computational fluid dynamics and related fields (e.g. air or water pollution). In this paper we study (as have many before us) certain aspects of an important and prototypical equation, the advection–diffusion equation, in 1, 2 and 3 dimensions in an Eulerian reference frame. Specifically, we will examine the stability of certain finite difference approximate solutions to the advection–diffusion (or convection–diffusion) equation

$$\partial\varphi/\partial t + \mathbf{u} \cdot \nabla\varphi = \nabla \cdot (\mathbf{K} \cdot \nabla\varphi) \tag{1}$$

where φ is the transported (advected and diffused) dependent variable, $\mathbf{u}$ is the (constant) advecting velocity vector, and $\mathbf{K}$ is the (constant) diffusivity tensor. Of central importance

0271–2091/84/090853–45$04.50

Received 14 January 1983
Revised 27 March 1984

will be the discretization resulting from the simplest second-order centred spatial differencing and first-order forward temporal differencing (Euler), often called FTCS (forward time, centred space).[1] For this case we will present a stability theorem based on the definition of stability which we believe to be most useful. The method is based on Fourier analysis in the manner of von Neumann and the results are well corroborated by actual computations.

Part of the motivation for this study lies in our use of the explicit Euler method to integrate the ordinary differential equations (ODEs) generated by the application of a (modified) Galerkin finite element method (FEM) using multilinear basis functions to the conservation equations for mass, momentum (Navier–Stokes equations), and energy in 2 and 3 dimensions.[2–4] This spatial discretization method generates a 9-point stencil in 2-D and a 27-point stencil in 3-D, the complete stability analysis of which has thus far proved intractable. As a first step, however, we have been successful in analysing the stability of the simpler FTCS scheme, which uses only a 5-point stencil in 2-D and a 7-point stencil in 3-D.

Another stimulus relates to the interesting history of stability predictions for this equation. Apparently beginning with Fromm[5] who incorrectly analysed the von Neumann stability equation (modified, as defined in Section 2.2) for FTCS in 2-D, the error was promulgated (1-D and 2-D) in the popular book by Roache.[1] The erroneous result, which places a severe restriction on the *spatial* discretization (i.e. the cell Reynolds (or Peclet) number cannot exceed unity) independently of the time-step size, would be a major deterrent to those interested in modelling advection-dominated flows by this method. Leonard[6] appears to be the first to have detected this error, at least for the 1-D case, and his results (which are a special case of the multi-dimensional results reported herein) are correct. (Earlier, however, Hirt[7] and Morton[8] had also presented the correct 1-D results.) More recently the confusion has again surfaced, beginning with a paper by Siemieniuch and Gladwell[9] hereafter referred to as S & G, in which they could not explain an important difference between their theoretical stability limits, obtained using a particular concept of stability which we call the matrix method, and their (1-D) numerical results—the latter were less stable than predicted by their theory. Another recent (2-D) attempt, this time using the Fourier method, was presented by Rigal.[10] Interestingly, he stated that the analytical work is 'not difficult', but 'somewhat tedious'; yet his results are, we believe, erroneous. In contrast, Mitchell and Griffiths[11] called the full analysis 'intractable'; unfortunately, however, an error in their ensuing simplified analysis resurrected a (2-D) cell Reynolds number constraint. Finally, the recent book by Lapidus and Pinder[12] interestingly reports the wrong result in one place (p. 187, 'after tedious calculation') and the correct one in another (p. 506). Considering these facts, and our own, thus far futile, attempts to complete the analysis for the more complex FEM discretization, it appears to us that the analytical work is indeed difficult.

Morton[13] has responded to the S & G paper and, among other things, explained their numerical results; he also made the following cogent remark, with which we agree: 'Unfortunately most of the analysis was based on the so-called matrix method, and an associated concept of stability, which is misleading both in theory and practice for such problems.' Griffiths *et al.*[14] also reacted to the S & G paper, in a different way, which also explained their observations. They showed that the error in the (1-D) FTCS scheme can become arbitrarily large in finite time if the matrix stability results are employed, even though the error does return to zero as $t \to \infty$. Another recent (1-D) discovery of the 'cell-Reynolds number error', and its correction, is given by Clancy,[15] who also presented supporting numerical results at large cell Reynolds number ($\sim$200). More recently, Griffiths[16] has shown, again in 1-D, that the (correct) results from the modified von Neumann method (for a particular set of boundary conditions) actually preclude error

growth as a function of time—in marked contrast to the looser and less conservative stability limits often obtained using the matrix method.

In the remainder of this paper we will analyse the stability of explicit Euler for several spatial discretization schemes (FTCS, a modified FTCS, and the simplest upwind scheme). The fundamental tool will be the von Neumann method, which 'ignores' boundary conditions, yet still usually yields the best results (instability is usually generated far from boundaries). However, we will also analyse some schemes using the same matrix method employed by S & G, including several boundary conditions, and comment on the differences. To conclude the section on 1-D, we will present a discussion of the behaviour to be expected when a numerical simulation is performed in which the stability limits are exceeded, inadvertently or otherwise. We believe that these results may be useful in interpreting certain anomalous behaviour when solving either non-linear or variable-coefficient advection–diffusion or even Navier–Stokes equations—they have been helpful to us. The remaining sections present the new results regarding the stability of FTCS and variants for both 2-D and 3-D, and some (partial) results for some finite element-based schemes.

2. THE ONE-DIMENSIONAL CASE

2.1. *Preliminaries*

The advection–diffusion equation (1) in the 1-D case, in a bounded domain, can be written as follows:

$$\partial\varphi/\partial t + u\,\partial\varphi/\partial x = K\,\partial^2\varphi/\partial x^2, \qquad 0 \le x \le 1, \quad t \ge 0 \tag{2}$$

where $K \ge 0$ and u are constant. The boundary conditions of interest initially (as in S & G) are $\varphi(0, t) = 1$, $\varphi_x(1, t) = 0$; the initial condition $\varphi(x, 0)$ is arbitrary, but given. The semi-discretized, second-order accurate, centred-difference form of this problem on a mesh with $\Delta x = 1/N$ is

$$\mathrm{d}\varphi_j/\mathrm{d}t = -u(\varphi_{j+1} - \varphi_{j-1})/2\Delta x + K(\varphi_{j+1} - 2\varphi_j + \varphi_{j-1})/\Delta x^2, \qquad 1 \le j \le N \tag{3}$$

with $\varphi_0 = 1$ and $\varphi_{N+1} = \varphi_{N-1}$ (the 'image point' method of approximating the derivative boundary condition). This has the form of a linear system of ordinary differential equations in $\mathrm{y} \equiv \{\varphi_j\}_{j=1}^N$:

$$\dot{\mathrm{y}} \equiv \mathrm{dy}/\mathrm{d}t = A\mathrm{y} + b \tag{4}$$

where A is a tridiagonal matrix and the vector b has zero components except for

$$b_1 = u/2\Delta x + K/\Delta x^2$$

Here, the eigenvalues of A all lie in the open left half plane (S & G) for $K > 0$. Hence, for any fixed N, the ODE system (4) is stable. For $K = 0$, the eigenvalues all lie on the imaginary axis and the ODE system is said to be weakly stable.

Next, we form the forward difference (Euler) time discretization of (4) to obtain the FTCS scheme:

$$\begin{aligned} \mathrm{y}^{(n+1)} &= \mathrm{y}^{(n)} + \Delta t(A\mathrm{y}^{(n)} + b) \\ &= E\mathrm{y}^{(n)} + \Delta t b \end{aligned} \tag{5}$$

where $E \equiv (I + \Delta t\, A)$, and $\mathrm{y}^{(n)} = \{\varphi_j^{(n)}\}$ is the discrete approximation to y at time $t_n = n\,\Delta t$.

This recursion can easily be accumulated to give

$$y^{(n)} = E^n y^{(0)} + (E^n - I)A^{-1}b \tag{6}$$

Thus the behaviour of the discrete solution depends entirely on the behaviour of powers of the tridiagonal matrix E. In particular, it is necessary that $E^n \to 0$ as $n \to \infty$ if a steady state solution of (4) is to be obtained. In terms of

the diffusion parameter,	$\alpha \equiv 2K\,\Delta t/\Delta x^2$
the Courant number (for $u \geq 0$),	$c \equiv u\,\Delta t/\Delta x$
and the grid Peclet number (for $u \geq 0$),	$P \equiv u\,\Delta x/2K = c/\alpha$

the non-zero elements of E are (reading across rows):

$$(\alpha + c)/2 = \alpha(1+P)/2, \qquad 1-\alpha, \qquad (\alpha - c)/2 = \alpha(1-P)/2$$

except for the last row, which has non-zero elements α, $1-\alpha$; in the above, $1-\alpha$ corresponds to the main diagonal. We will have need of these relations later.

2.2. *The von Neumann and matrix methods of stability analysis*

Historically, two different notions of stability have been applied to difference schemes such as (5). One, due to von Neumann, is based on a Fourier mode analysis. The other is based on a spectral radius analysis of the amplification matrix E. The paper by Morton[13] gives an enlightening, and we believe valid, comparison of the two, with an argument in favour of the Fourier method. (There appears to be an inconsequential error in that paper: the eigenvalues of E are *not* always distinct, but all coalesce when $P = 1$ ($\lambda h = 2$ in his notation). Thus the diagonalization preceding equation (8) there does not hold in this particular case. This coalescence was also noted by Price *et al.*[17] and S & G.) For detailed discussion on the concept of stability and on methods for its analysis, see Richtmeyer and Morton,[18] hereafter referred to as R & M.

The Fourier (or von Neumann) method applied to (5) consists of examining Fourier modes $\varphi_j^{(n)} = \xi^n e^{ijk\Delta x}$ for appropriate wave numbers k and associated wavelengths, $\lambda = 2\pi/|k|$. ($i \equiv \sqrt{-1}$). Strictly speaking, this analysis applies only to the problem with periodic boundary conditions rather than those posed, or, equivalently, for the initial value problem on the infinite x-axis; we shall say more on this later. Also, when analysing any single Fourier mode, we ignore the initial conditions actually posed, as they will be met only by a complete Fourier series. On substituting the above Fourier mode into the interior difference equations, a value for the complex amplification factor $\xi \equiv \varphi_j^{(n+1)}/\varphi_j^{(n)}$ is determined. As a function of the phase angle $\theta \equiv k\,\Delta x$, it is, in this case,

$$\xi = 1 - \alpha(1 - \cos\theta) - ic\sin\theta \tag{7}$$

Following Morton,[8,13] the (practical) von Neumann stability requirement that we employ is, for the stable ODE systems of interest here, $|\xi| \leq 1$. This is referred to as the modified von Neumann stability condition since the original, weaker condition, $|\xi| \leq 1 + O(\Delta t)$, often leads to unacceptably large errors (and would in this case[8]). For a thorough treatment of these issues, see R & M (p. 266) who originally introduced the stricter notion of 'practical stability conditions' based on the principle 'that no Fourier component of the approximation should be allowed to grow more rapidly than the most rapid possible growth of the exact solution'. Applied to the case of interest here, this leads to $|\xi| \leq 1$, since all modes of the exact solution

decay in time. Hereafter we will omit the adjectives 'practical' and 'modified' in the interest of brevity.

The choice of wave numbers k also deserves some discussion. The given mesh actually supports only some finite discrete set, such as

$$k = 2\pi n, \qquad n = n_0, n_0+1, \ldots, n_0+N-1 \tag{8}$$

where n_0 is any integer. Convenient choices for n_0 are (i) for N even, $n_0 = -N/2$ (a $2\Delta x$ wave), and (ii) for N odd, $n_0 = -(N-1)/2$ (nearly a $2\Delta x$ wave; a $2\Delta x$ wave cannot be represented exactly on the mesh for N odd). In fact, however, any vector $\{e^{ijk\Delta x}\}_{j=1}^{N}$ (with k an *arbitrary* real number) can be written as a linear combination of the (basis) vectors corresponding to the finite set (size N) of k values in (8); i.e. all other frequencies are 'aliased' to these, the only ones distinguishable on the mesh. Now the actual stability of the difference equations would require that $|\xi| \leq 1$ only for each of these *discrete* k values. But deriving conditions on Δt (in terms of the problem parameters) equivalent to this discrete maximum condition is usually much less tractable, and so the standard procedure is to impose the condition $|\xi| \leq 1$ for *all* real values of k, or equivalently for all θ, $-\pi \leq \theta \leq \pi$. This is therefore a slightly stricter condition than that required for any given finite N, but for large N the two conditions become essentially equivalent to within $O(1/N^2)$ in general. (See also Reference 19.) We will generally follow precedent (e.g., Reference 11) and continue to refer to the continuous k analysis as the von Neumann method. Finally we remark that the $k = 0$ mode (a constant vector) makes no contribution to the stability (or otherwise) of the difference scheme, since then $\xi = 1$ for any Δt; $\theta = 0$ is always neutrally stable.

In the case where ξ is given by (7), i.e. for FTCS, it can be shown[6–8,15] that $|\xi| \leq 1$ for all θ is equivalent to the pair of inequalities

$$c^2 \leq \alpha \leq 1 \tag{9}$$

These inequalities are necessary and sufficient for stability in the sense of von Neumann as defined above. This has the alternative equivalent forms

$$\alpha \leq 1 \quad \text{and} \quad \alpha \leq 1/P^2 \tag{10}$$

and

$$c \leq P \leq 1/c, \quad \text{or} \quad c \leq P \quad \text{and} \quad c \leq 1/P \tag{11}$$

In terms of the physical parameters of the problem, the 'diffusion limit' ($\alpha \leq 1$) is $\Delta t \leq \Delta x^2/2K$ and the 'advection–diffusion' limit ($c^2 \leq \alpha$) is $\Delta t \leq 2K/u^2$. The former is limiting when $P < 1$ and the latter governs when $P \geq 1$; in both cases $c \leq 1$ is a necessary consequence (i.e. it is a necessary, but not sufficient condition for stability). (Note that if we exclude the trivial case $\alpha = c = 0$, stability requires that $\alpha > 0$.)

The matrix method as used by S & G (and others) and herein, applied to (5), consists of computing (or at least getting an upper bound for) the spectral radius of E, $\rho(E)$, and defining the difference scheme to be stable if $\rho(E) \leq 1$ and eigenvalues of magnitude 1 (if any) are simple; in particular, it is stable if $\rho(E) < 1$. (In contrast with some authors, we do not allow $\rho(E) \leq 1$ unconditionally because this could allow unbounded growth of E^n, as we will demonstrate.)

An alternative notion of stability is to require that $\|E\| \leq 1$ for some (induced) matrix norm. This condition implies $\|E^n\| \leq 1$ and hence that the solution (see (6)) is uniformly bounded in n and N (as long as the initial and steady state vectors vary boundedly). Since $\rho(E) \leq \|E\|$, and this inequality is usually strict (it is here unless $P = 0$ and the norm is Euclidean), the matrix stability condition $\rho(E) < 1$ clearly cannot be trusted to give such boundedness.

Rather, it guarantees only that solution perturbations are damped to zero as $n \to \infty$ for *fixed* N and Δt[13,14] This distinction will be discussed further in Section 2.4.

As shown by S & G, a similarity transformation and some Gerschgorin analysis yield the following: for $0 \leq P \leq 1$, the eigenvalues, $\lambda(E)$, of E lie on the real interval

$$1-\alpha-\alpha\sqrt{(1-P^2)} \leq \lambda(E) \leq 1-\alpha+\alpha\sqrt{(1-P^2)}$$

whereas for $P>1$, the eigenvalues lie on a vertical interval with endpoints

$$1-\alpha-i\alpha\sqrt{(P^2-1)} \quad \text{and} \quad 1-\alpha+i\alpha\sqrt{(P^2-1)}$$

In the case $P \leq 1$, since the upper bound never exceeds 1, we can guarantee that $\rho(E)<1$ if

$$1-\alpha-\alpha\sqrt{(1-P^2)}>-1$$

or

$$\alpha<2/[1+\sqrt{(1-P^2)}] \tag{12}$$

In the case $P>1$, we have

$$\begin{aligned}\rho^2(E) &\leq (1-\alpha)^2+\alpha^2(P^2-1)\\ &= 1-2\alpha+c^2\end{aligned}$$

and thus $\rho(E)<1$ whenever $c^2<2\alpha$, or

$$\alpha<2/P^2 \tag{13}$$

Conditions (12) and (13) are sufficient to ensure that $\rho(E)<1$ but not necessary, because the eigenvalue bounds used are not sharp. However, a detailed analysis (the results of which will be shown later) shows that for large N the bounds are very close, so that the above upper bounds for α are nearly necessary as well as sufficient.

Comparing these results with those from the von Neumann method, it is seen that the matrix method always permits a larger time step for $P>0$; for $P \geq 1$ it is larger by a factor of 2, and for $P<1$ the factor is less than 2. It is just this (important) factor (between 1 and 2) that caused the confusion in the S & G numerical results and resulted in the follow-up papers by Morton[13] and Griffiths *et al.*[14] We will return to this point after considering the effect of different boundary conditions.

2.3. Stability results for other boundary conditions

Since the matrix A in (4) is (slightly) different for different boundary conditions, it follows that the (matrix) stability results may also differ. In this section, we summarize the results when the matrix method is applied to five types of boundary conditions: (i) Dirichlet, (ii) Neumann, (iii) Dirichlet/Neumann (the only case discussed thus far), (iv) periodic and (v) Dirichlet/Robin. In addition, we will present the corresponding spectra (eigenvalues and eigenfunctions) for the continuum problem, partly for completeness and partly as a check on the discrete results; i.e. in all cases, it can be shown that the discrete (matrix) results approach those from the continuum as $\Delta x \to 0$. Although relatively few of the following results are actually used in the sequel and although it is probably true that all of the results are already available (scattered through the literature), it seems useful to record them all in one place.

2.3.1. *Preliminaries.* To analyse stability via the spectral radius method, we begin by defining the eigenvalue problems associated with the continuum and the semi-discretized

systems, from (2) and (4), respectively. For the general case (inhomogeneous boundary conditions), a change of dependent variable (e.g. $\bar{\varphi} \equiv \varphi + \alpha + \beta x$ for appropriate α, β) recovers the homogeneous case (with a source term in general) and the stability results are the same.

Thus, for the continuum, we seek a solution to (2) in the form $\varphi(x, t) = \Phi(x)e^{-\lambda t}$ to give the continuous eigenvalue problem,

$$K\Phi'' - u\Phi' = -\lambda\Phi, \qquad 0 \le x \le l \quad (l = 1) \tag{14}$$

The corresponding approach applied to the semi-discretized system, (4), via $y(t) = Ye^{-\mu t}$ gives, for $b = 0$, the discrete eigenvalue problem

$$AY = -\mu Y \tag{15}$$

and the problem definitions are completed by specifying the (homogeneous) boundary conditions (which of course are also reflected in the elements of A) for the various cases. Stability limits on Δt will generally be of interest only when the ODE system (4) is stable, i.e. when each μ in (15) satisfies $\mathrm{Re}(\mu) \ge 0$ and eigenvalues with $\mathrm{Re}(\mu) = 0$ (if any) are simple. The equation corresponding to the general row of (15) (excluding the first and last, which vary with the boundary conditions) is

$$(K/\Delta x^2)(Y_{j+1} - 2Y_j + Y_{j-1}) - (u/2\Delta x)(Y_{j+1} - Y_{j-1}) = -\mu Y_j, \tag{16}$$

which is the discrete analogue of (14).

We now display the results ($\lambda, \Phi(x)$ for the continuum and μ, Y for the semi-discretized system) for the boundary conditions of interest and introduce the *global* Peclet number, defined here as $Pe \equiv ul/2K = Pl/\Delta x$.

2.3.2. *Dirichlet.* Here $\Phi(0) = \Phi(1) = 0$, $Y_0 = Y_{N+1} = 0$, and $\Delta x = 1/(N+1)$. The results are:

$$\lambda_n = K(n^2\pi^2 + Pe^2)/l^2 \tag{17a}$$

$$\Phi_n(x) = e^{Pex/l} \sin n\pi x/l; \qquad n = 1, 2, \ldots \tag{17b}$$

and

$$\mu_m = (2K/\Delta x^2)(1 - \sqrt{(1 - P^2)} \cos m\pi\, \Delta x/l) \tag{18}$$

$$Y_j^{(m)} = \begin{cases} \left(\dfrac{1+P}{1-P}\right)^{j/2} \sin jm\pi\, \Delta x/l, & \text{for } P < 1 \quad (19a) \\ (-i)^j \left(\dfrac{P+1}{P-1}\right)^{j/2} \sin jm\pi\, \Delta x/l, & \text{for } P > 1 \quad (19b) \end{cases}$$

where j, m range over $1, 2, \ldots, N$. Note that μ_m is real (like λ_m), and $Y_j^{(m)}$ is 'similar' to $\Phi_m(x)$ for $P < 1$, and μ_m is complex (unlike λ_m), with $Y_j^{(m)}$ also displaying a complex, ostensibly non-physical behaviour for $P > 1$. Finally, for $P = 1$, we have a bidiagonal matrix and the concomitant degenerate case,

$$\mu = 2K/\Delta x^2 \text{ for all } m \text{ (an eigenvalue of multiplicity } N) \tag{20a}$$

$$Y = (0, 0, \ldots, 0, 1)^T \text{ (only one eigenvector)} \tag{20b}$$

2.3.3. *Neumann.* Here $\Phi'(0) = \Phi'(1) = 0$ and, for the discrete system, the image point method is used at the first and last nodes, j ranges over $0, 1, 2, \ldots, N$, $\Delta x = 1/N$, and the size

of A is $(N+1)\times(N+1)$. The results are:

$$\lambda_0 = 0; \qquad \lambda_n = K(n^2\pi^2 + Pe^2)/l^2, \qquad n = 1, 2, \ldots \tag{21a}$$

$$\Phi_0(x) = 1; \qquad \Phi_n(x) = e^{Pex/l}[\cos n\pi x/l - (Pe/n\pi)\sin n\pi x/l]; \qquad n = 1, 2, \ldots \tag{21b}$$

and

$$\mu_0 = 0, \qquad \mu_m = (2K/\Delta x^2)(1-\sqrt{(1-P^2)}\cos m\pi\,\Delta x/l), \qquad m = 1, 2, \ldots, N-1 \tag{22a}$$

$$Y^{(0)} = (1, \ldots, 1)^T \tag{22b}$$

$$Y_j^{(m)} = \begin{cases} \left(\dfrac{1+P}{1-P}\right)^{j/2}(\cos jm\pi\,\Delta x/l - P\cot m\pi\,\Delta x/l \sin jm\pi\,\Delta x/l), & \text{for } P<1 \quad (22c) \\ (-i)^j\left(\dfrac{P+1}{P-1}\right)^{j/2}(\cos jm\pi\,\Delta x/l - P\cot m\pi\,\Delta x/l \sin jm\pi\,\Delta x/l), & \text{for } P>1 \quad (22d) \end{cases}$$

$j = 0, 1, 2, \ldots, N$ and $m = 1, 2, \ldots, N-1$

And finally,

$$\mu_N = 4K/\Delta x^2 \tag{23a}$$

$$Y_j^{(N)} = (-1)^j \tag{23b}$$

a $2\Delta x$ wave. For $P = 1$, the degenerate result is the same as for the Dirichlet case.

2.3.4. *Dirichlet/Neumann.* Here $\Phi(0) = \Phi'(1) = 0$ and, for the discrete system, $Y_0 = 0$ and the image point method is used at $x = 1$; here j ranges from 1 to N (the matrix size) and $\Delta x = 1/N$. The eigenvalues in this case are only available implicitly rather than in closed form:

$$\lambda_n = K(\gamma_n^2 + Pe^2)/l^2 \tag{24a}$$

where γ_n are the roots of

$$Pe\tan\gamma + \gamma = 0 \tag{24b}$$

$$\Phi_n(x) = e^{Pex/l}\sin\gamma_n x, \qquad n = 1, 2, \ldots \tag{24c}$$

Also

$$\mu_m = (2K/\Delta x^2)(1-\sqrt{(1-P^2)}\cos\psi_m) \tag{25a}$$

where the ψ_m are the N roots $(0 < \psi_m < \pi)$ of

$$P\tan N\psi_m + \tan\psi_m = 0 \tag{25b}$$

$$Y_j^{(m)} = \begin{cases} \left(\dfrac{1+P}{1-P}\right)^{j/2}\sin j\psi_m, & \text{for } P<1 \quad (25c) \\ (-i)^j\left(\dfrac{P+1}{P-1}\right)^{j/2}\sin j\psi_m, & \text{for } P>1 \quad (25d) \end{cases}$$

Again, for $P = 1$, the degenerate result is the same as for the Dirichlet case.

2.3.5. *Periodic.* Here $\Phi(0) = \Phi(1)$ and $\Phi'(0) = \Phi'(1)$, whereas for the discrete system we have $\Delta x = 1/N$, $j = 1, 2, \ldots, N$, and periodicity is enforced via $Y_0 = Y_N$ and $Y_{N+1} = Y_1$ (i.e. there are now entries in the $(1, N)$ and $(N, 1)$ positions of A). The results are

$$\lambda_n = 4n\pi K(n\pi + iPe)/l^2 \tag{26a}$$

$$\Phi_n(x) = e^{i2\pi nx/l}, \qquad n = 0, \pm1, \pm2, \ldots \tag{26b}$$

and

$$\mu_m = (2K/\Delta x^2)[(1-\cos 2m\pi\,\Delta x/l) + iP \sin 2m\pi\,\Delta x/l] \tag{27a}$$

$$Y_j^{(m)} = e^{i2\pi jm\,\Delta x/l}; \qquad m, j = 1, 2, \ldots, N \tag{27b}$$

and the latter apply for all values of $P \geq 0$.

The Dirichlet/Robin case is deferred until Section 2.3.8 for reasons which will become clear.

2.3.6. *Stability.* The (spectral radius) matrix method of stability analysis translates immediately to the following requirement for the eigenvalues of A (for all m):

$$|1-\mu_m\,\Delta t| \leq 1 \text{ and, whenever } |1-\mu_m\,\Delta t| = 1,\ \mu_m \text{ must be simple} \tag{28}$$

If $\mu_m = 0$, it must be a simple eigenvalue, and there is no corresponding Δt constraint. If μ_m is real and non-zero, it must be positive and (28) becomes

$$\Delta t < 2/\mu_m \quad \text{or} \quad \Delta t \leq 2/\mu_m \quad \text{for} \quad \mu_m \text{ simple} \tag{29}$$

If μ_m is complex, $\mu_m = \mu_m^{(r)} + i\mu_m^{(i)}$, then $\mu_m^{(r)}$ must be greater than zero and (28) becomes

$$\Delta t < 2\mu_m^{(r)}/[(\mu_m^{(r)})^2 + (\mu_m^{(i)})^2], \quad \text{or} \quad \Delta t \leq 2\mu_m^{(r)}/[(\mu_m^{(r)})^2 + (\mu_m^{(i)})^2] \quad \text{for} \quad \mu_m \text{ simple} \tag{30}$$

For all cases except Dirichlet/Robin (see Section 2.3.8), the conditions for the stability of (4) hold.

For the Dirichlet and Dirichlet/Neumann (mixed) boundary condition cases, it is easy to see, for the practical case $N \gg 1$, that these stability limits are

$$\Delta t \leq \frac{\Delta x^2/K}{1+\sqrt{(1-P^2)}}, \qquad \text{for } P < 1 \ (\mu_m \text{ is real}) \tag{31}$$

$$\Delta t \leq \Delta x^2/P^2K = 4K/u^2, \qquad \text{for } P > 1 \ (\mu_m \text{ is complex}) \tag{32}$$

and

$$\Delta t < \Delta x^2/K, \qquad \text{for } P = 1 \ (\mu_m \text{ is real, but not simple}) \tag{33}$$

For the Neumann case, we obtain, in addition to the above results,

$$\Delta t \leq \Delta x^2/2K \tag{34}$$

for any value of P; this corresponds to the $2\Delta x$ wave.

Finally, for the periodic case we first present the precise results, from (27):

$$\Delta t \leq \begin{cases} \Delta x^2/2K, & \text{for } N \text{ even and } P \leq 1 \quad (35a) \\ \dfrac{\Delta x^2/K}{(1+P^2)+(1-P^2)\cos \pi/N}, & \text{for } N \text{ odd and } P \leq 1 \quad (35b) \end{cases}$$

and

$$\Delta t \leqq \frac{4K/u^2}{(1+\cos 2\pi/N)+(1-\cos 2\pi/N)/P^2}, \qquad \text{for } P > 1 \tag{35c}$$

For $N \gg 1$ these are easily seen to be:

$$\Delta t \leq \Delta x^2/2K, \qquad \text{for } P \leq 1 \tag{36a}$$

and

$$\Delta t \leq 2K/u^2, \qquad \text{for } P > 1 \tag{36b}$$

which are equivalent to (9)–(11).

2.3.7. *Discussion.* In the two cases where at least one boundary condition is Dirichlet, the matrix results for the upper bound on Δt are larger than those from the von Neumann method, by a factor of $2/[1+\sqrt{(1-P^2)}]$ for $P \leq 1$ and by a factor of 2 for $P \geq 1$. To this extent then, Dirichlet boundary conditions are 'stabilizing'. The problem with this result, as already alluded to, as shown by Griffiths *et al.*,[14] and as will be further exemplified in Section 2.4, is that it is misleading; these looser limits on Δt can result in *significant* growth of the numerical 'solution' before the boundary conditions finally 'stabilize' it.

For the Neumann case, the matrix results are better, but still not strict enough since they too permit too large a Δt for $P > 1$ and a factor of 2 larger Δt when $P^2 > 2$ (the $2\Delta x$ wave leads to the correct result for $P \leq 1$).

Finally, for periodic boundary conditions, the results are in excellent agreement with those from the (continuous) von Neumann analysis which is, of course, not surprising. The matrix results for this case are the most conservative of all the above matrix results and the von Neumann result is only slightly ($O(1/N^2)$) more conservative than this. See also Reference 19 for further elaboration of this point. For the case of pure diffusion ($P = 0$), it is noteworthy in all cases that the matrix results agree with those from von Neumann analysis. It is the first spatial derivative term (advection) that causes the 'problem'; the system is no longer self-adjoint.

A final remark here is that the von Neumann method assumes that some 'energy' is contained in every Fourier mode whereas the matrix method basically tests the stability of each eigenvector. This distinction leads to the rare but possible case where the von Neumann results could be much too conservative; e.g. consider Dirichlet boundary conditions with the initial data specified to be the first eigenvector and $P = 0$. If $N \gg 1$, the true stability limit (to $O(1/N^2)$ from (18) and (29)) is $\Delta t \leq 2l^2/\pi^2 K$ whereas that from the von Neumann method is (still) $\Delta t \leq l^2/2N^2K$. These special cases are ignored in the general stability analysis by assuming 'arbitrary' initial data (which also includes the random perturbations associated with round-off error).

2.3.8 *Robin boundary conditions.* Finally we present a summary of results for one other type of 'mixed' boundary condition which is employed in practice (especially for pure diffusion): Dirichlet at the left ($\varphi = 0$ at $x = 0$) and Robin (or 'boundary condition of the third kind') at the right ($\partial\varphi/\partial x + h\varphi = 0$ at $x = l$, where $h \geq 0$ is a given constant). It is a special case in so many ways that is presented separately. Although the associated continuum eigenvalue problem is straightforward, that corresponding to the semi-discrete equations is very complicated; for this reason (in part), the matrix stability analysis is supplemented by one using more modern and powerful methods for stability analysis. Also, this case will be seen to be exceptional regarding stability: (1) under some conditions, it is less stable than a von Neumann analysis would predict (the Robin boundary condition can truly be destabilizing) and (2) under other conditions ($P > 1$ and h 'too large'), it is *unconditionally* unstable; i.e. the ODEs from FTCS are themselves unstable.

The latter problem (among others) is avoided if the Robin boundary condition is implemented via the finite element method (as a natural boundary condition) rather than the image point method usually associated with FTCS. Thus, we begin by presenting the two different ODEs associated with the last node (N): the image point method yields, with $H \equiv h\,\Delta x$,

$$(\Delta x^2/2K)\dot{\varphi}_N = \varphi_{N-1} - [1 + H(1-P)]\varphi_N \tag{37}$$

whereas the FEM-derived equation is

$$(\Delta x^2/2K)\dot{\varphi}_N + P(\varphi_N - \varphi_{N-1}) = \varphi_{N-1} - (1+H)\varphi_N \tag{38}$$

both of which imply $\partial\varphi/\partial x + h\varphi = 0$ as $\Delta x \to 0$. However, the very form of the 'finite difference' equation should be suspect at once if $P > 1$ and, in particular, if $H(P-1)$ is large, since then the ODE matrix has a large positive coefficient on the diagonal; indeed, this suspicion is warranted, as we will show. A further advantage of the FEM equation, which is borne out in practice, is that for $K = h = 0$ it yields a proper outflow equation consistent with the associated hyperbolic PDE (whereas the FTCS equation yields the spurious result $\dot{\varphi}_N = 0$). Thus (38) is especially useful as a passive outflow equation in the common case of advection-dominated flow ($P \gg 1$), especially when $h = 0$, where it is the FEM version of the Dirichlet/Neumann boundary condition discussed earlier via the image point method.

The spectrum for the continuous problem is

$$\lambda_n = K(\gamma_n^2 + Pe^2)/l^2 \tag{39a}$$

$$\Phi(x) = e^{Pex/l} \sin \gamma_n x \tag{39b}$$

where

$$\tan \gamma + \gamma/(hl + Pe) = 0 \quad \text{yields} \quad \gamma_n, \qquad n = 1, 2, \ldots \tag{39c}$$

The complete discrete results are too complex to state. Rather, we shall present an alternative stability analysis which gives the same results as the ($N \gg 1$) matrix method (which we have also analysed) when each is supplemented by the von Neumann condition.

Since we are primarily interested in the effect of the Robin boundary condition at $x = l$, we shall ignore the Dirichlet condition at $x = 0$ by reposing the problem on the quadrant $\{-\infty < x < 1, t > 0\}$. This device will have little discernible effect on the stability conditions, provided that N is sufficiently large; i.e. there are sufficiently many grid points in $(0 < x < 1)$ to avoid any significant interactions between the two boundaries. Discretization of (3) then leads to the equations

$$\varphi_j^{n+1} = (1/2)[\alpha(1+P)\varphi_{j-1}^n + 2(1-\alpha)\varphi_j^n + \alpha(1-P)\varphi_{j+1}^n] \tag{40}$$

for $-\infty < j \leq N$ and, employing first a simple centred difference approximation (image point) of the Robin boundary condition, we have

$$\varphi_{N+1}^{n+1} = \varphi_{N-1}^{n+1} - 2H\varphi_N^{n+1} \tag{41}$$

(Note that (40) at $j = N$ with (41) at time index n give exactly the forward time difference form of the ODE (37).)

The stability analysis we present is based on the work of Osher.[20] Though it is closely related to the normal mode analysis developed by Kreiss and his coworkers for hyperbolic equations (see, e.g. Reference 21), it has the advantage that it does not rely on consistency with differential equations of a particular type.

We seek a solution of (40) and (41) in the form

$$\varphi_j^n = C\xi^n \Phi_j \tag{42}$$

where C is a constant, ξ is the analogue of the amplification factor in von Neumann analyses, and Φ is a bounded function of j in the sense that

$$\Delta x \sum_{-\infty}^{N+1} (\Phi_j)^2 < \infty \tag{43}$$

Substituting (42) into (40) and (41) gives, respectively,

$$\alpha(1+P)\Phi_{j-1} + 2(1-\alpha-\xi)\Phi_j + \alpha(1-P)\Phi_{j+1} = 0 \tag{44}$$

and

$$\Phi_{N+1} + 2H\Phi_N - \Phi_{N-1} = 0 \tag{45}$$

Appealing now to the results of Godunov and Ryabenkii (see, e.g. R & M), necessary conditions for the stability of (40) and (41) are:

1. The system (40) should be stable in the sense of von Neumann, i.e. (9) must be satisfied.
2. Any non-trivial bounded solution to (44) and (45) must have $|\xi| \leq 1$. (If (40) and (41) were written in matrix-vector form, this condition would be equivalent to the requirement that the spectral radius of E should not exceed unity.)

To analyse when condition 2 holds, we note that Φ_j satisfies a three point difference equation and therefore has a general solution of the form $\Phi_j = A\kappa_1^j + B\kappa_2^j$ where $z = \kappa_1, \kappa_2$ are the two roots of the equations (from (44) and (45))

$$\alpha(1-P)z^2 + 2(1-\alpha-\xi)z + \alpha(1+P) = 0 \tag{46}$$

$$z^2 + 2Hz - 1 = 0 \tag{47}$$

Equation (47) implies that the product $\kappa_1\kappa_2 = -1$ so that we may stipulate $|\kappa_1| \geq 1 \geq |\kappa_2|$ (with strict inequality for $H \neq 0$) and, to obtain a bounded solution we must have $B = 0$ so that we only consider the larger root κ_1. From (47) we find $\kappa_1 = -H - \sqrt{(1+H^2)}$ and hence, from (46), we obtain

$$\xi = 1 - \alpha[1 - HP + \sqrt{(1+H^2)}] \tag{48}$$

Applying the necessary condition that $|\xi| \leq 1$ leads to

(i) If $P \leq 1$: $0 \leq \alpha \leq 2/[1 - HP + \sqrt{(1+H^2)}]$ (49a)

(ii) If $P > 1$: $H \leq 2P/(P^2-1)$ (49b)

and these have to be taken in conjunction with (10) from the von Neumann condition.

Remarks

(i) If $P = 1$, (49a) still applies but is less strict than the von Neumann condition, $\alpha \leq 1$.

(ii) For $P > 1$, the ODEs themselves ($N \gg 1$) are unstable if $H > 2P/(P^2-1)$; if $H < 2P/(P^2-1)$, they are stable, and the difference equations are then also stable for $\alpha \leq 1$.

We now turn to the derivation of sufficient conditions for stability. Specializing the results of Osher[20] to the present situation we find these to be:

1. The scheme should be stable in the sense of von Neumann; (10) must be satisfied.
2. For each $|\xi| > 1$, the roots of (46) should be distinct. It is easy to verify that for $0 < \alpha \leq 1$, $\alpha P^2 \leq 1$, $P \neq 1$, this condition holds. The degenerate cases $\alpha = 0$ and $\alpha > 0$, $P = 1$ are easily analysed separately and are stable as long as the von Neumann conditions hold.
3. $\xi \equiv 1$ should not be a solution of (46), i.e. we require $\alpha \neq 0$.
4. $\kappa_1^2 + 2H\kappa_1 - 1 = 0$ and $z = \kappa_1$ imply $|\xi| < 1$. Except for the case $|\xi| = 1$, the analysis follows that for the necessary conditions given above. In this way we derive (49) with strict equality omitted. If these conditions are satisfied, the scheme (40), (41) is stable.

We now examine the FEM version of the boundary equation (38), discretized in the obvious way,

$$\varphi_N^{n+1} = [1 - \alpha(1+H+P)]\varphi_N^n + \alpha(1+P)\varphi_{N-1}^n \tag{50}$$

This system ((40) for $-\infty < j < N$ and (50)) is equivalent to (40) for $-\infty < j \leq N$ with the boundary condition,

$$(1-P)\varphi_{N+1}^{n+1} + 2(H+P)\varphi_N^{n+1} - (1+P)\varphi_{N-1}^{n+1} = 0 \tag{51}$$

where the case $P = 1$ is temporarily excluded.

The analysis of (40) and (51) now closely follows the previous case with (47) replaced by

$$(1-P)z^2+2(H+P)z-(1+P)=0 \tag{52}$$

We now find that $\kappa_1=[P+H+\sqrt{(1+2HP+H^2)}]/(P-1)$ and the analogue of (48) is

$$\xi=1-\alpha[1+\sqrt{(1+2HP+H^2)}] \tag{53}$$

Necessary conditions for this scheme are therefore the von Neumann conditions (10) together with

$$0\leq\alpha\leq 2/[1+\sqrt{(1+2HP+H^2)}] \tag{54}$$

The specification of sufficient conditions for stability follows in an obvious way as in the preceding case and leads to (54) with strict equality omitted. The case $P=1$ can be shown to be stable, if (54) is satisfied, by a direct study of the difference equations.

Remarks

(i) There is never unconditional instability in this case.

(ii) For pure diffusion ($P=0$), both schemes give $\alpha\leq 2/[1+\sqrt{(1+H^2)}]$ as necessary and sufficient for stability. For this case, Gerschgorin analysis ensures stability if $\alpha\leq 2/(2+H)$. In at least several references[12,22,23] this more conservative Gerschgorin sufficient condition was presented as the stability requirement.

(iii) The stability analysis presented above can also be applied to the boundary conditions treated earlier by the matrix method, and would lead to (9) for all cases.

As a final remark concerning the Robin boundary condition, we point out that only the case $hl\leq O(1)$, giving $H\ll 1$, is 'practical' in the following sense: If $H\geq O(1)$, then $h(=H/\Delta x)\gg 1$ and the Robin boundary condition is actually very close (effectively) to the Dirichlet boundary condition ($\varphi=0$) and should be so replaced rather than adding unnecessary and expensive (small Δt is required) additional *stiffness* to the ODEs. And in this practical case, it is seen that (unless P is so large that $HP=O(1)$) the original von Neumann results come quite close to describing the true stability limits.

2.4. *Example of matrix method failure*

Initially, only the original (Dirichlet/Neumann) boundary condition case will be considered here, for simplicity. Recall that the actual discrete FTCS approximations, ignoring round-off errors, are given in vector form by (6), which involves the $N\times N$ tridiagonal matrix E. Thus the asymptotic nature of these values is governed entirely by that of the matrix powers, E^n. When Δt is limited by the matrix method stability condition (see equations (12) and (13)), we are guaranteed that the spectral radius of E satisfies $\rho(E)<1$, and therefore that $E^n\to 0$ as $n\to\infty$. However, this is a statement about the large n limit for fixed matrix order N, and says nothing about the sizes of the elements of E^n as both n and N vary. In fact, for a fixed but large value of N, these sizes can become extremely large before approaching zero as $n\to\infty$. (See also Reference 14.) This is the basic failure of the matrix method.

We can illustrate this flaw quite concretely by studying E^n in the simple case $P=1$. Here the matrix method stability bound is $\alpha<2$, whereas the von Neumann stability bound is $\alpha\leq 1$. The discrepancy, by a factor of 2, is as large as it ever gets. This choice also makes E^n easier to analyse, as E is now lower triangular and bidiagonal (and the ensuing analysis now also applies to the case with Dirichlet boundary conditions):

$$E=\begin{bmatrix} 1-\alpha & \searrow & \\ \alpha & \searrow & \searrow \end{bmatrix}$$

This case also illustrates the need to require more than just $\rho(E) \leq 1$, to ensure that $E^n \to 0$. For if $\alpha = 2$, we have $\rho(E) = 1$, but because 1 is a multiple eigenvalue, E^n has elements of magnitude $2n$ (among others of possibly greater size) for all $n > 1$. (See the Appendix for details.) Thus in this case E^n does not approach zero for any fixed N (see also Reference 24).

As we are interested in how large the elements of E^n can get, we first define

$$m_N(\alpha, n) \equiv \max_{i,j} |(E^n)_{ij}|$$

i.e. the maximum element magnitude in E^n. Thus for $\alpha < 2$, $m_N(\alpha, n)$ approaches zero as $n \to \infty$, but can get as large as

$$M_N(\alpha) \equiv \max_{n \geq 1} m_N(\alpha, n)$$

Figure 1 shows $m_N(\alpha, n)$ vs. n for $\alpha = 1.2$, $N = 40$ and 100; the peak value is roughly 10^6 for $N = 40$ and 10^{16} for $N = 100$. On the same plot is also shown the von Neumann (monotonic) growth curve, given by $(1{\cdot}4)^n$; the close proximity of the two results, during the growth phase, is interesting.

The Appendix contains a detailed analysis of the function $M_N(\alpha)$. In what follows, only the results will be stated.

When $\alpha \leq 1$, we find that $M_N(\alpha) \leq 1$ independent of N, and in fact the norms $\|E^n\|_\infty$ are also uniformly bounded by 1 (see also Reference 16). (Here $\|A\|_\infty \equiv \max_i \sum_j |\alpha_{ij}|$.) Uniform boundedness of E^n (in an appropriate norm), uniformly in n and N, is also stated by Morton[13] to be the essential stability property needed in the Lax equivalence theorem relating stability to convergence.

For any fixed α with $1 < \alpha < 2$, and for any $N \geq 2$, we find that

$$M_N(\alpha) \leq \left(\frac{\alpha}{2-\alpha}\right)^{N-1} \tag{55}$$

a bound which is obviously very large if $N \gg 1$ and α is close to 2. The estimates of the $m_N(\alpha, n)$ used for this result have their maximum value roughly at $n = n^* \equiv (N-1)/(2-\alpha)$

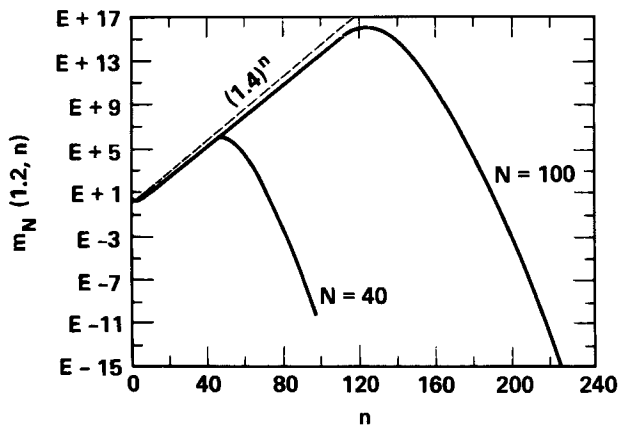

Figure 1. $m_N(\alpha, n)$ vs. n for $\alpha = 1{\cdot}2$ and $N = 40$, 100. The dashed line shows the corresponding growth of the von Neumann amplification factor, $|\xi|^n$

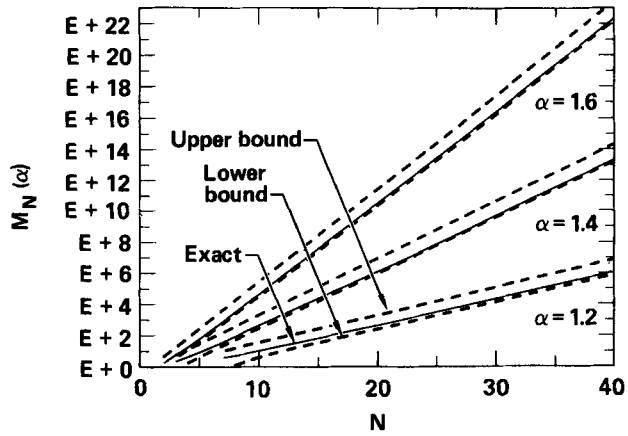

Figure 2. $M_N(\alpha)$ vs. N for three values of $\alpha > 1$. The dashed lines show the upper and lower bounds

for large N, and this suggests that the same is true for the $m_N(\alpha, n)$, themselves. Moreover, for large N and n near n^*, the estimates of the individual element magnitudes are maximized at the corner element $(E^n)_{N,1}$. This information leads to a good lower bound, namely:

$$M_N(\alpha) > 0{\cdot}335 \frac{[1-d(\alpha)/(N-1)]}{\sqrt{[(\alpha-1)(N-1)]}} \left(\frac{\alpha}{2-\alpha}\right)^{N-1} \tag{56}$$

where $d(\alpha) \equiv (2-\alpha)(5/2-\alpha)/(\alpha-1)$, which is valid for all $N \geq 1{\cdot}5/(\alpha-1)$. Figure 2 shows plots of $M_N(\alpha)$ for $\alpha = 1{\cdot}2$, $1{\cdot}4$ and $1{\cdot}6$, and these upper and lower bounds as functions of N. Note that the lower bound is quite close to $M_N(\alpha)$.

The ratio of the upper bound in (55) to the lower bound in (56) is clearly much smaller than either of the bounds themselves when N is large. Thus either bound can be taken as a good approximation to $M_N(\alpha)$ for $1 < \alpha < 2$ and all sufficiently large N, but we will use that in (55) because of its simplicity. It is clear then that for any fixed α with $1 < \alpha < 2$, the elements of E^n can be arbitrarily large, with a maximum magnitude growing exponentially with N as N gets large. When this happens, we can expect the computed transient FTCS solution to have grossly inaccurate values, even though they may (depending solely on Δx) approach an accurate steady state in the limit as $n \to \infty$.

The occurrence of numerically unstable or grossly inaccurate answers for FTCS when $1 < \alpha < 2$ was observed by S & G, but not explained by them. An explanation was given by Morton[13] and Griffiths *et al.*[14] and the present analysis provides an even more quantitative explanation, at least for the special case $P = 1$. In the notation of S & G, the advection coefficient is λ, the diffusion coefficient is 1, Δx is called h, and $r = \Delta t/h^2$, so that the case $P = 1$, $1 < \alpha < 2$ corresponds in their notation to $\alpha = \lambda h/2 = 1$, $1/2 < r < 1$, and the mesh size N is given by $N = 1/h = \lambda/2$. Figure 5 of S & G has plots, in terms of λh and r, of the regions that were found empirically to be stable or in which 'solutions obtained appeared stable but were very inaccurate'. Plots are given for $\lambda = 10$, 100 and 877·9. Along the line $\lambda h = 2(P = 1)$, the boundary between the stable and inaccurate regions appears to lie at about

$$r = 0{\cdot}72\ (\alpha = 1{\cdot}44), \qquad \text{for } \lambda = 10 (N = 5)$$
$$r = 0{\cdot}58\ (\alpha = 1{\cdot}16), \qquad \text{for } \lambda = 100 (N = 50)$$
$$r = 0{\cdot}5\ (\alpha = 1), \qquad \text{for } \lambda = 877{\cdot}9 (N = 439)$$

By comparison, if we define this boundary to be where $M_N(\alpha)$ reaches (say) 1000, and use the approximation in (55), we find that the corresponding values of α are

$$\alpha = 1{\cdot}70 \qquad (N = 5)$$
$$\alpha = 1{\cdot}07 \qquad (N = 50)$$

and

$$\alpha = 1{\cdot}008 \qquad (N = 439)$$

in reasonable agreement with the results of S & G.

Similarly, we can explain the numerical results of Griffiths *et al.*,[14] who give tables of $\|E^n\|_\infty$, together with some analytical upper and lower bounds for these norms. With regard to cases where very large values of this norm occur, there is relatively little distinction between the matrix norm $\|A\|_\infty$ and the maximum element magnitude studied herein, as the two can differ by at most a factor of N, and the values of $M_N(\alpha)$ in question are much larger than N. In our notation, Griffiths *et al.*[14] tabulated $\|E^n\|_\infty$ and its computed maximum value over n, for selected values of P, α and N. For $P = 1$, they consider only $\alpha = 1{\cdot}2$, and for the values $N = 10$, 20 and 40, their computed values of $\max \|E^n\|_\infty$ are 23, 1030 and $2{\cdot}6 \times 10^6$, respectively. By comparison, our estimates of $M_N(\alpha)$ from (55), with the same α and N values, are 38, 2217, and $7{\cdot}4 \times 10^6$, respectively. The agreement is within a factor of 3 in all cases. Griffiths *et al.*[14] also discuss the fact that large values of $\|E^n\|_\infty$ imply the possibility of very inaccurate computed transient solutions.

Finally we note that an example very close to this had been used previously by R & M (p. 152); although they were actually considering pure advection ($K = 0$) via upwinding, the final equations are essentially the same. (Pure advection with upwinding is equivalent to FTCS with $P = 1$.) More importantly, they point out that this is the type of matrix (not a normal matrix, which describes E for all $P > 0$) for which the spectral radius concept of stability is inappropriate.

We now discard the matrix method and in the remainder of the paper use the (continuous) von Neumann method almost exclusively.

2.5. Additional results from the von Neumann method

The von Neumann results from the FTCS case are easily extended to two related difference schemes. Although not new, these 1-D results are presented for completeness, since similar (and new) results will also be derived in 2-D and 3-D.

2.5.1. *Modified FTCS.* Although the correct FTCS stability limits are now known, they are still too restrictive for advection-dominated simulations, i.e. Δt must be too small (from $c \leq 1/P$) to be practical when $P \gg 1$. Thus we present a modified FTCS scheme and point out its significant advantages in cost-effectiveness. In a later section we will extend this scheme to multi-dimensions.

If the forward Euler method is applied to (2) prior to spatial discretization, a Taylor series analysis of the resulting scheme shows that the local time truncation error is responsible for reducing the effective diffusivity so that the equation actually being solved looks more like

$$\partial\varphi/\partial t + u\,\partial\varphi/\partial x = (K - u^2\,\Delta t/2)\,\partial^2\varphi/\partial x^2 \tag{57}$$

where higher order derivatives have been neglected. (This implies that $2K/u^2$ will be an approximate upper stability bound on Δt for *any* spatial discretization scheme which is sufficiently accurate (e.g. it does not apply to schemes such as upwinding which increase the

effective diffusivity via the advection term) and that the actual equation effectively being solved is closer to the *pure* advection equation the closer Δt is to this value; and, for FTCS, numerical experiments support this interpretation.[4]) This observation is one basis (for others, see Reference 1) for at least considering the following scheme for generating an approximate solution to (2):

$$(\varphi_j^{(n+1)} - \varphi_j^{(n)})/\Delta t + u(\varphi_{j+1}^{(n)} - \varphi_{j-1}^{(n)})/2\Delta x = (K + u^2\,\Delta t/2)(\varphi_{j+1}^{(n)} - 2\varphi_j^{(n)} + \varphi_{j-1}^{(n)})/\Delta x^2 \tag{58}$$

which we refer to as modified FTCS. In the absence of physical diffusion ($K = 0$) this is called Leith's method (e.g., see Reference 1); it is also equivalent, for the special case considered herein, to the Lax–Wendroff method (e.g., see Reference 25), again for $K = 0$.

Since (58) is equivalent, with K replaced by $K + u^2\,\Delta t/2$, to FTCS, the stability results for the latter may be applied directly to the modified FTCS scheme; e.g. by replacing α by $\alpha + c^2$ in (9). Here and henceforth we retain the original definitions of α and P, in terms of K. The results are $c^2 \leq \alpha + c^2 \leq 1$. The left inequality gives $\alpha \geq 0$ and the right one yields, using $\alpha = c/P$,

$$c \leq 2P/[1 + \sqrt{(1 + 4P^2)}] = [\sqrt{(1 + 4P^2)} - 1]/2P \tag{59a}$$

or, equivalently,

$$\alpha \leq 2/[1 + \sqrt{(1 + 4P^2)}] = [\sqrt{(1 + 4P^2)} - 1]/2P^2 \tag{59b}$$

It is noteworthy that modified FTCS is stable when $K = 0$ if $c \leq 1$ ($P = \infty$ in (59a)) whereas FTCS is unconditionally unstable in the absence of diffusion. In addition to significantly enlarging the stability limit for large P (specifically, for $P > \sqrt{2}$), the numerical phase speed ($K = 0$) is more accurate, especially for $c \to 1$. Overall, this scheme has much to recommend it over FTCS. For a discussion of phase and damping error of this scheme, see References 3 and 4.

2.5.2. *Upwinded advection.* If the advection operator in the spatial discretization is changed, for $u \geq 0$, to $u(\varphi_j - \varphi_{j-1})/\Delta x$, we have the well known, but controversial first-order (in space) upwind difference scheme. Since this scheme may also be derived from FTCS, by replacing K by $K + u\,\Delta x/2$, the necessary and sufficient conditions for stability of this scheme are also contained in (9), with α replaced by $\alpha + c$. The result (from $\alpha \leq 1$ in (9)) is

$$c \leq P/(1 + P) \quad \text{or} \quad \alpha \leq 1/(1 + P) \tag{60}$$

The other inequality in (9) is non-limiting, because the corresponding bound on c or α is always larger than the one above. Again $K = 0$ is permissible; stability then requires $c \leq 1$, as for modified FTCS. (If the matrix method were applied to this case, the results would be

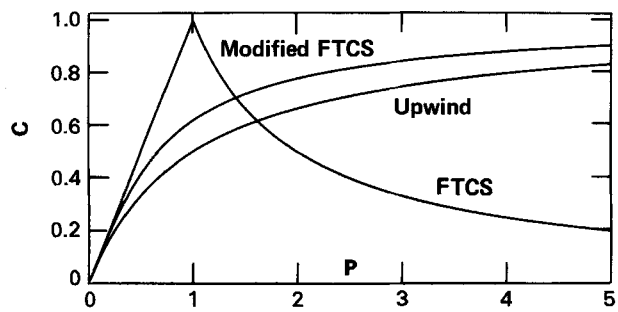

Figure 3. Stability limits for three schemes in the form of Courant number vs. grid Peclet number. The schemes are unstable if C lies above the curves

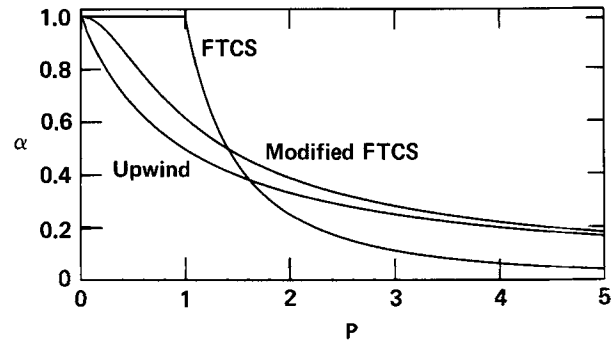

Figure 4. Stability limits for three schemes in the form of the diffusion parameter (α) vs. grid Peclet number. The schemes are unstable if α lies above the curves

$c<2$; this is the case critically discussed by R & M (p. 152).) This scheme, however, has little to recommend it because it is much too diffusive.[26]

Figures 3 and 4 show the stability results, as c and α vs. grid Peclet number, for the three schemes analysed above. (These results were previously obtained by Morton.[8]) FTCS only has a stability advantage over modified FTCS for $P<\sqrt{2}$; it is especially restrictive for advection–dominated flow ($P \gg 1$), which is probably one reason that it is rarely used in practice for this case.

2.6. *Behaviour of the most unstable mode*

It is of some interest to enquire about the behaviour of a numerical 'solution' if the stability limit is exceeded. Such information is useful, not only to further our understanding of the behaviour of a numerical algorithm, but also to provide clues and insight into the numerical results obtained in more difficult situations (e.g. non-constant or non-linear coefficients and/or a variable mesh, multi-dimensions, etc.). To this end then, we present a discussion of the detailed behaviour, in both time and space, when FTCS (and variants) is operated in the unstable regime for the case of periodic boundary conditions.

2.6.1. *Analysis for FTCS.* The starting point for the analysis is the complex amplitude coefficient from the von Neumann analysis, equation (7). From this equation we easily derive $F(z) \equiv |\xi|^2 - 1$ where $z \equiv \cos k\,\Delta x$:

$$F(z) = (1-z)[\alpha^2(1-z) + c^2(1+z) - 2\alpha], \quad -1 \le z \le 1 \tag{61}$$

and the FTCS scheme is unstable when $F>0$.

We will also need the following definitions:

(i) the most unstable mode is that whose wavelength ($\lambda_m = 2\pi/k_m$) generates the maximum value of $|\xi|$.
(ii) $|\xi_m| \equiv |\xi(\lambda_m)|$ is the growth rate of the most unstable mode.
(iii) the temporal period of the most unstable mode is denoted by τ.
(iv) The phase speed (u_p) is the apparent translational speed of the wave (which is just u in the continuum); it is given by $u_p \equiv \lambda_m/\tau$ for the most unstable mode.
(v) If Δt_c is the critical time step (neutrally stable, $|\xi_m| = 1$), the measure of the size of the instability will be ε, defined by $\Delta t \equiv (1+\varepsilon)\,\Delta t_c$.

Remarks

(i) Even though many wave numbers could be unstable for a given ε, the most unstable wave would *ultimately* dominate all others in most practical calculations (one involving essentially all of the possible Fourier modes). This of course means that the results to be presented may actually be misleading for certain short-time computations, in which the initial conditions may well cause a different unstable mode to dominate the solution.

(ii) The instability parameter need not be small; we will consider $0 \leqq \varepsilon < \infty$.

(iii) Wherever ambiguity may arise regarding positive or negative wave numbers, we shall use the convention that assumes $k > 0$; the final results are the same for either sign of k.

Since the behaviour of the most unstable mode depends in a crucial way on the value of the grid Peclet number (even more so than does the stability), it is convenient to classify and discuss it according to the value of P.

(i) $P \leq 1$. For this case we have $\Delta t_c = \Delta x^2/2K$ and thus,

$$\alpha = 2K\,\Delta t/\Delta x^2 = 2K\,\Delta t_c(1+\varepsilon)/\Delta x^2 = 1+\varepsilon$$

and

$$c = u\,\Delta t/\Delta x = u\,\Delta t_c(1+\varepsilon)/\Delta x = u\,\Delta x(1+\varepsilon)/2K = P(1+\varepsilon)$$

Inserting these into (61) for $P<1$ gives a simple parabola, concave upward, and a little analysis reveals that $F(z)$ first becomes positive at $z=-1$ for $\varepsilon>0$ and that $F(-1)$ is always the largest value of $F(z)$ on $-1 \leq z \leq 1$ for $\varepsilon > 0$. Since this corresponds to $k\,\Delta x = \pi$, we have $\lambda_m = 2\Delta x$ (the ubiquitous $2\Delta x$ wave) for this case. Inserting $\theta = k_m\,\Delta x = \pi$ and $\alpha = 1+\varepsilon$ into (7) gives $\xi_m = -(1+2\varepsilon)$ and a growth rate $|\xi_m| = 1+2\varepsilon$. Since then $\xi_m^n = (-1)^n\,|\xi_m|^n$, the $2\Delta x$ instability is manifest as a wave with a $2\Delta t$, period, i.e. $\tau = 2\Delta t$. This common instability is also probably the easiest to detect since it grows rapidly and 'everything' oscillates as fast as possible.

For $P=1$, $F(z)$ is linear and *all* modes become simultaneously unstable (longer waves are stable for small ε when $P<1$), with the $2\Delta x$ wave again showing the largest growth rate $(1+2\varepsilon)$ and a $2\Delta t$ period.

Finally, the phase speed is $u_p = \lambda_m/\tau = \Delta x/\Delta t$, and thus the wave moves one grid point per time step. The relative phase speed is $u_p/u = 1/c = 1/[P[(1+\varepsilon)]$, which can be much greater than 1 for small P (and ε not too large). The fact that a $2\Delta x$ wave is even *moving* is perhaps paradoxical, but we believe that this is a more appropriate interpretation than that which construes it to be stationary. If, however, $u=0$ (i.e. pure diffusion is being studied), we too revert to the interpretation that the (most unstable) $2\Delta x$ wave is stationary (we then regard ξ as purely real and abandon the notion of $u_p = \lambda_m/\tau$), as indeed is *any* wave; in this case, the standing wave will decay monotonically for $\alpha < 1/2$ but will display a $2\Delta t$ period for $\alpha > 1/2$ and will do so unstably if $\alpha > 1$. Finally we remark that the actual discrete solution is, of course, indifferent to the manner of interpretation.

(ii) $1 < P^2 < 2$. For $P>1$ we have $\Delta t_c = 2K/u^2$ and obtain, in a similar manner as above, $\alpha = (1+\varepsilon)/P^2$ and $c = (1+\varepsilon)/P$ which again yields a parabola for $F(z)$, this time concave downward and an interior maximum could occur. Then, setting $F'(z_m) = 0$ yields the (potentially) most unstable mode,

$$z_m = \cos\theta_m = \cos k_m\,\Delta x = \cos(2\pi\,\Delta x/\lambda_m) = (P^2-1-\varepsilon)/[(P^2-1)(1+\varepsilon)] \qquad (62)$$

where we require $|z_m| \leq 1$. The function $\lambda_m(\varepsilon)$ decreases monotonically from $\lambda_m \simeq (\pi \Delta x/P)\sqrt{[2(P^2-1)/\varepsilon]}$ for $\varepsilon \ll 1$ (but see Section 2.6.3), passes through $\lambda_m = 4\Delta x$ at $\varepsilon = P^2-1$, and reaches $\lambda_m = 2\Delta x$ at $\varepsilon = \varepsilon_c \equiv 2(P^2-1)/(2-P^2)$ (the value at which z_m reaches -1). For $\varepsilon > \varepsilon_c$, there is no maximum of $F(z)$ in the interior $|z|<1$; rather, the maximum now always occurs at $z=-1$ and thus, $\lambda_m = 2\Delta x$ is the most unstable mode for all $\varepsilon \geq \varepsilon_c$ (even though all modes are then unstable).

Inserting the above values of α and c into (7) with θ_m given by (62) gives

$$\xi_m = \frac{P^2-1-\varepsilon}{P^2-1} - \frac{i}{P^2-1}\sqrt{\{\varepsilon[(P^2-1)(2+\varepsilon)-\varepsilon]\}} \tag{63}$$

which is valid when $\varepsilon \leq \varepsilon_c$. The growth rate, from (63), is

$$|\xi_m| = \sqrt{\left(1+\frac{\varepsilon^2}{P^2-1}\right)}, \qquad \text{for } \varepsilon \leq \varepsilon_c \tag{64a}$$

and (from (7) with $\theta = \pi$)

$$|\xi_m| = 2(1+\varepsilon)/P^2 - 1, \qquad \text{for } \varepsilon > \varepsilon_c \tag{64b}$$

Equation (63) can also be expressed as $\xi_m = |\xi_m|\, e^{-i\psi}$ where ψ is the phase angle, given by

$$\tan\psi = \frac{\sqrt{\{\varepsilon[(P^2-1)(2+\varepsilon)-\varepsilon]\}}}{P^2-1-\varepsilon} \tag{65}$$

The period of the oscillation associated with ξ_m^n is then $2\pi/|\psi|$ time steps, or

$$\tau = 2\pi\, \Delta t/|\psi| \tag{66}$$

The period, $\tau(\varepsilon)$, decreases monotonically from $\tau \simeq \pi\, \Delta t\sqrt{[2(P^2-1)/\varepsilon]}$ for $\varepsilon \ll 1$ (but see Section 2.6.3) and passes through $4\Delta t$ at $\varepsilon = P^2-1$. Thus, at $\varepsilon = P^2-1$, the most unstable mode has a $4\Delta x$ wavelength and a $4\Delta t$ period. τ continues to decrease with increasing ε, finally reaching $2\Delta t$ at $\varepsilon = \varepsilon_c$, the same point that λ_m attains $2\Delta x$. For $\varepsilon \geq \varepsilon_c$, the most unstable wave has a length of $2\Delta x$ and a period of $2\Delta t$.

Finally, the relative phase speed, $u_p/u = \lambda_m/u\tau$, decreases like $1-\varepsilon/3$ for $\varepsilon \ll 1$, passes through $1/P$ at $\varepsilon = P^2-1$ (the $4\Delta x$, $4\Delta t$ wave), and is given by $u_p/u = 1/c = P/(1+\varepsilon)$ for $\varepsilon \geq \varepsilon_c$ (for $\varepsilon < \varepsilon_c$, u_p is obtained from (62) and (66)); the latter is equivalent to $u_p = \Delta x/\Delta t$ and again the $2\Delta x$ wave moves one grid point per time step with a $2\Delta t$ period.

(iii) $2 \leq P^2 < \infty$. For this advection-dominated case, the maximum of $F(z)$ remains in the interval $|z| \leq 1$ for all ε so that (62)–(66) (except (64b) when $\varepsilon_c = \infty$) are always applicable and there are no $2\Delta x$, $2\Delta t$ waves (except for $P^2 = 2$ and $\varepsilon \to \infty$). Asymptotically, as $\varepsilon \to \infty$, we have $\lambda_m \to 2\pi\, \Delta x/\cos^{-1}[-1/(P^2-1)] \geq 2\Delta x$ and $\tau/\Delta t \to 2\pi/\tan^{-1}[-\sqrt{(P^2-2)}] \geq 2$, where it is important to note that the argument of the inverse tangent is in the second quadrant of the complex plane (for ξ, cf. (63)).

(iv) $P = \infty$. For the pure advection ($K = 0$) case, although all modes are unstable for *any* Δt, the $4\Delta x$ wave is the most unstable. Its growth rate is $|\xi_m| = \sqrt{(1+c^2)}$ and its period is $\tau = 2\pi\, \Delta t/\tan^{-1} c$ which varies from $\tau \simeq 2\pi\, \Delta t/c = 2\pi\, \Delta x/u$ for $c \ll 1$, through $\tau = 8\Delta t$ at $c = 1$ and finally to $\tau \simeq 4\Delta t$ for $c \gg 1$. The relative phase speed is: $u_p/u = (2/\pi c)\tan^{-1} c$ which decreases monotonically from $2/\pi$ at $c = 0$ toward zero at large c.

The most unstable wavelength is plotted in Figure 5 and its growth rate, period, and phase speed (for $P \geq 1$) are shown in Figures 6–8, where it is interesting to note the small growth rates and large periods when $P \gg 1$, for which the critical time step is very small. Apparently,

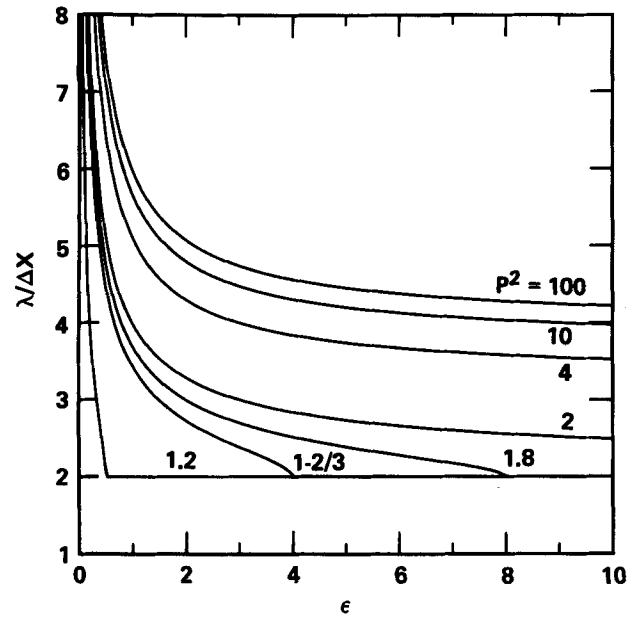

Figure 5. Most unstable wavelength for FTCS vs. Δt for several grid Peclet numbers

since $c \ll 1$ for these cases, very many times steps may be required before unstable behaviour would be clearly detected.

It follows from the above discussion that the behaviour of FTCS in the general case (arbitrary initial data) can be expected to be quite complex when $P > 1$ and $\varepsilon > 0$. The instantaneous waveform will be a combination of many different unstable waves and the dominance of a $2\Delta t$ oscillation will be rare. Only at very large times (for which amplitudes

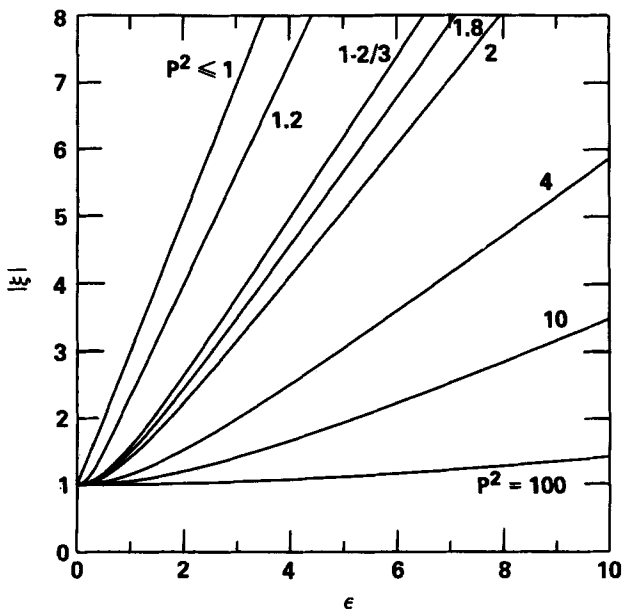

Figure 6. Growth rate of the most unstable mode for FTCS vs. Δt for several grid Peclet numbers

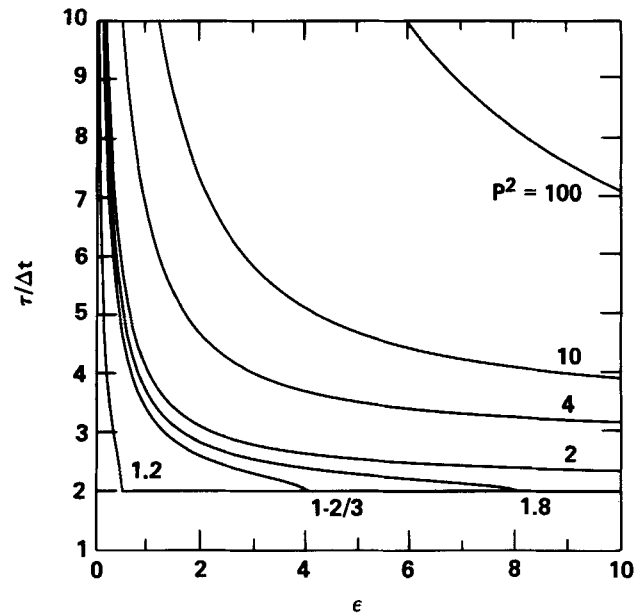

Figure 7. Period of the most unstable mode for FTCS vs. Δt for several grid Peclet numbers

could be ridiculously large) will the behaviour be predictable; it will then agree with the graphs as the most unstable mode finally prevails.

2.6.2. *Variants of FTCS.* For both modified FTCS and upwinded advection, it is straightforward to determine that $F(z)$, see (61), is (for any Δt of interest here) a parabola of the same general character as that for FTCS when $P \leq 1$; but now it is true for all values of P. Thus, the most unstable mode is the shortest wave, $\lambda_m = 2\Delta x$, with the shortest period,

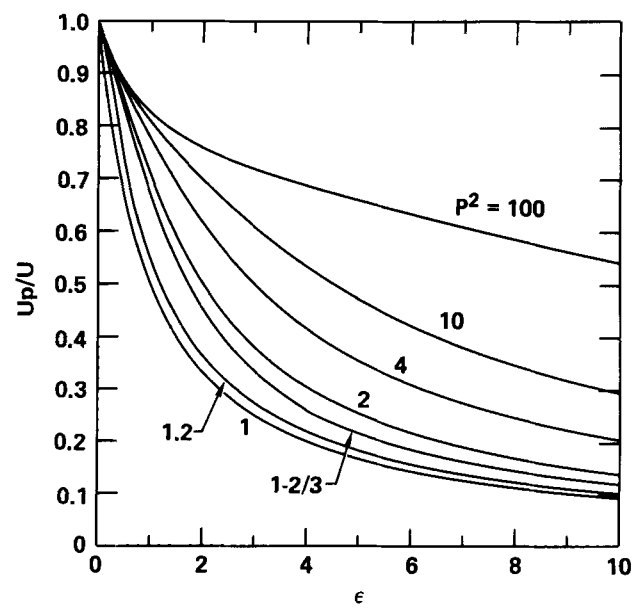

Figure 8. Phase speed of the most unstable mode for FTCS vs. Δt for several grid Peclet numbers

$\tau = 2\Delta t$, and the phase speed is $u_p = \Delta x/\Delta t$ or $u_p/u = 1/c$. Unstable behaviour should thus be simpler to detect than for FTCS, especially for modified FTCS, which has a larger growth rate (especially for large P) when unstable; it is given by

$$|\xi| = 1 + 2\varepsilon[1 + c_c^2(1+\varepsilon)] \tag{67}$$

where $c_c = 2P/[1+\sqrt{(1+4P^2)}]$ is the critical Courant number. For the upwind scheme it is $|\xi| = 1 + 2\varepsilon$, as for FTCS when $P \leq 1$.

2.6.3. *Modifications for discrete systems.* As noted earlier, the matrix results are in very good agreement with the (continuous k) von Neumann analysis when $N \gg 1$ and periodic boundary conditions are employed. Here we note the principal differences between these two analyses for the most unstable mode of FTCS, the matrix results now being the precise ones. (See also Reference 19.)

If $P = 1$, the results of the two analyses are identical. If $P \neq 1$, the critical Δt is very slightly larger than predicted by the von Neumann method for some cases; the critical wavelength and period are also affected:

(i) $P < 1$. If N is odd, the most unstable mode has a wavelength of $\lambda/\Delta x = 2/(1-\Delta x)$ rather than 2, and, from (35b),

$$\Delta t_c/\Delta t_{vN} = \frac{2}{(1+P^2)+(1-P^2)\cos \pi/N} \tag{68a}$$

$$\simeq 1 + (1-P^2)\pi^2/4N^2, \quad \text{for} \quad N \gg 1 \tag{68b}$$

where $\Delta t_{vN} = \Delta x^2/2K$ is the von Neumann result. If N is even, both Δt_c and the most unstable wavelength agree with von Neumann. In both cases, the period and phase speed agree with von Neumann.

(ii) $P > 1$. For this case, the critical Δt is, from (35c),

$$\Delta t_c/\Delta t_{vN} = \frac{2}{(1+\cos 2\pi/N)+(1-\cos 2\pi/N)/P^2} \tag{69a}$$

$$\simeq 1 + (1-1/P^2)\pi^2/N^2 \quad \text{for} \quad N \gg 1 \tag{69b}$$

where $\Delta t_{vN} = 2K/u^2$. Probably the most important difference is that the most unstable mode (at Δt_c; from (27a) and (30)) actually has $\lambda_m = N\,\Delta x = 1$, the longest (finite) resolvable wave, rather than $\lambda_{vN} \to \infty$, with a period given by $\tau/\Delta t_c \simeq NP$ (for $N \gg 1$) rather than $\tau_{vN} \to \infty$. For finite N the precise result is

$$\tau/\Delta t_c = \frac{2\pi}{\tan^{-1}\left[\dfrac{2P \sin 2\pi/N}{(P^2-1)+(P^2+1)\cos 2\pi/N}\right]} \tag{70}$$

For larger Δt, successively shorter waves (with shorter periods), become the most unstable.

2.7. *Numerical results*

We have numerically verified essentially all of the theory presented above for two cases and periodic boundary conditions: FTCS and modified FTCS. For example, we verified (FTCS) that the most unstable wave has $\lambda = 2\Delta x$ for $P \leq 1$, whereas a long wave, given (approximately) by (62), grows fastest for $P > 1$. Critical Δt's, growth rates and periods were also in accord with the theory. We appeal to the numerical results of S & G to show that the von Neumann theory is also quite useful for other boundary conditions.

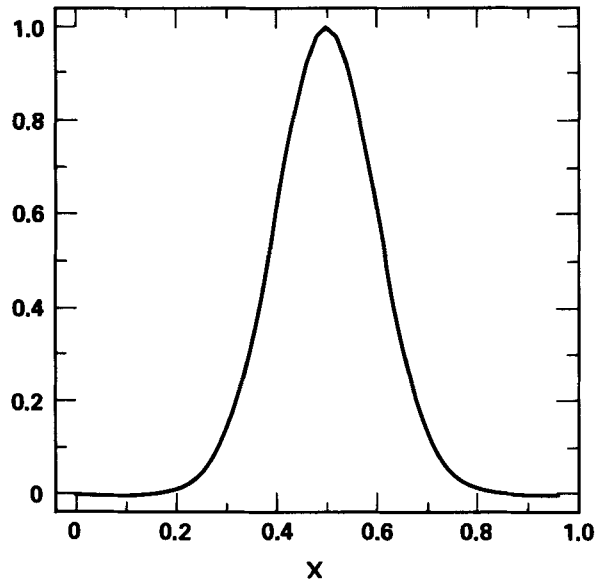

Figure 9. Inital condition for all of the following Figures

We will present results at the critical $\Delta t(\varepsilon = 0)$ and in the unstable regime (at $\varepsilon = 1$) for FTCS and modified FTCS for $P = 10$, an advection–dominated situation. Again, Δt is given by $\Delta t = (1 + \varepsilon)\,\Delta t_c$, where Δt_c is obtained either from the von Neumann results or those from the (periodic) matrix result. We take $N = 50(\Delta x = 0{\cdot}02)$ and $u = 1$, giving $K = 0{\cdot}001$. At $t = 0$ we place a Gaussian wave with $\sigma = 0{\cdot}1 = 5\Delta x$ on the unit span (see Figure 9). For FTCS the von Neumann theory predicts $\Delta t_c = 0{\cdot}002$ and for $\varepsilon = 1$, $\lambda_m/\Delta x \simeq 5{\cdot}97$, $\tau/\Delta t \simeq 36{\cdot}15$, and

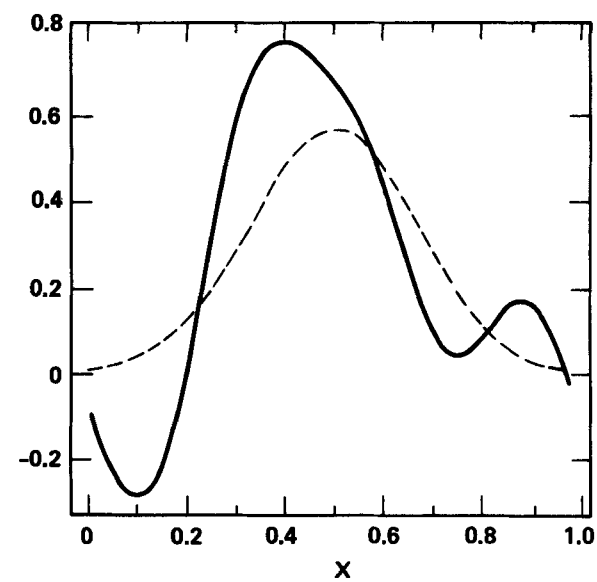

Figure 10. FTCS solution for $\varepsilon = 0$ at $t \simeq 10$. the dashed line represents an analytic solution here and in all remaining Figures

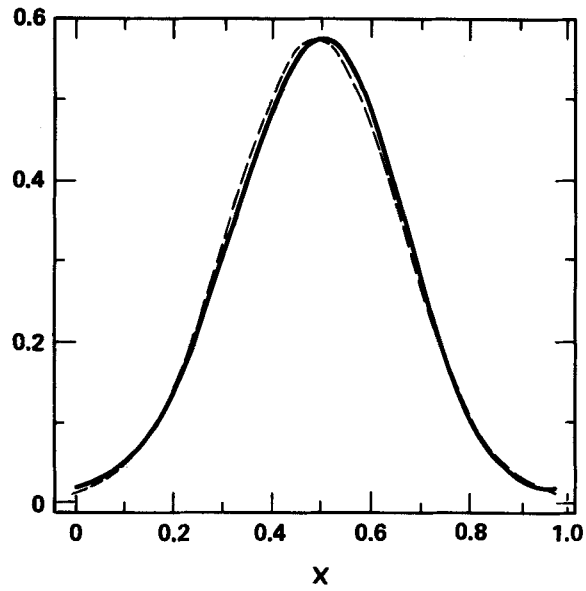

Figure 11. Modified FTCS solution for $\varepsilon = 0$ at $t \simeq 10$

$|\xi_m| \simeq 1{\cdot}00504$. The matrix theory (more precise in this case) yields $\Delta t_c \simeq 0{\cdot}002008$, and for $\varepsilon = 1$, $\lambda_m/\Delta x \simeq 6{\cdot}25$ (the eighth mode), $\tau/\Delta t \simeq 37{\cdot}07$, and $|\xi_8| \simeq 1{\cdot}00508$. In addition, the first 12 modes (wavelengths from $50\Delta x$ to $25/6\Delta x$) are unstable and the last 38 are stable; the growth rates of modes 7 and 9 are within 0·038 and 0·005 per cent, respectively, of that for mode 8, so we may expect that many time steps will be required to clearly see mode 8. For modified FTCS, the von Neumann theory gives $\Delta t_c \simeq 0{\cdot}019$ ($c_c \simeq 0{\cdot}951$) and, of course, $\lambda_m/\Delta x = \tau_m/\Delta t = 2$ for $\varepsilon > 0$. Finally, from (67), the growth rate at $\varepsilon = 1$ is $|\xi_m| \simeq 6{\cdot}62$.

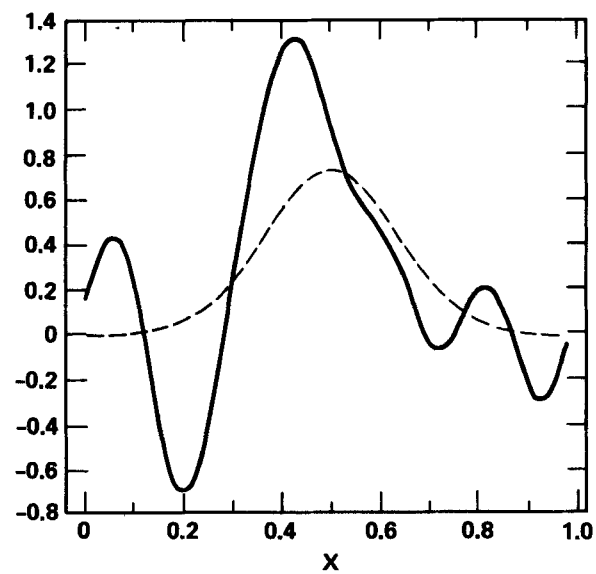

Figure 12. FTCS solution for $\varepsilon = 1$ at $t \simeq 4$

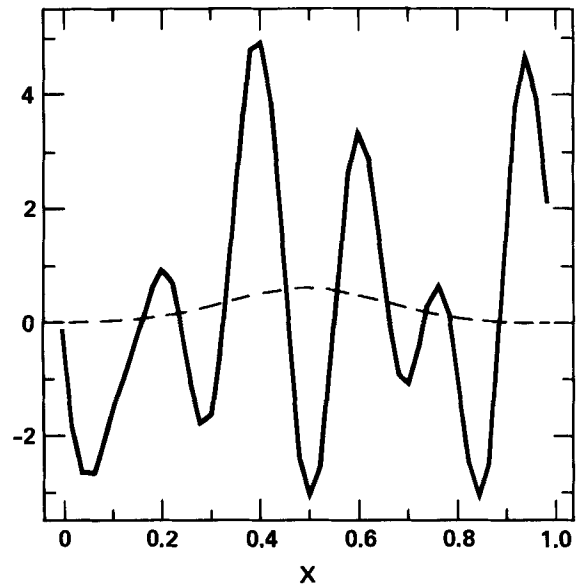

Figure 13. Same as Figure 12 except $t \simeq 8$

The results at neutral stability ($\varepsilon = 0$) are shown at $t = 10$ (10 'round trips' via the periodic boundary conditions) in Figures 10 and 11 for FTCS and modified FTCS, respectively. (The infinite span exact solution is shown for comparison as dashed lines; it is a good approximation to the periodic case as long as $\varphi(x = 0, t = \text{integer}) \ll 1$.) The improvement in accuracy resulting from the modified FTCS is quite striking (and at an almost 10-fold larger time step); unfortunately, equivalent accuracy is (probably) rarely attainable in 'real-world' simulations (variable velocities on variable grids in multi-dimensions, etc.). For FTCS, the

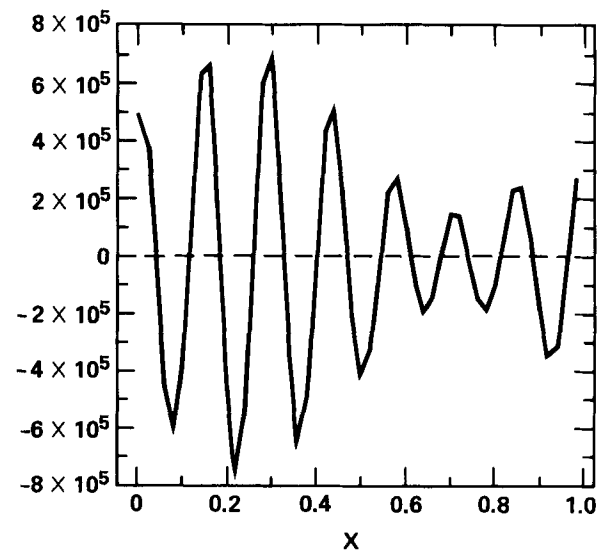

Figure 14. Same as Figure 12 except $t \simeq 20$

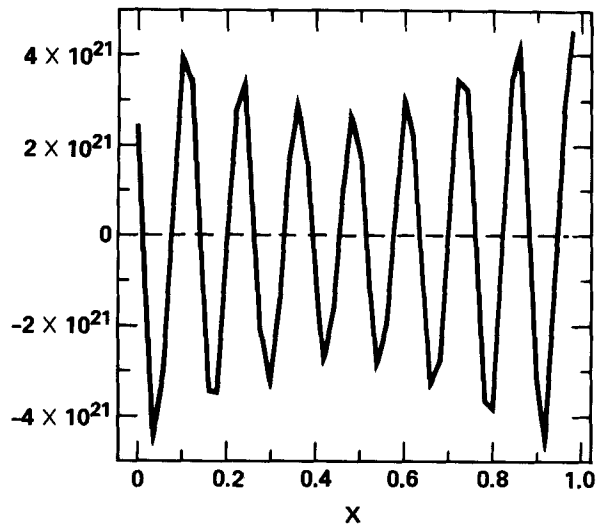

Figure 15. Same as Figure 12 except $t \simeq 50$

longest wave ($\lambda = 50\,\Delta x$) is neutrally stable and its presence is detectable. For modified FTCS, however, the $2\Delta x$ wave is neutrally stable and, since its initial amplitude is small, it is virtually undetectable.

The next series of Figures (12–16) shows the evolution of the unstable behaviour for FTCS at $\varepsilon = 1$ ($\Delta t = 0{\cdot}004016$). Although unstable behaviour is present at all times, the shape of the initial waveform causes the longer wavelengths to dominate at early time (their initial amplitude is larger). For instance, at $t = 8$, the fifth mode ($m = 5$) seems dominant, $m \simeq 7$ at $t = 20$, with $m = 8$ prevailing by $t = 30$, at which time the number of time steps is predictably

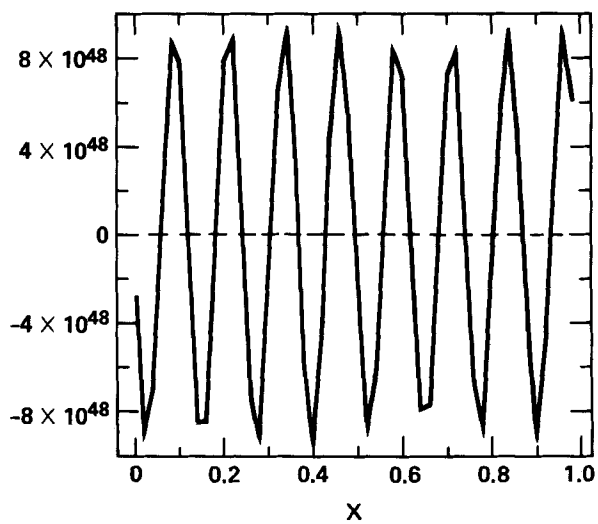

Figure 16. Same as Figure 12 except $t \simeq 100$

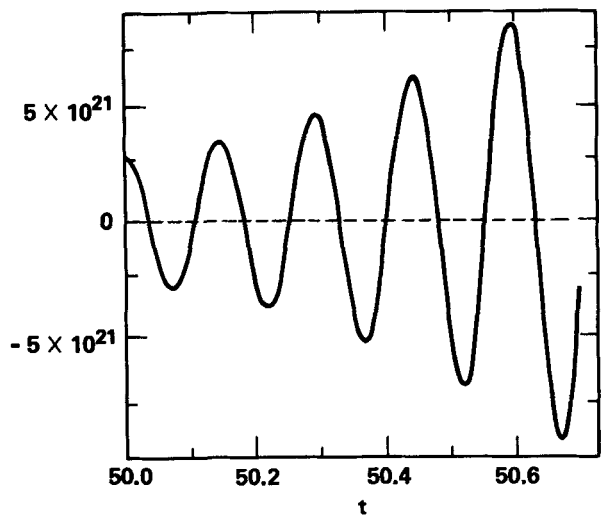

Figure 17. Time history of central node ($j = 25$) for FTCS with $\varepsilon = 1$ at $t \simeq 50$

large, ~7500. The overall evolution of the unstable behaviour is, as predicted by the theory, quite complex for FTCS. Figure 17 shows the time history of the central node ($j = 25$) near $t = 50$, showing a period of $\sim 37{\cdot}2\,\Delta t$. The observed growth rate, obtained by $|\xi| = [\max(\varphi_2)/\max(\varphi_1)]^{1/\Delta n}$ at times t_1 and t_2, where Δn is the number of steps between t_1 and t_2, is $\sim 1{\cdot}0051$ for $t_1 = 50$, $t_2 = 100$.

For modified FTCS with $\varepsilon = 1$, results are shown in Figures 18 and 19 at $t \simeq 1$ and $t \simeq 4$. The rapid growth of the $2\Delta x$ wave is rather obvious (and the period is $2\Delta t$); here the empirical growth rate, determined as above, is $|\xi| \simeq 6{\cdot}56$ between these two times, again in good agreement with the von Neumann prediction.

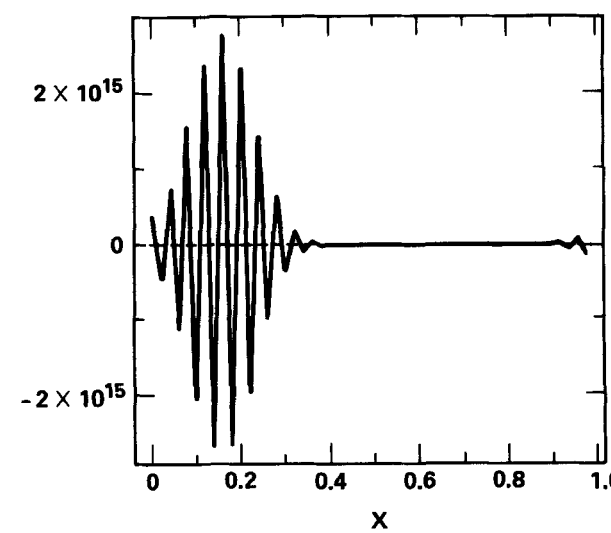

Figure 18. Modified FTCS solution for $\varepsilon = 1$ at $t \simeq 1$

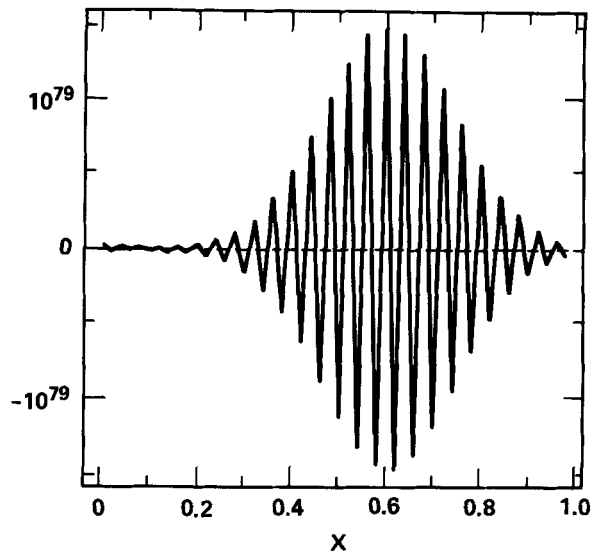

Figure 19. Same as Figure 18 except $t \simeq 4$

3. MULTI-DIMENSIONAL CASES

3.1. *Von Neumann stability theorem*

Here we are concerned with the constant coefficient advection–diffusion equation in M dimensions,

$$\frac{\partial \varphi}{\partial t} + \sum_{m=1}^{M} u_m \frac{\partial \varphi}{\partial x_m} = \sum_{m=1}^{M} K_m \frac{\partial^2 \varphi}{\partial x_m^2} \tag{71}$$

on the M-dimensional cube $0 \leq x_m \leq 1$, $t \geq 0$, with K_m non-negative. Although only $M = 1, 2$ and 3 are of practical interest, the results are quite general, and in a sense even simpler to express and prove for general M. In the light of our results in 1-D, we will perform only a Fourier analysis of the corresponding difference equations, and in this analysis the boundary conditions are implicitly assumed to be periodic. Equivalently, we may take $-\infty < x_m < \infty$ and address the purely initial value problem.

Consider the discretization that is centred in space and forward Euler in time (FTCS):

$$\frac{\varphi_j^{(n+1)} - \varphi_j^{(n)}}{\Delta t} + \sum_{m=1}^{M} u_m \frac{\Delta_m \varphi_j}{2\Delta x_m} = \sum_{m=1}^{M} K_m \frac{\delta_m^2 \varphi_j}{\Delta x_m^2} \tag{72}$$

Here j represents a multi-index $(j_1, j_2, \ldots, j_M)$, Δ_m is the central first difference operator with respect to the mth co-ordinate index j_m, and δ_m^2 is the central second difference operator with respect to j_m (the obvious generalizations of the differencing in (3)). For each m, we define diffusion parameters

$$\alpha_m = 2K_m \, \Delta t / \Delta x_m^2 \tag{73}$$

Courant numbers,

$$c_m = u_m \, \Delta t / \Delta x_m \tag{74}$$

and grid Peclet numbers,

$$P_m = u_m \Delta x_m / 2K_m = c_m/\alpha_m \tag{75}$$

Following the von Neumann (Fourier) analysis as in 1-D, we analyse the Fourier modes

$$\varphi_j^{(n)} = \xi^n \exp\left[i \sum_{m=1}^{M} j_m \theta_m\right] \tag{76}$$

with arbitrary phase angles $\theta_m = k_m \Delta x_m$, $-\pi \le \theta_m \le \pi$, where k_m is the component of the wave number vector, $\mathbf{k}$, in the m-direction. Implicitly invoking periodic boundary conditions, and initial conditions corresponding to the same Fourier mode, we substitute (76) into (72) and obtain

$$\xi - 1 + i \sum_{m=1}^{M} c_m \sin\theta_m = \sum_{m=1}^{M} \alpha_m (\cos\theta_m - 1) \tag{77}$$

We define the difference scheme (72) to be stable if $|\xi| \le 1$ for all θ_m. One of the main contributions of this paper is the following

Theorem. The scheme (72) is stable (in the von Neumann sense) if and only if

$$\sum_{m=1}^{M} \alpha_m \le 1 \tag{78}$$

and

$$\sum_{m=1}^{M} \frac{c_m^2}{\alpha_m} \le 1 \tag{79}$$

Remarks

(i) These conditions obviously imply the Courant–Friedrich–Lewy conditions $|c_m| \le 1$, but the latter are by no means sufficient for stability.

(ii) Inequality (79) is interpreted to imply that $c_m^2 \le \alpha_m$, even if $\alpha_m = 0$. This allows one to deduce the 1-D result from the theorem, for example.

(iii) An equivalent form for (79) is $\sum c_m P_m \le 1$, and another is $\sum P_m^2 \alpha_m \le 1$. From the latter, we see that inequality (78) prevails (is more restrictive) when all $P_m < 1$, whereas (79) prevails when all $P_m > 1$. Otherwise, both (78) and (79) are required.

Proof. We first write, from (77),

$$|\xi|^2 = \left[1 - \sum \alpha_m (1 - \cos\theta_m)\right]^2 + \left[\sum c_m \sin\theta_m\right]^2 \tag{80}$$

Necessity

We are given $|\xi|^2 \le 1$ for all choices of the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)^T$. For the case of all $\theta_m = \pi$, we have

$$|\xi|^2 = \left(1 - 2\sum \alpha_m\right)^2 \le 1$$

and this requires that (78) hold. For the limiting case $\theta_m \to 0$ with $|\theta_m| \le \theta$ for all m, we can write

$$\begin{aligned} |\xi|^2 &= \left[1 - \sum \alpha_m \theta_m^2/2 + O(\theta^4)\right]^2 + \left[\sum c_m \theta_m + O(\theta^3)\right]^2 \\ &= 1 - \sum \alpha_m \theta_m^2 + \left(\sum c_m \theta_m\right)^2 + O(\theta^4) \\ &= 1 - \boldsymbol{\theta}^T(\alpha - \mathbf{c}\mathbf{c}^T)\boldsymbol{\theta} + O(\theta^4) \end{aligned}$$

where $\alpha = \text{diag}\,(\alpha_1, \ldots, \alpha_M)$ and $\mathbf{c} = (c_1, \ldots, c_M)^T$. Thus, in order to have $|\xi|^2 \le 1$ for all $\boldsymbol{\theta}$,

the symmetric matrix $\beta \equiv \alpha - \mathbf{c}\mathbf{c}^{\mathrm{T}}$ must be non-negative definite (positive semi-definite). In particular, the diagonal elements $\alpha_m - c_m^2$ must be non-negative. Thus $c_m^2 \leq \alpha_m$, and if any $\alpha_m = 0$, then that $c_m = 0$ and the mth dimension can be dropped from the problem. So we may assume that all $\alpha_m > 0$. If we then set

$$\gamma = \alpha^{-1/2} = \mathrm{diag}\,(\alpha_1^{-1/2}, \ldots, \alpha_M^{-1/2})$$

we have

$$\beta = \alpha^{1/2}(I - \gamma\mathbf{c}\mathbf{c}^{\mathrm{T}}\gamma)\alpha^{1/2}$$

and the matrix

$$\beta' = I - \gamma\mathbf{c}\mathbf{c}^{\mathrm{T}}\gamma = I - (\gamma\mathbf{c})(\gamma\mathbf{c})^{\mathrm{T}} = I - \mathbf{d}\mathbf{d}^{\mathrm{T}}$$

where $\mathbf{d} \equiv \gamma\mathbf{c}$, must also be non-negative definite. From the associated quadratic form,

$$\mathbf{z}^{\mathrm{T}}\beta'\mathbf{z} = \mathbf{z}^{\mathrm{T}}\mathbf{z} - (\mathbf{d}^{\mathrm{T}}\mathbf{z})^2$$

we see that this is true if and only if $\mathbf{d}^{\mathrm{T}}\mathbf{d} \leq 1$. Since $\mathbf{d}^{\mathrm{T}}\mathbf{d} = \sum c_m^2/\alpha_m$, inequality (79) follows.

Sufficiency

Now assume (78) and (79). If all the $\alpha_m > 0$, then the Cauchy–Schwartz inequality gives, for arbitrary $\boldsymbol{\theta}$,

$$\begin{aligned}(\sum c_m \sin\theta_m)^2 &\leq [\sum (|c_m|/\sqrt{\alpha_m})(\sqrt{\alpha_m}\,|\sin\theta_m|)]^2 \\ &\leq [\sum c_m^2/\alpha_m][\sum \alpha_m \sin^2\theta_m] \\ &\leq \sum \alpha_m \sin^2\theta_m \end{aligned} \tag{81}$$

If any $\alpha_m = 0$, then (79) implies $c_m = 0$, and so (81) follows by summing first only over those m for which $\alpha_m > 0$. Inserting (81) into (80), and denoting $1 - \cos\theta_m$ by z_m, we have

$$\begin{aligned}|\xi|^2 &\leq (1 - \sum \alpha_m z_m)^2 + \sum \alpha_m [1 - (1 - z_m)^2] \\ &= 1 - 2\sum \alpha_m z_m + (\sum \alpha_m z_m)^2 + \sum \alpha_m (2z_m - z_m^2) \\ &= 1 - \sum \alpha_m z_m^2 + (\sum \alpha_m z_m)^2 \end{aligned} \tag{82}$$

Again using Cauchy–Schwartz, the last term in (82) is

$$(\sum \sqrt{\alpha_m}\,\sqrt{\alpha_m}\,z_m)^2 \leq (\sum \alpha_m)(\sum \alpha_m z_m^2) \leq \sum \alpha_m z_m^2$$

using (78). It thus follows that $|\xi|^2 \leq 1$. QED

3.2. *Applications to 2-D*

3.2.1. *FTCS.* The above Theorem applies directly to the FTCS scheme in 2-D upon setting $M = 2$. In terms of the physical problem parameters, (78) is

$$\Delta t \leq \frac{1}{2K_1/\Delta x_1^2 + 2K_2/\Delta x_2^2} \tag{83}$$

and (79) is

$$\Delta t \leq \frac{1}{u_1^2/2K_1 + u_2^2/2K_2} \tag{84}$$

In (83), the shortest resolvable wave is the most unstable, whereas in (84) the longest resolvable wave is the most unstable.

3.2.2. *Modified FTCS.* The modified FTCS difference equations in 2-D are obtained by replacing each K_m by

$$\bar{K}_m = K_m + u_m^2 \Delta t/2, \quad m = 1, 2$$

or, equivalently, replacing α_m by

$$\bar{\alpha}_m = \alpha_m + c_m^2.$$

(In multi-dimensions, as in 1-D, we retain the definitions of α_m and P_m in terms of the original K_m.) The (von Neumann) stability conditions are now

$$\bar{\alpha}_1 + \bar{\alpha}_2 = \alpha_1 + c_1^2 + \alpha_2 + c_2^2 \le 1 \tag{85}$$

$$c_1^2/\bar{\alpha}_1 + c_2^2/\bar{\alpha}_2 = c_1^2/(\alpha_1 + c_1^2) + c_2^2/(\alpha_2 + c_2^2) \le 1 \tag{86}$$

In general, inequality (85) has the form $Q(\Delta t) \le 0$, where Q is a quadratic which has real roots, one positive and one negative. Thus (85) is found to be equivalent to

$$\Delta t \le \Delta t_1 \equiv \frac{\sqrt{\left[\left(\frac{K_1}{\Delta x_1^2} + \frac{K_2}{\Delta x_2^2}\right)^2 + \left(\frac{u_1^2}{\Delta x_1^2} + \frac{u_2^2}{\Delta x_2^2}\right)\right]} - \left(\frac{K_1}{\Delta x_1^2} + \frac{K_2}{\Delta x_2^2}\right)}{u_1^2/\Delta x_1^2 + u_2^2/\Delta x_2^2} \tag{87}$$

Inequality (86) leads to

$$c_1^2 c_2^2 \le \alpha_1 \alpha_2$$

or

$$c_1 c_2 \le 1/P_1 P_2$$

or

$$\Delta t \le \Delta t_2 \equiv 2\sqrt{(K_1 K_2)}/u_1 u_2 \tag{88}$$

If either $u_m = 0$, inequality (86) holds trivially. We can state this as a corollary to the Theorem:

Corollary 1

The modified FTCS scheme in 2-D is stable if and only if

$$\alpha_1 + \alpha_2 + c_1^2 + c_2^2 \le 1 \tag{89}$$

and

$$c_1^2 c_2^2 \le \alpha_1 \alpha_2 \tag{90}$$

or, equivalently, if and only if $\Delta t \le \min(\Delta t_1, \Delta t_2)$, where Δt_1 and Δt_2 are given by (87) and (88). (For special cases with $u_m = 0$, use the limiting values of Δt_1 and Δt_2).

In contrast to the situation in 1-D, *both* inequalities (89) and (90) are needed as necessary and sufficient conditions for stability. To see this, we can insert Δt_2 into $Q(\Delta t)$, and find

$$\begin{aligned} Q(\Delta t_2) &= 4(u_1^2/\Delta x_1^2 + u_2^2/\Delta x_2^2) K_1 K_2/u_1^2 u_2^2 \\ &\quad + 4(K_1/\Delta x_1^2 + K_2\,\Delta x_2^2)\sqrt{(K_1 K_2)}/u_1 u_2 - 1 \\ &= (R_1 + 1/R_1 + R_2 + 1/R_2)/P_1 P_2 - 1 \end{aligned}$$

where

$$R_1 \equiv u_1\,\Delta x_2/u_2\,\Delta x_1$$

$$R_2 \equiv (\Delta x_2/\Delta x_1)\sqrt{(K_1/K_2)}$$

Clearly $Q(\Delta t_2)$ can be made either positive or negative by various choices of the parameters, and so Δt_2 can be either larger or smaller than Δt_1. However, in any case we have $R_1+1/R_1\geq 2$ and $R_2+1/R_2\geq 2$, so that $Q(\Delta t_2)\geq 4/P_1P_2-1$. So when $P_1P_2\leq 4$, we have $Q(\Delta t_2)\geq 0$, and inequality (90) follows from inequality (89). Thus we have

Corollary 2

If $P_1P_2\leq 4$, the modified FTCS scheme in 2-D is stable if and only if $\alpha_1+\alpha_2+c_1^2+c_2^2\leq 1$, or equivalently if and only if $\Delta t\leq \Delta t_1$ (given by (87)).

Remarks

(1) Unlike the 1-D counterpart, the 2-D modified FTCS scheme cannot be stabilized if $K_1=K_2=0$, as noted earlier by Leith.[27]

(2) Nevertheless, this scheme does display some stability advantages over FTCS.

3.2.3. *Upwinded advection.* Consider the simplest upwind differencing in 2-D. For definiteness, suppse both $u_m\geq 0$. Then the corresponding difference equations are obtained by replacing K_m by

$$\bar{K}_m = K_m + u_m\,\Delta x_m/2, \qquad m=1,2$$

or equivalently, replacing α_m by

$$\bar{\alpha}_m = \alpha_m + c_m$$

The (von Neumann) stability conditions now become

$$\bar{\alpha}_1+\bar{\alpha}_2=\alpha_1+\alpha_2+c_1+c_2\leq 1$$

and

$$c_1^2/\bar{\alpha}_1+c_2^2/\bar{\alpha}_2=c_1^2/(\alpha_1+c_1)+c_2^2/(\alpha_2+c_2)\leq 1$$

An equivalent form for this pair of inequalities is

$$\alpha_1(1+P_1)+\alpha_2(1+P_2)\leq 1$$

and

$$\alpha_1P_1^2/(1+P_1)+\alpha_2P_2^2/(1+P_2)\leq 1$$

But since $P_m^2/(1+P_m)<1+P_m$, the second inequality always follows from the first. Thus we have

Corollary 3

The forward-time upwinded-space difference scheme in 2-D is stable if and only if

$$\alpha_1(1+P_1)+\alpha_2(1+P_2)\leq 1 \tag{91}$$

or, equivalently, if and only if

$$\Delta t\leq 1/(2K_1/\Delta x_1^2+2K_2/\Delta x_2^2+u_1/\Delta x_1+u_2/\Delta x_2) \tag{92}$$

Remarks

(1) If the difference scheme is properly modified for general u_m (to remain upwinded), the above result can be expressed more generally by using $|u_m|$ in place of u_m above.

(2) Unlike modified FTCS, this scheme *can* be used for 'pure' advection ($K_m=0$), requiring $c_1+c_2\leq 1$ for stability. The resulting numerical diffusion, however, is usually large enough to render the results highly inaccurate.

3.3. *Application to 3-D*

3.3.1. *FTCS.* Application of the theorem to $M = 3$ is straightforward. Again we interpret (79) to imply $c_m^2 \leq \alpha_m$, and so the analogous 2-D theorem is a special case in which $\alpha_m = c_m = 0$ for one value of m. In terms of the physical problem parameters, these stability limits are given by the obvious extension of (83) and (84) to 3-D.

3.3.2. *Modified FTCS.* the modified FTCS difference equations in 3-D are obtained, as in 2-D, by replacing K_m by

$$\bar{K}_m = K_m + u_m^2 \Delta t/2, \qquad m = 1, 2, 3$$

or by replacing α_m by $\bar{\alpha}_m = \alpha_m + c_m^2$. The stability conditions become

$$\sum_1^3 (\alpha_m + c_m^2) \leq 1 \tag{93}$$

and

$$\sum_1^3 c_m^2/(\alpha_m + c_m^2) \leq 1 \tag{94}$$

As before, (93) has the form $Q(\Delta t) \leq 0$, where Q is a quadratic in Δt with one positive root, Δt_1, given by the obvious extension of (87).

Inequality (94), when cleared of fractions, reduces to

$$2c_1^2c_2^2c_3^2 + \alpha_1 c_2^2 c_3^2 + \alpha_2 c_1^2 c_3^2 + \alpha_3 c_1^2 c_2^2 \leq \alpha_1\alpha_2\alpha_3$$

or

$$2c_1c_2c_3P_1P_2P_3 + P_1P_2c_1c_2 + P_1P_3c_1c_3 + P_2P_3c_2c_3 \leq 1$$

(There is no simple analogue of the 2-D constraint (88).)

If no u_m is zero, this has the form $C(\Delta t) \leq 0$, where C is a cubic polynomial,

$$C(\Delta t) = 2A(\Delta t^3 + B\,\Delta t^2) - 1 \tag{95}$$

where

$$A = \prod_1^3 u_m P_m/\Delta x_m = (1/8) \prod_1^3 u_m^2/K_m > 0$$

and

$$B = (1/2) \sum_1^3 \Delta x_m/P_m u_m = \sum_1^3 K_m/u_m^2 > 0$$

Since $C(0) = -1$ and $C'(\Delta t) > 0$ for $\Delta t > 0$, $C(\Delta t)$ has a single positive root, say Δt_2, and inequality (94) is equivalent to $\Delta t \leq \Delta t_2$. If any one of the $u_m = 0$, then (94) reduces to the analogous inequality, (88), in 2-D. The bound in that inequality is also the limiting value, as $u_m \to 0$, of the general Δt_2, from (95), and is the positive root of the quadratic to which $C(\Delta t)$ degenerates as $u_m \to 0$. If two of the $u_m = 0$, (94) holds trivally, and we may use the limiting value $\Delta t_2 = \infty$.

We can thus state

Corollary 4

The modified FTCS scheme in 3-D is stable if and only if

$$\sum_1^3 (\alpha_m + c_m^2) \leq 1 \tag{96}$$

and

$$\sum_1^3 c_m^2/(\alpha_m + c_m^2) \leq 1 \tag{97}$$

or, equivalently, if $\Delta t \leq \min(\Delta t_1, \Delta t_2)$, where

$$\Delta t_1 = \frac{\sqrt{[(\sum_1^3 K_m/\Delta x_m^2)^2 + \sum_1^3 u_m^2/\Delta x_m^2]} - \sum_1^3 K_m/\Delta x_m^2}{\sum_1^3 u_m^2/\Delta x_m^2}$$

and Δt_2 is the unique positive root of the cubic $C(\Delta t)$ in (95). (For special cases $u_m = 0$, use the limiting values of Δt_1 and Δt_2).

Remark

As in 2-D, pure advection ($K_m = 0$) cannot be stabilized with this modified FTCS scheme.

3.3.3. *Upwinded advection.* As in 2-D, we consider only the simplest upwind differencing in 3-D. For definiteness, suppose each $u_m \geq 0$. Then the difference equations are obtained by replacing K_m by

$$\bar{K}_m = K_m + u_m \Delta x_m/2, \qquad m = 1, 2, 3$$

or equivalently by replacing each α_m by $\bar{\alpha}_m = \alpha_m + c_m$. The stability conditions become

$$\sum_1^3 \bar{\alpha}_m = \sum_1^3 (\alpha_m + c_m) \leq 1$$

and

$$\sum_1^3 c_m^2/\bar{\alpha}_m = \sum_1^3 c_m^2/(\alpha_m + c_m) \leq 1$$

An equivalent form for these is

$$\sum_1^3 \alpha_m (1 + P_m) \leq 1$$

and

$$\sum_1^3 \alpha_m P_m^2/(1 + P_m) \leq 1$$

But since $P_m^2/(1+P_m) < 1 + P_m$, the second inequality always follows from the first. Thus we have

Corollary 5

The forward-time upwinded-space difference scheme in 3-D is stable if and only if

$$\sum_1^3 \alpha_m (1 + P_m) \leq 1 \tag{98}$$

or, equivalently,

$$\Delta t \leq 1 \Big/ \left(2 \sum_1^3 K_m/\Delta x_m^2 + \sum_1^3 u_m/\Delta x_m\right) \tag{99}$$

The same remarks made for the corresponding 2-D case apply here.

4. EXTENSION TO A FINITE-ELEMENT-BASED METHOD

Having successfully analysed several multi-dimensional FTCS-related schemes which we do not use (nor recommend), we now briefly discuss a scheme which we use and advocate, both

for advection–diffusion and Navier–Stokes simulations. Our derivation of it is based on (lumped mass) finite elements (with multilinear basis functions) but this is not an absolute requirement (E.g., see Reference 28.).

Using compass point notation (NE = north-east, etc.), the semi-discrete version of (1) in 2 dimensions (for a full, but symmetric diffusivity tensor K_{ij}) is described by the following 9-point 'stencil' on a regular mesh (here φ_0 is located at the centre of the stencil):

$$\begin{aligned}
&\dot{\varphi}_0 + \frac{u_1}{2\Delta x_1(2+\gamma)}[(\varphi_{SE}-\varphi_{SW})+\gamma(\varphi_E-\varphi_W)+(\varphi_{NE}-\varphi_{NW})] \\
&\quad + \frac{u_2}{2\Delta x_2(2+\gamma)}[(\varphi_{NW}-\varphi_{SW})+\gamma(\varphi_N-\varphi_S)+(\varphi_{NE}-\varphi_{SE})] \\
&= \frac{K_{11}}{\Delta x_1^2(2+\gamma)}[(\varphi_{SE}-2\varphi_S+\varphi_{SW})+\gamma(\varphi_E-2\varphi_0+\varphi_W)+(\varphi_{NE}-2\varphi_N+\varphi_{NW})] \\
&\quad + \frac{K_{22}}{\Delta x_2^2(2+\gamma)}[(\varphi_{NW}-2\varphi_W+\varphi_{SW})+\gamma(\varphi_N-2\varphi_0+\varphi_S)+(\varphi_{NE}-2\varphi_E+\varphi_{SE})] \\
&\quad + \frac{K_{12}}{2\Delta x_1\,\Delta x_2}(\varphi_{NE}-\varphi_{NW}+\varphi_{SW}-\varphi_{SE})
\end{aligned} \tag{100}$$

where $\gamma = 2$, 4 or ∞ for the cases of interest herein. The approximation using the Galerkin finite element method (GFEM) is that corresponding to $\gamma = 4$. The case $\gamma = 2$ corresponds to a modified GFEM wherein one-point quadrature is used to evaluate the Galerkin integrals[4,29] and is the case of most current interest to us. Finally, setting γ to ∞ (and discarding the cross-diffusion terms) recovers the conventional finite difference stencil corresponding to FTCS.

As in one dimension, a Taylor series analysis in time of the forward Euler time integration method for (1) can be performed to show[4] that this scheme 'reduces' the effective diffusivity tensor from K_{ij} to $(K_{ij} - u_i u_j\,\Delta t/2)$. Thus, as in one dimension, this observation leads directly to our improved scheme, which we claim is an *appropriate* multidimensional generalization of the one-dimensional modified FTCS scheme (or, for pure advection, an appropriate generalization of Leith's method): namely replace K_{ij} by $K_{ij} + u_i u_j\,\Delta t/2$ in (100). When this is done, the same sort of cost-effectiveness mentioned earlier for 1-D carries over to 2-D and 3-D; i.e. for advection-dominated flows, the time step limitation appears to be basically a 'Courant (or CFL) condition'.

Although the von Neumann stability analysis of the forward Euler time discretization of (100), and its 3-D analogue, have thus far proved intractable in the general case, we do have partial (proved) results that are worth reporting (which we do without proof in view of their 'special case' nature):

1. For $2 \le \gamma \le 6$, $u = 0$ (pure diffusion), and $K_{ij} = 0$ for $i \ne j$ (no cross-diffusion), the necessary and sufficient stability conditions in M spatial dimensions are

$$\alpha_m \le 1, \qquad m = 1, 2, \ldots, M \tag{101}$$

This scheme is thus more stable than FTCS which requires $\sum \alpha_m < 1$. Also for this pure diffusion case and $\gamma > 6$, the necessary and sufficient stability conditions are slightly more complicated than (101).

2. For $\gamma = 2$ and pure advection (with $K_{ij} = u_i u_j\,\Delta t/2$), the necessary and sufficient stability

condition is

$$\sum_{m=1}^{M} c_m^2 \leq 1 \tag{102}$$

(This is also a necessary condition for any value of γ.)

3. For $\gamma = \infty$ and pure advection (with $K_{ij} = u_i u_j \Delta t/2$), the necessary and sufficient stability condition is

$$\sum_{m=1}^{M} c_m^{2/3} \leq 1 \tag{103}$$

This scheme (in 2-D only) is a familiar Lax–Wendroff method;[30] (Lax and Wendroff proved only the sufficient conditions $c_m \leq 1/\sqrt{8}$). The unique feature of the $\gamma = \infty$ scheme is that the stencil is the simplest one possible that gives centred, second-order spatial difference approximations.

4. For $\gamma = 2$ we have the following necessary conditions for stability in 2-D.

(i) The 1-D requirements, (9)–(11), apply separately in each direction.

(ii)
$$\Delta t \leq (1 - K_{12}^2/K_{11}K_{22})/(u_1^2/2K_{11} + u_2^2/2K_{22} - u_1 u_2 K_{12}/K_{11}K_{22}) \tag{104}$$

where $K_{12}^2 < K_{11}K_{22}$ is required in order that the problem be well-posed (even in the continuum).

5. If we replace K_{ij} by $K_{ij} + u_i u_j \Delta t/2$ (the recommended scheme), the above necessary conditions (in 4) are changed to those in which the 1-D results of modified FTCS, (59), apply separately in each direction. (104) is always satisfied in this case, since it degenerates to $K_{12}^2 < K_{11}K_{22}$.

Further remarks

(i) This 'balancing tensor diffusivity' method ($K_{ij} \rightarrow K_{ij} + u_i u_j \Delta t/2$) is not new. It has been previously discussed and applied to the solution of the incompressible Navier–Stokes equations by Dukowicz and Ramshaw,[31] albeit with a somewhat different spatial discretization. (Recall that our derivation suggests that its utility is essentially independent of spatial discretization.)

(ii) Unlike modified FTCS, discussed in Section 3, the balancing tensor diffusivity method can easily be stabilized when $K_{ij} = 0$ (as discussed above). This is the reason we call it an appropriate generalization of Leith's method.

(iii) The 3-D version of (100) is, for $\gamma = 2$, obtainable via a tensor product of the 'averaging coefficients' (1/4, 1/2, 1/4), just like the 2-D version, and leads to a 27-point stencil.

(iv) Since the derivation of this scheme is based on the time-dependent equations, the full utility of it is also limited to transient simulations. If a steady state is approached, the balancing tensor diffusivity method leads to a streamline upwinding of the advection terms; i.e. there is additional artificial (numerical) diffusivity in the streamline direction (but, importantly, *not* in the 'crosswind direction', which has been found to be quite deleterious[32] and present in large measure if upwinded advection is employed in the manner adopted in Sections 3.2.3 and 3.3.3). See also Reference 4 for additional steady-state analysis and numerical results.

To conclude this section, we compare the relative cost of three of the schemes discussed in this paper—at least when they are used for pure advection ($K_{ij} = 0$) simulations (the only

Table I

Scheme	$\tan \theta_i$	$\Delta t_{\min}$
1	$\left(\frac{1/\Delta x_i}{\sum_j 1/\Delta x_j}\right)^{1/2}$	$\frac{1}{\lvert\mathbf{u}\rvert \sum_j 1/\Delta x_j}$
2	$\frac{1/\Delta x_i}{(\sum_j 1/\Delta x_j^2)^{1/2}}$	$\frac{1}{\lvert\mathbf{u}\rvert (\sum_j 1/\Delta x_j^2)^{1/2}}$
3	See footnote*	$\frac{1}{\lvert\mathbf{u}\rvert} \min_j (\Delta x_j)$

* The critical direction for Scheme 3 is that with $\min(\Delta x_i)$.

case for which we have complete results):

Scheme 1 is the Lax–Wendroff method, discussed above.

Scheme 2 is the upwind method as derived from FTCS by replacing K_i by $u_i \Delta x_i/2$ ($u_i \geq 0$).

Scheme 3 is (see reference 4) that of (100) with K_{ij} replaced by $u_i u_j \Delta t/2$ and setting $\gamma = 2$ (as discussed above).

From the necessary and sufficient conditions for stability of these schemes ((103) for Scheme 1,

$$\sum_{m=1}^{M} c_m \leq 1$$

for Scheme 2 (see Section 3), and (102) for Scheme 3), we can answer the following legitimate questions related to the cost of each scheme: (1) Is there a particular direction of the velocity vector that results in the smallest allowable time step? (2) If so, what is this direction and what is the minimum Δt? the answer to question (1) is yes (except for Scheme 3 in the case where Δx_i is the same in all spatial directions) and that to question (2) is given in Table I, wherein $\tan \theta_i \equiv u_i/|\mathbf{u}|$ and θ_i define the critical direction for $\mathbf{u}$.

Remarks

(i) $\Delta t_1 \leq \Delta t_2 \leq \Delta t_3$, where Δt_i denotes $\Delta t_{\min}$ for Scheme i. The stability differences in Δt are largest for a uniform grid ($\Delta x_i = \text{constant}$), wherein $\Delta t_1 = \Delta t_3/M$ and $\Delta t_2 = \Delta t_3/\sqrt{M}$ where $\Delta t_3 = \Delta x/|\mathbf{u}|$.

(ii) The direction of the wave number vector for which the Δt bound is minimal turns out to be the same θ_i as above.

(iii) This measure of relative cost should not be confused with cost-effectiveness, which necessarily raises issues that are beyond the scope of this paper.

5. CONCLUSIONS

(i) The von Neumann method is generally the best single technique for analysing the stability of difference schemes. It should always be part of a stability analysis, even if other techniques are also employed.

(ii) The von Neumann stability results are necessary conditions regardless of boundary conditions. For periodic boundary conditions, they are also sufficient (cf. also Reference 11.) For other boundary conditions, the von Neumann results augmented by those from a Godunov–Ryabenkii type of analysis will yield the necessary and sufficient stability conditions.

(iii) When stability results from the matrix method are less strict than those from the von Neumann method, they should not be relied upon since errors could become very large before finally decaying to zero. Typically in these cases, instability begins far from boundaries and it may require a long time before the boundary conditions can finally 'stabilize' the numerical solution.

(iv) For advection–dominated flows, Robin boundary conditions should not be implemented using the image point method of finite differences.

(v) The behavioural transition of FTCS in 1-D as P passes through unity (from below) is quite marked: the discrete spectrum changes from real to complex, and the inequality describing stability (von Neumann sense) completely changes its form, as does the behaviour of the most unstable mode. Only for $P<1$ does the discrete spectrum resemble that of the continuum.

(vi) Correct stability results have been obtained in 2-D and 3-D for FTCS and two variants of it. For FTCS the necessary and sufficient conditions for stability are:

$$\Delta t \leq 1 \Big/ \sum_{j=1}^{M} (2K_j/\Delta x_j^2)$$

and

$$\Delta t \leq 1 \Big/ \sum_{j=1}^{M} (u_j^2/2K_j)$$

where $M = 1, 2$ or 3 is the spatial dimensionality.

(vii) Balancing tensor diffusivity (modified FTCS in 1-D) appears to be as useful for advection–diffusion as it is for the limiting case of pure advection, in which limit it is an appropriate multi-dimensional generalization of Leith's (or the Lax–Wendroff) method. It has also proved useful for solving the Navier–Stokes equations.

ACKNOWLEDGEMENTS

We are grateful to Craig D. Upson of Lawrence Livermore National Laboratory, whose skill with the CRAY-1 made the supporting numerical experiments simple and fun. We also thank the referee who forced us to clarify our own thinking; his comments have greatly improved the paper. This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

APPENDIX. MATRIX POWER BOUNDS FOR 1-D FTCS

In Section 2.4, results are given concerning powers of the $N \times N$ matrix

$$E = \begin{bmatrix} 1-\alpha & & & \\ \alpha & \searrow & & \\ & \searrow & \searrow & \\ & & \searrow & \end{bmatrix}$$

The derivations of those results are given here.

We define, for $0 \leq \alpha < 2$,

$$m_N(\alpha, n) = \max_{i,j} |(E^n)_{ij}|$$

and

$$M_N(\alpha) = \max_{n \geq 1} m_N(\alpha, n)$$

If we define L to be the matrix with 1 on the first subdiagonal and 0 elsewhere, we have $E=(1-\alpha)I+\alpha L$ and so

$$E^n=\sum_{k=0}^{n}\binom{n}{k}\alpha^k(1-\alpha)^{n-k}L^k$$

where $\binom{n}{k}$ denotes a binomial coefficient. We note that L^k has 1 on the kth subdiagonal, in positions $(i, j)=(j+k, j)$, if $k\leq N-1$, and 0 elsewhere, and that $L^N=0$. Thus the upper limit in the sum above should really be $\min(n, N-1)$. It follows that the magnitudes of the non-zero elements in E^n are

$$e_k(\alpha, n)=\binom{n}{k}\alpha^k(\alpha-1)^{n-k},\quad 0\leq k\leq \min(n, N-1)$$

One easy observation is that for $0\leq\alpha\leq 1$, the elements of E^n are all non-negative, and the sum over any row or any column is bounded by

$$\sum_{0}^{n}\binom{n}{k}\alpha^k(1-\alpha)^{n-k}=1$$

Thus both $M_N(\alpha)$ and $\|E^n\|_\infty$ are bounded by 1 when $\alpha\leq 1$.

For $1<\alpha<2$, we can obtain good estimates with the help of Stirling's formula, in the form

$$1<m!/[m^m e^{-m}\sqrt{(2\pi m)}]<12/11$$

for $m\geq 1$. From this we find that for $n\geq 2$ and $1\leq k\leq n-1$

$$\binom{n}{k}\Big/\left[\frac{n^n}{k^k(n-k)^{n-k}}\sqrt{\left(\frac{n}{k(n-k)}\right)}\right]$$

is bounded below by $(11/12)^2/\sqrt{(2\pi)}>0\cdot 335$, and above by $(12/11)/\sqrt{(2\pi)}<0\cdot 436$. We also have

$$\sqrt{[n/k(n-k)]}\leq\sqrt{[n/(n-1)]}\leq\sqrt{2}$$

Thus if we define

$$f_n(k)=n^n(\alpha/k)^k[(\alpha-1)/(n-k)]^{n-k},\qquad 0<k<n$$

we have

$$0\cdot 335\, f_n(k)\sqrt{[n/k(n-k)]}<e_k(\alpha, n)<0\cdot 62\, f_n(k)$$

for $n\geq 2$ and $1\leq k\leq n-1$. We will need to check separately the cases

$$e_0(n, \alpha)=(\alpha-1)^n<1$$
$$e_n(n, \alpha)\alpha^n,\qquad \text{when } n\leq N-1$$
$$m_N(\alpha, 1)=\alpha$$

Upper bounds

We need to study $f_n(k)$, and for this we regard k as a continuous real variable. We have

$$\log f_n=n\log n+k\log(\alpha/k)+(n-k)\log[(\alpha-1)/(n-k)]$$
$$f_n'/f_n=\log(\alpha/k)-1-\log[(\alpha-1)/(n-k)]+1=\log[\alpha(n-k)/k(\alpha-1)]$$

The last expression (in brackets) decreases monotonically on $0<k<n$, and so $f_n'(k)$ has a single zero at $k=k^*$ given by

$$\alpha(n-k^*)=k^*(\alpha-1)$$

or

$$k^*=\alpha n/(2\alpha-1)$$

Note that $(2/3)n<k^*<n$ for $1<\alpha<2$. From the relations

$$\alpha/k^*=(\alpha-1)/(n-k^*)=(2\alpha-1)/n$$

we see that $f_n(k)$ has a unique maximum on $0<k<n$ of

$$f_n(k^*)=n^n[(2\alpha-1)/n]^n=(2\alpha-1)^n$$

This value of f_n is relevant only if $k^*\le N-1$, which is equivalent to

$$n\le N^*\equiv(2-1/\alpha)(N-1)$$

Note that $N^*>N-1\ge 1$. the information obtained so far can be collected as follows:

Case 1. $n=1$. Here $m_N(\alpha, n)=\alpha$.

Case 2. $2\le n\le N^*$ and $n\le N-1$. Here the relevant values of k are $0,\ldots,n$. We have

$$e_0(n,\alpha)=(\alpha-1)^n<1$$
$$e_n(n,\alpha)=\alpha^n$$

and for $1\le k\le n-1$,

$$e_k(n,\alpha)<0{\cdot}62\,f_n(k)\le 0{\cdot}62\,f_n(k^*)=0{\cdot}62(2\alpha-1)^n$$

Thus

$$m_N(\alpha,n)\le\max[\alpha^n, 0{\cdot}62(2\alpha-1)^n]\le\max[\alpha^{N^*}, 0{\cdot}62(2\alpha-1)^{N^*}]$$

Case 3. $2\le n\le N^*$ and $n>N-1$. Here only $k\le N-1$ is relevant, and for $1\le k\le N-1$ we have

$$e_k(n,\alpha)<0{\cdot}62\,f_n(k^*)=0{\cdot}62(2\alpha-1)^n$$

Also $e_0(n,\alpha)=(\alpha-1)^n$; but from

$$(2\alpha-1)/(\alpha-1)=2+1/(\alpha-1)>3$$

we conclude that $0{\cdot}62(2\alpha-1)^n>(\alpha-1)^n$ for all $n\ge 2$. Thus

$$m_N(\alpha,n)<0{\cdot}62(2\alpha-1)^n\le 0{\cdot}62(2\alpha-1)^{N^*}$$

Case 4. $n>N^*$. Here $n>N^*>N-1$, $k^*>N-1$, and $f_n(k)$ is monotone increasing on $0<k\le N-1$. Thus $e_0(\alpha,n)<1$ and for $1\le k\le N-1$,

$$e_k(\alpha,n)<0{\cdot}62\,f_n(k)\le 0{\cdot}62\,f_n(N-1)$$
$$m_N(\alpha,n)<\max[1, 0{\cdot}62\,f_n(N-1)]$$

We now must study the function

$$g(n)\equiv f_n(N-1)=n^n[\alpha/(N-1)]^{N-1}[(\alpha-1)/(n-N+1)]^{n-N+1}$$

where n will be regarded as continuous and greater than $N-1$. We have

$$\log g=n\log n+(N-1)\log[\alpha/(N-1)]+(n-N+1)\log[(\alpha-1)/(n-N+1)]$$

and

$$g'/g=\log n+1+\log[(\alpha-1)/(n-N+1)]-1=\log[n(\alpha-1)/(n-N+1)]$$

The last expression (in brackets) is monotone decreasing as n increases, and g' has a single zero at $n = n^*$ given by

$$n^*(\alpha - 1) = n^* - N + 1, \quad \text{or} \quad n^* = (N-1)/(2-\alpha)$$

Note that $n^* > N-1$. Thus $g(n)$ has a unique maximum on $n > N-1$ of

$$g(n^*) = (n^*)^{n^*}[\alpha/(N-1)]^{N-1}(1/n^*)^{n^*-N+1} = [\alpha n^*/(N-1)]^{N-1} = [\alpha/(2-\alpha)]^{N-1}$$

To compare $(2\alpha - 1)^{N^*}$ with $g(n^*)$ we must examine

$$a(\alpha) \equiv (2\alpha - 1)^{(2-1/\alpha)} \quad \text{and} \quad b(\alpha) \equiv \alpha/(2-\alpha)$$

We have $a(1) = b(1) = 1$, and

$$a'/a = [(2 - 1/\alpha)\log(2\alpha - 1)]' = \alpha^{-2}\log(2\alpha - 1) + 2/\alpha$$
$$c(\alpha) \equiv \alpha^2 a'/a = \log(2\alpha - 1) + 2\alpha$$
$$b'/b = 1/\alpha - 1/(2-\alpha) = 2/[\alpha(2-\alpha)]$$

and

$$d(\alpha) \equiv \alpha^2 b'/b = 2\alpha/(2-\alpha)$$

Next, observe that $c(1) = d(1) = 2$, and for $\alpha > 1$

$$c'(\alpha) = 2/(2\alpha - 1) + 2 < c'(1) = 4$$

and

$$d'(\alpha) = 4/(2-\alpha)^2 > d'(1) = 4$$

Thus for $1 < \alpha < 2$, we have $d' > c'$, $d > c$, $\log b > \log a$, and $b > a$. It follows that

$$b^{N-1} = [\alpha/(2-\alpha)]^{N-1} > a^{N-1} = (2\alpha - 1)^{N^*}$$

Now $M_N(\alpha)$ is the maximum of the $m_N(\alpha, n)$ as n goes through the four cases above. Using the last of the above inequalities, together with $\alpha^{N^*} > \alpha$ and $f_n(N-1) \le g(n^*)$, we find

$$M_N(\alpha) \le \max\{\alpha^{N^*}, 0{\cdot}62[\alpha/(2-\alpha)]^{N-1}\}$$

A simpler bound results from increasing the 0·62 to 1 and using

$$\alpha^{N^*} < (2\alpha - 1)^{N^*} < [\alpha(2-\alpha)]^{N-1}$$

namely,

$$M_N(\alpha) < [\alpha/(2-\alpha)]^{N-1}$$

The value of this bound arose as a bound on $m_N(\alpha, n)$ with $n = n^*$, and for this value of n it arose as the bound on the lower left corner element of E^n, i.e. $e_{N-1}(\alpha, n)$.

Lower bounds

In an attempt to get a lower bound on $M_N(\alpha)$, we restrict our attention to the particular $e_k(\alpha, n)$ which gave rise to the upper bound, namely the corner element $e_{N-1}(\alpha, n)$ with n at or near n^*. To use the lower bound on this element obtained earlier, we must keep $N - 1 \le n - 1$, or $n \ge N \ge 2$. Thus we can write

$$M_N(\alpha) \ge \max_{n \ge N} m_N(\alpha, n) \ge \max_{n \ge N} e_{N-1}(\alpha, n) > \max_{n \ge N} 0{\cdot}335\, h(n)$$

where

$$h(n) \equiv g(n)\sqrt{\{n/[(N-1)(n-N+1)]\}}$$

and $g(n)$ is as before. From this information, the best available lower bound would be the result of finding the maximum of $h(n)$ over $n \geq N$. This problem appears intractable however, and we will settle for the value $h(n^*)$, on the grounds that $h(n)$ and $g(n)$ differ relatively little, and the maximum of $g(n)$ is at $n = n^* = (N-1)/(2-\alpha)$. We must ensure that $n^* \geq N$, by requiring that $N \geq 1/(\alpha - 1)$. Note that

$$h(n^*) = g(n^*)/\sqrt{[(2-\alpha)(n^* - N + 1)]} = [\alpha/(2-\alpha)]^{N-1}/\sqrt{[(\alpha-1)(N-1)]}$$

If n^* were an integer, we could take $0{\cdot}335\, h(n^*)$ itself as a lower bound on $M_N(\alpha)$. In general, however, we will take $0{\cdot}335\, h(n)$ with the integer $n = [n^*] + 1$. (Here $[x]$ denotes the greatest integer less than or equal to x.) To account for this small shift in argument, we must analyze $h(n)$ in the interval $n^* \leq n \leq n^* + 1$. We have

$$\log h = \log g + (1/2) \log n - (1/2) \log [(N-1)(n - N + 1)]$$

and

$$h'/h = \log [n[\alpha - 1)/(n - N + 1)] + 1/2n - 1/[2(n - N + 1)]$$

For $n^* \leq n \leq n^* + 1$, we have

$$|1/2n - 1/[2(n - N + 1)]| = (N-1)/[2n(n - N + 1)] < 1/[2(n^* - N + 1)]$$

and the argument of the log in h'/h satisfies

$$\begin{aligned} |n(\alpha-1)/(n-N+1) - 1| &= |[n(\alpha-2) + (N-1)]/(n-N+1)| \\ &= |(n-n^*)(\alpha-2)|/(n-N+1) \leq (2-\alpha)/(n-N+1) \\ &\leq (2-\alpha)/(n^*-N+1) \end{aligned}$$

Note that

$$1/(n^* - N + 1) = (2-\alpha)/[(\alpha-1)(N-1)]$$

So if we define

$$d_1(\alpha) = (2-\alpha)/[2(\alpha-1)]$$

and

$$d_2(\alpha) = (2-\alpha)^2/(\alpha-1)$$

we have, for $n^* \leq n \leq n^* + 1$ and $N \geq 1/(\alpha - 1)$

$$h'/h = h_1 + \log(1 + h_2)$$

with

$$|h_1| \leq d_1/(N-1), \quad |h_2| \leq d_2/(N-1)$$

Thus in this interval,

$$h'(n)/h(n) \geq -d_1/(N-1) + \log [1 - d_2/(N-1)]$$

Now fix $n = [n^*] + 1$ and integrate the above inequality from n^* to n, to get

$$\log [h(n)/h(n^*)] \geq -d_1/(N-1) + \log (1 - d_2/(N-1))$$

or

$$\begin{aligned} h(n)/h(n^*) &\geq [1 - d_2/(N-1)] \exp [-d_1/(N-1)] \\ &> [1 - d_2/(N-1)][1 - d_1/(N-1)] > 1 - (d_1 + d_2)/(N-1) \end{aligned}$$

Define

$$d(\alpha) = d_1 + d_2 = (2-\alpha)(5-2\alpha)/[2(\alpha-1)]$$

Then for $N \geq 1/(\alpha - 1)$ and $N - 1 \geq d(\alpha)$, we have

$$h(n) > [1 - d(\alpha)/(N-1)] h(n^*)$$

for the integer $n = [n^*] + 1$. The minimum allowed value of N here is the larger of $1/(\alpha - 1)$

and

$$1+d(\alpha)=(\alpha^2-7\alpha/2+4)/(\alpha-1)$$

The quadratic $\alpha^2-7\alpha/2+4$ has its maximum in $1\leq\alpha\leq2$ at $\alpha=1$, namely 3/2. Thus it suffices to restrict N to

$$N\geq1{\cdot}5/(\alpha-1)$$

For any such N, we have shown that for $1<\alpha<2$,

$$M_N(\alpha)>\frac{0{\cdot}335[1-d(\alpha)/(N-1)]}{\sqrt{[(\alpha-1)(N-1)]}}[\alpha/(2-\alpha)]^{N-1}$$

with $d(\alpha)=(2-\alpha)(5/2-\alpha)/(\alpha-1)$.

Clearly, for large N, the factor which dominates this lower bound is the (exponentially growing) expression

$$[\alpha/(2-\alpha)]^{N-1}$$

which serves as the upper bound. To illustrate this numerically, for $\alpha=1{\cdot}5$ and $N=50$, note that the upper bound is $3^{49}\simeq2{\cdot}4\times10^{23}$, whereas the lower bound is $\simeq1{\cdot}6\times10^{22}$.

REFERENCES

1. P. Roache, *Computational Fluid Dynamics*, Hermosa Publishers, P.O. Box 8172, Albuquerque, NM, 1972.
2. S. T. Chan, P. M. Gresho, R. L. Lee and C. D. Upson, 'Simulation of three-dimensional, time-dependent, incompressible flows by a finite element method', *Proc. AIAA 5th Comp. Fl. Dyn. Conf.*, Palo Alto, Calif., 22–23 June (1981).
3. P. M. Gresho and C. D. Upson, 'Current progress in solving the time-dependent, incompressible Navier–Stokes equations in three-dimensions by (almost) the FEM', *Proc., Fourth Int. Conf. on Finite Elements in Water Resources*, Hannover, Germany, 21–25 June (1982).
4. P. M. Gresho, S. T. Chan, C. D. Upson and R. L. Lee, 'A modified finite element method for solving the time-dependent, incompressible Navier–Stokes equations', *Int. J. Num. Meth. Fluids* (in press).
5. J. Fromm, 'The time dependent flow of an incompressible viscous fluid', *Methods of Comp. Phys.*, **3,** 345–382 (1964).
6. B. P. Leonard, 'Note on the von Neumann stability of the explicit FTCS convective diffusion equation', *Appl. Math. Mod.*, **4,** 401, October (1980).
7. C. W. Hirt, 'Heuristic stability theory for finite difference equations', *J. Comp. Phys.*, **2,** 339 (1968).
8. K. W. Morton, 'Stability and convergence in fluid flow problems', *Proc. Roy. Soc. London, A*, **323,** 237–253 (1971).
9. J. Siemieniuch and I. Gladwell, 'Analysis of explicit difference methods for a diffusion–convection equation', *Int. J. Num. Meth. Eng.*, **12,** 899–916 (1978).
10. A. Rigal, 'Stability analysis of explicit finite difference schemes for the Navier–Stokes equations', *Int. J. Num. Meth. Eng.*, **14,** 617–628 (1979).
11. A. R. Mitchell and D. F. Griffiths, *The Finite Difference Method in Partial Differential Equations*, Wiley, Chichester, England, 1980.
12. L. Lapidus and G. Pinder, *Numerical Solution of Partial Differential Equations*, Wiley, New York, 1982.
13. K. W. Morton, 'Stability of finite difference approximations to a diffusion–convection equation', *Int. J. Num. Meth. Eng.*, **15,** 677–683 (1980).
14. D. F. Griffiths, I. Christie and A. R. Mitchell, 'Analysis of error growth for explicit difference schemes in conduction–convection problems', *Int. J. Num. Meth. Eng.*, **15,** 1075–1081 (1980).
15. R. M. Clancy, 'A note on finite differencing of the advection–diffusion equation', *Monthly Weather Rev.*, **109,** 1807–1809 (1981).
16. D. F. Griffiths, 'The stability of finite difference approximations to non-linear partial differential equations', *Numerical Analysis Report NA/51*, Dept. Math. Sci., University of Dundee, Scotland, 1981.
17. H. Price, R. Varga and J. Warren, 'Application of oscillation matrices to diffusion convection equations', *J. Math. Phys.*, **45,** 301–311 (1966).
18. R. Richtmeyer and K. W. Morton, *Difference Methods for Initial Value Problems*, Interscience Publishers, New York, 1967.
19. S. Paolucci and D. R. Chenoweth, 'A note on the stability of the explicit finite differenced transport equation', *J. Comp. Phys.*, **47,** (3), 489–496 (1982).

20. S. Osher, 'Systems of difference equations with general homogeneous boundary conditions', *Trans. Amer. Math. Soc.*, **137,** 177–201 (1969).
21. K. W. Morton, 'Initial value problems by finite difference and other methods', in D. A. H. Jacobs (Ed.) *The State of the Art in Numerical Analysis*, Academic Press, 1977, pp. 699–756.
22. G. D. Smith, *Numerical solution of Partial Differential Equations*, Oxford Univ. Press, London/New York, 1965.
23. J. J. Trapp and J. Ramshaw, 'A simple heuristic method for analyzing the effect of boundary conditions on numerical stability', *J. Comp. Phys.*, **20,** 238–242 (1976).
24. P. Keast and A. R. Mitchell, 'On the instability of the Crank–Nicholson formula under derivative boundary conditions', *The Computer J.*, **9,** 110–114 (1966).
25. A. J. Gadd, 'A numerical advection scheme with small phase speed errors', *Quart. J.R. Met. Soc.*, **104,** 583–594 (1978).
26. P. M. Gresho and R. L. Lee, 'Don't suppress the wiggles—they're telling you something', *Comp. & Fluids*, **9,** (2), 223–255 (1981).
27. C. Leith, 'Numerical simulation of the Earth's atmosphere', *Meth. in Comp. Phys.*, **4,** 1–28 (1965).
28. P. Smolarkiewicz, 'The multidimensional Crowley advection scheme', *Monthly Weather Rev., Monthly Weather Review*, **110,** 12, 1968–1983 (1982).
29. S. T. Chan and P. M. Gresho, 'Solution of the multi-dimensional, incompressible Navier–Stokes equations using low-order finite elements and one-point quadrature', *Proc. 4th Int. Sym. on Finite Elements in Flow Problems*, Tokyo, Japan, 26–29 July (1982).
30. P. Lax and B. Wendroff, 'Difference schemes for hyperbolic equations with high order of accuracy', *Commun. Pure & Appl. Math.* **XVII,** 381–398 (1964).
31. J. Dukowicz and J. Ramshaw, 'Tensor viscosity method for convection in numerical fluid dynamics', *J. Comp. Phys.*, **32,** 71–79 (1979).
32. A. Brooks and T. J. R. Hughes, 'Streamline upwind–Petrov–Galerkin formulation for convection–dominated flows with particular emphasis on the incompressible Navier–Stokes equations', *Comp. Meth. Appl. Mech. Eng.*, **32,** 199–259 (1982).