JOURNAL OF COMPUTATIONAL PHYSICS 64, 334–367 (1986)

An Accuracy Analysis of Boundary Conditions for the Forced Shallow Water Equations

M. G. G. FOREMAN

Institute of Ocean Sciences, Sidney, British Columbia V8L 4B2, Canada

Received December 6, 1984; revised July 19, 1985

The accuracy of numerical boundary conditions for the linearized one-dimensional shallow water equations is determined by extending the dispersion (Fourier) analysis originally presented by C. K. Chu and A. Sereny (*J. Comput. Phys.* **15**, 476 (1974)). Accuracy is measured by calculating steady state amplitudes of incoming and outgoing numerical waves at a boundary as functions of a forcing frequency. Radiating, closed, and forced/radiating boundary conditions are studied in combination with two numerical schemes: the Richardson–Sielecki explicit finite difference scheme and the Galerkin finite element method with piecewise linear basis functions and Crank–Nicolson time-stepping. Contamination of the finite element solution with short waves is seen to vary with the choice of extraneous boundary conditions. An example of L. N. Trefethen's (*J. Comput. Phys.* **49**, 199 (1983)) interpretation of the B. Gustafsson, H. Kreiss, and A. Sundström (*Math. Comput.* **26**, 649 (1972)) instability is also given for the finite element method.

INTRODUCTION

It is well known that the implementation of boundary conditions for the numerical solution of hyperbolic partial differential equations (PDEs) can affect both accuracy and stability. Boundary conditions may introduce instabilities to a numerical method which is stable on a periodic domain (i.e., Cauchy stable). They may also affect the accuracy of a stable solution directly, by generating inaccurate reflections, and indirecly, by generating undesirable short waves which contaminate the solution. Consequently, numerical methods which are both accurate and stable for the Cauchy problem may be less attractive when combined with inappropriate boundary conditions. In this study, it is shown that the dispersion (Fourier) analyses which are commonly used to measure accuracy for the Cauchy problem can be extended to study the accuracy of forced initial boundary value problems.

The accuracy of boundary conditions is often determined by examining truncation errors. Gustafsson [4] showed that boundary and initial approximations may be one order of accuracy lower than the interior approximations without decreasing the overall accuracy. Sköllermo [5, 6] extended this result by developing a technique for the total error analysis of a finite difference scheme, taking into account initial approximations, boundary conditions, and the interior

0021-9991/86 $3.00

approximation. Since a small truncation error constant may, for some waves, make a boundary scheme competitive which is formally not of the right order, she studied boundary condition accuracy indirectly. In particular, she measured the number of meshpoints per wavelength that are needed to compute each Fourier component of the solution to some preassigned relative accuracy. Both Sloan [7] and Gottlieb and Turkel [8] have used Sköllermo's analysis to compare numerical boundary conditions. However, Gottlieb and Turkel found that the approach was useful only for eliminating the least accurate conditions.

In this study, accuracy is measured by calculating steady state amplitudes of incoming and outgoing waves at a boundary as continuous functions of a forcing frequency. This approach extends the reflection analysis presented by Chu and Sereny [1] and is closely related to the recent work of Trefethen [2, 9–11], Halpern [12], and Higdon [13]. Chu and Sereny calculated reflection coefficients at the one discrete time Δt for three extraneous boundary conditions when used in combination with a solid wall condition and a two-step Lax–Wendroff scheme. Trefethen used *reflection equations* and *reflection matrices* (which become reflection coefficients in the 1×1 case) to interpret the numerous instabilities that arise with difference models of linear one-dimensional hyperbolic PDEs with one or two boundaries or interfaces. (An example of such an instability is given in Section 5.) Reflection coefficients were also used by Halpern to obtain error estimates of the reflected energy arising from absorbing boundary conditions for discretizations of the one-dimensional wave equation, and by Higdon to construct absorbing boundary conditions for the standard second-order centered difference approximation to the two-dimensional wave equation.

The following accuracy analysis is applied to the one-dimensional linearized shallow water equations with constant depth and two boundaries. Two particular problems are studied. In problem P1, periodic forcing is applied at one end of the channel and a radiation condition is imposed at the other. In problem P2, one end of the channel is closed while the other end simultaneously specifies an incoming wave and radiates all outgoing waves. Boundaries such as these are common in tidal and storm surge models. For both problems, only the steady state solution is considered. However, unlike Rudy and Strikwerda [14], who evaluated boundary conditions for both their effect on the accuracy of the solution and the rate of convergence to steady state, in this study boundary conditions are only considered for their accuracy.

Many mathematical expressions have been developed for radiating or absorbing boundary conditions (e.g., [15–17]). This study does not examine these expressions per se. Rather, it concentrates on the numerical implementation of a specific radiation condition that is exact for the frictionless shallow water equations. In particular, several implementations of this condition are studied in combination with two numerical schemes for the channel interior.

The first scheme is an explicit finite difference method whose two-dimensional extension, often referred to as the Richardson–Sielecki (RS) scheme, is commonly used in tidal and storm surge models (e.g., [18, 19]). Spatial and temporal

staggering of the variables for this scheme mean that many implementations of the radiation of forced/radiation condition are possible. In Section 3, the relative accuracy of four implementations is determined and the Orlanski [20] radiation condition is briefly discussed.

The second scheme, henceforth referred to as FEM1, is a Galerkin finite element method which combines piecewise linear basis functions and Crank–Nicolson time-stepping. Coincident variables for this scheme mean that the physical boundary conditions have obvious numerical implementations. However, additional boundary conditions are required in order to fully specify the numerical problem. It is well known (e.g., [21, 8]) that these additional conditions can affect both the accuracy and stability of the numerical solution. With FEM1, the accuracy effects are twofold. Not only can the physical properties of a boundary be modelled inaccurately, but the solution can be severely contaminated with short waves. Platzman [22], Walters and Carey [23], and Walters [24] have discussed the generation of these waves in conjunction with closed or forced boundaries. In this study, it is shown that short wave contamination also arises with radiating boundaries and varies with the choice of additional boundary conditions. In Section 4, four pairs of additional conditions are examined to determine which pair most accurately represents the boundary physics and which pair is most successful in minimizing the generation of short waves.

As discussed by Gustafsson [25], it is important to differentiate between stability and convergence to a steady state solution. Stability, both in the Lax–Richtmyer and GKS sense (Gustafsson, Kriess, and Sundström [3]), together with consistency ensures that the solution of a time-dependent difference model converges, as the mesh size approaches zero, to the correct solution of the differential equation at each fixed time t. Hence a stable difference model may admit solutions that grow in time provided this growth does not get worse as the mesh is refined. On the other hand, a model which reaches a steady state solution must not admit any growing solutions. However, this does not mean that such a model is stable. Gustafsson [25] gives an example of numerical method which would be expected to produce a steady state solution, yet is GKS unstable. In order to provide a more restrictive stability definition, he also presents sufficient conditions such that a GKS stable method, when written in the form

$$v^{n+1} = Qv^n$$

for some difference operator Q, has all the eigenvalues of Q inside the unit circle. This condition guarantees convergence to a steady state and is stronger than P-stability (Yee, Beam, and Warming [26]), which requires GKS stability and no eigenvalues outside the unit circle.

In the subsequent analysis, it is only assumed that the eigenvalues of Q are inside the unit circle. Stability is not essential for the analysis, though there is little point in determining boundary condition accuracy if the associated method is unstable. Consequently, this analysis complements Gustafsson's work in the particular case of

the shallow water equations. Given several numerical boundary conditions which satisfy the Gustafsson conditions, the following approach can determine their relative accuracy.

1. Boundary Conditions for the Shallow Water Equations

Assuming constant depth and linear friction, the one-dimensional linearized shallow water equations can be expressed in matrix form as

$$\frac{\partial \mathbf{w}}{\partial t} = Q_1 \frac{\partial \mathbf{w}}{\partial x} + Q_2 \mathbf{w} \tag{1.1a}$$

where

$$Q_1 = \begin{pmatrix} 0 & -h \\ -g & 0 \end{pmatrix}, \qquad Q_2 = \begin{pmatrix} 0 & 0 \\ 0 & -\tau \end{pmatrix} \tag{1.1b}$$

$$\mathbf{w} = \begin{pmatrix} z \\ u \end{pmatrix} \tag{1.1c}$$

and

$z(x, t)$ = elevation above mean sea level,

$u(x, t)$ = velocity,

h = depth,

g = gravity,

τ = coefficient of linear bottom friction.

These equations are to be solved on the interval $x \in [0, 1]$ for $t > 0$. Initial conditions are

$$\mathbf{w}(x, 0) = \mathbf{f}(x) \tag{1.1d}$$

for some function $\mathbf{f}$.

Since the eigenvalues of Q_1 are $\pm(gh)^{1/2}$, exactly one boundary condition should be specified at each end of the channel [27]. It will be assumed that the left boundary is either closed or radiating and the right boundary is either forced or forced/radiating.

Closed and radiating boundaries will be represented as

$$u(0, t) = 0 \tag{1.2}$$

and

$$u(0, t) = -\left(\frac{g}{h}\right)^{1/2} z(0, t) \tag{1.3}$$

respectively. Equation (1.3) is the precise relationship (when $\tau = 0$) between elevation and velocity for a leftward travelling wave [28]. It ensures no reflection at the left boundary by setting the incoming characteristic variable to zero.

The pure driving condition at the right boundary will have the form

$$u(1, t) = F(t) \tag{1.4}$$

for some function $F(t)$. The driving/radiation condition will have the form

$$u(1, t) = \left(\frac{g}{h}\right)^{1/2} z(1, t) + f(F(t)), \tag{1.5}$$

where f is some function and $F(t)$ now specifies only the inward component of the solution at the right boundary. The combined condition is designed to generate leftward waves and radiate rightward waves.

When reexpressed in terms of characteristic variables, all these boundary conditions have the form required by Kreiss [27].

2. Development of the Analysis

The following boundary condition analysis is based on separability of the spatial and temporal components of the steady state numerical solution. In this section, sufficient conditions for separability are found and the analysis approach is outlined.

Assume that the shallow water equations (1.1) and a pair of well-posed boundary conditions are solved with a finite difference or finite element method. The complete set of difference equations and numerical boundary conditions can be expressed in matrix form as

$$A\mathbf{X}^{n+1} = B\mathbf{X}^n + \mathbf{X}_D F(n+1). \tag{2.1a}$$

A and B are matrices defining the difference operations and numerical boundary conditions, $\mathbf{X}$ is the vector of discretized variables, and $\mathbf{X}_D$ is the vector which locates the forced variable at the right boundary. A is nonsingular. The forcing function is assumed to be

$$F(n) = \mathrm{Re}[ae^{i(n\omega\Delta t - \phi)}] \tag{2.1b}$$

where $0 < \omega\Delta t < \pi$ and $a > 0$. ω, a, and ϕ are referred to as the frequency, amplitude, and phase, respectively.

Repeated substitution into (2.1a) gives

$$\mathbf{X}^{n+1} = (A^{-1}B)^{n+1}\mathbf{X}^0 + \mathrm{Re}\left[ae^{i[(n+1)\omega\Delta t - \varphi]} \left(\sum_{l=0}^{n} e^{i(l-n)\omega\Delta t}(A^{-1}B)^{n-l} \right) A^{-1}\mathbf{X}_D \right] \tag{2.2}$$

where $\mathbf{X}^0$ is the vector of initial conditions. But

$$\left(\sum_{l=0}^{n} e^{i(l-n)\omega\Delta t}(A^{-1}B)^{n-l} \right) C = I - (e^{-i\omega\Delta t}A^{-1}B)^{n+1} \tag{2.3a}$$

where

$$C = [I - e^{-i\omega\Delta t}(A^{-1}B)]. \tag{2.3b}$$

So when C is invertible, (2.3a) can be written as

$$\sum_{l=0}^{n} e^{i(l-n)\omega\Delta t}(A^{-1}B)^{n-l} = [I - (e^{-i\omega\Delta t}A^{-1}B)^{n+1}]\, C^{-1}. \tag{2.4}$$

Under what conditions is C invertible? Assume that C is singular. Then for some vector $\mathbf{x} \neq 0$,

$$C\mathbf{x} = 0. \tag{2.5}$$

This implies

$$B\mathbf{x} = e^{i\omega\Delta t}A\mathbf{x} \tag{2.6a}$$

and

$$\lambda(\omega\,\Delta t) = e^{i\omega\Delta t} \tag{2.6b}$$

is an eigenvalue of the matrix $A^{-1}B$. Therefore, provided λ is not an eigenvalue of $A^{-1}B$, C is invertible.

When the driving frequency and time step are chosen so that $\lambda(\omega\,\Delta t)$ is not an eigenvalue of $A^{-1}B$, (2.2) can be rewritten as

$$\mathbf{X}^{n+1} = (A^{-1}B)^{n+1}\mathbf{X}^0 + \mathrm{Re}\{ae^{i[(n+1)\omega\Delta t - \phi]}(I - (e^{-i\omega\Delta t}A^{-1}B)^{n+1})\mathbf{Y}\} \tag{2.7}$$

where

$$\mathbf{Y} = C^{-1}A^{-1}\mathbf{X}_D. \tag{2.8}$$

$\mathbf{X}^{n+1}$ converges to a steady state when $(A^{-1}B)^{n+1}$ converges. Conditions for this are as follows.

THEOREM [29]. $\lim_{n\to\infty}(A^{-1}B)^n = L$ *(a constant matrix) if and only if*

(i) $|\lambda| \leqslant 1$ *for all eigenvalues of* $A^{-1}B$,

(ii) *if* $|\lambda| = 1$ *then* $\lambda = 1$,

(iii) *the Jordan block associated with each eigenvalue* $\lambda = 1$ *has dimension* 1×1.

THEOREM [29]. $\lim_{n\to\infty}(A^{-1}B)^n = 0$ *if and only if* $|\lambda| < 1$ *for all eigenvalues of* $A^{-1}B$.

The steady state solution for $\mathbf{X}^{n+1}$ is complicated when $A^{-1}B$ has at least one eigenvalue equal to unity. However, if it is assumed that all eigenvalues are strictly inside the unit circle (as would follow from Gustafsson's conditions [25]), then

$$\mathbf{X}^{n+1} = \mathrm{Re}[ae^{i[(n+1)\omega\Delta t - \phi]}\mathbf{Y}] \tag{2.9}$$

is the steady state solution. Notice that its spatial and temporal components are separable. The spatial profile of the steady state solution is contained in the vector $\mathbf{Y}$ and the temporal component has the same frequency as the forcing function.

The precise form of $\mathbf{Y}$ can be found by extending conventional dispersion (Fourier) analyses for the Cauchy problem. Assume the separable steady state solution has the form

$$\begin{pmatrix} z_j^n \\ u_j^n \end{pmatrix} = \begin{pmatrix} \zeta_0 \\ \mu_0 \end{pmatrix} \lambda^n \kappa^j \tag{2.10}$$

where

$$\lambda = e^{i\omega\Delta t} \tag{2.11}$$

and κ is a complex number. (This same substitution is made in the normal mode stability analysis [3, 30, 2] to form the resolvent equations.) For (2.10) to be a non-trivial solution of the interior difference equations, a characteristic equation must be satisfied. This equation is a polynomial in λ and κ. Assume that in terms of κ, the polynomial has order m. Then for a specific value of λ there are m roots, $\kappa_1, \ldots, \kappa_m$. If each root has multiplicity one, the general numerical solution is

$$\begin{pmatrix} z_j^n \\ u_j^n \end{pmatrix} = \lambda^n \sum_{l=1}^{m} \begin{pmatrix} \zeta_l \\ \mu_l \end{pmatrix} \kappa_l^j \tag{2.12}$$

for some complex coefficients $\zeta_1, \ldots, \zeta_m$, $\mu_1, \ldots, \mu_m$. In the case of multiple roots, the general solution is given by Trefethen [2, Eq. (2.7)]. Precise values for these coefficients are calculated by solving a system of equations determined by the boundary conditions and the interior difference equations.

As described in [2, 10], each term in (2.12) can be identified with a wave whose direction is determined by its group velocity. When the group velocity is nonzero, these waves are either incoming or outgoing at a boundary. Consequently, once

(2.12) is fully determined and the wave directions are identified, reflection characteristics can be calculated for each boundary and the accuracy of the boundary conditions can be assessed as a function of the driving $\omega\,\Delta t$. The following two sections illustrate the procedure in more detail.

3. The Richardson–Sielecki Scheme

Many numerical methods for solving the one-dimensional shallow water equations (such as FEM1) have their elevation and velocity variables located at the same spatial point. Since only one physical condition must be specified at each boundary, two additional conditions are required by such methods in order to fully specify the numerical problem. Methods such as the RS scheme stagger z and u spatially (see Fig. 1). Only one variable is then located at each boundary and the need for extra conditions is avoided. However, spatial staggering also means that (1.3) and (1.5) must be implemented with some type of extrapolation. The particular choice can affect both accuracy and stability. In this section, four implementations that produce steady state solutions are examined for their relative accuracy. Orlanski's [20] radiation condition is also briefly discussed.

In the domain interior, the RS equations are

$$z_j^{n+1/2} = z_j^{n-1/2} - \frac{h\,\Delta t}{\Delta x}\,(u_{j+1}^n - u_j^n) \tag{3.1a}$$

$$\left(1+\frac{1}{2}\tau\,\Delta t\right) u_j^{n+1} = \left(1-\frac{1}{2}\tau\,\Delta t\right) u_j^n - \frac{g\,\Delta t}{\Delta x}\,(z_j^{n+1/2} - z_{j-1}^{n+1/2}). \tag{3.1b}$$

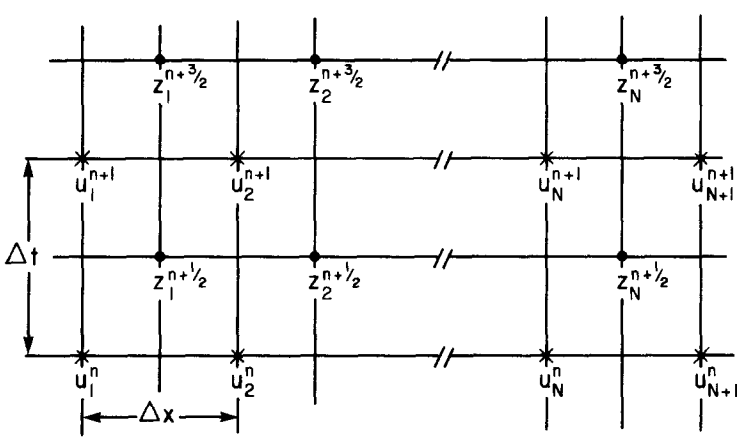

FIG. 1. One-dimensional RS grid.

Initial conditions are assumed to be

$$u_j^0 = g_1(j) \tag{3.1c}$$

$$z_j^{1/2} = g_2(j) \tag{3.1d}$$

for some functions g_1 and g_2. The RS scheme is Cauchy stable [31] when

$$f_2 = (gh)^{1/2} \frac{\Delta t}{\Delta x} \leqslant 1. \tag{3.2}$$

f_2 is commonly referred to as the Courant number.

Closed and driving boundaries are easily implemented with the RS scheme.

$$u_1^n = 0 \tag{3.3a}$$

and

$$u_{N+1}^n = F(n) \tag{3.3b}$$

respectively simulate a closed left boundary and a driving right boundary. However, temporal and spatial staggering of z and u give rise to many implementations of the radiation condition (1.3). For problem P1, four implementations will be studied in combination with the driving condition (3.3b).

The first implementation

$$u_1^{n+1} = -\left(\frac{g}{h}\right)^{1/2} z_1^{n+1/2} \tag{3.4a}$$

is commonly employed with the RS scheme [18]. It uses zeroth-order space-time extrapolation

$$z_{1/2}^{n+1} = z_1^{n+1/2} \tag{3.4b}$$

to calculate the z value coincident with u_1^{n+1}. The second and third implementations are higher-order versions of (3.4a). They are first-order space-time extrapolation

$$u_1^{n+1} = -2\left(\frac{g}{h}\right)^{1/2} z_1^{n+1/2} - u_2^n, \tag{3.5}$$

and second-order space-time extrapolation

$$u_1^{n+1} = -\left(\frac{g}{h}\right)^{1/2} (3z_1^{n+1/2} + z_2^{n-1/2}) - 3u_2^n. \tag{3.6}$$

The fourth implementation combines linear spatial extrapolation with phase velocity. Its development is illustrated in Fig. 2. Assume that the numerical wave

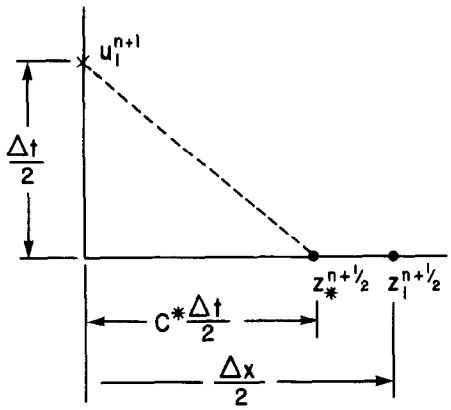

FIG. 2. Schematic for boundary condition (3.7).

has the phase speed C^*. Then in the time $\frac{1}{2}\Delta t$, elevation $z_*^{n+1/2}$ travels $\frac{1}{2}C^*\,\Delta t$ and is coincident with u_1^{n+1}. Setting

$$r = \tfrac{1}{2}(1 - C^*\,\Delta t/\Delta x), \tag{3.7a}$$

linear extrapolation for $z_*^{n+1/2}$ gives

$$z_*^{n+1/2} = (1+r)\,z_1^{n+1/2} - rz_2^{n+1/2}. \tag{3.7b}$$

The radiation condition is then simply a refinement of (3.4a) to

$$u_1^{n+1} = -\left(\frac{g}{h}\right)^{1/2} z_*^{n+1/2}. \tag{3.7c}$$

The analysis in Section 2 is easily applied to the RS scheme. Define the vector $\mathbf{X}$ as

$$\mathbf{X}^{n+1} = (u_1^{n+1}, z_1^{n+1/2}, u_2^{n+1}, z_2^{n+1/2}, \ldots, z_N^{n+1/2}, u_{N+1}^{n+1}). \tag{3.8}$$

Then Eqs. (3.1), (3.3b), and one of either (3.4a), (3.5), (3.6), or (3.7) can be expressed in the form (2.1a) with matrix $A = I$, the identity matrix. For fixed values of N, f_2, and

$$f_1 = \frac{\tau\,\Delta x}{(gh)^{1/2}} \tag{3.9}$$

assume that all the eigenvalues of B have been shown (either numerically or with Gustafsson's conditions) to lie inside the unit circle. Then the steady state numerical solution has the separable form (2.9). In particular, assume

$$z_j^{n+1/2} = \zeta_0 \lambda^{n+1/2} \kappa^j \tag{3.10a}$$

$$u_j^{n+1} = \mu_0 \lambda^{n+1} \kappa^{j-1/2} \tag{3.10b}$$

with λ defined by (2.11). A nontrivial solution for (3.1a) and (3.1b) requires that the characteristic equation

$$\kappa^2 - 2\kappa\{1 + \tfrac{1}{2}[\lambda - 2 + 1/\lambda + \tfrac{1}{2}f_1 f_2(\lambda - 1/\lambda)]/f_2^2\} + 1 = 0 \tag{3.11}$$

be satisfied. For a specific value of λ there are two roots, κ_1 and κ_2, whose product is 1. Assuming $\kappa_1 \neq \kappa_2$, these roots are designated as follows:

(i) when both roots are real, κ_2 has the larger magnitude, i.e.,

$$|\kappa_2| = r > 1 > \frac{1}{r} = |\kappa_1|, \tag{3.12a}$$

(ii) otherwise, κ_2 has positive argument, i.e.,

$$\kappa_2 = re^{ik_2 \Delta x}, \tag{3.12b}$$

where $r > 0$ and $0 < k_2 \Delta x < \pi$. The general numerical solution is then

$$z_j^{n+1/2} = e^{i(n+1/2)\omega \Delta t}[\zeta_1 r^{-j} e^{-ijk_2 \Delta x} + \zeta_2 r^j e^{ijk_2 \Delta x}] \tag{3.13a}$$

$$u_j^{n+1} = e^{i(n+1)\omega \Delta t}[\mu_1 r^{-(j-1/2)} e^{-i(j-1/2)k_2 \Delta x} + \mu_2 r^{j-1/2} e^{i(j-1/2)k_2 \Delta x}] \tag{3.13b}$$

for some complex coefficients ζ_1, ζ_2, μ_1, and μ_2.

This solution may be interpreted (see [2]) as two waves with spatially varying amplitude profiles. The first wave has wavenumber k_2 and propagates rightward as n increases. When $r > 1$, its amplitude decreases with propagation. The second wave has wavenumber $-k_2$ and propagates leftward. When $r > 1$, its amplitude also decreases with propagation. The propagation speed of each wave is called its phase velocity, C. However, in studying wave reflections, leftward and rightward waves should be defined in terms of their group velocity G, rather than their phase velocity, since G is approximately the speed of energy propagation (e.g., see Lemma 5.1 in [10]). *Leftward* and *rightward* waves are therefore defined to have $G < 0$ and $G > 0$, respectively. This definition is consistent with Trefethen [2, 10].

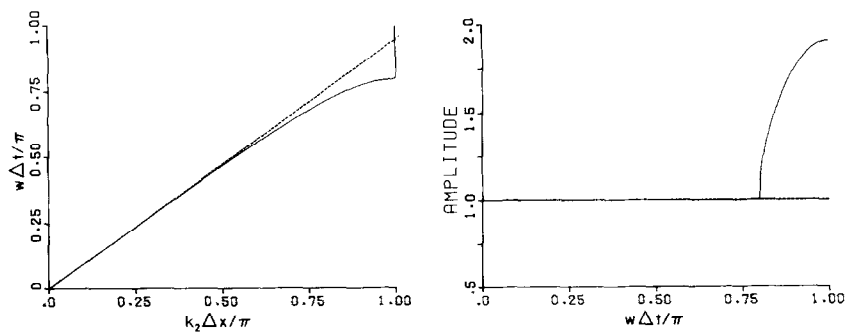

FIG. 3. Amplitude and phase of κ_2 for $f_1 = 0.0$ and $f_2 = 0.95$. Solid line, RS scheme; dashed line, analytic solution.

Figure 3 shows the amplitude and phase of κ_2 as a function of the driving frequency $\omega\,\Delta t$ when $f_1 = 0.0$ and $f_2 = 0.95$. The diagram on the right plots $|\kappa_2|$ versus $\omega\,\Delta t$ and indicates how wave amplitudes change with propagation. The diagram on the left is referred to as a dispersion curve. Nondimensional phase and group speeds are calculated as

$$\frac{C}{(gh)^{1/2}} = -\frac{\omega\,\Delta t}{f_2 k\,\Delta x} \tag{3.14a}$$

and

$$\frac{G}{(gh)^{1/2}} = -\frac{1}{f_2}\frac{\partial(\omega\,\Delta t)}{\partial(k\,\Delta x)}, \tag{3.14b}$$

where in this case $k = k_2$. For small frequencies, both speeds are close to their analytic values of -1.0. Notice that when $0 < k_2\,\Delta x < \pi$, C and G have the same sign. This is true for all values of (f_1, f_2) and means that when κ_2 and κ_1 are defined by (3.12b), they are associated with waves whose group velocities are leftward and rightward, respectively. Equation (3.12a) only arises when $f_1 = 0$ and both roots are negative. With reference to Table II in [10], the solutions associated with κ_2 and κ_1 are still referred to as leftgoing and rightgoing, respectively.

In Fig. 3, notice that $2\Delta x$ waves ($k_2\,\Delta x = \pi$) arise when $\omega\,\Delta t$ is larger than the cutoff value, ω_c, of 2.5. Such waves are similar to the $4\Delta x$ waves that Vichnevetsky [32] prodicts for an unstaggered finite difference gird. κ_2 amplitudes are seen to increase dramatically when $\omega\,\Delta t > \omega_c$. This means that the amplitude profile due to κ_2 decreases away from the right boundary, while the profile due to κ_1 increases. Usually ζ_2 is much larger that ζ_1 and the resultant $2\Delta x$ wave has an amplitude profile that decreases to the left. Vichnevetsky [32] observes that the amplitude of these *evanescent* waves decays in space at a rate which increases monotonically with the excess of frequency above the cutoff.

Reflection coefficients are calculated from the amplitudes of the leftward and rightward waves at the boundaries. At time step n, the leftward and rightward u waves at the left boundary are given by $\mu_2\lambda^n\kappa^{1/2}$ and $\mu_1\lambda^n\kappa^{-1/2}$, respectively, where κ is now used in place of the κ_2 defined by (3.12). For any left boundary condition, the reflection coefficient for u is therefore

$$R_L = \left(\frac{\mu_1}{\mu_2\kappa}\right). \tag{3.15}$$

At the right boundary, the leftward and rightward u waves are given by $\mu_2\lambda^n\kappa^{N+1/2}$ and $\mu_1\lambda^n\kappa^{-(N+1/2)}$, respectively. Imposing boundary condition (3.3b) and assuming that $F(n)$ has the form (2.1b), the reflected portion of the leftward wave is $\lambda^n(\mu_2\kappa^{N+1/2} - ae^{-i\phi})$ and the reflection coefficient for the right boundary is

$$R_R = \frac{\mu_2\kappa^{N+1/2} - ae^{-i\phi}}{\mu_1\kappa^{-(N+1/2)}}. \tag{3.16}$$

But (3.3b) implies $R_R = -1.0$, regardless of the value of N. So (3.3b) is a perfect reflector of outgoing waves.

The reflection coefficient for radiation condition (3.4a) is calculated as follows. Substituting (3.13) into (3.4a) yield

$$\lambda^{1/2}(\mu_1 \kappa^{-1/2} + \mu_2 \kappa^{1/2}) + \left(\frac{g}{h}\right)^{1/2} (\zeta_1 \kappa^{-1} + \zeta_2 \kappa) = 0. \tag{3.17}$$

Making the same substitution in the continuity equation (3.1a) for $j = 1, N$ implies

$$\zeta_1 = -\frac{h\,\Delta t}{\Delta x}\mu_1 \left(\frac{\kappa^{-1/2} - \kappa^{1/2}}{\lambda^{1/2} - \lambda^{-1/2}}\right) \tag{3.18}$$

and

$$\zeta_2 = -\frac{h\,\Delta t}{\Delta x}\mu_2 \left(\frac{\kappa^{1/2} - \kappa^{-1/2}}{\lambda^{1/2} - \lambda^{-1/2}}\right). \tag{3.19}$$

Substituting these values into (3.17) then yields

$$R_L = -\left\{1 - f_2\left(\frac{\kappa - 1}{\lambda - 1}\right)\right\} \Big/ \left\{1 - f_2\left(\frac{\kappa^{-1} - 1}{\lambda - 1}\right)\right\}. \tag{3.20}$$

Notice that this result is independent of the right boundary condition and N, the number of grid points.

Reflection coefficients for boundary conditions (3.5), (3.6), and (3.7c) with $C^* \Delta t/\Delta x = f_2$ are calculated similarly. They are

$$R_L = -\left\{1 + \frac{\kappa}{\lambda} - 2f_2\left(\frac{\kappa - 1}{\lambda - 1}\right)\right\} \Big/ \left\{1 + \frac{1}{\lambda\kappa} - 2f_2\left(\frac{\kappa^{-1} - 1}{\lambda - 1}\right)\right\}, \tag{3.21}$$

$$R_L = -\left\{1 + \frac{3\kappa}{\lambda} - f_2\left(\frac{\kappa - 1}{\lambda - 1}\right)\left(3 + \frac{\kappa}{\lambda}\right)\right\} \Big/ \left\{1 + \frac{3}{\lambda\kappa} - f_2\left(\frac{\kappa^{-1} - 1}{\lambda - 1}\right)\left(3 + \frac{1}{\lambda\kappa}\right)\right\}, \tag{3.22}$$

and

$$R_L = -\left\{1 - f_2\left(\frac{\kappa - 1}{\lambda - 1}\right)[(1 + r) - r\kappa]\right\} \Big/ \left\{1 - f_2\left(\frac{\kappa^{-1} - 1}{\lambda - 1}\right)[(1 + r) - r\kappa^{-1}]\right\}, \tag{3.23}$$

respectively.

Reflection coefficients whose absolute value is zero denote an outgoing wave that is transmitted through the boundary without any reflection. In such cases the associated radiation condition is exact. The relative accuracy of boundary conditions (3.4a), (3.5), (3.6), and (3.7) can therefore be measured by examining the magnitude of their reflection coefficients. In general, these magnitudes vary with f_1, f_2, and $\omega\,\Delta t$.

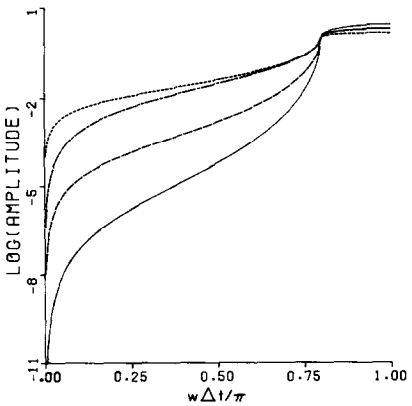

FIG. 4. Reflection coefficient amplitudes for the RS scheme with $f_1 = 0.0$, $f_2 = 0.95$, and left boundary conditions calculated using: short dashed line, constant space-time extrapolation (3.4a); long dashed line, linear space-time extrapolation (3.5); solid line, quadratic space-time extrapolation (3.5); long-short dashed line, linear spatial extrapolation (3.7) with $C^* \Delta t/\Delta x = f_2$.

Figure 4 plots the reflection coefficient amplitudes arising from (3.20), (3.21), (3.22), and (3.23). Long waves have coefficient amplitudes that are very close to zero. This means that they are almost completely absorbed by the boundary. These amplitudes increase with $\omega \Delta t$, indicating less absorption (or greater reflection) as the wavelength decreases. Reflection coefficients also increase beyong the cutoff frequency and have amplitudes larger than 1. This means that the reflected wave is larger than the incident wave. Such behaviour could conceivably cause instability if the reflected wave did not decrease in amplitude as it moved away from the boundary.

As would be expected, higher orders of space-time extrapolation produce more

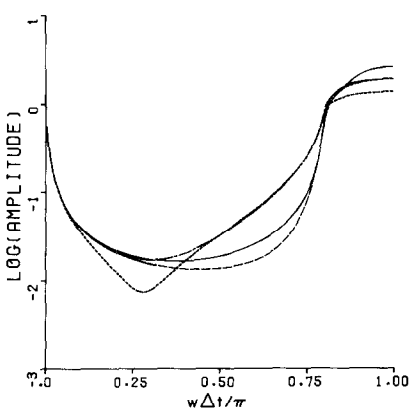

FIG. 5. Reflection coefficient amplitudes for the RS scheme with $f_1 = 0.05$ and $f_2 = 0.95$. Notation as in Fig. 4.

accurate implementations of radiation condition (1.3). Linear spatial extrapolation coupled with phase velocity is seen to be more accurate than constant space-time extrapolation, but not as accurate as linear space-time extrapolation.

When $\tau \neq 0$, the reflection coefficients change dramatically. Figure 5 is similar to Fig. 4 but has $f_1 = 0.05$. In all cases, the left boundary condition is no longer most effective with long waves. This is to be expected since when $\tau > 0$, the elevation and velocity of a travelling wave are no longer in phase [28]. Consequently, a boundary condition such as (1.3) which specifies a one-point scalar relationship between z and u cannot be expected to effectively absorb waves at the boundary.

In order to obtain more accurate boundary conditions in the presence of friction, (1.3) should either be refined or replaced with another radiation condition. Verboom and Slob [33] suggest refinements which are constructed to be most accurate for particular values of the parameter τ/ω. Unfortunately, they also report instabilities with some implementations.

The radiation condition presented by Orlanski [20] is a promising alternative to (1.3). It is a two-stage process whereby a discretized version of the Sommerfeld radiation condition is first used to estimate the phase speed near a radiating boundary at time step n. This estimate is then inserted in a consistent discretization to calculate boundary values at time step $n+1$. Although the resulting condition is nonlinear and cannot be analyzed with the previous approach, it is easily verified that the condition does not reflect any wave solutions of the form (3.10), even when $\tau \neq 0$. Unfortunately, stability constraints require that the phase speed estimates must be bounded. When an estimate exceeds the bound, Orlanski resets the estimate to the limiting value. With this refinement, perfect radiation is no longer guaranteed, and indeed does not seem to occur for travelling waves. Orlanski's boundary condition was included in the model tests whose results are summarized in Table I. Although it is more accurate than the four implementations of (1.3) when $f_1 = 0.05$, $f_2 = 0.95$, $N = 10$, and $\omega \Delta t = 0.76183622$, there is a reflected wave. Camerlengo and O'Brien [34] present a variation of Orlanski's condition wherein the phase speed is not estimated but set to the constant $\Delta x/\Delta t$. The resulting condition may be viewed as a variation of (3.7).

Problem P2 combines numerical implementations of the driving/radiation condition (1.5) with the numerical closed boundary condition (3.3a). Implementations of the driving/radiation condition are derived by writing u_{N+1}^{n+1} and $z_N^{n+1/2}$ in terms of their leftward and rightward components

$$\begin{aligned} u_{N+1}^{n+1} &= u_{\mathrm{L}}^{n+1} + u_{\mathrm{R}}^{n+1} \\ z_N^{n+1/2} &= z_{\mathrm{L}}^{n+1/2} + z_{\mathrm{R}}^{n+1/2}. \end{aligned} \tag{3.24}$$

The first implementation assumes constant space-time extrapolation for the rightward components

$$u_{\mathrm{R}}^{n+1} = \left(\frac{g}{h}\right)^{1/2} z_{\mathrm{R}}^{n+1/2}, \tag{3.25}$$

and (3.19) for the leftward components. The condition is

$$u_{N+1}^{n+1} - \left(\frac{g}{h}\right)^{1/2} z_N^{n+1/2} = u_L^{n+1}\left[1 + f_2\left(\frac{1-\kappa^{-1}}{\lambda-1}\right)\right]. \tag{3.26}$$

In practice, the real part of the right side of (3.26) is the driving condition. The complex multiplier for u_L^{n+1} therefore has the effect of changing the amplitude and phase of the original forcing function.

The second implementation combines (3.19) with linear space-time extrapolation of the outgoing wave components. In this case, the driving/radiation condition becomes

$$\begin{aligned} u_{N+1}^{n+1} + u_N^n - 2\left(\frac{g}{h}\right)^{1/2} z_N^{n+1/2} \\ = u_L^{n+1}\left\{1 + (2f_2 - 1)\left(\frac{1-\kappa^{-1}}{\lambda-1}\right)\right\} + u_L^n\left\{1 + \left(\frac{1-\kappa^{-1}}{\lambda-1}\right)\right\}. \end{aligned} \tag{3.27}$$

The third implementation calculates the outgoing wave components by combining linear extrapolation in time with a phase speed estimate. Assuming the numerical phase speed C^* and setting

$$r = \frac{\Delta x}{C^* \Delta t}, \tag{3.28a}$$

this driving/radiation condition is

$$\begin{aligned} \left(\frac{1+r}{2}\right) u_{N+1}^{n+1} - \left(\frac{r-1}{2}\right) u_{N+1}^n - \left(\frac{g}{h}\right)^{1/2} z_N^{n+1/2} \\ = u_L^{n+1}\left\{\left(\frac{1+r}{2}\right) + f_2\left(\frac{1-\kappa^{-1}}{\lambda-1}\right)\right\} - \left(\frac{r-1}{2}\right) u_L^n. \end{aligned} \tag{3.28b}$$

Right boundary reflection coefficients for P2 are calculated similarly to those for the left boundary in P1. Those arising from (3.26) and (3.27) are equal to (3.20) and (3.21), respectively. Consequently, they too are independent of N and the boundary condition at the other end of the channel. The reflection coefficient arising from (3.28b) is also independent of N. However, it is not identical to (3.23) because the latter used spatial rather than temporal extrapolation. However, with $f_2 = 0.95$, $f_1 = 0.0$, and $r = f_2^{-1}$, the two coefficients are very close.

When (3.19) is not used in the derivation of the preceding conditions, reflection coefficient magnitudes can become higher. For example, when

$$z_L^{n+1/2} = -(h/g)^{1/2} u_L^n \tag{3.29}$$

is assumed instead of (3.19), a reflection coefficient that is dependent on N arises.

TABLE I

Reflection Coefficients at the Left Boundary for the RS Solution of Problem P1

Radiation condition	Result source	$f_1 = 0.0$		$f_1 = 0.05$	
		Amplitude	Phase	Amplitude	Phase
(3.4a): constant space-time extrapolation	Analysis	0.01145075	1.97288352	0.00872593	−0.15186741
	Model	0.01145074	1.97288350	0.00872593	−0.15186751
(3.5): linear space-time extrapolation	Analysis	0.00013112	0.80417439	0.01794716	−0.73455338
	Model	0.00013112	0.80417498	0.01794716	−0.73455347
(3.6: quadratic space-time extrapolation	Analysis	0.00000150	−0.36453475	0.01946156	−0.34215353
	Model	0.00000150	−0.36457008	0.01946156	−0.34215357
(3.7): spatial/ phase speed extrapolation	Analysis	0.00455887	0.78673915	0.01856583	−0.52312655
	Model	0.00455887	0.78673906	0.01856652	−0.52312658
Orlanski	Model	0.00960697	1.04989808	0.00212887	0.82618428

TABLE II

Reflection Coefficients at Both Boundaries for the RS Solution of Problem P2

Driving radiation condition	Result source	Right boundary		Left boundary	
		Amplitude	Phase	Amplitude	Phase
(3.26): constant	Analysis	0.01145075	1.97288352	1.00000000	3.14159265
space-time extrapolation	Model	0.01145078	1.97288537	0.99999999	3.14159263
(3.27): linear	Analysis	0.00013112	0.80417439	1.00000000	3.14159265
space-time extrapolation	Model	0.00013110	0.80411099	1.00000000	−3.14159268
(3.28): C and linear	Analysis	0.00436959	3.14159266	1.00000000	3.14159265
extrapolation in time	Model	0.00436962	−3.14158791	0.99999998	3.14159261

[redacted] oscillates around the curve arising from (3.20) in Fig. 4.

The preceding analysis results were partially confirmed with numerical model tests. Both problems P1 and P2 were tested with all the preceding boundary conditions, $N = 10$, and the driving frequency $\omega \Delta t = 0.76183622$. Parameter values for the model runs were identical to those in the analysis, and the model computations were done in double precision on a SPERRY 1100/60. Each run lasted for 500 time steps (approximately 60 cycles) so that the solution would be reasonably close to a steady state (if it did indeed converge). Least squares analyses of the model results were then used to calculate the coefficients of the leftward and rightward waves, as predicted by (3.13).

In order to determine if the coefficients were converging, four least squares fits were made over the successive time step ranges [401, 425], [426, 450], [451, 475], and [476, 500]. Residuals decreased with each successive fit, and the fitted values seemed to be converging. Reflection coefficients for both the left and right boundaries were then calculated from the fitted μ_1 and μ_2 values. In all cases (except Orlanski), the reflection coefficient amplitudes from the fourth fit were identical to at least 6 decimal places with the analysis results. These results are summarized in Tables I and II.

4. The Galerkin Finite Element Method

This section examines the accuracy of several boundary conditions for the Galerkin finite element method which combines piecewise linear basis functions with Crank–Nicolson time stepping. Although a wide variety of time-stepping schemes can be used with this finite element method, Crank–Nicolson time-stepping is chosen here because it is (generally) the most accurate linear two-step method for the Cauchy problem (1.1) [28].

Unlike the RS scheme, the discrete z_j^n and u_j^n variables for FEM1 are not staggered in either time or space. In the domain interior, the FEM1 difference equations are

$$\frac{1}{6}\left[(z_{j-1}^{n+1} - z_{j-1}^{n}) + 4(z_j^{n+1} - z_j^n) + (z_{j+1}^{n+1} - z_{j+1}^{n})\right] + \left(\frac{h\,\Delta t}{4\Delta x}\right)\left[u_{j+1}^{n+1} - u_{j-1}^{n+1} + u_{j+1}^{n} - u_{j-1}^{n}\right] = 0 \tag{4.1a}$$

$$\frac{1}{6}\left[(u_{j-1}^{n+1} - u_{j-1}^{n}) + 4(u_j^{n+1} - u_j^n) + (u_{j+1}^{n+1} - u_{j+1}^{n})\right] + \left(\frac{g\,\Delta t}{4\Delta x}\right)\left[z_{j+1}^{n+1} - z_{j-1}^{n+1} + z_{j+1}^{n} - z_{j-1}^{n}\right] + \frac{1}{12}\tau\,\Delta t\left[u_{j-1}^{n+1} + u_{j-1}^{n} + 4(u_j^{n+1} + u_j^n) + u_{j+1}^{n+1} + u_{j+1}^{n}\right] = 0. \tag{4.1b}$$

Initial conditions

$$u_j^0 = g_1(j) \tag{4.2a}$$

$$z_j^0 = g_2(j) \tag{4.2b}$$

are assumed for some functions g_1 and g_2.

Because there is no staggering of the variables, the numerical implementation of boundary conditions for problem P1 is straightforward.

$$u_1^n = -\left(\frac{g}{h}\right)^{1/2} z_1^n \tag{4.3a}$$

and

$$u_N^n = F(n) \tag{4.3b}$$

approximate the absorbing left and driving right boundaries, respectively. They are called *physical* conditions. However, since (4.1a) and (4.1b) can only be applied to grid points $j = 2, N-1$, two other conditions are required in order that the system of equations which determines the numerical solution at each time step has full rank. These conditions are normally applied at the boundaries and are called *extraneous* or *computational* boundary conditions.

Extraneous boundary conditions are commonly constructed by either extrapolating variables from the grid interior or approximating a governing equation at the boundary with a one-sided difference expression. With a finite element scheme, the latter type arises naturally if basis functions and the Galerkin condition are applied at the boundary points. The particular choice of extraneous condition can affect both the stability and accuracy of the numerical solution. Gunzburger [35] has shown that hyperbolic systems can becomes unstable when the Galerkin method is applied at the boundary. Gottlieb, Gunzburger, and Turkel [36] have shown that in general extrapolation of outgoing characteristic variables is stable for both finite difference and semi-discrete Galerkin solutions of linear hyperbolic systems of equations in one dimension. In comparative studies, Chu and Sereny [1] and Sloan [7] found that characteristic extrapolations also produced more accurate numerical solutions.

In this section, three types of characteristic extrapolation are compared with the extraneous boundary conditions that arise from applying the Box scheme to the governing equations. Other pairs of extraneous conditions were briefly examined but rejected when it was discovered that they would not produce a steady state solution. For example, an eigensolution with amplitude greater than 1.0 arises when either constant or linear spatial extrapolation is applied to z.

The four pairs of extraneous boundary conditions are:

(i) zeroth-order spatial extrapolation of the leftward characteristic variable at the left boundary and the rightward characteristic variable at the right boundary

$$-\left(\frac{g}{h}\right)^{1/2} z_1^n + u_1^n = -\left(\frac{g}{h}\right)^{1/2} z_2^n + u_2^n \tag{4.4a}$$

$$\left(\frac{g}{h}\right)^{1/2} z_N^n + u_N^n = \left(\frac{g}{h}\right)^{1/2} z_{N-1}^n + u_{N-1}^n, \tag{4.4b}$$

(ii) constant space-time extrapolation of the outgoing characteristic variables

$$-\left(\frac{g}{h}\right)^{1/2} z_1^{n+1} + u_1^{n+1} = -\left(\frac{g}{h}\right)^{1/2} z_2^n + u_2^n \tag{4.5a}$$

$$\left(\frac{g}{h}\right)^{1/2} z_N^{n+1} + u_N^{n+1} = \left(\frac{g}{h}\right)^{1/2} z_{N-1}^n + u_{N-1}^n, \tag{4.5b}$$

(iii) linear spatial extrapolation of the outgoing characteristic variables

$$-\left(\frac{g}{h}\right)^{1/2} (z_1^n - 2z_2^n + z_3^n) + u_1^n - 2u_2^n + u_3^n = 0 \tag{4.6a}$$

$$\left(\frac{g}{h}\right)^{1/2} (z_N^n - 2z_{N-1}^n + z_{N-2}^n) + u_N^n - 2u_{N-1}^n + u_{N-2}^n = 0, \tag{4.6b}$$

(iv) Box scheme applied to the momentum equation at the left boundary and the continuity equation at the right boundary

$$(u_1^{n+1} - u_1^n) + (u_2^{n+1} - u_2^n) + \frac{g\,\Delta t}{\Delta x}[(z_2^n - z_1^n) + (z_2^{n+1} - z_1^{n+1})]$$

$$+ \frac{1}{2}\tau\,\Delta t(u_1^{n+1} + u_1^n + u_2^{n+1} + u_2^n) = 0 \tag{4.7a}$$

$$(z_N^{n+1} - z_N^n) + (z_{N-1}^{n+1} - z_{N-1}^n) + \frac{h\,\Delta t}{\Delta x}[(u_N^n - u_{N-1}^n) + (u_N^{n+1} - u_{N-1}^{n+1})] = 0. \tag{4.7b}$$

Other applications of the Box scheme were examined but found to have eigensolutions with amplitude equal to 1.0. Specifically, $\lambda = 1$ is an eigenvalue of $A^{-1}B$ when the Box scheme is applied to the continuity equation at both boundaries, and $\lambda = -1$ is an eigenvalue when the Box scheme is applied to the momentum equation at both boundaries. In both instances, these eigenvalues exist for all values of f_1, f_2, and N. On the other hand, numerical computations with extraneous boundary conditions (4.7a), (4.7b) and selected values of f_1, f_2, and N consistently found that all eigenvalues of $A^{-1}B$ lay inside the unit circle.

The boundary condition analysis for FEM1 proceeds as in Section 2. Define the vector $\mathbf{X}$ as

$$\mathbf{X} = (u_1^n, z_1^n, u_2^n, z_2^n, \ldots, u_N^n, z_N^n). \tag{4.8}$$

Equations (4.1), (4.3), and one of (4.4), (4.5), (4.6), or (4.7) can then be expressed in the form (2.1). Assuming that all the eigenvalues of $A^{-1}B$ are inside the unit circle, the steady state numerical solution then has the separable form (2.9). In particular, assume

$$\begin{pmatrix} z_j^n \\ u_j^n \end{pmatrix} = \begin{pmatrix} \zeta_0 \\ \mu_0 \end{pmatrix} \lambda^n \kappa^j \tag{4.9}$$

where λ is given by (2.11). Equation (4.1) has nontrivial solutions of this form when

$$(\lambda-1)^2(\kappa^2+4\kappa+1)^2+\frac{1}{2}\tau\,\Delta t(\lambda^2-1)(\kappa^2+4\kappa+1)^2 - \frac{9}{4}gh\left(\frac{\Delta t}{\Delta x}\right)^2(\lambda+1)^2(\kappa^2-1)^2=0. \tag{4.10}$$

This characteristic equation has four roots for each value of λ. If the roots are distinct, the numerical solution is

$$\begin{pmatrix} z_j^n \\ u_j^n \end{pmatrix} = \lambda^n \sum_{l=1}^{4} \begin{pmatrix} \zeta_l \\ \mu_l \end{pmatrix} \kappa_l^j, \tag{4.11}$$

where the coefficients ζ_l, μ_l are determined by the boundary conditions and (4.1). Provided each of the four κ roots has a nonzero argument, $k\Delta x$, (4.11) describes four travelling waves with wavenumbers $-k$. Two wavenumbers are positive and correspond to waves with phase velocity $C>0$, while the other two are negative and correspond to waves with $C<0$.

Figure 6 illustrates the relationship between $\lambda = e^{i\omega\Delta t}$ and $\kappa = re^{ik\Delta x}$ for the parameter values $f_2 = 1.0$ and $f_1 = 0.0$. The diagram on the left is the dispersion curve for waves with $k\Delta x>0$. Equation (3.14) therefore implies $C<0$. The analogous curve for waves with $C>0$ is simply the reflection about the $\omega\Delta t$ axis.

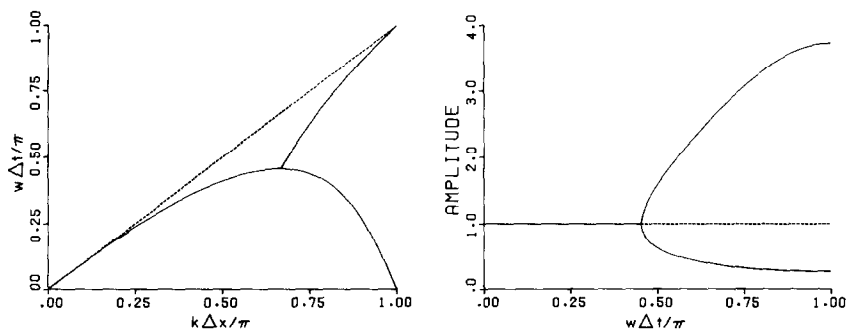

FIG. 6. Amplitude and phase of κ for $f_1 = 0.0$ and $f_2 = 1.0$. Solid line, FEM1; dashed line, analytic solution.

Notice that $\omega \Delta t$ values less than the cutoff frequency ω_c are associated with two wevenumbers. As was pointed out by Platzman [22], this suggests that the response to forcing will be a short wavelength noise component as well as the longer wavelength that is physically appropriate to the forcing frequency. The diagram on the right plots $|\kappa|$ as a function of the forcing frequency. When $\omega \Delta t < \omega_c$ both associated κ values have amplitude unity. (This is no longer true when $\tau > 0$.) When $\omega \Delta t > \omega_c$, the two κ values have the same wavenumbers but different amplitudes. If the coefficients ζ_l and μ_l indicate that the κ value with the larger magnitude dominates, an evanescent signal emanating from the right boundary will decay as it propagates leftward. This is predicted by Vichnevetsky [32], and was also seen with the RS scheme. However, unlike the fixed wavenumbers associated with the RS and Vichnevetsky evanescent signals, these wavenumbers vary with $\omega \Delta t$.

The dispersion curve in Fig. 6 shows that the shorter wave associated with

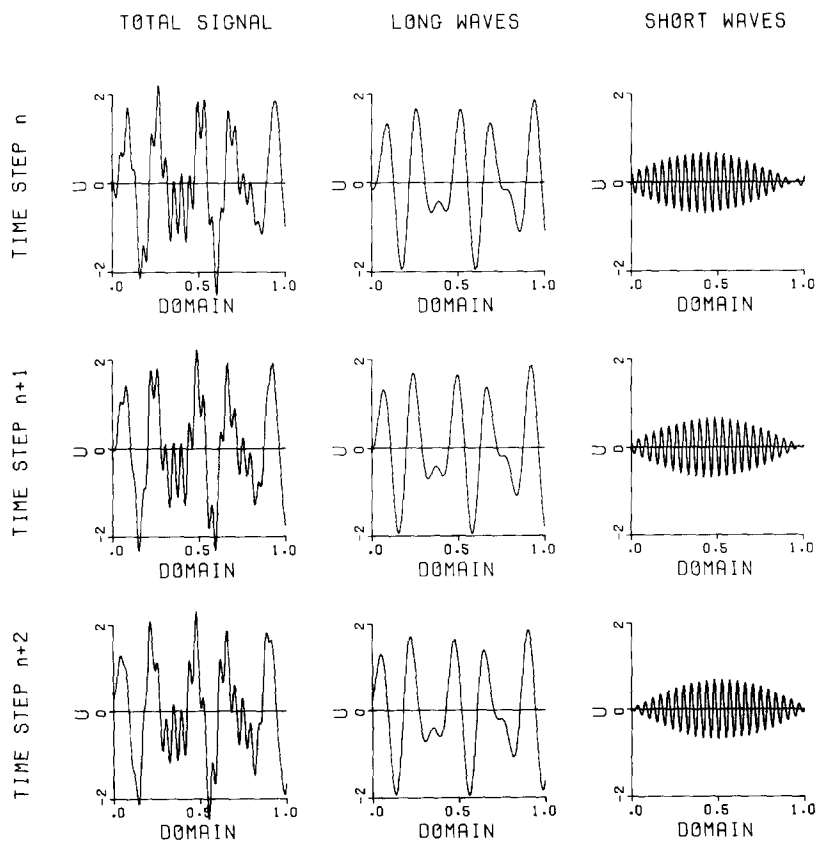

FIG. 7. Steady state u values at three consecutive time steps for the FEM1 solution of problem P1 with extraneous boundary conditions (4.4). $f_1 = 0.0$, $f_2 = 1.0$, $N = 50$. The right boundary is forced with two sinusoidal functions of equal amplitude and frequencies $\omega \Delta t = 0.57334066$ and $\omega \Delta t = 0.85608400$.

$\omega \Delta t < \omega_c$ has $C < 0$ yet $G > 0$. This means that wave energy travels in the opposite direction to the wave itself. This phenomenon is illustrated in Fig. 7. The steady state solution arising from two forcing frequencies is seen to comprise two long waves and two short waves. Both long waves have negative phase and group velocity and are seen to move leftward. However, the short wave envelope moves rightward, indicating $G > 0$, whereas the short carrier wave moves leftward, indicating $C < 0$.

The calculation of reflection coefficients can be much more complicated with FEM1 than with the RS scheme because energy may be transferred between short and long waves. This is illustrated in Fig. 8. Early u values for the FEM1 solution of P1 with $\omega \Delta t = 0.30630528$ and extraneous boundary conditions (4.4) are shown. Periodic forcing at the right boundary seemingly generates only a long leftward wave. However, a smaller short (rightward) wave is generated when the leftward wave arrives at the left boundary. The numerical radiation condition is therefore transmitting most of the long wave energy and reflecting the remainder in the form of a short wave. An analysis of the steady state solution for this problem confirms that there is negligible energy in both the long rightward and short leftward waves. It also implies that at the right boundary, the short rightward wave is reflected as a long leftward wave.

Although energy transfers between long and short waves can be calculated with reflection matrices [9], this will not be done here. Instead, boundary condition accuracy will be determined simply by calculating the coefficients ζ_l and μ_l in (4.11). Ideally, short wave coefficients should be zero and long wave coefficients should have values that are physically appropriate for the problem.

The calculation of the ζ_l and μ_l coefficients for FEM1 is similar to that for the RS

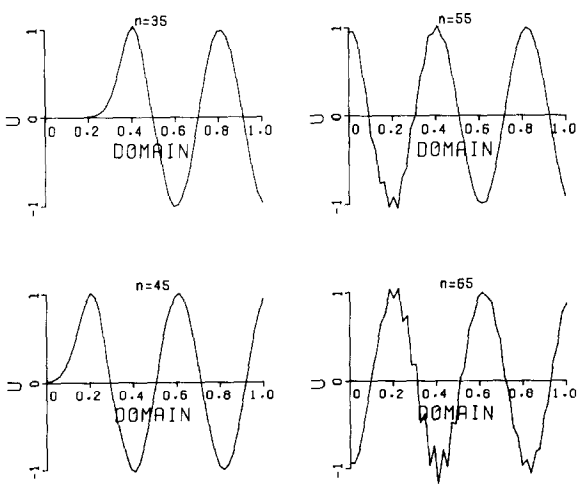

FIG. 8. Early u solutions for the FEM1 solution of problem P1 with extraneous boundary conditions (4.4), $f_1 = 0.0$, $f_2 = 1.0$, $N = 50$, and $\omega \Delta t = 0.30630528$.

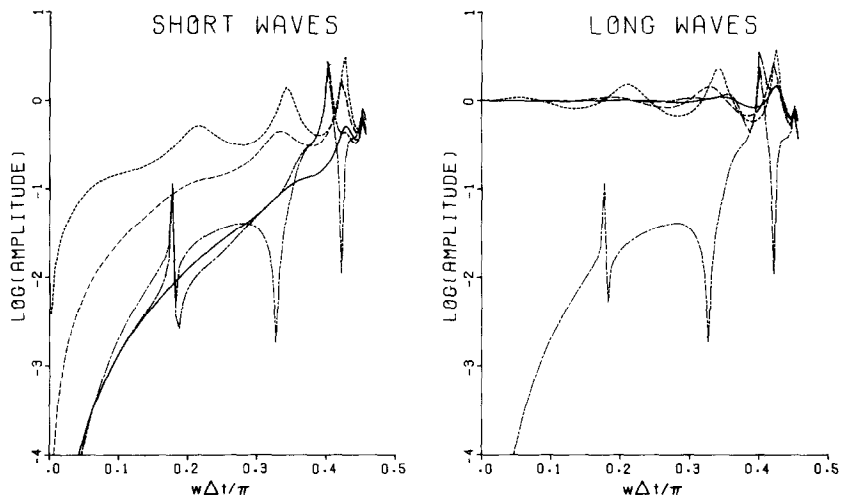

FIG. 9. Wave amplitudes (for u) for the FEM1 solution of problem P1. $\omega\,\Delta t < \omega_c$, $f_1 = 0.0$, $f_2 = 1.0$, $N = 10$. Short dashed line, constant spatial extrapolation of the characteristic variables (4.4); solid line, constant space-time extrapolation of the characteristic variables (4.5); long-short dashed line, Box scheme applied to the continuity equation (4.7); long dashed line, linear spatial extrapolation of the characteristic variables (4.6).

scheme. Eight equations are now required to determine the eight unknowns. Boundary conditions such as those given by (4.3) and (4.4) produce four equations. The discrete continuity (or momentum) equation (4.1a) yields another four, since it must be satisfied for each of the four types of waves. In most instances, the resultant matrix equation is nonsingular and ζ_l, μ_l can be found. However, if the matrix is singular, these coefficients cannot be determined. This would suggest that an assumption has been violated. Specifically, either not all the κ values are distinct or not all the eigenvalues of $A^{-1}B$ are inside the unit circle.

Figure 9 shows the short and long wave amplitudes at the left boundary for the four pairs of extraneous conditions when $f_1 = 0$, $f_2 = 1.0$, and $N = 10$. Only values for $\omega\,\Delta t \leqslant \omega_c$ are shown. For all boundary conditions and driving frequencies, the matrix equations were nonsingular. Although wave amplitudes are now dependent on N, results with $N = 10, 11, 50$ suggest that this parameter does not affect the relative performance of the four pairs of extraneous conditions.

Only the Box scheme conditions assigned nonzero amplitudes to the short leftward and long rightward waves. (In particular, the same amplitude was assigned to each.) For the other three extraneous conditions, the short wave amplitude shown in Fig. 9 refers to the rightward wave and the long wave amplitude refers to the leftward wave. Since the left boundary is radiating and the amplitude of the forced wave is 1.0, ideal boundary conditions should assign all this energy to the longer leftward wave. Clearly these conditions are not ideal. For small $\omega\,\Delta t$, short wave amplitudes are close to zero and the amplitudes of the long leftward wave are

close to 1. However, as $\omega\,\Delta t$ increases, so do the amplitudes of the short wave(s). For example, when $\omega\,\Delta t = 0.2425\pi$, amplitudes of the short and long waves arising from (4.4) are 0.396 and 1.004, respectively.

Notice that for all four pairs of boundary conditions, both amplitude patterns oscillate as $\omega\,\Delta t$ increases. Sharp peaks are an exaggeration of this phenomenon. They occur when $e^{i\omega\Delta t}$ is very close to an eigenvalue of the matrix $A^{-1}B$ (as defined in Section 2). For example, with the Box scheme conditions two eigenvalues of $A^{-1}B$ are $\lambda = 0.999972e^{i.559166}$ and $\lambda = 0.994212e^{i1.26839}$. Consequently, the forcing terms $\lambda = e^{i.559166}$ and $\lambda = e^{i1.26839}$ are almost eigenvalues. Were they eigenvalues, a separable steady state solution could not be assumed and the foregoing analysis could not be applied.

Though it is not shown in Fig. 9, all four numerical waves have the same length when $\omega\,\Delta t > \omega_c$. Phase and group velocities for each wave also have the same sign. At the right boundary, virtually all the leftward energy is assigned to the wave associated with $|\kappa| > 1$. This is further confirmation that the forced oscillation has a spatial decay of the type discussed by Vichnevetsky [32].

Figure 9 shows that constant space-time extrapolation of the characteristic variables is the most accurate pair of extraneous conditions. For most driving frequencies, its short wave amplitudes are closer to zero and its long leftward wave amplitudes are closer to one. This scheme also seems to exhibit less oscillation in the amplitude profiles.

The implementation of physical boundary conditions for problem P2 is also straightforward. A closed left boundary and a driving/radiating right boundary are respectively represented as

$$u_1^n = 0 \tag{4.12a}$$

and

$$u_N^n = \left(\frac{g}{h}\right)^{1/2} z_N^n + 2F(n). \tag{4.12b}$$

The driving/radiation condition is derived by assuming that the leftward and rightward wave components at the right boundary are related as

$$u_L = -\left(\frac{g}{h}\right)^{1/2} z_L^n = F(n) \tag{4.13a}$$

and

$$u_R^n = \left(\frac{g}{h}\right)^{1/2} z_R^n. \tag{4.13b}$$

Extraneous boundary conditions (4.4), (4.5), and (.46) were implemented for P2 as they were with P1, but the Box scheme application was reversed. This ensured

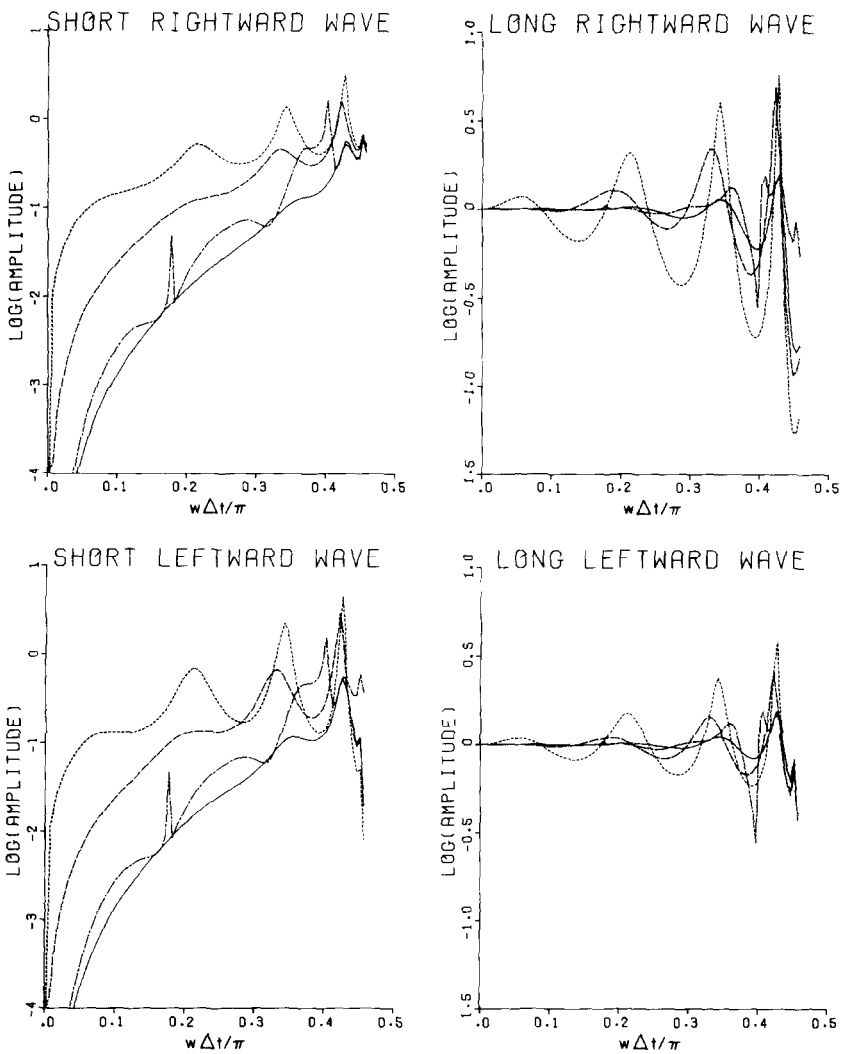

FIG. 10. Wave amplitudes (for u) for the FEM1 solution of problem P2. Parameter values and notation as in Fig. 9.

that the momentum equation remained coupled with the radiation condition, and the continuity equation remained with the boundary condition that is closed with respect to outward waves. Figure 10 shows the relative accuracy of the extraneous conditions for the same parameter values as in Fig. 9. Except for the Box scheme, amplitudes of the long leftward and short rightward waves are the same as those with P1. With the Box scheme, the two long waves have the same amplitude and the two short waves have the same amplitude. For the other three sets of

TABLE III

μ_l Coefficient Amplitudes for the FEM1 Solution of Problem P1

Artificial boundary conditions	Result	μ_1 (short rightward)	μ_2 (long rightward)	μ_3 (long leftward)	μ_4 (short leftward)
Constant spatial extrapolation of the characteristic variables	Analysis	0.00000000	0.00000000	1.00433997	0.39638618
	Model	0.00000117	0.00000015	1.00434043	0.39638522
Constant space-time extrapolation of the characteristic variables	Analysis	0.00000000	0.00000000	1.00468976	0.02382211
	Model	0.00000000	0.00000000	1.00468975	0.02382211
Box scheme applied to the continuity/momentum equations	Analysis	0.03369668	0.03369668	0.97651825	0.01704874
	Model	0.03345572	0.03344803	0.97623990	0.01749702
Linear spatial extrapolation of the characteristic variables	Analysis	0.00000000	0.00000000	0.89377854	0.13922122
	Model	0.00000000	0.00000000	0.89377853	0.13922122

TABLE IV

μ_l Coefficient Amplitudes for the FEM1 Solution of Problem P2

Artificial boundary conditions	Result	μ_1 (short rightward)	μ_2 (long rightward)	μ_3 (long leftward)	μ_4 (short leftward)
Constant spatial extrapolation of the characteristic variables	Analysis	0.34443534	0.87270999	1.00433997	0.39638618
	Model	0.34440992	0.87271939	1.00433959	0.39639822
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Box scheme applied to the continuity/momentum equations	Analysis	0.04859875	0.95326214	0.95326214	0.04859875
	Model	0.04855210	0.95320658	0.95329311	0.04854550
Linear spatial extrapolation of the characteristic variables	Analysis	0.13459286	0.86406518	0.89377854	0.13922122
	Model	0.13459287	0.86406516	0.89377853	0.13922122

extraneous conditions, the amplitude ratio of the long rightward wave to the long leftward wave is the same as the amplitude ratio of the short leftward wave to the short rightward wave. However, each ratio varies with $\omega \Delta t$.

Constant space-time extrapolation of the characteristic variables is again most accurate. Its short wave amplitudes are generally smaller and its long wave amplitudes are generally closer to the driving amplitude 1.0. Listed in terms of decreasing accuracy, the next best extraneous conditions are the Box scheme, linear spatial extrapolation of the characteristic variables, and constant spatial extrapolation of the characteristic variables.

The preceding analysis results were partially confirmed with numerical tests similar to those for the RS scheme. Both problems P1 and P2 were tested with all four pairs of extraneous boundary conditions and the parameter values $f_1 = 0.0$, $f_2 = 1.0$, and $N = 10$. All computations were done in double precision and again, only one driving frequency, $\omega \Delta t = 0.70685835$, was tested. Each test was run for 1000 time steps. Least squares analyses over the successive time steps ranges [801, 850], [851, 900], [901, 950], and [951, 1000] were used to calculate the ζ_l and μ_l coefficients of (4.11). In all cases, residuals decreased with each successive fit, and the fitted coefficients seemed to be converging. However, convergence was much slower than with the RS scheme because in several tests (with the Box scheme conditions) eigenvalues of $A^{-1}B$ which lay just inside the unit circle caused the transient solutions to decay more slowly.

Tables III and IV compare the numerical and analysis results. In all cases, the μ_l coefficient amplitudes obtained in the fourth fit were identical to at least 3 decimal places with those predicted by the analysis. Were the models run longer, agreement would be closer.

5. An Example of Trefethen's Instability

Most stability theory for finite difference models of hyperbolic initial boundary value problems is based on the classic yet complex paper by Gustafsson, Kreiss, and Sundström [3]. Their normal mode analysis for stability involves substitutions similar to those in the previous analyses, and checks for nontrivial solutions associated with eigenvalues whose magnitudes are not less than unity. Trefethen [2, 10] has recently shown that the GKS perturbation test for unstable "generalized eigensolutions" has a physical interpretation in terms of group velocity. In particular, he shows that GKS instability amounts to spontaneous radiation of energy from the boundary into the problem domain. His main result is a necessary condition for stability which involves checking the signs of the group velocities corresponding to eigenvalue solutions with modulus unity.

FEM1 can have an instability of this type. Reexpressing (1.1) in terms of characteristic variables and assuming $\tau = 0$, the FEM1 equations for the leftward characteristic variable become

$$\tfrac{1}{6}[(w_{j-1}^{n+1} - w_{j-1}^{n}) + 4(w_j^{n+1} - w_j^n) + (w_{j+1}^{n+1} - w_{j+1}^n)]$$
$$= \tfrac{1}{4}(gh)^{1/2}(\Delta t/\Delta x)[w_{j+1}^n - w_{j-1}^n + w_{j+1}^{n+1} - w_{j-1}^{n+1}]. \tag{5.1}$$

Assuming the separable solution

$$w_j^n = \alpha_0 \lambda^n \kappa^j, \tag{5.2}$$

the characteristic equation for (5.1) is

$$(\lambda - 1)(1 + 4\kappa + \kappa^2) - \tfrac{3}{2} f_2(\lambda + 1)(\kappa^2 - 1) = 0. \tag{5.3}$$

For each λ, there are two values of κ, namely, κ_1 and κ_2. The general numerical solution therefore has the form

$$w_j^n = \lambda^n(\alpha_1 \kappa_1^j + \alpha_2 \kappa_2^j). \tag{5.4}$$

Consider the pair of boundary conditions

$$w_N^n + w_{N-1}^n = 0 \tag{5.5a}$$

$$w_1^n = w_3^n. \tag{5.5b}$$

The former condition is consistent with the well-posed [27] analytical condition $w = 0$, while the latter is a form of constant spatial extrapolation. Assuming the general solution (5.4), (5.5a) implies

$$\alpha_1 \kappa_1^{N-1}(1 + \kappa_1) + \alpha_2 \kappa_2^{N-1}(1 + \kappa_2) = 0. \tag{5.6a}$$

while (5.5b) implies

$$\alpha_1 \kappa_1(\kappa_1^2 - 1) + \alpha_2 \kappa_2(\kappa_2^2 - 1) = 0. \tag{5.6b}$$

Consider the particular solution $\lambda = 1$, $\kappa_1 = 1$, $\kappa_2 = -1$. Then (5.6b) is satisfied for all α_1 and α_2, whereas (5.6a) requires $\alpha_1 = 0$. The $2\Delta x$ wave

$$w_j^n = \alpha_2(-1)^j \tag{5.7}$$

is therefore a nontrivial solution to (5.1) and (5.5). If its group velocity is positive, Trefethen's interpretation of GKS theory would imply an instability at the left boundary.

The group velocity is found by assuming the travelling wave solution

$$w_j^n = \alpha_0 e^{i(jk\Delta x + n\omega \Delta t)} \tag{5.8}$$

for (5.1). The resultant dispersion relationship is

$$\tan\left(\frac{1}{2}\omega \Delta t\right) = \frac{f_2}{2}\left(\frac{3 \sin k \Delta x}{2 + \cos k \Delta x}\right), \tag{5.9}$$

and group velocities are calculated using (3.14b). The group velocity for $2\Delta x$ waves (i.e., $k\,\Delta x = \pm\pi$) is $3(gh)^{1/2}$. Consequently, at the left boundary, energy from a $2\Delta x$ wave radiates into the problem domain. Since this energy does not arise from the reflection of a wave with negative group velocity (rather it radiates spontaneously), Trefethen's theory predicts instability.

With $N = 10$, 20, 40, 80 and $f_2 = 1.0$, it can be shown numerically that no other eigensolutions or generalized eigensolutions (nontrivial solutions with $|\lambda| \geq 1$) are supported by boundary conditions (5.5). Instability therefore arises solely from the generalized eigensolution (5.7). This (mild) instability was confirmed with a test model which assumed random initial conditions and no forcing. As in Gustafsson [25], the accumulation of rounding errors was accelerated by adding a small random number to each w_j at each time step.

6. Summary and Conclusions

The preceding analysis has demonstrated a valid approach for evaluating the relative accuracy of numerical boundary conditions. Although only two numerical methods for solving the one-dimensional shallow water equations were examined, the concepts are sufficiently general that they could be applied to other methods and other one-dimensional forced hyperbolic equations. In fact, the analysis should also be extendable to two-dimensional forced equations where angles of incidence to the boundary will be another factor affecting accuracy [13].

Although the preceding analyses were more illustrative than comprehensive, some conclusions can be drawn from the results. With the RS scheme, boundary condition accuracy was seen to be independent of N and the boundary condition at the other end of the channel. When $\tau = 0$, both radiating and driving/radiating boundaries became more accurate as higher orders of space-time extrapolation were applied to the outgoing characteristic variables. Accuracy was highest with long waves and deteriorated as the wavenumber increased. However, the same was not true when a $\tau > 0$ meant that z and u were no longer in phase. Boundary conditions ceased to be most accurate for long waves and higher orders of space-time extrapolation did not necessarily produce greater accuracy. This may not be true for variations of (1.3) that include friction (e.g., [33]), or numerical implementations based on other mathematical expressions of the radiation condition.

The FEM1 analysis illustrated some of the short wave problems that can arise with a Galerkin finite element method which uses piecewise linear basis functions. Similar difficulties can be expected for any numerical scheme whose dispersion curve is not monotonic. Short waves were seen to be generated not only with forced and closed boundaries, but also with radiating boundaries. Short wave contamination was also seen to vary with the choice of extraneous boundary conditions.

The FEM1 analysis indicated that extraneous boundary conditions which are constructed using constant space-time extrapolation of the outgoing characteristic

variables are more accurate than those constructed from the Box scheme, or from constant or linear spatial extrapolation of the characteristic variables. All four conditions seemed to perform equally well (or poorly) in representing the boundary physics and minimizing the generation of short waves. And although boundary condition accuracy did depend on N, the overall relative accuracy of the four conditions did not seem to be affected by changes in this parameter.

ACKNOWLEDGMENTS

I thank Professor J. M. Varah, Dr. R. F. Henry, and Dr. A. F. Bennett for helpful comments and discussions; Dr. L. N. Trefethen and Dr. R. L. Higdon for sending preprints of their work; and the referees for their constructive criticism of an earlier version of this paper.

REFERENCES

1. C. K. CHU AND A. SERENY, *J. Comput. Phys.* **15**, 476 (1974).
2. L. N. TREFETHEN, *J. Comput. Phys.* **49**, 199 (1983).
3. B. GUSTAFSSON, H. KREISS, AND A. SUNDSTRÖM, *Math. Comput.* **26**, 649 (1972).
4. B. GUSTAFSSON, *Math. Comput.* **29**, 396 (1975).
5. G. SKÖLLERMO, Department of Computer Sciences Report No. 62, Uppsala University, 1975 (unpublished).
6. G. SKÖLLERMO, *Math. Comput.* **33**, 11 (1979).
7. D. M. SLOAN, *Int. J. Numer. Methods Engrg.* **15**, 1113 (1980).
8. D. GOTTLIEB AND E. TURKEL, *J. Comput. Phys.* **26**, 181 (1978).
9. L. N. TREFETHEN, in *Proceedings of the AMS–SIAM 1983 Summer Seminar on Large-Scale Computations in Fluid Mechanics*, edited by S. Osher (to appear).
10. L. N. TREFETHEN, *Commun. Pure Appl. Math.* **37**, 329 (1984).
11. L. N. TREFETHEN, NASA Contractor Report 172319, ICASE Report No. 84-11, 1984 (unpublished). (Also to appear in *Math. Comput.*)
12. L. HALPERN, *Math. Comput.* **38**, 415 (1982).
13. R. L. HIGDON, "Absorbing Boundary Conditions for Difference Approximations to the Multi-dimensional Wave Equation," to appear.
14. D. H. RUDY AND J. C. STRIKWERDA, *Comput. & Fluids* **9**, 327 (1981).
15. A. F. BENNETT, *J. Atmos. Sci.* **33**, 176 (1976).
16. B. ENGQUIST AND A. MAJDA, *Math. Comput.* **31**, 629 (1977).
17. L. N. TREFETHEN AND L. HALPERN, "New Families of One-Way Wave Equations and Absorbing Boundary Conditions," to appear.
18. R. F. HENRY AND N. S. HEAPS, *J. Fish Res. Board Can.* **33**, 2362 (1976).
19. R. A. FLATHER, *Mem. Soc. R. Sci. Liège, Ser. 6* **10**, 141 (1976).
20. I. ORLANSKI, *J. Comput. Phys.* **21**, 251 (1976).
21. B. GUSTAFSSON AND H. KREISS, *J. Comput. Phys.* **30**, 333 (1979).
22. [illegible]
23. R. A. WALTERS AND G. F. CAREY, *Comput. & Fluids* **11**, 51 (1983).
24. R. A. WALTERS, *Int. J. Numer. Methods Fluids* **3**, 591 (1983).
25. B. GUSTAFSSON, *J. Comput. Phys.* **48**, 270 (1982).
26. H. C. YEE, R. M. BEAM, AND R. F. WARMING, "Stable Boundary Approximations for a Class of Implicit Schemes for the One-Dimensional Inviscid Equations of Gas Dynamics," AIAA Computational Fluid Dynamics Conference, Palo Alto, CA, 1981.

27. H.-O. KREISS, *Commun. Pure Appl. Math.* **23**, 277 (1970).
28. M. G. G. FOREMAN, *J. Comput. Phys.* **51**, 454 (1983).
29. N. J. PULLMAN, *Matrix Theory and Its Applications, Selected Topics* (Dekker, New York, 1976).
30. R. M. BEAM, R. F. WARMING, AND H. C. YEE, *J. Comput. Phys.* **48**, 200 (1982).
31. R. F. HENRY, *J. Comput. Phys.* **41**, 389 (1981).
32. R. VICHNEVETSKY, *Math. Comput. Simulation* **22**, 98 (1980).
33. G. K. VERBOOM AND A. SLOB, in *Proceedings of the Fifth International Conference on Finite Elements in Water Resources*, edited by J. P. Laible *et al.* (Springer-Verlag, Berlin, 1984).
34. A. L. CAMERLENGO AND J. J. O'BRIEN, *J. Comput. Phys.* **35**, 12 (1980).
35. M. D. GUNZBURGER, *Math. Comput.* **31**, 661 (1977).
36. D. GOTTLIEB, M. GUNZBURGER, AND E. TURKEL, *SIAM J. Numer. Anal.* **19**, 671 (1982).