JOURNAL OF COMPUTATIONAL PHYSICS 52, 290–312 (1983)

An Analysis of the "Wave Equation" Model for Finite Element Tidal Computations

M. G. G. FOREMAN

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

Received July 14, 1982; revised March 18, 1983

The one-dimensional linearized version of the "wave equation" model developed by Gray and Lynch for solving the shallow water equations is analysed. Similarities with a mixed interpolation finite element method are discovered, and the proposed time-stepping methods are shown to be a subset of a much larger class. Using dispersion and asymptotic analyses, particular time-stepping methods which most accurately represent wave propagation and wave amplitude growth are determined for both the lumped and unlumped approaches.

INTRODUCTION

Recently Gray and Lynch [4, 5] introduced a finite element method for solving the shallow water equations. Rather than working with the governing equations in conservation form, their "wave equation" scheme involves transforming the continuity equation to a second-order partial differential equation. The revised system of equations is then solved with a Galerkin finite element method, piecewise-linear basis functions, and centred time stepping. Through propagation factor analyses they show that the resultant numerical method is more accurate than several alternatives, and avoids the troublesome accumulation of $2\Delta x$ waves which often occurs with finite element schemes. Their numerical tests confirm these results.

In this discussion, the one-dimensional linearized version of the "wave equation" method (WEM), and the lumped "wave equation" method (LWEM) are studied further. Accuracy is examined through a dispersion or Fourier (phase/space) analysis which includes group velocity. This approach is developed and illustrated in [2]. The importance of group velocity in numerical methods is surveyed by Trefethen [7].

This paper is divided into seven sections. Section 1 shows that the WEM conditionally incorporates the same spatial discretization attributes as the mixed interpolation approach of Williams and Zienkiewicz [9]. Section 2 specifies its characteristic roots and stability restrictions. Section 3 generalizes the time-stepping methods proposed by Lynch and Gray by showing them to be a subset of the second-order linear two-step methods which can be developed for their governing equations. Section 4 applies these generalized two-step methods to both the lumped and unlumped spatial discretizations, and determines the particular methods with the most

0021-9991/83 $3.00

accurate wave propagation, and wave amplitude growth (or decay) characteristics. These results are confirmed with asymptotic analyses in Section 5, and numerical tests in Section 6. Section 7 summarizes and briefly discusses the results.

1. An Analysis of the Spatial Discretization

The linearized, one-dimensional, constant depth versions of the governing equations solved by Lynch and Gray [5] are

$$\frac{\partial^2 z}{\partial t^2} + \tau \frac{\partial z}{\partial t} - gh \frac{\partial^2 z}{\partial x^2} = 0, \tag{1a}$$

$$\frac{\partial u}{\partial t} + \tau u + g \frac{\partial z}{\partial x} = 0, \tag{1b}$$

where $z(x, t)$ is the elevation above mean sea level, $u(x, t)$ the velocity, $h(x)$ the mean sea depth, g the gravity, and τ the linear bottom friction coefficient. After employing a Galerkin finite element method with piecewise-linear basis functions, the unlumped spatially discretized system of ordinary differential equations (ODEs) becomes

$$\left(\frac{\partial^2}{\partial t^2} + \tau \frac{\partial}{\partial t}\right)\left(\frac{1}{6} z_{j-1} + \frac{2}{3} z_j + \frac{1}{6} z_{j+1}\right) - \frac{gh}{\Delta x^2}(z_{j+1} - 2z_j + z_{j-1}) = 0, \tag{2a}$$

$$\left(\frac{\partial}{\partial t} + \tau\right)\left(\frac{1}{6} u_{j-1} + \frac{2}{3} u_j + \frac{1}{6} u_{j+1}\right) + \frac{g}{2\,\Delta x}(z_{j+1} - z_{j-1}) = 0, \tag{2b}$$

where depth and Δx are assumed constant, and z_j, u_j are the time-dependent variable values at node j. Assuming travelling wave solutions of the form

$$\begin{pmatrix} z_j(t) \\ u_j(t) \end{pmatrix} = \begin{pmatrix} z_0 \\ u_0 \end{pmatrix} \exp i(kj\,\Delta x + \omega t), \tag{3}$$

where k is wave number and ω is frequency, the dispersion relationships arising from (2) are

$$\omega = i\tau \tag{4a}$$

and

$$\omega = i\,\tfrac{1}{2}\tau \pm \left[\frac{6gh}{\Delta x^2}\left(\frac{1 - \cos(k\,\Delta x)}{2 + \cos(k\,\Delta x)}\right) - (\tfrac{1}{2}\tau)^2\right]^{1/2}. \tag{4b}$$

The latter expression is identical to the dispersion relationship for the mixed inter-

polation approach recently described by Williams and Zienkiewicz [9]. They solve the momentum equation (1b) and the customary continuity equation

$$\frac{\partial z}{\partial t} + \frac{\partial (hu)}{\partial x} = 0 \tag{5}$$

with a Galerkin finite element method, piecewise-linear basis functions for approximating $u(x, t)$, and piecewise constant for $z(x, t)$.

Equivalence of these two approaches is revealing. The principal dispersion relationship has not changed when the order of the continuity equation is increased from 1 to 2, and the order of the approximating basis function for z is increased from 0 to 1. This suggests that similar relationships may exist between other finite element approaches. However, equivalence of the principal dispersion relationships does not imply that the numerical solutions will be identical. A secondary or parasitic relationship is present with the WEM and in some circumstances, it will affect the numerical results.

Relative accuracy of the phase and group velocity associated with (4b) is shown in [2] for the case $\tau = 0$. For the one-dimensional equations, it produces one of the better approximations to the analytic solution. Unfortunately, efforts to extend the mixed interpolation formulation to triangular elements in two dimensions have proven difficult because of discontinuities in $z(x, t)$ at the interelement nodes [8]. However, since the WEM has been extended to two dimensions, in some sense it may be regarded as equivalent to a successful mixed interpolation extension.

2. Numerical Eigenvalues of the "Wave Equation" Method

Lynch and Gray solved their system of ODEs with the following time stepping approximations which are centred for all values of the parameter θ:

$$\frac{\partial}{\partial t} z_j(n\,\Delta t) \simeq \frac{z_j^{n+1} - z_j^{n-1}}{2\,\Delta t}, \tag{6a}$$

$$\frac{\partial^2}{\partial t^2} z_j(n\,\Delta t) \simeq \frac{z_j^{n+1} - 2z_j^n + z_j^{n-1}}{\Delta t^2}, \tag{6b}$$

$$z_j(n\,\Delta t) \simeq \tfrac{1}{2}\theta(z_j^{n+1} + z_j^{n-1}) + (1-\theta)\, z_j^n. \tag{6c}$$

In order to relax the stability constraints, the friction term in the momentum equation is treated as in (6c), but with the separate weighting parameter α.

Applying these approximations to (2) and assuming nontrivial travelling wave solutions requires satisfaction of one of the following two quadratics:

$$\lambda^2(1 + \alpha\tau\,\Delta t) + 2\tau\,\Delta t(1-\alpha)\,\lambda - (1 - \alpha\tau\,\Delta t) = 0, \tag{7a}$$

$$\lambda^2[1 + \tfrac{1}{2}(\theta E^2 + \tau\,\Delta t)] + \lambda[-2 + (1-\theta)\,E^2] + 1 + \tfrac{1}{2}(\theta E^2 - \tau\,\Delta t) = 0, \tag{7b}$$

where

$$E^2 = 6gh \left(\frac{\Delta t}{\Delta x}\right)^2 \left(\frac{1 - \cos(k\,\Delta x)}{2 + \cos(k\,\Delta x)}\right) \tag{7c}$$

and

$$\lambda = \exp(i\omega \Delta t). \tag{7d}$$

Constant depth, constant Δx and Δt are also assumed. The product of these quadratics is the characteristic polynomial. The roots are also eigenvalues of the amplification matrix resulting from a linear stability analysis of the numerical method [6]. Consequently, they will be referred to as either roots, or eigenvalues, throughout this paper.

The eigenvalues for (7) are

$$\lambda_{1,2} = \frac{-\tau\,\Delta t(1-\alpha) \pm [1 + (\tau\,\Delta t)^2(1-2\alpha)]^{1/2}}{1 + \alpha\tau\,\Delta t} \tag{8a}$$

and

$$\lambda_{3,4} = \frac{1 - \frac{1}{2}(1-\theta)\,E^2 \pm i[E^2 + \frac{1}{4}(2\theta - 1)\,E^4 - (\frac{1}{2}\tau\,\Delta t)^2]^{1/2}}{1 + \frac{1}{2}(\theta E^2 + \tau\,\Delta t)}. \tag{8b}$$

λ_1 and λ_2 are parasitic and arise from the spatially discretized solution (4a). They are independent of wave number and thus have zero group velocity. If they are real valued and positive, they also have zero phase velocity. λ_3 and λ_4 are the principal numerical eigenvalues. When their imaginary parts are nonzero, they represent progressive and retrogressive waves.

A necessary condition for the stability of any numerical method is

$$|\lambda| \leqslant 1.0 + O(\Delta t) \tag{9}$$

for all eigenvalues, and all $k\,\Delta x$ in the range of $(0, \pi]$. This is the von Neumann condition. When the exact solution does not grow exponentially, the $O(\Delta t)$ term is usually omitted [6]. Since $h(x)$ is assumed constant and $\tau \geqslant 0$, this is the case here.

For the WEM eigenvalues in (8), stability is ensured with the following restrictions on θ and α:

(i) for the parasitic roots,

$$\alpha \geqslant \tfrac{1}{2}; \tag{10a}$$

(ii) for the propagating principal roots,

$$\theta \geqslant \frac{-\Delta x^2}{6gh\,\Delta t^2}; \tag{10b}$$

(iii) for the nonpropagating principal roots,

$$\theta \geqslant \frac{1}{2}\left(1 - \frac{\Delta x^2}{3gh\,\Delta t^2}\right). \tag{10c}$$

With nonzero friction, (10b) can be made less restrictive. However, this is not essential since the constraint imposed by (10c) dominates. Conditions (i) and (iii) are given in [5].

An analysis of the LWEM yields similar results. With the following substitution for (7c)

$$E^2 = 2gh(\Delta t/\Delta x)^2(1 - \cos(k\,\Delta x)), \tag{11}$$

Eqs. (7) and (8) remain valid. Stability of the parasitic root is again dictated by (10a) while the counterparts to (10b) and (10c) are

$$\theta \geqslant \frac{-\Delta x^2}{2gh\,\Delta t^2} \tag{12a}$$

and

$$\theta \geqslant \frac{1}{2}\left(1 - \frac{\Delta x^2}{gh\,\Delta t^2}\right), \tag{12b}$$

respectively. As with the WEM, condition (12b) overrides (12a).

3. Two-Step Methods for Solving the ODEs in Time

A simple ODE corresponding to (2a) is

$$\frac{\partial^2 y}{\partial t^2} + \frac{\partial y}{\partial t} = f(y) \tag{13}$$

and a general two-step method which may be used to solve it has the form

$$\begin{aligned} &c_2 y^{n+2} + c_1 y^{n+1} + c_0 y^n + \Delta t(a_2 y^{n+2} + a_1 y^{n+1} + a_0 y^n) \\ &\quad = \Delta t^2(b_2 f^{n+2} + b_1 f^{n+1} + b_0 f^n). \end{aligned} \tag{14}$$

Requiring second-order accuracy (i.e., a truncation error of $O(\Delta t^3)$) and assuming Gear's normalization [3],

$$b_0 + b_1 + b_2 = 1 \tag{15}$$

leaves only three coefficients to be specified freely. Choosing them to be $a_2, b_2,$ and b_1, the others are

$$\begin{aligned} &c_2 = c_0 = 1, \qquad a_0 = a_2 - 1, \qquad b_0 = 1 - b_1 - b_2, \\ &c_1 = -2, \qquad\quad a_1 = 1 - 2a_2. \end{aligned} \tag{16}$$

Third-order methods have the additional constraints

$$a_2 = \tfrac{1}{2}, \quad b_0 = b_2, \tag{17}$$

while fourth-order accuracy is not possible with this set of two-step methods.

For solving the simple ODE,

$$\frac{\partial y}{\partial t} = f(y) + g(y) \tag{18}$$

with the two-step method

$$\begin{aligned} &a_2\, y^{n+2} + a_1\, y^{n+1} + a_0\, y^n \\ &\quad = \Delta t(b_2 f^{n+2} + b_1 f^{n+1} + b_0 f^n + d_2 g^{n+2} + d_1 g^{n+1} + d_0 g^n) \end{aligned} \tag{19}$$

similar calculations lead to the following constraints for second-order accuracy:

$$\begin{aligned} &a_0 = a_2 - 1, \qquad b_0 = \tfrac{1}{2} - a_2 + b_2, \qquad d_0 = \tfrac{1}{2} - a_2 + d_2, \\ &a_1 = 1 - 2a_2, \qquad b_1 = \tfrac{1}{2} + a_2 - 2b_2, \qquad d_1 = \tfrac{1}{2} + a_2 - 2d_2. \end{aligned} \tag{20}$$

In this case third-order methods also require

$$b_2 = d_2 = \tfrac{1}{2}a_2 - \tfrac{1}{12} \tag{21}$$

and fourth-order accuracy occurs with $(a_2, b_2, d_2) = (\frac{1}{2}, \frac{1}{6}, \frac{1}{6})$.

The ODEs in (2) can be solved with the preceding two-step methods. Simultaneously requiring at least second-order accuracy for both equations, and insisting on a consistent approximation for the first derivative (i.e., $a_2, a_1, a_0, b_2, b_1, b_0$ are the same for each method) leads to the following combined restrictions:

$$\begin{aligned} &c_2 = c_0 = 1, \qquad a_0 = a_2 - 1, \qquad b_0 = \tfrac{1}{2} - a_2 + b_2, \qquad d_0 = \tfrac{1}{2} - a_2 + d_2, \\ &c_1 = -2, \qquad\quad a_1 = 1 - 2a_2, \qquad b_1 = \tfrac{1}{2} + a_2 - 2b_2, \qquad d_1 = \tfrac{1}{2} + a_2 - 2d_2. \end{aligned} \tag{22}$$

Notice that the particular case

$$a_2 = \tfrac{1}{2}, \qquad b_2 = \tfrac{1}{2}\theta, \qquad d_2 = \tfrac{1}{2}\alpha, \tag{23}$$

makes (14) third-order accurate and is precisely the subset of time-stepping methods

proposed by Lynch and Gray. Fourth-order accuracy for (19) and third-order accuracy for (14) is obtained with the additional constraint $b_2 = d_2 = \frac{1}{6}$. The highest-order time-stepping method for both the WEM and the LWEM therefore occurs with $\theta = \alpha = \frac{1}{3}$. However, (10a) indicates that it will be unstable.

4. A Dispersion Analysis

In [2] it was seen that the highest-order time-stepping method may not be the one which produces the most accurate phase velocity, group velocity, or wave amplitude. In order to determine which time-stepping method is the most accurate, a dispersion analysis is now performed for the two-parameter class of second-order two-step methods given by (22). The methods proposed by Lynch and Gray are a subset of this class.

Define

$$\tilde{s}_j = \tfrac{1}{6}(s_{j-1} + 4s_j + s_{j+1}), \tag{24a}$$

$$\hat{s}_j = s_{j+1} - 2s_j + s_{j-1}, \tag{24b}$$

$$\Delta s_j = s_{j+1} - s_{j-1}, \tag{24c}$$

where s can be either z or u. Application of a second-order two-step method to solve (2) then produces the system of equations

$$\begin{aligned} &\tilde{z}_j^{n+2} - 2\tilde{z}_j^{n+1} + \tilde{z}_j^n + \tau\,\Delta t(a_2\tilde{z}_j^{n+2} + a_1\tilde{z}_j^{n+1} + a_0\tilde{z}_j^n) \\ &\quad = gh\left(\frac{\Delta t}{\Delta x}\right)^2 (b_2\hat{z}_j^{n+2} + b_1\hat{z}_j^{n+1} + b_0\hat{z}_j^n), \end{aligned} \tag{25a}$$

$$\begin{aligned} a_2\tilde{u}_j^{n+2} + a_1\tilde{u}_j^{n+1} + a_0\tilde{u}_j^n = -\Delta t\Bigg[&\tau(d_2\tilde{u}_j^{n+2} + d_1\tilde{u}_j^{n+1} + d_0\tilde{u}_j^n) \\ &+ \frac{g}{2\,\Delta x}(b_2\,\Delta z_j^{n+2} + b_1\,\Delta z_j^{n+1} + b_0\,\Delta z_j^n)\Bigg], \end{aligned} \tag{25b}$$

where the restrictions imposed by (22) are assumed but have not been included.

For nontrivial travelling wave solutions to (25) one of the following two quadratics must be satisfied:

$$a_2\lambda^2 + a_1\lambda + a_0 + \tau\,\Delta t(d_2\lambda^2 + d_1\lambda + d_0) = 0 \tag{26a}$$

or

$$\lambda^2 - 2\lambda + 1 + \tau\,\Delta t(a_2\lambda^2 + a_1\lambda + a_0) + E^2(b_2\lambda^2 + b_1\lambda + b_0) = 0. \tag{26b}$$

For the WEM, E^2 and λ are defined by (7c) and (7d), respectively. For the LWEM, E^2 is defined by (11).

The root $\lambda = 1$ can be troublesome for it represents an undamped nonpropagating wave. If energy is transferred to a wave number which has this eigenvalue as a solution, it will simply accumulate. Consequently, accuracy of the numerical solution can be severely affected. The root $\lambda = -1$ is equally undesirable for the associated waves are also undamped and flip sign from one time step to the next. Energy can accumulate here as well. In fact, any short waves with real roots of magnitude slightly less than 1.0 may be equally troublesome. Provided their magnitudes are larger than those of the desired longer waves, these short waves will decay more slowly (or grow more rapidly) and eventually dominate the calculations.

The occurrence of $\lambda = \pm 1$ for $2\Delta x$ waves (i.e., when $k\Delta x = \pi$) is a common problem with finite element schemes (e.g., [8]). However, it can be avoided in this case. From (22) and (26), it is seen that $2\Delta x$ waves have the solution $\lambda = 1$ only when $\tau = 0$ and the parasitic eigenvalue is dominant. And for specified values of τ and E^2, $\lambda = -1$ is a $2\Delta x$ solution to (26a) or (26b) only for certain values of (a_2, b_2, d_2). Therefore with nonzero friction and a judicious choice of these parameters, the generalized "wave equation" method given by (25) can avoid the troublesome accumulation of $2\Delta x$ waves.

Dispersion relationships and phase and group velocities can be calculated from (26). Three associated functions can then be defined to measure accuracy of the numerical solution; one for each wave amplitude growth, phase velocity, and group velocity. The reader is referred to [2] for these definitions and general background to the subsequent discussion. Velocity accuracy measures are simply relative errors between the dominant numerical wave and the analytic wave. (A parasitic wave is dominant when the parasitic eigenvalue is larger than the principal eigenvalue.) Negative values denote waves travelling too slowly while zero values are optimal. For example, -0.01 denotes a numerical velocity which is 1% too slow. Amplitude measures are ratios denoting the growth (or decay) factor per time step of the dominant numerical wave relative to the analytic wave. Values greater than the optimum of 1. signify a solution which will decay too slowly or grow too rapidly.

Figure 1 illustrates the dispersion curves, eigenvalue amplitudes, and phase and group velocities for four two-step methods. Values for the analytic and spatially discretized solutions are also included. Results are parameterized in terms of

$$f_1 = \frac{\tau \Delta x}{(gh)^{1/2}} \tag{27a}$$

and

$$f_2 = (gh)^{1/2} \frac{\Delta t}{\Delta x}. \tag{27b}$$

The latter parameter is commonly referred to as the Courant number.

All four methods are from the subset proposed by Lynch and Gray (i.e., they satisfy (22) and (23)). The first three methods have not been lumped while the fourth

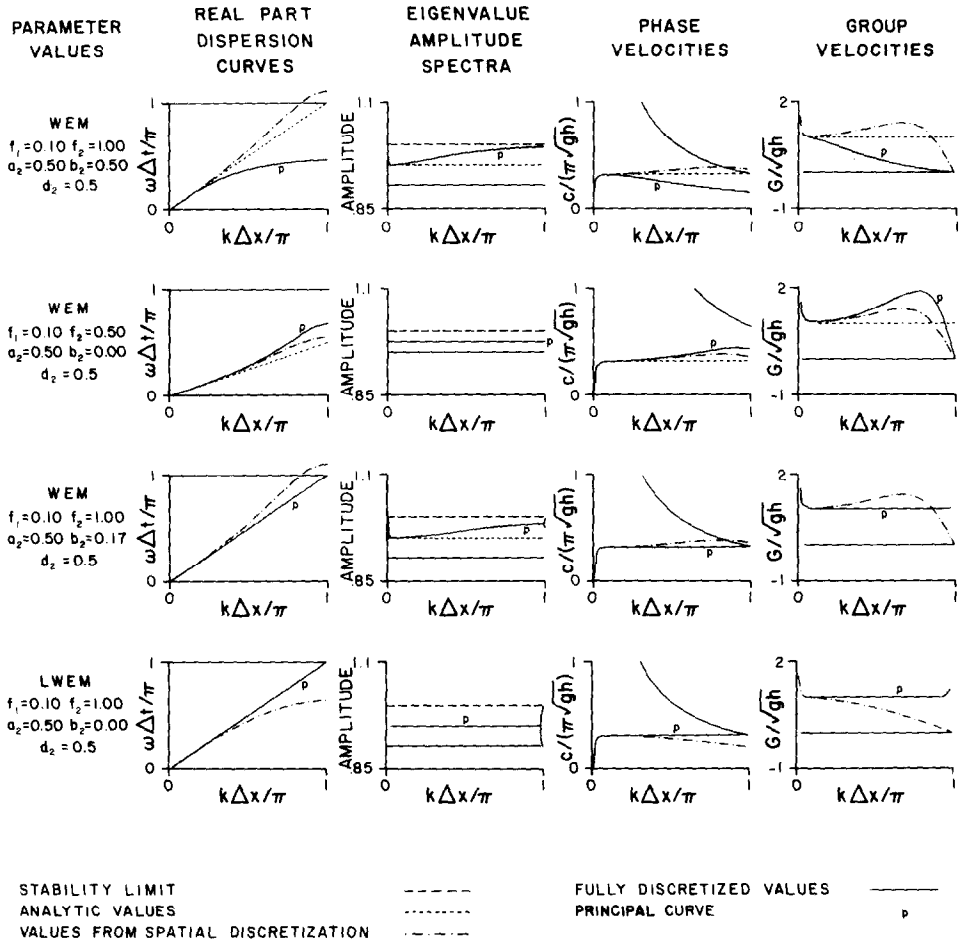

FIG. 1. Dispersion curves, eigenvalue amplitudes, and phase and group velocities for four "wave equation" models.

has. Stability of the parasitic eigenvalue is ensured for all four methods by choosing $d_2 = \frac{1}{2}\alpha = \frac{1}{2}$. The first and second methods are the fully implicit and fully explicit cases introduced in [4]. The third method produces third-order accuracy for (14) and second-order accuracy for (19). Its principal dispersion curve and phase and group velocities are seen to approximate the analytic values closely. In fact, wave propagation inaccuracies which were introduced by the spatial discretization have been effectively cancelled by the time-stepping method. If for this method d_2 were also equal to $\frac{1}{6}$, (19) would become fourth-order accurate but unstable. Specifically, the amplitude of the parasitic eigenvalue would now exceed 1.0 for all $k\,\Delta x$.

The fourth method is the explicit LWEM. It should be more economical in both storage requirements and computation time than the other three methods.

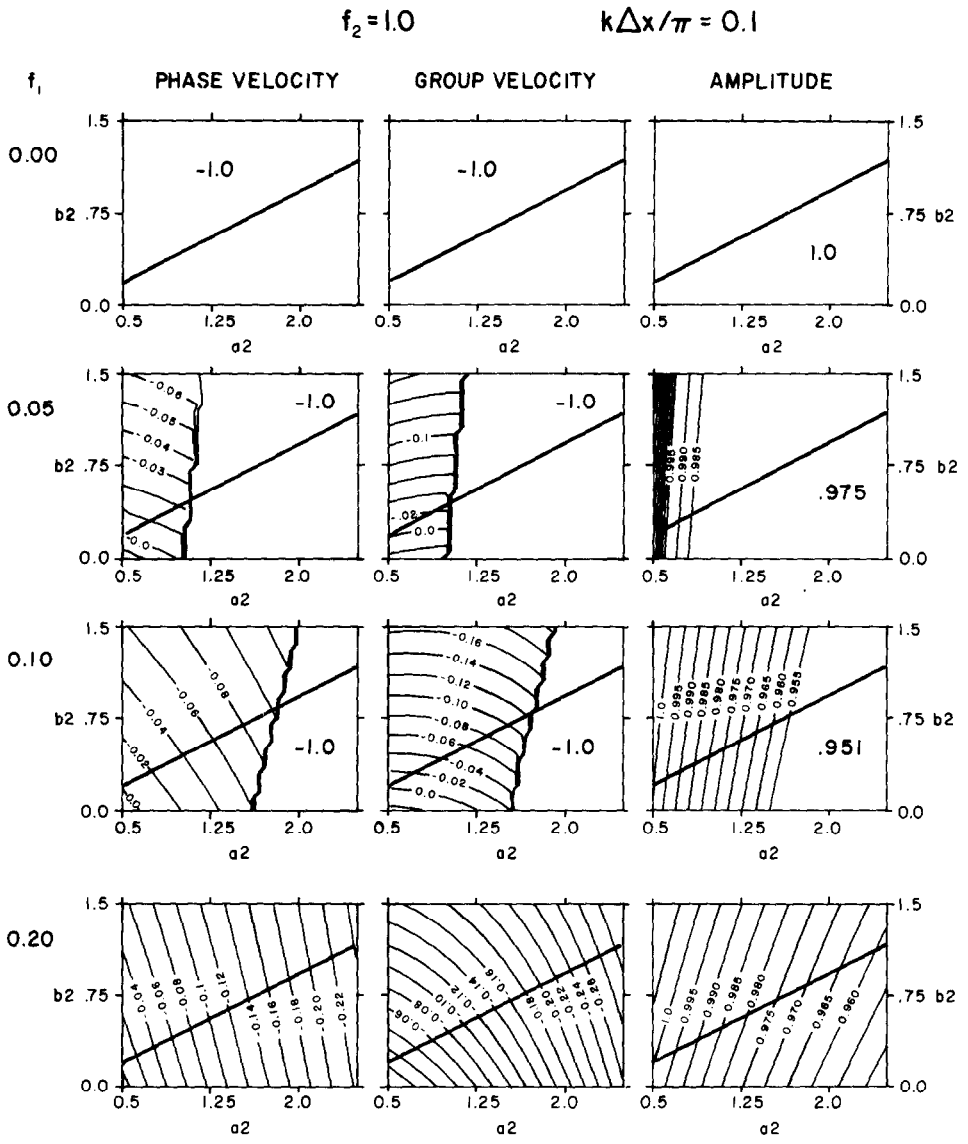

FIG. 2. Accuracy measure values for the WEM as functions of (a_2, b_2) for $f_2 = 1.0$, $d_2 = 0.5$, and $k\,\Delta x/\pi = 0.1$.

Surprisingly, this economy does not correspond to a loss in accuracy. Its wave propagation characteristics are seen to be as accurate as those of the third example, while its eigenvalue amplitude is more accurate.

Figure 2 shows accuracy measure values for the WEM as functions of the two-step parameters a_2 and b_2. d_2, f_2, and $k\,\Delta x/\pi$ are fixed at 0.5, 1.0, and 0.1 respectively,

while f_1 assumes four increasing values. In all instances, the stability region is bounded from below by the heavy solid line and to the left by $a_2 = \frac{1}{2}$. Large regions which have not been contoured have constant accuracy measure values that are due to a dominant parasitic eigenvalue. In particular, the roots of (26a) are real valued thereby making both the phase and group velocity zero and their corresponding accuracy measures -1. For $f_1 = 0.0$, the parasitic eigenvalue $\lambda = 1$ is dominant everywhere except along the line $a = \frac{1}{2}$. For larger f_1, larger values of a_2 are required before the parasitic roots dominate.

A similar plot with $f_2 = 0.5$ exhibits many of the same features. In general, the stability region becomes less restrictive (i.e., the lower stability boundary drops) and the parasitic eigenvalue becomes dominant for slightly larger values of a_2. The most notable characteristic of both plots is that all lines of optimal accuracy either lie very close to, or cross the line $a_2 = \frac{1}{2}$. (Optimal values for the velocity and amplitude measures are 0.0 and 1.0, respectively.) This phenomenon seems to be independent of the value $k\,\Delta x/\pi = 0.1$ for it also occurs with $f_1 = 0.1$ and the $k\,\Delta x/\pi$ values 0.05, 0.2 and 0.4.

In light of the previous two-step method development, greater accuracy with $a_2 = \frac{1}{2}$ is not surprising. It substantiates the desirability of third order accuracy for (14). It also suggests that one can restrict the search for an optimal method to the subset originally proposed by Lynch and Gray. In subsequent discussions it is therefore assumed that the parameters b_2 and θ, and d_2 and α are related through (23).

Figure 3 shows a series of revised accuracy measure contours for the WEM and the case $f_2 = 1.0$ and $a_2 = d_2 = \frac{1}{2}$. Fixing a_2 permits its replacement along the horizontal axis with $k\,\Delta x$. The accuracy measures are revised in the sense that they are calculated only from the principal eigenvalue. The small regions where the parasitic eigenvalue dominates have been shaded, but the accuracy measure values do not reflect this dominance. A concentration of contour lines near $k\,\Delta x = 0$ has not been shown because nonzero friction does not permit a wave solution there. The associated accuracy measure values are therefore meaningless. Constant uncontoured values in the lower right corner of the plot arise because the principal eigenvalue is real valued and unstable.

Two important points are evident from Fig. 3. The first is that except for very small wave numbers, the single value $b_2 = \frac{1}{6}$ $(\theta = \frac{1}{3})$ produces optimal accuracy for both the phase and group velocity. Moreover, it remains optimal for all $k\,\Delta x$ and is virtually independent of f_1. The second point is that except for small wave numbers, $b_2 = 0$ produces optimal accuracy for the wave amplitudes. It is also independent of $k\,\Delta x$ and f_1, although for $f = 0.0$, it does occur over a large region.

Figure 4 is a similar plot with $f_2 = 0.5$. The optimal b_2 value now differs slightly for the two velocity measures. From the approximate optimum of $b_2 = 0.42$ for $f_1 = 0.0$ and small $k\,\Delta x$, the measure values decrease slightly with increasing $k\,\Delta x$, and increase slightly with increasing f_1. The amplitude measures however remain optimal with $b_2 = 0$.

Figure 5 has identical parameter values to those of Fig. 3, but is for the LWEM. Figures 5 and 3 are remarkably similar. The amplitude measures are virtually iden-

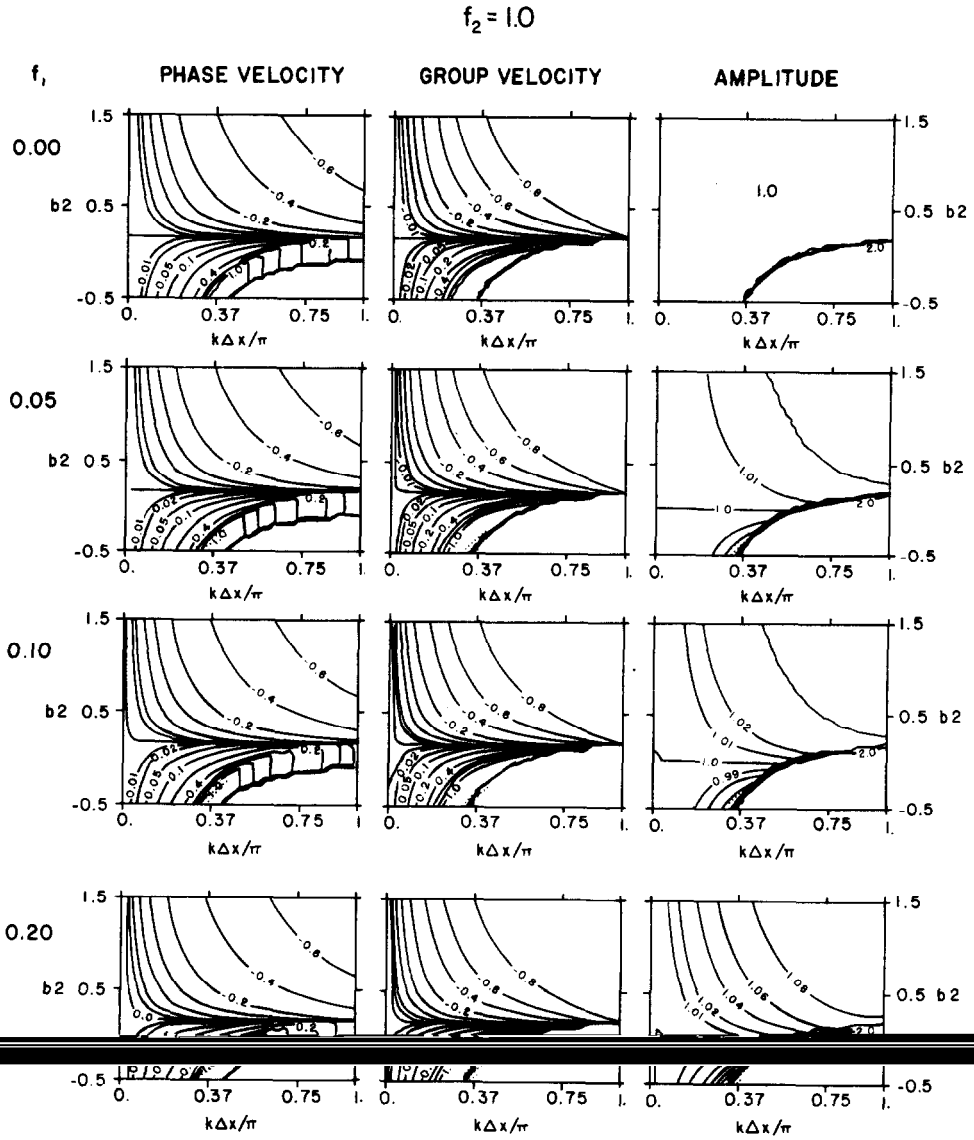

FIG. 3. Accuracy measure values for the WEM as functions of $(b_2, k\,\Delta x)$ for $a_2 = d_2 = 0.5$ and $f_2 = 1.0$. Shaded regions denote dominance of the parasitic eigenvalue.

tical while the velocity measures seem only to differ by a vertical shift. Provided this result extends to other values of f_1 and f_2, it has two important implications. The first is that lumping has not affected wave amplitude accuracy. The second is that by

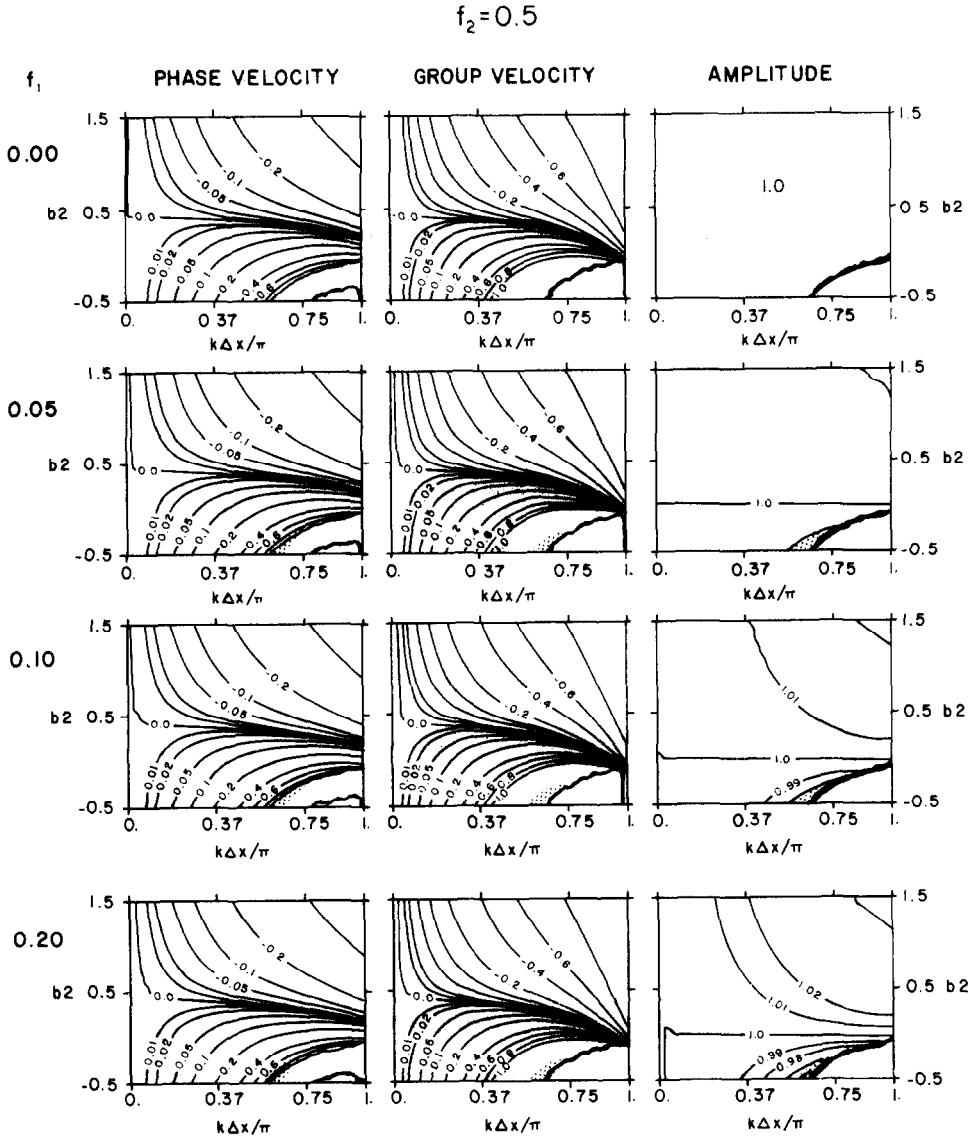

FIG. 4. Accuracy measure values for the WEM as functions of $(b_2, k\,\Delta x)$ for $a_2 = d_2 = 0.5$ and $f_2 = 0.5$. Shaded regions denote dominance of the parasitic eigenvalue.

simply choosing a different time-stepping method, any wave propagation accuracy with the WEM is also possible with the LWEM. These hypotheses are confirmed in Section 5.

Combining the restriction $a_2 = \frac{1}{2}$ with (26) and (22) has the following implications

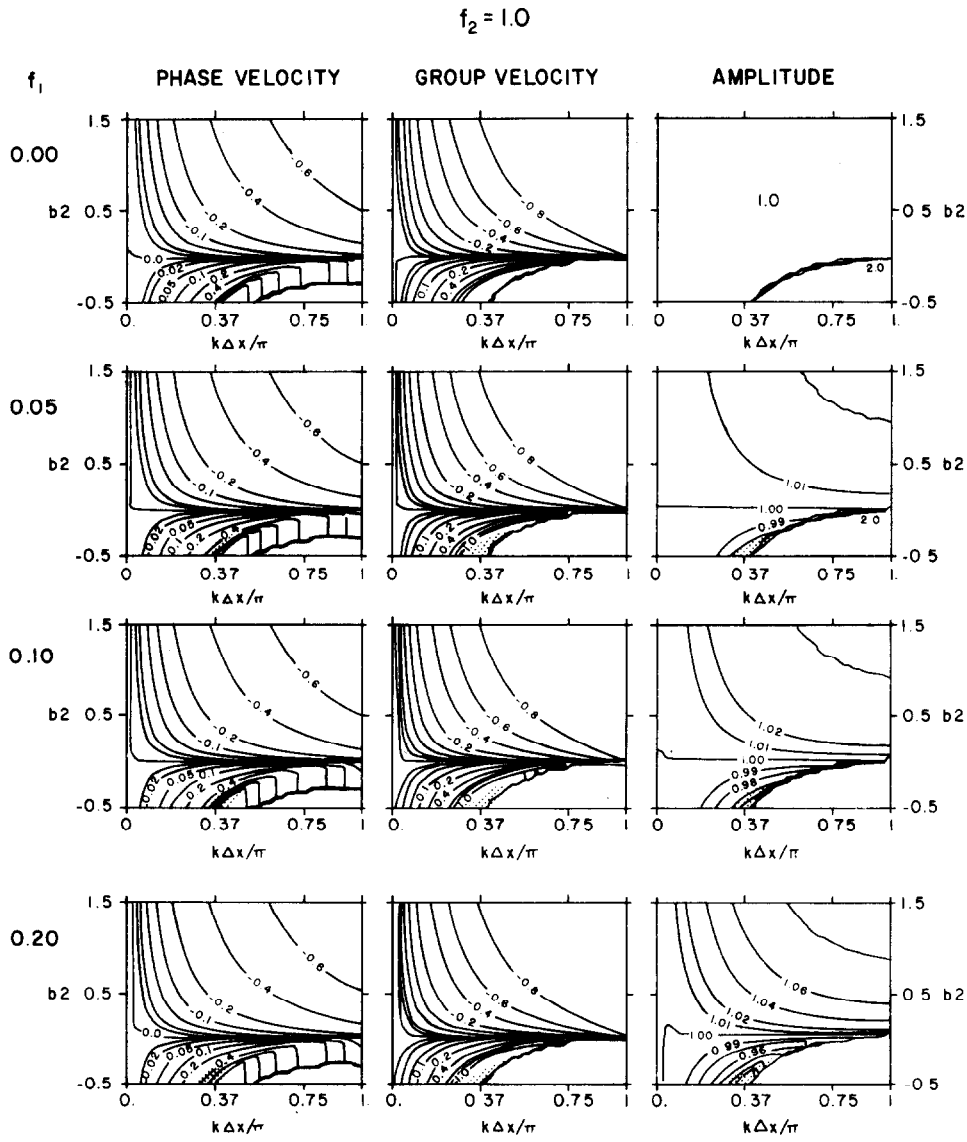

FIG. 5. Accuracy measure values for the LWEM as functions of $(b_2, k\,\Delta x)$ for $a_2 = d_2 = 0.5$ and $f_2 = 1.0$. Shaded regions denote dominance of the parasitic eigenvalue.

for $2\,\Delta x$ waves: $\lambda = 1$ does not satisfy (26b) and only satisfies (26a) when $\tau = 0$; $\lambda = -1$ satisfies (26a) when either $d_2 = \frac{1}{4}$ or $\tau\,\Delta t = 0$, and satisfies (26b) when

$$b_2 = \tfrac{1}{4}(1 - (1/3f_2^2)) \qquad \text{for the WEM,} \tag{28a}$$

and

$$b_2 = \tfrac{1}{4}(1 - (1/f_2^2)) \qquad \text{for the LWEM.} \tag{28b}$$

Therefore with $d_2 > \frac{1}{4}$ ($\theta > \frac{1}{2}$) and $\tau \neq 0$, $2\,\Delta x$ waves should accumulate only when the time-stepping method parameter b_2, and the Courant number f_2, are related by (28). Since $d_2 \geqslant \frac{1}{4}$ is required for stability, and bottom friction is usually included in a model, these first two restrictions are normally satisfied.

5. An Asymptotic Analysis

The preceding accuracy measure analysis suggests that for small $k\,\Delta x$, it is possible to choose a value of b_2 or θ which produces optimal accuracy for phase and group velocity, or wave amplitude growth. It also implies that an accuracy loss through lumping can be avoided. In this section, these hypotheses are confirmed with an asymptotic expansion for small $k\,\Delta x$. Since numerical models are usually designed so that desired wavelengths are at least $20\,\Delta x$ (i.e., $k\,\Delta x \leqslant 0.314$), such an expansion is valid. Desired waves which are significantly shorter in some model regions suggest the need for a mesh refinement. Short waves (e.g., $2\,\Delta x$ waves) that have been generated by boundary conditions, interfaces, and round-off errors may exist in a model and may be important insofar as they can contaminate the desired waves. However, it is not important that they be modelled accurately, only that their growth be controlled.

The analytic dispersion relationship for (5) and (1b) is

$$\omega = i\,\tfrac{1}{2}\tau \pm [ghk^2 - (\tfrac{1}{2}\tau)^2]^{1/2}. \tag{29}$$

Therefore

$$\mathrm{Re}(\omega\,\Delta t)^2 = f_2^2\,(k\,\Delta x)^2 - (\tfrac{1}{2}\tau\,\Delta t)^2. \tag{30}$$

Assuming a nonzero imaginary part for the complex root

$$\lambda = r\exp(i\phi) \tag{31}$$

of the general quadratic

$$a\lambda^2 + b\lambda + c = 0 \tag{32}$$

implies

$$\phi = \arcsin \zeta^{1/2} \tag{33a}$$

where

$$\zeta = 1 - (b^2/4ac). \tag{33b}$$

Provided $\zeta < 1$, the associated power series expansion [1] is

$$\phi = \zeta^{1/2} + \tfrac{1}{6}\zeta^{3/2} + \tfrac{3}{40}\zeta^{5/2} + O(\zeta^{7/2}). \tag{34}$$

Therefore

$$\phi^2 = \zeta + \tfrac{1}{3}\zeta^2 + \tfrac{8}{45}\zeta^3 + O(\zeta^4). \tag{35}$$

Applying these results to the quadratic for the principal eigenvalues arising from the WEM, (7b), yields

$$\zeta = \frac{E^2 - E^4(\frac{1}{4} - \frac{1}{2}\theta) - (\frac{1}{2}\tau\,\Delta t)^2}{(1 + \frac{1}{2}\theta E^2)^2 - (\frac{1}{2}\tau\,\Delta t)^2}. \tag{36}$$

Setting $x = k\,\Delta x$, an asymptotic expansion of E^2 for small x is

$$E^2 = f_2^2 x^2 \left(1 + \frac{1}{12}x^2 + \frac{1}{360}x^4\right) + O(x^8) \tag{37}$$

and a similar expansion for E^4 is

$$E^4 = f_2^4 x^4 (1 + \tfrac{1}{6}x^2) + O(x^8). \tag{38}$$

Substituting these expansions into (36) yields

$$\zeta = f_2^2 x^2 + f_2^4 x^4(-\tfrac{1}{2}\theta + (1/12f_2^2) - \tfrac{1}{4}) + O(x^6) - (\tfrac{1}{2}\tau\,\Delta t)^2(1 + O(x^2) + O((\tau\,\Delta t)^2)) \tag{39}$$

and substituting (39) into (35) produces

$$\begin{aligned}\phi^2 = {} & f_2^2 x^2 + f_2^4 x^4[-\tfrac{1}{2}\theta + \tfrac{1}{12}(1 + (1/f_2^2))] + O(x^6) \\ & - (\tfrac{1}{2}\tau\,\Delta t)^2(1 + O(x^2) + O((\tau\,\Delta t)^2)).\end{aligned} \tag{40}$$

But in this case, $\phi = \omega_r\,\Delta t$ where ω_r is the frequency arising from the principal numerical eigenvalue. Matching (40) with (30), it then follows that ω_r will be a good approximation to $\mathrm{Re}(\omega)$ when

$$\theta = \tfrac{1}{6}(1 + (1/f_2^2)). \tag{41}$$

For the f_2 values 1 and $\frac{1}{2}$, (41) predicts that the best approximation to the analytic dispersion relationship will occur for $\theta = \frac{1}{3}$ and $\frac{5}{6}$, respectively. These same values should also produce the most accurate phase and group velocities. This is confirmed by Figs. 3 and 4.

An asymptotic analysis of the LWEM follows similarly. The expansion of E^2 for small x now becomes

$$E^2 = f_2^2 x^2 \left(1 - \frac{1}{12}x^2 + \frac{1}{360}x^4\right) + O(x^8) \tag{42}$$

and the best representation of the analytic dispersion relationship is attained with

$$\theta = \tfrac{1}{6}(1 - (1/f_2^2)). \tag{43}$$

Denoting the optimal parameter values of (41) and (43) by θ^*, both the lumped and unlumped versions of (40) can be reexpressed as

$$\phi^2 = f_2^2 x^2 + \tfrac{1}{2} f_2^4 x^4 (\theta^* - \theta) + f_2^6 x^6 \left[\frac{1}{4} (\theta^* - \theta)^2 + \frac{1}{240} \left(1 - \frac{1}{f_2^4} \right) \right] + O(x^8) + O((\tau \Delta t)^2). \tag{44}$$

This explains the similar contour patterns in Figs. 3 and 5. Around their respective θ^* values, both the WEM and LWEM have the same accuracy deterioration for ϕ^2. Furthermore, the best time-stepping method for the lumped scheme produces the same wave propagation accuracy (to $O((k \Delta x)^8)$) as the best time-stepping method for the unlumped scheme.

A similar asymptotic analysis reveals the optimal value of θ for wave amplitude accuracy. In this case, the analytic eigenvalue amplitude for a propagating wave is

$$|\lambda| = \exp(-\tfrac{1}{2} \tau \Delta t) \tag{45}$$

and its counterpart for a propagating principal numerical eigenvalue is

$$|\lambda| = \left(\frac{1 + (\theta/2) E^2 - (\tau \Delta t/2)}{1 + (\theta/2) E^2 + (\tau \Delta t/2)} \right)^{1/2}. \tag{46}$$

The time-stepping parameter value

$$\theta = 0 \tag{47}$$

now produces highest accuracy since it matches terms to $O((\tau \Delta t)^3)$. This value has further advantages. It denotes an explicit time-stepping method. So when combined with the lumped approach, it is most economical with regard to storage requirements and computation time. It also makes (46) independent of $k \Delta x$, and identical for both the WEM and LWEM. Consequently, the optimal accuracy associated with $\theta = 0$ is not lost in switching from the WEM to the LWEM. These results are substantiated by Figs. 3–5.

Figure 6 illustrates the stability regions and the most accurate time-stepping methods for both the WEM and LWEM. Values for f_2 and θ should be chosen so that the resultant numerical method is stable. The particular choice will be a compromise between accuracy and time-step size. Large values of Δt (or f_2) result in less computation cost but are usually less accurate. The choice $(\theta, f_2) = (\frac{1}{3}, 1)$ provides the largest stable Δt with optimal wave propagation accuracy for the WEM. The similar choice for the LWEM, $(\theta, f_2) = (0, 1)$, is also most accurate for wave amplitude.

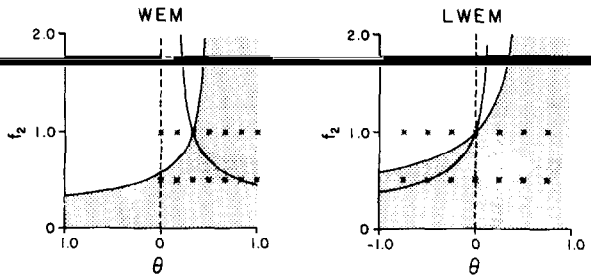

FIG. 6. Stability regions and lines of optimal accuracy for the unlumped and lumped approaches. Asterisks denote methods used in the test problems, shaded areas denote stability, and the most accurate methods for wave propagation and wave amplitude decay are shown with solid and dashed lines, respectively.

6. NUMERICAL TESTS

The analysis of Section 4 was further confirmed with numerical tests similar to the first series reported in [2]. Depth, Δx, and Δt were constant through each test and the additional complication of boundary conditions was avoided by choosing a ring as the test domain. The test conditions therefore correspond to the assumptions underlying the dispersion analysis. All tests were initial value problems where a single progressive wave was studied as it travelled around the ring. Such tests permit validation of the amplitude and phase velocity accuracy measure functions. Further experiments with two progressive waves were not performed but could be expected to produce a validation of group velocity accuracy similar to the demonstration in [2].

Six test problems were selected, each with f_1, f_2, and $k\,\Delta x/\pi$ values corresponding to one of the plots in Fig. 3 or 4. Wavelength and depth were chosen so that the resultant problem would be realistic for semidiurnal tides along a one-dimensional continental shelf.

Each test problem was run for approximately 10 periods and solved with seven different second-order two-step methods. Analytic values for $z(x, t)$ and $u(x, t)$ at times 0 and Δt were used as initial conditions. All methods had a_2 and d_2 fixed at $\frac{1}{2}$ and so were characterized solely by their b_2 values. The (θ, f_2) pairs for the test problems are shown with asterisks in Fig. 6. In each test, the amplitude and phase lag of the wave were calculated at the end of each period and compared to the analytic results. The amplitude change per time step and the non-dimensional phase velocity were also calculated and compared to the values predicted by a dispersion analysis of the numerical method.

Results of the WEM tests are given in Table I. A run was judged unstable when the absolute value of the first elevation point became greater than 10 times the initial amplitude. All unstable methods are predicted by (10). Methods which produce the most accurate representations of wave amplitude decay and phase velocity are designated. For all tests, they confirm the predictions in Figs. 3 and 4.

All discrepancies between the analysis and model results were less than 1%.

TABLE I

Numerical Test Results

		Problem number					
		1			2		
		f_1 0.10	f_2 1.0	$k\Delta x/\pi$ 0.4	f_1 0.10	f_2 1.0	$k\Delta x/\pi$ 0.1
Two-step method parameter: $b_2 =$	Source of results	$\lvert\lambda\rvert$		$C/(gh)^{1/2}$	$\lvert\lambda\rvert$		$C/(gh)^{1/2}$
0.0	Analysis	0.95119[a]		1.16840	0.95119[a]		0.99610
	Model	Unstable			Unstable		
0.08333	Analysis	0.95741		1.07407	0.95158		0.99205
	Model	Unstable			Unstable		
0.16667	Analysis	0.96223		0.99981[a]	0.95197		0.98806
	Model	0.96223		0.99981	0.95197		0.98810
0.25	Analysis	0.96607		0.93929	0.95235		0.98411
	Model	0.96638		0.93961	0.95237		0.98415
0.33333	Analysis	0.96920		0.88870	0.95272		0.98022
	Model	0.96921		0.88872	0.95276		0.98028
0.41667	Analysis	0.97180		0.84555	0.95309		0.97636
	Model	0.97009		0.84379	0.95314		0.97649
0.5	Analysis	0.97399		0.80818	0.95345		0.97256
	Model	0.96840		0.80720	0.95350		0.97276
Analytic solution	Analysis	0.95123		0.99921	0.95123		0.98725
	Model	0.95123		0.99921	0.95123		0.98725

[a] Most accurate value.

Relatively large values can be traced to the initial conditions. For all test methods, the $z(x, t)$ and $u(x, t)$ values specified at time Δt are inconsistent, in varying degrees, with the numerical behaviour of the progressive wave. Consequently, energy is assigned to the other numerical waves. Interference of these waves then causes the numerical results to differ from those predicted by the dispersion analysis. For example, in test 1 with $b_2 = \frac{1}{2}$, the retrogressive wave is initially assigned an amplitude which is 13% that of the progressive wave. The stationary parasitic waves receive no energy.

for the WEM

and parameter values

3			4			5			6		
f_1	f_2	$k\,\Delta x/\pi$	f_1	f_2	$k\,\Delta x/\pi$	f_1	f_2	$k\,\Delta x/\pi$	f_1	f_2	$k\,\Delta x/\pi$
0.00	1.0	0.2	0.05	0.5	0.4	0.05	0.5	0.1	0.20	0.5	0.2

$\lvert\lambda\rvert$	$C/(gh)^{1/2}$	$\lvert\lambda\rvert$	$C/(gh)^{1/2}$	$\lvert\lambda\rvert$	$C/(gh)^{1/2}$	$\lvert\lambda_1\rvert$	$C/(gh)^{1/2}$
1.00000[a]	1.03464	0.98758[a]	1.08719	0.98758[a]	1.00203	0.95119[a]	1.00882
Unstable		0.98681	1.08811	0.98756	1.00205	0.95118	1.00876
1.00000[a]	1.01688	0.98802	1.06663	0.98760	1.00100	0.95159	1.00462
Unstable		0.98771	1.06613	0.98759	1.00101	0.95159	1.00463
1.00000[a]	1.00000[a]	0.98844	1.04719	0.98763	0.99997	0.95199	1.00048
1.00000	1.00000	0.98832	1.04754	0.98762	0.99997	0.95199	1.00054
1.00000[a]	0.98394	0.98882	1.02878	0.98765	0.99894	0.95238	0.99638
1.00015	0.98397	0.98849	1.02870	0.98765	0.99895	0.95238	0.99646
1.00000[a]	0.96862	0.98919	1.01131	0.98768	0.99792	0.95276	0.99233
1.00013	0.96917	0.98916	1.01119	0.98768	0.99793	0.95276	0.99242
1.00000[a]	0.95400	0.98952	0.99470[a]	0.98770	0.99690[a]	0.95313	0.98834[a]
0.99982	0.95437	0.98955	0.99470	0.98770	0.99691	0.95315	0.98842
1.00000[a]	0.94003	0.98984	0.97889	0.98773	0.99588	0.95351	0.98439
1.00018	0.94004	0.99002	0.97913	0.98773	0.99590	0.95353	0.98448
1.00000	1.00000	0.98758	0.99980	0.98758	0.99683	0.95123	0.98725
1.00000	1.00000	0.98758	0.99980	0.98758	0.99683	0.95123	0.98725

These same six problems were also solved with the LWEM. The b_2 values for the time-stepping methods were now chosen as -0.375, -0.25, -0.125, 0, 0.125, 0.25, and 0.375. They are illustrated in Fig. 6. As predicted by (12b), the first three methods were unstable when solving the first three problems. Of the remaining stable methods, $b_2 = 0$ was most accurate for both wave amplitude and phase velocity. For problems 4–6, $b_2 = 0$ was most accurate for amplitude while $b_2 = -0.25$ was most accurate for phase velocity. These results validate (43) and (47). As with the results in Table I, the maximum discrepancy between the analysis and model results was less than 1%.

7. Summary and Conclusions

The preceding analysis has determined the following features of the one-dimensional linearized "wave equation" finite element method:

(i) a similarity to the mixed interpolation approach discussed by Williams and Zienkiewicz;

(ii) a superset for the second-order time-stepping methods proposed by Lynch and Gray;

(iii) the time-stepping methods which most accurately approximate the analytic dispersion relationship, and the analytic wave amplitude decay factor, for both the WEM and LWEM;

(iv) a choice of time-stepping methods which avoids loss of accuracy through lumping.

In particular the analysis indicates that an explicit LWEM with $f_2 = 1$ is the best "wave equation" method since:

(i) it is the stable LWEM which combines the largest Δt with optimal accuracy,

(ii) it produces a diagonal matrix for the matrix equations which must be solved at each time step [5], and is thus the most economical with respect to computation time and storage requirements,

(iii) it combines in one method, the same accuracy as the best unlumped methods for wave propagation and wave amplitude growth.

Unfortunately, the explicit LWEM also has a major disadvantage; it may have problems with $2\,\Delta x$ waves. This is evident from (28). With $f_2 = 1.0$ and the optimal values given by (41) or (43), $\lambda = -1$ is a $2\,\Delta x$ eigenvalue for both the WEM and LWEM. Therefore any $2\,\Delta x$ waves introduced into the model will accumulate rather than decay, and flip sign from one time step to the next.

The third example of Fig. 1 shows that an extension of this same problem can exist for all short waves. Its amplitude curve for the principal eigenvalue increases monotonically with increasing $k\,\Delta x$. (When $k\,\Delta x = \pi$, both the progressive and retrogressive principal roots are real valued. One of them equals -1.) This implies that short waves decay more slowly (or grow more rapidly) in time than long waves. Short waves are therefore favoured by the numerical method. The relative energy in short waves can thus be expected to increase with each time step and may eventually contaminate the numerical solution. The fourth example in Fig. 1 avoids a monotonically increasing amplitude curve but unfortunately still permits the $2\Delta x$ solution $\lambda = -1$. In order to avoid $2\,\Delta x$ problems with the LWEM and still retain the economy of an explicit method, f_2 must be chosen less than 1. This will increase the number of time steps in a run and reduce the phase and group velocity accuracy of all waves.

Since f_2 usually varies throughout a numerical model, choosing a time-stepping

method which depends on this parameter may seem impractical. However, f_2 can be made constant by designing the spatial mesh so that

$$\Delta x = c(h)^{1/2} \tag{48}$$

for some constant c. Intuitively, this is not an unreasonable strategy. Constant frequency (e.g., tidal) waves have their wave numbers increase as they enter shallow water. If $k\,\Delta x$ were maintained constant throughout such transitions then the same wave sampling rate would exist everywhere in the model. Using the analytic dispersion relationship for constant depth (29), a first approximation to uniform sampling is attained through (48). Such a choice also implies that the stability constraints (10c) and (12b) are not determined by spatial elements in deep regions of the model where there may be little variation in the numerical solution. Such would be the case if Δx were constant throughout.

Apart from stability considerations, parasitic eigenvalues have been ignored in the preceding analysis. They can pose problems when for some wave numbers, their magnitudes are greater than those of the principal eigenvalues. In such cases, they grow more rapidly, or decay more slowly, and eventually dominate their principal counterparts. Ideally we would like to choose a value of d_2 such that the parasitic eigenvalues are always subdominant. This is not possible in general since their magnitudes depend on $\tau\,\Delta t$. As demonstrated in Figs. 3–5, with small values of $\tau\,\Delta t$ and $d_2 = \frac{1}{2}$, parasitic eigenvalues are generally subdominant. When considered as functions of a positive $\tau\,\Delta t$, minimal parasitic eigenvalue amplitudes occur when

$$d_2 = \tfrac{1}{2}\alpha = \tfrac{1}{4}(1 + 1/(\tau\,\Delta t)^2) \tag{49}$$

and have the value

$$|\lambda| = \left|\frac{1 - \tau\,\Delta t}{1 + \tau\,\Delta t}\right|. \tag{50}$$

These d_2 values coincide with the switchover from a real to a complex eigenvalue. For small $\tau\,\Delta t$ amplitudes vary only slightly with d_2. So an optimal choice is not

choice guarantees a smaller parasitic eigenvalue for both the WEM and LWEM. And the same dominance is ensured for negative θ, provided

$$\theta \geqslant -1/12f_2^2 \tag{51a}$$

and

$$\theta \geqslant -1/4f_2^2 \tag{51b}$$

for the unlumped and lumped approaches respectively. In fact, Figs. 3–5 suggest that these conditions may be overly restrictive.

ACKNOWLEDGMENTS

I thank Professor J. M. Varah for his advice and support throughout this work, and the referees for their constructive criticism of an earlier version of this paper.

REFERENCES

1. M. ABRAMOWITZ AND I. A. STEGUN, "Handbook of Mathematical Functions," Dover, New York, 1965.
2. M. G. G. FOREMAN, *J. Comput. Phys.* **52** (1983), 290.
3. C. W. GEAR, "Numerical Initial Value Problems in Ordinary Differential Equations," Prentice-Hall, Englewood Cliffs, N.J., 1971.
4. W. G. GRAY AND D. R. LYNCH, *Advan. Water Resources* **1** (1977), 83.
5. D. R. LYNCH AND W. G. GRAY, *Comput. and Fluids* **7** (1979), 207.
6. R. D. RICHTMEYER AND K. W. MORTON, "Difference Methods for Initial-Value Problems," Wiley–Interscience, New York, 1967.
7. L. N. TREFETHEN, *SIAM Rev.* **24** (1982), 113.
8. R. A. WALTERS AND G. F. CAREY, *Comput. and Fluids* **11** (1983), 51.
9. R. T. WILLIAMS AND O. C. ZIENKIEWICZ, *Int. J. Numer. Methods Fluids* **1** (1981), 81.