DISPERSION ANALYSES
OF
FINITE ELEMENT SOLUTIONS
OF THE
SHALLOW WATER EQUATIONS

By

MICHAEL GEORGE GARVIN FOREMAN

B.Sc., Queen's University, 1971

M.Sc., University of Victoria, 1973

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department of Mathematics

Institute of Applied Mathematics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June 1984

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date August 9, 1984.

Supervisor: .....

ABSTRACT

This thesis investigates the accuracy and stability of finite element solutions of the shallow water equations. The method of investigation is referred to as *a dispersion analysis.* It compares numerical phase velocities, group velocities, and wave amplification factors to their analytic counterparts.

Chapter 1 discusses the shallow water equations, finite element and finite difference methods, and reviews previous work. The advantages and disadvantages of *a dispersion analysis* are also discussed.

Chapters 2 and 3 are restricted to numerical solutions of the one dimensional linearized shallow water equations. The phase and group velocities of eight spatial discretizations are calculated and examined for their relative merits. The most accurate two-step time-stepping methods are found for three finite element spatial discretizations; the *wave equation* model of Gray and Lynch, the Galerkin method with linear basis functions, and the Galerkin method which combines quadratic basis functions for velocity with linear functions for elevation. It is also shown that with an appropriate time-stepping method, lumping the *wave equation* model need not cause an accuracy loss.

Chapter 4 extends the analysis to the linearized two dimensional equations. Finite element solutions are computed for two configurations of triangular elements. Two finite element methods, Thacker's method and the *lumped wave equation* model, are shown to be cost competitive and as accurate as the Richardson-Sielecki explicit finite difference method. The analysis also suggests that finite element meshes comprised of equilateral triangles most accurately represent phase and group velocity.

Chapter 5 extends the one dimensional dispersion analysis to include boundary conditions. The stability and relative accuracy of several absorbing boundary conditions are examined. Accuracy is evaluated through the calculation of reflection coefficients. An unstable boundary condition of the type examined by Trefethen is also found.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am grateful to many people for help and advice given in the course of this work. They include Professor John Morris, Professor Paul LeBlond, Professor Lawrence Mysak, Dr. Andrew Bennett, and Dr. Richard Thomson for carefully reading the manuscript and suggesting many improvements; Dr. Lloyd Trefethen, Dr. Daniel Lynch, and Dr. Roy Walters for discussing their work and bringing many new references to my attention; the reviewers of [Fo83], [Fo83b], and [Fo84] for their constructive criticism of earlier versions of these papers; Coralie Wallace for assisting with the figures and tables; and fellow graduate student Tom Nicol for sharing ideas and helping with numerical computations at UBC.

I also wish to express special thanks to two people. One is my supervisor, Professor James Varah, whose guidance, encouragement, and friendship have made this research a most pleasurable learning experience. The second is Dr. Falconer Henry, who as Head of the Numerical Modelling Section and my supervisor at the Institute of Ocean Sciences, introduced me to numerical modelling, encouraged and supported my educational leave, provided many valuable suggestions during the research, and allowed me to pursue the research as part of my work at Patricia Bay. I am indebted to both Jim and Falconer; without their support and guidance this study could not have been carried out.

Technologically, this thesis was produced with the typesetting system TEX. I am grateful to Dr. Larry Roberts for bringing TEX to UBC and for patiently answering my many questions on its use. Additional thanks in this regard also go to Dr. Afton Cayford, Vince Manis, and Josef Roehrl, and to Terry Coatta for assisting with the TEX entry.

Most of the thesis research and computing were done at the Institute of Ocean Sciences, Patricia Bay. I thank Dr. C.R. Mann, Director-General, and Dr. J.F. Garrett, Division Head of Ocean Physics, for the opportunity to pursue this work at IOS. I also thank the Natural Sciences and Engineering Research Council for partial financial support during my educational leave.

Last, but not least, I thank Karina for listening to my thesis chatter and patiently enduring my 'mathemagics' while she tended to the house and garden.

1. INTRODUCTION

1.1 Motivation and Objectives

Waves in the ocean arise from many types of forcing and dynamics, and have a broad range of wavelengths and periods. LeBlond and Mysak [Le78] classify oceanic waves into five basic types according to their restoring forces. Sound waves arise from the compressibility of the ocean. Capillary waves are dominated by surface tension acting between two different fluids, such as air and water. Gravity waves occur through the restoring action of buoyancy on water particles displaced from equilibrium levels. Inertial waves arise from the Coriolis force which acts at right angles to a velocity vector, and is due to the rotation of the Earth. Finally, planetary or Rossby waves arise from variations in the equilibrium potential vorticity due to changes in the Coriolis parameter or the fluid depth.

Detailed studies of waves in the ocean (or the atmosphere) are based on mathematical descriptions of fluid motion on the surface of a rotating Earth. These motions are governed by conservation laws for mass and momentum, an equation of state, and the laws of thermodynamics. The shallow water equations are a particular mathematical description for waves whose amplitudes are much smaller than their wavelengths, and whose wavelengths are much longer than the depth of fluid over which they are travelling [St57]. They are frequently encountered in both oceanographic and atmospheric problems. Solutions to the shallow water equations are of two types; often referred to as waves of the first and second class [Le78]. These types are characterized by the relative size of ω , the wave frequency, and f, the Coriolis parameter. Waves of the first class are gravity waves, for which $\omega > f$, and for which rotation plays only a modifying role. Waves of the second class are planetary waves, for which $\omega \ll f$. They would not exist without rotation.

In most instances there is no hope of obtaining analytic solutions to the shallow water

equations. Complexities due to the nonlinear terms, bottom topography, and an irregular coastline mean that one must resort to numerical approximation techniques. Of these, finite difference methods (FDMs) and finite element methods (FEMs) are the most common. FDMs have been used for many years to solve the shallow water equations (e.g., Hansen [Ha62, Ha66], Leendertse [Le67], Heaps [He69], Crean [Cr76], Henry and Heaps [He76]). However it is only within the last decade that FEMs have also become popular (e.g., Wang and Connor [Wa75], Walters and Cheng [Wa79]). A discussion of the principles underlying each method and their respective advantages and disadvantages will follow in Section 1.3. Generally, FEMs provide a better resolution of the flow domain but are more costly to implement and to execute (Weare [We76]). Measures to reduce the FEM cost have been investigated, but they are usually accompanied by a loss of accuracy in the numerical solution (e.g., Strang and Fix [St73], Mullen and Belytschko [Mu82]). One objective of this thesis will be to investigate such compromises. In particular, it will be shown that FEMs can be as economic and as accurate as FDMs.

A second objective will be to determine which FEM is best. Many methods are available since typically, each combines a spatial discretization with a time-stepping or spectral method. The spatial discretization is determined by the particular finite element approach (e.g., Galerkin), the approximating basis functions, and the size, shape, and configuration of the spatial elements. The spatial discretization has the effect of reducing the governing partial differential equations (PDEs) to a system of ordinary differential equations (ODEs) in time. These ODEs can then be solved by one of many methods discussed in texts such as Gear [Ge71] or Lambert [La73].

In this study, the accuracy of a FEM is determined by comparing the amplitudes and velocities of numerical and analytic plane wave solutions. Although this may seem to be a natural approach, it does have limited application. The calculation of plane wave solutions usually requires that the governing PDEs be linear and have constant coefficients. (Periodic boundary conditions are usually assumed too, but it will be seen in Chapter 5 that this assumption is not necessary.) Since the shallow water equations have both

nonlinear terms and nonconstant coefficients, simplifications are required. Solutions similar to plane waves do exist when at least one coefficient, the depth, is linear. However, these solutions are only discussed briefly in Section 2.9. Everywhere else in this thesis, it will be assumed that the shallow water equations are linearized and all coefficients are constant. In particular, both the ocean depth, h, and the Coriolis parameter, f, will be assumed constant. Unfortunately, under these assumptions all planetary wave solutions reduce to steady currents. In order to have propagating planetary waves, either h or f must be nonconstant. Neither of these cases will be studied here, but they certainly warrant future attention.

Since the assumption of constant f/h eliminates propagating Rossby waves, only gravity wave solutions of the shallow water equations are considered in this thesis. Tides, storm surges, and tsunamis are the most common examples of these waves. Tides are generated by the simultaneous action of the moon's gravitational force, the sun's gravitational force, and the revolution about one another of the earth and moon, and the earth and sun [Po78]. Due to the periodic nature of their forcing, astronomical tides are consistent and predictable. Both numerical models and time series methods are commonly used to predict tidal elevations and currents. Storm surges and tsunamis, on the other hand, have irregular forcing. Storm surges are generated by strong winds and atmospheric pressure gradients. Tsunamis are usually caused by earthquakes or events connected with them (e.g., landslides), but they may also arise from man-made nuclear explosions or the explosions of volcanic islands [Mu77]. Both storm surges and tsunamis cause coastal flooding beyond the normal tidal ranges. In extreme cases, they can result in extensive loss of life and property. Consequently, their predictability is very important. Numerical models of the shallow water equations can provide accurate forecasts of the intensity and timing of these events when accurate forcing and initial conditions are available.

1.2 The Shallow Water Equations

The shallow water equations are derived from general dynamic equations which describe the conservation of mass (continuity) and momentum in an incompressible, non-

diffusive fluid (LeBlond and Mysak [Le78, pages 8-10]). Assumptions required for this derivation include uniform density, hydrostatic pressure, fluid velocities that are vertically homogeneous, and a fluid depth that is much smaller than the horizontal scale of motion. Under these conditions, a two dimensional description of fluid flow can be obtained by integrating the continuity and hydrostatic pressure equations through the fluid depth, and substituting the integrated pressure into the momentum equations. With bottom friction and atmospheric forcing specified as in Lynch and Gray [Ly79], the resultant shallow water equations in Cartesian coordinates are [Le78, page 128]

$$\frac{\partial z}{\partial t} + \frac{\partial u(z+h)}{\partial x} + \frac{\partial v(z+h)}{\partial y} = 0 \qquad (1.2.1a)$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - fv + g\frac{\partial z}{\partial x} + \tau u = F_x \qquad (1.2.1b)$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + fu + g\frac{\partial z}{\partial y} + \tau v = F_y \qquad (1.2.1c)$$

where

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

$u(x,y,t)$ = x component of velocity,

$v(x,y,t)$ = y component of velocity,

$f(x,y)$ = Coriolis parameter,

g = gravity,

$\tau(x,y,t)$ = bottom friction parameter,

= $g(u^2+v^2)^{1/2}/(C^2h)$ for the Chezy dissipation model,

C = Chezy coefficient,

$F_x(x,y,t)$ = x component of force due to wind stress and atmospheric pressure gradient,

$F_y(x,y,t)$ = y component of force due to wind stress and atmospheric pressure gradient.

(1.2.1a) is the continuity equation while (1.2.1b) and (1.2.1c) are the momentum equations. Apart from the bottom friction and forcing terms, the atmospheric shallow water equations are similar [Ha80,Na79].

In order to solve (1.2.1) for a specific space-time domain, the force components, initial conditions, and boundary conditions must be specified. Often, the solution is required to be in dynamic equilibrium. (This is sometimes referred to as a *steady state* solution.) This means that provided transient solutions die away, the initial conditions will not influence the final results. However initial conditions are important, for a good choice will accelerate the convergence to equilibrium.

Common boundary conditions for the shallow water equations are:

i) A solid land boundary through which no flow is permitted. Mathematically, this condition is $\mathbf{u} \cdot \mathbf{n} = 0$ where $\mathbf{u} = (u, v)$ and $\mathbf{n}$ is a unit vector normal to the shoreline.

ii) A specified or forcing boundary where $\mathbf{u}$ and/or z is known.

iii) An open or radiating boundary through which waves are to propagate freely without reflection.

Sometimes ii) and iii) are combined so that inward waves are specified and outward waves pass freely through the boundary. The analysis in Chapter 5 will consider all these conditions. Analyses in Chapters 2, 3, and 4 assume a periodic domain thereby avoiding the need for boundary conditions. Initial-value problems on unbounded domains are called *Cauchy problems.*

1.3 A Review of FDMs and FEMs

In order to solve a system of PDEs such as (1.2.1), FDMs approximate each partial derivative by a divided difference. The domain of the problem is usually fitted with a rectangular grid. Approximating values for each of the dependent variables are then obtained at discrete points within the mesh by solving, at each point, the difference equations corresponding to the original PDEs. The accuracy of the approximating solution depends on the resolution of the rectangular grid and the order, or truncation error, associated with the divided differences.

For most time dependent problems, the FDM solution is found by *stepping* through

discrete time levels and at each one calculating the dependent variable values at all the spatial mesh points. If at a new time level each independent variable can be calculated individually, that is solely from values at previous time levels, the FDM is said to be explicit. If however, several new values are linked so that a system of equations must be solved, the FDM is said to be implicit.

Excellent references for the application of FDMs to initial-value problems and initial boundary value problems are Richtmyer and Morton [Ri67] and Kreiss and Oliger [Kr73], respectively.

When applied to solving the shallow water equations, FDMs have several advantages over most FEMs. They are simple conceptually, easy to program, and depending on their specific type, usually quick to solve. The rectangular grid also means that the *no flow* boundary conditions are easily implemented by setting $u(x, y, t)$ or $v(x, y, t)$ to zero.

However FDMs also have disadvantages. One disadvantage is the coarse boundary approximation that results from fitting the domain with rectangles. A finer mesh would reduce the inaccuracy; however, in so doing the number of discrete grid points, and hence the computational effort, would be increased. Another disadvantage arises from using uniformly sized rectangles over the entire domain. Although nested rectangles have been successful in some models (e.g., Greenberg [Gr76]), many problems can develop across the grid change boundary [Su79] if continuity and momentum conditions are not properly matched. The desire for a nonuniform grid arises from the fact that gravity waves have wavelengths that are approximately proportional to the square root of the depth. In order to maintain the same spatial sampling rate per wavelength in all depths of water, Δx should therefore vary with $h^{1/2}$.

FEMs avoid both these disadvantages because they are not restricted to rectangular grid approximations of the domain. They permit elements of any size or shape which, for ease of computation, are usually chosen to be triangles or quadrilaterals. The sides of these elements need not be straight lines. Thus the boundary can be fitted much more accurately and element sizes can be made roughly proportional to $h^{1/2}$. FEMs also approximate each

dependent variable in the shallow water equations with a linear combination of continuous basis functions. This means that unlike the FDM solution, the FEM solution is continuous throughout the flow domain.

However, FEMs also have disadvantages. They are more difficult to program, generally more costly to solve (depending on the solution approach), and pose more difficulties in implementing the *no flow* condition. When the normal velocity is specified at a boundary node, the velocity in the tangential direction must still be calculated [Gr77]. Because the boundaries of a finite element domain are not, in general, parallel to one of the coordinate axes, the specified velocity must be resolved into its $u(x,y,t)$ and $v(x,y,t)$ components and a momentum equation must be solved in the tangential direction. Gray [Gr77], and Walters and Cheng [Wa80] illustrate this procedure.

FEMs have an entirely different approach than FDMs. Set $\mathbf{V}= (z,u,v)$ and rewrite (1.2.1) in matrix form as

$$\mathcal{L}\mathbf{V} = \frac{\partial \mathbf{V}}{\partial t} + A(\mathbf{V})\frac{\partial \mathbf{V}}{\partial x} + B(\mathbf{V})\frac{\partial \mathbf{V}}{\partial y} + C\mathbf{V} + \mathbf{F} = \mathbf{0} \tag{1.3.1a}$$

where $\mathcal{L}$ is an operator, and

$$A(\mathbf{V}) = \begin{pmatrix} u & z+h & 0 \\ g & u & 0 \\ 0 & 0 & u \end{pmatrix} \qquad B(\mathbf{V}) = \begin{pmatrix} v & 0 & z+h \\ 0 & v & 0 \\ g & 0 & v \end{pmatrix} \tag{1.3.1b}$$

$$C = \begin{pmatrix} 0 & h_x & h_y \\ 0 & \tau & -f \\ 0 & f & \tau \end{pmatrix} \qquad \mathbf{F} = \begin{pmatrix} 0 \\ F_x \\ F_y \end{pmatrix}. \tag{1.3.1c}$$

Then for some sufficiently differentiable and suitably chosen basis functions $\{\phi_i(x,y)\}_{i=1}^N$, the FEM approach is to approximate $\mathbf{V}(x,y,t)$ by

$$\hat{\mathbf{V}}(x,y,t) = \sum_{i=1}^{N} \mathbf{a}_i(t)\phi_i(x,y). \tag{1.3.2a}$$

The values of $\{\mathbf{a}_i(t)\}_{i=1}^N$ are chosen so that the residual $\mathcal{L}\hat{\mathbf{V}}$ is minimized.

The basis functions could be time dependent and nonseparable, that is,

$$\hat{\mathbf{V}}(x,y,t) = \sum_{i=1}^{N} \mathbf{a}_i\phi_i(x,y,t). \tag{1.3.2b}$$

However there is usually little advantage to this. The time domain is simply $t > 0$ and in most applications there is no need for a discretization that has some time intervals larger than others. In fact, with time dependent and nonseparable basis functions, a huge system of equations that includes all time levels would have to be solved at considerable expense. With a separable time dependency, the resultant equations become a system of ODEs that can be solved much more cheaply and with a variety of techniques.

There are various types of FEMs depending on how $\mathcal{L}\hat{\mathbf{V}}$ is minimized. Three common ones are:

i) Collocation: The set $\{\mathbf{a}_i(t)\}_{i=1}^N$ is determined by requiring

$$\mathcal{L}\hat{\mathbf{V}}(\xi_i) = 0 \tag{1.3.3a}$$

for N specified points $\xi_i = (x_i, y_i)$.

ii) Least squares:

$$S = \int_D (\mathcal{L}\hat{\mathbf{V}})^2 \, dx \, dy \tag{1.3.3b}$$

is minimized with respect to the set $\{\mathbf{a}_i(t)\}_{i=1}^N$. D is the spatial domain of the problem.

iii) Galerkin: The residual $\mathcal{L}\hat{\mathbf{V}}$ is required to be orthogonal to all basis functions; that is

$$\int_D \mathcal{L}\hat{\mathbf{V}}\phi_i \, dx \, dy = 0 \qquad i = 1, N. \tag{1.3.3c}$$

Petrov-Galerkin methods are a generalization of type iii) wherein $\mathcal{L}\hat{\mathbf{V}}$ is required to be orthogonal to a set of weighting functions $\{\psi_i(x, y)\}_{i=1}^N$, rather than the basis functions. Galerkin FEMs are probably the most common type.

All these approaches lead to a $3N$ by $3N$ system of ODEs of the form

$$M\frac{\partial \mathbf{s}}{\partial t} = P(\mathbf{s})\mathbf{s} + \mathbf{Q}(t) \tag{1.3.4}$$

where

$\mathbf{s}$= $(z_1, u_1, v_1, \ldots, z_N, u_N, v_N)$ = the vector of variable values at N specified discrete points in the domain D,

M= a global matrix which, for linear PDEs, is usually a function of only the element geometry,

$P(\mathbf{s})$ = a nonlinear matrix which is a function of both the element geometry and $\mathbf{s}$,

$\mathbf{Q}(t)$ = a vector of time dependent forcing components.

These ODEs can be solved with a wide variety of time-stepping methods. The resultant system of fully discrete equations will be explicit only when M is diagonal and the time-stepping method is explicit.

A standard reference for FEMs is Strang and Fix [St73]. Pinder and Gray [Pi77] is also most useful for hydrological problems. Lapidus and Pinder [La82] is an excellent new volume discussing the numerical solution of PDEs with both FDMs and FEMs.

In these investigations, only Galerkin FEMs and triangular elements will be considered. Basis functions will be either piecewise linear or piecewise quadratic functions with local support. Each basis function is associated with a discrete point (node) in the domain and is defined so that it has the value unity at its associated node, and zero at all other nodes. This is illustrated for one dimension in Fig. 1.1.

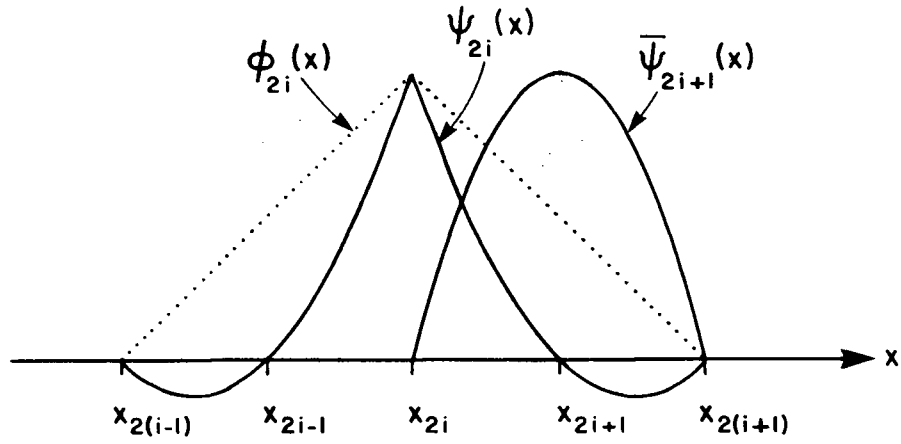

Fig. 1.1. Linear and quadratic basis functions in one dimension.

The linear basis functions are

$$\begin{aligned}
\phi_{2i}(x) &= (x - x_{2i-2})/(x_{2i} - x_{2i-2}) && x \in [x_{2i-2}, x_{2i}] \\
&= (x_{2i+2} - x)/(x_{2i+2} - x_{2i}) && x \in [x_{2i}, x_{2i+2}] \\
&= 0 && \textit{elsewhere.}
\end{aligned}$$

The quadratic basis functions are

$$\begin{aligned}
\bar{\psi}_{2i+1}(x) &= (x - x_{2i+2})(x - x_{2i})/((x_{2i+1} - x_{2i})(x_{2i+1} - x_{2i+2})) && x \in [x_{2i}, x_{2i+2}] \\
&= 0 && \textit{elsewhere,} \\
\psi_{2i}(x) &= (x - x_{2i-1})(x - x_{2i-2})/((x_{2i} - x_{2i-1})(x_{2i} - x_{2i-2})) && x \in [x_{2i-2}, x_{2i}] \\
&= (x - x_{2i+1})(x - x_{2i+2})/((x_{2i} - x_{2i+1})(x_{2i} - x_{2i+2})) && x \in [x_{2i}, x_{2i+2}] \\
&= 0 && \textit{elsewhere.}
\end{aligned}$$

Theoretically, a higher order basis function should yield more accurate results. However, Cullen [Cu76] shows that for the one dimensional linearized shallow water equations with constant depth and a uniform mesh, piecewise linear basis functions are fourth order accurate whereas piecewise quadratics are only second order. Cullen explains this surprising result by viewing the FEM as a computational medium consisting of a discrete number of points. If the points are not equally spaced, or if they are different in character (e.g., one point is a vertex of a triangular mesh and another is the midpoint of a side), waves will see non-uniformities in the medium. Spurious reflection and refraction will then occur, since the problem is physically similar to propagation through a irregular medium. Thus with linear basis functions and regular spacing, the computational medium is uniform and more accurate. However, it becomes non-uniform and less accurate with either irregular spacing or quadratic basis functions.

In practical applications, the spatial grid is seldom regular. Consequently, quadratic basis functions have higher order accuracy. They also have other attributes. Walters and Cheng [Wa79] demonstrate that with this choice, smooth curved-sided elements can be used at all shoreline boundaries. Such elements reduce the mass conservation and flow problems that result when implementing the $\mathbf{u} \cdot \mathbf{n} = 0$ condition for a boundary approximation that has corners, and hence discontinuities in $\mathbf{n}$.

Finally, it should be mentioned that the same type of basis function need not be used for all variable approximations. Hood and Taylor [Ho74], in their studies of the Navier-Stokes equations, recommend using basis functions for the pressure variables that are one order less than those used for the velocities. Their rationale does not extend to the

shallow water equations unless there are diffusion terms (e.g., $\partial^2 u/\partial x^2$) in the momentum equations. Walters and Cheng [Wa79,Wa80] include these terms to approximate molecular and Reynolds stresses, and have successfully used quadratic basis functions for the velocities and linear basis functions for the elevations.

1.4 An Analysis Based on Dispersion

An irony of the numerical approximation process is that the detailed behaviour of the finite difference (or element) formulas is generally a good deal more complicated than that of the differential equations they model [Tr82b]. For example, spurious numerical solutions may arise that have no physical basis and, the principal numerical solutions may not have the same physical properties as their analytic counterparts. These difficulties are usually unimportant provided the difference scheme is convergent (i.e., the numerical solution, everywhere at time t, approaches the solution to the differential equation as the time step size Δt approaches zero). For a FDM, such convergence will occur when the difference model is consistent and stable [Ri67]. This is the Lax Equivalence Theorem. Consequently, the behaviour of FDMs traditionally reduces to estimating truncation errors by Taylor expansions, in order to determine consistency and asymptotic accuracy, and to some kind of investigation of stability.

A similar analysis can be adopted for Galerkin FEMs since consistency and stability are equivalent to convergence here also [St73]. However the nature of the finite element approach and wavelike properties of the solution mean that other analysis techniques can also be used. Assuming a specific spatial discretization, a traditional analysis of ODE methods for solving (1.3.3) involves investigating the absolute stability region and calculating the truncation errors. Stability of a prospective time-stepping method is determined by insuring that the spectrum of the Jacobian of the ODE system, when scaled by Δt, lies in the absolute stability region of the method. Accuracy is determined by evaluating the local truncation error, or by comparing the principal root of the method's characteristic polynomial to the exponential function (which is the analytic result). This approach is discussed in Gear [Ge71] and used by Praagman [Pr79] for analysing finite element solutions

of the shallow water equations.

Another popular technique for evaluating numerical methods which solve hyperbolic PDEs was developed by Leendertse [Le67]. It is based on *propagation factors.* These are ratios of the computed wave to the analytic wave after the time it takes for the analytic wave to propagate one wavelength. Gray and Lynch [Gr77b] apply this analysis to finite element solutions of the shallow water equations. In one dimension, they assess various time-stepping schemes in combination with a Galerkin FEM and piecewise linear basis functions.

The analysis technique adopted here basically amounts to measuring the accuracy of numerical wave amplitudes, phase velocities, and group velocities. We shall call it *a dispersion analysis.* Hyperbolic equations are said to be dispersive if they have plane wave solutions whose velocities are wavelength dependent. Consequently, a packet consisting of several wavelengths will disperse as it propagates. Dispersion analyses are simply extensions of conventional Fourier analyses (e.g., Mesinger and Arakawa [Me76]) to include group velocity. From the amplification factors [Ri67] of the numerical method, dispersion relationships, phase velocities, group velocities, and amplitude decay factors are calculated and used to determine accuracy and stability.

Even though a hyperbolic PDE may be nondispersive, all FDM and FEM approximations of it are dispersive [He75]. This suggests that FDMs and FEMs may be viewed as not just mathematical corruptions of an ideal problem, but as media with analysable properties of their own [Tr82b]. In particular, as dispersive media FEMs and FDMs will turn out to have many of the same features as solid crystals [Br53].

Dispersion and propagation factor analyses reveal more about numerical inaccuracy than an examination of truncation errors. As discussed by Trefethen [Tr82], a numerical wave may have significant pointwise differences from the correct solution (i.e., have a large trunctation error), yet still be qualitatively correct. For example, the numerical phase velocity may simply be too slow. Dispersion analyses, rather than propagation factor analyses, were chosen for this study because they include an examination of the numerical

group velocity.

Group velocity is important in all wave problems since it describes the speed and direction of energy propagation. For tsunamis, the group velocity is vital since the wave packet speed rather than that of an individual wave determines the arrival time [Mu77]. Although shallow water waves have virtually the same phase and group velocity, their numerical model representations may not. It is therefore important to study the properties of both in assessing the merits of a numerical scheme. As will be seen, a method which most accurately represents phase velocity may not be best for group velocity.

Recent work by Trefethen [Tr83] has also linked group velocity to the stability theory of Gustafsson, Kreiss, and Sundstrom [Gu72], (henceforth GKS). In particular, he shows that if a FDM together with its boundary conditions can support a set of waves at the boundary with group velocities pointing into the domain, then the method is unstable. Since the GKS normal mode analysis for stability involves substitutions similar to those for plane wave solutions, it seems likely that some aspects of GKS stability could be investigated if the dispersion analysis were extended to include boundaries.

This analysis approach has other advantages. Calculations to determine accuracy and stability are closely correlated and expressed in terms of amplitude, phase velocity, and group velocity. These concepts are more familiar to the physical oceanographer than stability regions, truncation errors, and propagation factors. Furthermore, the same analysis technique can be used to evaluate a method both before and after the ODE is solved. That is, the analysis can assess the merits of the spatial discretization as well as the time-stepping method.

However, as discussed in Section 1.1, the analysis does have limited application. Usually it requires that the PDEs be linear, and have constant coefficients and periodic boundary conditions. A constant time step and a regular mesh configuration are usually assumed as well. Although few problems are this simple, it is important to understand numerical behaviour in such a setting before introducing the additional complexities of boundary conditions, nonlinear terms, and varying coefficients. In Chapter 5 it is shown that the

dispersion analysis can be extended to include boundary conditions, and in Section 2.9, analyses for non-constant coefficients are discussed.

1.5 Review of Previous Work

Perhaps the earliest finite element shallow water model was developed by Grotkop [Gr72,Gr73]. ([Gr72] was translated by Henry [He78].) He simulated tides in the North Sea with triangular elements and space-time linear basis functions.

Norton et al. [No73] followed with a model comprised of triangular elements, implicit time-stepping, and mixed basis functions; quadratic functions for the velocities and linear functions for the elevations. However, solving the full nonlinear equations with the Newton-Raphson method made the scheme uneconomical. A similar model (King et al. [Ki75]) required unrealistically high values of viscosity in order to produce reasonable numerical solutions.

Connor [Co74] and Wang [Wa75] employed triangles, linear basis functions, and experimented with several time-stepping schemes. A split scheme which calculated elevations and velocities at alternating time levels was found to be best. However short wavelength noise was evident in their results.

Taylor and Davis [Ta75] also tested several time-stepping schemes before combining the trapezoidal rule with cubic quadrilateral elements in a North Sea model. Their results agreed reasonably well with Leendertse's [Le67] finite difference model, but their velocities contained short wavelength noise.

Adey [Ad74] also found spurious short wave oscillations in his linear-triangular model of the Solent estuary. Severe difficulties with stability were also encountered but overcome by adding bottom friction and viscous type terms.

Partridge [Pa76] and Brebbia [Br76] also required large friction and smoothing to stabilize their North Sea model. Again short wavelength noise was present.

Gray [Gr77] developed a model which used a leapfrog time scheme and quadratic quadrilateral elements. With Simpson's rule quadrature, he obtained time-invariant diagonal (easily solved) matrices. He also developed a method for determining an average *no*

flow condition at nodes corresponding to discontinuous boundary approximations. However his solutions, especially with irregular geometry and nonconstant depth, contained short wavelength noise.

Pinder and Gray [Pi77] further examined trapezoidal time differencing with the one dimensional linearized shallow water equations. They found that with linear basis functions and no friction, the scheme was *neutrally stable* (the eigenvalues of the amplification matrix had modulus exactly equal to one) and had no distortion of wave amplitude, but did introduce a phase lag. With friction, the scheme became unconditionally stable and exhibited errors in both amplitude and phase. The same qualitative results were also found with quadratic basis functions for velocity and linear functions for elevation. They also showed that it was possible to produce an inconsistent scheme (i.e., the numerical solution does not converge to the analytic solution as Δt and Δx approach 0) if care was not taken to keep the FEM equations centered in time.

Several spectral and pseudospectral methods (e.g., Kawahara et al [Ka78], Pearson and Winter [Pe77], Jamart and Winter [Ja80], Le Provost et al. [Le81]) based on a finite element spatial discretization have also been developed. They assume a Fourier (harmonic) series time dependent solution and avoid time-stepping. Such methods are computationally efficient but with nonlinear terms in the equations, they can only be used in periodically-forced (tidal) problems. They will not be considered here.

Kawahara et al. [Ka78b,Ka80] developed tsunami and storm surge models using linear-triangle elements and explicit Lax-Wendroff time-stepping. The FEM was made explicit by *mass lumping* (summing all row elements and placing them in the diagonal position). All test results were reported to be in good agreement with either true data, or analytic or finite difference values.

Walters and Cheng [Wa79,Wa80] modified the King model so that smooth sided elements could be used at shoreline boundaries. This meant that unique normal vectors could be defined at all boundary nodes. Lateral stress terms were approximated as diffusion terms and their coefficients were given realistic values. They found that the precise specification

of inflow and outflow boundary conditions was crucial for continuity conservation. They also found their curve sided boundary elements to be superior to the straight sided ones with normal vectors calculated using Gray's method. Their best results were obtained with centered implicit time-stepping. A tidal model of San Francisco Bay produced results which compared favourably with existing data. However small spurious oscillations were present.

The most thorough investigation of FEMs for the shallow water equations was done by Gray and Lynch. In [Gr77b], they first examined ten time-stepping methods for solving the one dimensional equations with constant depth, linear basis functions, and equally sized elements. By calculating and plotting propagation factors and later distribution factors [Ly80], they were able to select three relatively efficient schemes that seemed most likely to avoid small wavelength oscillations and accurately model all other (especially long) wavelengths. Analytic test problem solutions [Ly78] were calculated for the linearized two dimensional equations with zero Coriolis force and a power law depth profile ($h(x,y) = h_0 x^n$ for any real number n). The model domain was either rectangular, or a truncated conic section, with tidal forcing on one side and land boundaries on the others. A spatially variant wind stress was also permitted. The three selected time-stepping methods were tested with these problems.

The leap frog scheme and the semi-implicit scheme both exhibited short wavelength noise in the tests [Gr79]. Their most promising scheme, the so-called *wave equation* model, will be discussed in detail in Chapters 3 and 4. Its test results were not only close to the analytic values, but also free of small wavelength oscillations. However, as yet this method has not been applied in a real setting.

A pseudo-FEM for the solving the shallow water equations has also been presented by Thacker [Th78a, Th78b]. He calculated *finite differences* over triangles and used explicit time-stepping. Experimental tests showed this technique to be cheaper but less accurate than a linear FEM defined on the same grid and solved with the same time-stepping. However, it was claimed that a similar level of accuracy could still be obtained more cheaply

by refining the grid for the *finite difference* approach. Thacker's scheme is investigated further in Section 4.4.

This literature review is confined to the most significant papers. Many models, too numerous to mention, have appeared since 1980. Nevertheless, it is still not clear which finite element spatial discretization is the best for solving the shallow water equations. All discretizations have both advantages and disadvantages. Recent investigations by Platzman [Pl81], Walters and Carey [Wa83], and Walters [Wa83b] have examined the spurious modes generated by FEMs, and the best choices of basis function and triangularization. However there are still unresolved problems.

The application of dispersive wave theory to difference models does not have many predecessors. Although many authors (e.g., [Me76], [Th78b]) have calculated numerical dispersion relationships, few (apparently) have simultaneously looked at phase velocity, group velocity, and wave amplitude accuracy. Warming and Hyett [Wa74] analysed the accuracy and stability of FDMs through *modified equations*. Aside from roundoff error, these equations represent the actual PDE solved when a numerical solution is computed with a FDM. They provide a natural resolution of both amplitude (dissipation) and phase (dispersion) errors. Chin and Hedstrom [Ch75,He75,Ch78] have also applied wave theory arguments to analyse many aspects of solution behaviour and stability. In particular, in [Ch79] they presented a linear wave analysis (including group velocity) of a simplified Galerkin method for solving hyperbolic equations.

Vichnevetsky and his colleagues have also analysed the wave propagation of both principal and parasitic waves, and wave behaviour at boundaries. Vichnevetsky [Vi80] shows that zero group velocity characterizes a cutoff frequency beyond which wave solutions exhibit a spurious amplitude decay. Vichnevetsky and Peiffer [Vi75] demonstrate that spurious $2\Delta x$ waves (waves of length twice the spatial grid interval), generated by mesh refinement or near-discontinuities in the exact solution, travel at the group speed. Most of this work is now summarized in [Vi82]. Schoenstadt [Sc80] and Williams [Wi81] examined phase velocity, group velocity, and amplitude-related coefficients in their evaluation of sev-

eral numerical methods for solving the atmospheric shallow water equations. In many of these studies though, only the effects of the spatial discretization were considered.

Although Trefethen's work [Tr82, Tr82b, Tr83] was only recently available, it has heavily influenced this thesis. This is particularly evident in Chapter 5 where the dispersion analysis is extended to include boundary conditions and linked to stability. However his work also provides an excellent survey of the relevance of group velocity in numerical schemes. Among the important points that he discusses are the following:

i) although wave crests travel at the phase velocity, wave packets travel at the group velocity,

ii) energy travels at the group velocity,

iii) group speed is the only meaningful speed for studying parasitic numerical solutions,

iv) instability of an initial boundary value problem is related to the possibility that at a boundary, incoming waves with positive group velocity may be generated spontaneously rather than from the reflection of outgoing waves with negative group velocity,

v) zero group velocity defines a cutoff frequency for transmission through an interface.

In brief, he demonstrates that there is more to the inaccuracy of a numerical scheme than its truncation error.

1.6 Outline and Summary

In Chapter 2 the dispersion analysis technique is developed for the one dimensional linearized shallow water equations. After specifying their analytic solutions, the terms dispersion relationship, phase velocity, and group velocity are defined. The analysis then begins with calculations of the phase and group velocities arising from eight finite element and finite difference spatial discretizations, and discussions of their relative merits. The class of two-step methods for solving an ODE is then introduced and used in combination with a specific spatial discretization, the Galerkin FEM with linear basis functions. Dominant phase and group velocities, and dominant wavenumbers are also defined and

illustrated. Accuracy measure or error functions are then used to find the most accurate two-step time-stepping method for this FEM. The same analysis is repeated with the mixed interpolation FEM which combines quadratic basis functions for velocity with linear functions for elevation. Numerical tests and truncation errors are used to validate these accuracy measure functions. Section 2.9 summarizes and briefly discusses the results.

Chapter 3 looks at the one-dimensional linearized version of the *wave equation* model developed by Gray and Lynch [Gr77b, Ly79]. Similarities with other spatial discretizations 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. It is also shown that with a judicious choice of time-stepping method, no loss in wave propagation accuracy need occur through lumping. Numerical tests confirm these results.

Chapter 4 extends the dispersion analysis to two dimensions. One FDM, the Richardson-Sielecki explicit scheme, and four FEMs are examined. The FEMs include Thacker's method and the three methods studied in Chapters 2 and 3. Two configurations of triangular elements are assumed for the spatial discretization. All the *wave equation* model results of Chapter 3 are seen to extend to two dimensions. Particular emphasis is given to comparing the relative cost and accuracy of all the methods. Accuracy is determined by comparing numerical and analytic plane wave solutions. Cost is measured as the number of computations per unit of real time and per unit of model area. Two of the FEMs, the lumped *wave equation* model and Thacker's method, are shown to be cost competitive and as accurate as the Richardson-Sielecki explicit FDM. Though not extensive, the finite element analyses also suggest that meshes consisting of equilateral triangles most accurately represent phase and group velocity.

Chapter 5 extends the one dimensional analysis to include boundary conditions. Specifically, the problem of a one dimensional channel with periodic forcing at one end and a closed or radiating boundary at the other is analysed. The chosen boundary conditions

are first shown to be well-posed. For zero friction, it is also shown that the radiating conditions are equivalent to the absorption conditions of Engquist and Majda [En77]. The analysis technique is then developed for the Richardson-Sielecki FDM, and the stability and relative accuracy of several numerical boundary conditions are studied. Accuracy is determined by comparing reflection coefficients. The GKS stability of one set of boundary conditions in combination with the Galerkin FEM with linear basis functions and Crank-Nicolson time-stepping is analysed next. An example of unstable boundary conditions of the Trefethen type [Tr83] is also given. Finally, the stability and relative accuracy of five sets of boundary conditions are studied for that same FEM.

Each chapter has its own introduction, and its own summary and discussion. Hopefully this format will allow readers to survey highlights without getting lost in the details.

Published accounts corresponding roughly to Chapters 2, 3, and 4 can be found in [Fo83], [Fo83b], and [Fo84].

2. ONE DIMENSIONAL DISPERSION ANALYSES

2.1 Introduction

In this chapter, the dispersion analysis is introduced for numerical methods that solve the one dimensional, linearized, shallow water equations on an infinite channel of constant depth. Such a problem describes the propagation of long gravity waves in a canal. Although this is a very simple application of the shallow water equations, it is important to understand numerical behaviour in such a setting before introducing the additional complexities of boundary conditions and varying depth, and before moving on to more realistic two dimensional problems. Two dimensional dispersion analyses will be considered in Chapter 4, and the effects of boundary conditions will be examined in Chapter 5. Some implications of nonconstant depth will be discussed briefly in Section 2.9.

The accuracy of a numerical method will be measured by comparing numerical wave amplitudes, phase velocities, and group velocities to their analytic counterparts. We call this approach *a dispersion analysis*. Hyperbolic PDEs are said to be dispersive if they have plane wave solutions whose velocities are wavelength dependent. This means that a packet consisting of several wavelengths will disperse as it propagates. Even though a hyperbolic PDE may be nondispersive, all FDM and FEM approximations of it are dispersive [He75]. Consequently, measuring the dispersive properties of numerical waves is always a valid technique for determining the accuracy of a numerical scheme.

Dispersion analyses are simply extensions of conventional Fourier analyses (e.g., [Me76]) to include group velocity. From the amplification factors [Ri67] of the numerical method, dispersion relationships, phase velocities, group velocities, and amplitude decay factors are calculated and used to determine accuracy and stability. Phase velocity and amplitude decay, in the guise of dispersion and dissipation, are often studied in the analysis of a

numerical method. However, the recent work of Vichnevetsky [Vi82] and Trefethen [Tr82] has demonstrated that the numerical group velocity should also be considered. Group velocity is important in all wave problems since it describes the speed of energy propagation. For tsunamis, the group velocity is vital since the wave packet speed rather than that of an individual wave determines the arrival time [Mu77]. Although shallow water waves have virtually the same phase and group velocity, their numerical model representations may not. It is therefore important to study the properties of both in assessing the merits of a numerical scheme.

The primary focus of this chapter is accuracy of the numerical solution. Little attention is given to program storage requirements for the methods, or the economy of their numerical calculations. In two dimensions, these are probably the most important criteria for selecting a numerical method. Therefore a complete evaluation of a numerical method should include not only the accuracy considerations studied here but also cost estimates of its implementation and execution. This will be done in Chapter 4.

This chapter is divided into nine sections. Section 2.2 specifies the one dimensional linearized shallow water equations and their analytic solution. It also defines the terms dispersion relationship, phase velocity, and group velocity. Section 2.3 calculates the phase and group velocities arising from eight finite element and finite difference spatial discretizations and discusses their relative merits. Section 2.4 introduces the class of two-step methods for solving an ODE. Section 2.5 applies these methods to the system of ODEs that arise from the Galerkin FEM with linear basis functions. Dominant phase and group velocities, and dominant wavenumbers are also defined and illustrated. Section 2.6 defines three accuracy measure or error functions and uses them to determine which two-step method is the most accurate. Section 2.7 then repeats the analysis for the Galerkin FEM with linear and quadratic basis functions. Section 2.8 verifies the accuracy analysis of Section 2.6 with numerical tests and a truncation error analysis. Finally, Section 2.9 summarizes and briefly discusses the results.

2.2 Analytic Results

The one dimensional linearized shallow water equations are

$$\frac{\partial z}{\partial t} + \frac{\partial (hu)}{\partial x} = 0 \tag{2.2.1a}$$

$$\frac{\partial u}{\partial t} + g\frac{\partial z}{\partial x} + \tau u = 0 \tag{2.2.1b}$$

where

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

$u(x,t)$ = velocity,

$h(x)$ = mean sea depth,

g = gravity

τ = linear bottom friction coefficient.

These equations will be solved for some initial conditions

$$z(x,0) = s_1(x) \tag{2.2.2a}$$

$$u(x,0) = s_2(x). \tag{2.2.2b}$$

The spatial domain may be viewed either as the infinite line $(-\infty, \infty)$, or as a ring. Although such a domain is not realistic, it does simplify the analysis. Realistic boundary conditions will be included in the dispersion analyses of Chapter 5. An initial-value problem on a domain without a boundary is called a *Cauchy problem.*

For linear hyperbolic problems, dispersive waves are usually recognized by the existence of elementary solutions in the form of travelling waves

$$\begin{pmatrix} z(x,t) \\ u(x,t) \end{pmatrix} = \begin{pmatrix} \zeta_0 \\ \mu_0 \end{pmatrix} e^{i(kx-\omega t)}. \tag{2.2.3}$$

ω is frequency and k is wavenumber. The distance between successive wave crests is the wavelength

$$L = 2\pi/k. \tag{2.2.4}$$

A complex exponential form for travelling waves can be assumed because only linear homogeneous equations are considered throughout this thesis. Results for computations in real

arithmetic then follow by taking real parts, or by adding a complex wave to its conjugate. The use of $-\omega t$ rather than $+\omega t$ in the exponential is designed so that the formulas for phase and group velocity do not require a minus sign; see (2.2.10) and (2.2.13).

Assume a constant depth. Substituting (2.2.3) into (2.2.1) and removing common factors yields the system of equations

$$\begin{pmatrix} -i\omega & ihk \\ igk & -i\omega+\tau \end{pmatrix}\begin{pmatrix} \zeta_0 \\ \mu_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{2.2.5}$$

In order that the solutions be nontrivial, the matrix determinant must be zero. This implies

$$\omega^2 + i\omega\tau - ghk^2 = 0. \tag{2.2.6}$$

When each root of this polynomial is expressed in the form

$$\omega_j = W_j(k,\tau,g,h) \qquad j = 1,2 \tag{2.2.7}$$

for some functions W_j, it is said to define a ***dispersion relationship*** [Wh74]. For (2.2.1) the dispersion relationships are

$$\begin{aligned} \omega_j &= -i\tfrac{1}{2}\tau \pm (ghk^2 - (\tfrac{1}{2}\tau)^2)^{1/2} \\ &= -i\tfrac{1}{2}\tau \pm \Omega, \end{aligned} \tag{2.2.8}$$

and the travelling wave solutions are

$$z(x,t) = \zeta_0 e^{-\frac{1}{2}\tau t + i(kx \pm \Omega t)} \tag{2.2.9a}$$

$$u(x,t) = \begin{cases} -\zeta_0 (g/h)^{1/2} e^{-\frac{1}{2}\tau t + i(kx+\Omega t+\theta)} \\ \zeta_0 (g/h)^{1/2} e^{-\frac{1}{2}\tau t + i(kx-\Omega t-\theta)} \end{cases} \tag{2.2.9b}$$

$$\text{where} \quad \theta = \arctan(\tfrac{1}{2}\tau, \Omega). \tag{2.2.9c}$$

From (2.2.3) it is clear that waves with wavenumber k travel at the velocity

$$C = \frac{\Omega}{k}. \tag{2.2.10}$$

This is called the phase velocity. However the propagation of a wave packet containing several wavenumbers is more complicated [Tr82]. Assume initial distributions s_1, s_2 so that a wave packet propagates rightward according to the dispersion relationship (2.2.8). Also

assume that $S_1(k)$ is the Fourier transform of $s_1(x)$. Then at time $t > 0$, the elevation (ignoring normalization factors) is

$$z(x,t) = e^{-\frac{1}{2}\tau t} \int_{-\infty}^{\infty} S_1(k) e^{i(kx-\Omega t)}\, dk$$
$$= e^{-\frac{1}{2}\tau t} \int_{-\infty}^{\infty} S_1(k) e^{it(kx/t-\Omega)}\, dk. \qquad (2.2.11)$$

Suppose x/t is held fixed as $t \to \infty$. This corresponds to a frame of reference that is moving rightward at a fixed velocity $x/t = constant$. After a long time, what is seen? As t increases, the integrand in (2.2.11) oscillates more and more rapidly so that contributions to the integral from adjacent subintervals nearly cancel. Assuming that $S_1(k)$ is smooth, which is the case when s_1 is localized, such cancellation takes place everywhere except where the phase $t(kx/t-\Omega)$ is stationary. These points of stationary phase are characterized by

$$\frac{\partial}{\partial k}\left(\frac{kx}{t} - \Omega\right) = 0 \qquad (2.2.12a)$$

or
$$\frac{\partial \Omega}{\partial k} = \frac{x}{t}. \qquad (2.2.12b)$$

As $t \to \infty$, only wavenumbers that satisfy this equation are seen. Consequently, the energy associated with a wavenumber k moves asymptotically at the group velocity

$$G = \frac{\partial \Omega}{\partial k}. \qquad (2.2.13)$$

More rigourous derivations of group velocity can be found in [Br60,Wh74,Li78].

Waves whose phase velocity C is not independent of k are said to be *dispersive*. The rightward phase and group velocities arising from (2.2.8) are

$$C = (gh)^{1/2}[1 - \tfrac{1}{4}\tau^2/(ghk^2)]^{1/2} \qquad (2.2.14a)$$

$$G = (gh)^{1/2}[1 - \tfrac{1}{4}\tau^2/(ghk^2)]^{-1/2}. \qquad (2.2.14b)$$

When $\tau = 0$, $C = G$ and these waves are nondispersive. However when $\tau > 0$, $G > C$. Individual waves will therefore seem to be created at the leading edge of the packet and disappear at the trailing edge.

Although analytic waves may be nondispersive, all discrete models of them are dispersive [He75,Tr82]. All numerical representations of shallow water waves are dispersive. However, they may not disperse correctly. As it will be seen, many numerical shallow water waves have $C > G$ when $\tau > 0$.

2.3 An Analysis of Spatial Discretizations

Dispersion relationships may also be calculated for the system of ODEs that arise from spatial discretizations of (2.2.1). For example, with constant grid spacing Δx and $z_j = z(j\Delta x, t)$, the spatially discretized equations for a Galerkin FEM with piecewise linear basis functions are

$$\frac{1}{6}\frac{\partial}{\partial t}(z_{j-1} + 4z_j + z_{j+1}) + \frac{h}{2\Delta x}(u_{j+1} - u_{j-1}) = 0 \qquad (2.3.1a)$$

$$\frac{1}{6}\left(\frac{\partial}{\partial t} + \tau\right)(u_{j-1} + 4u_j + u_{j+1}) + \frac{g}{2\Delta x}(z_{j+1} - z_{j-1}) = 0. \qquad (2.3.1b)$$

If nontrivial travelling wave solutions of the form

$$\begin{pmatrix} z_j \\ u_j \end{pmatrix} = \begin{pmatrix} \zeta_0 \\ \mu_0 \end{pmatrix} e^{i(kj\Delta x - \omega t)} \qquad (2.3.2)$$

are now assumed, dispersion relationships can be calculated as they were for the analytic solution. They are

$$\omega = -i\tfrac{1}{2}\tau \pm \left[\frac{gh}{(\Delta x)^2}\left(\frac{3\sin k\Delta x}{2 + \cos k\Delta x}\right)^2 - (\tfrac{1}{2}\tau)^2\right]^{1/2}. \qquad (2.3.3)$$

A consequence of the spatial discretization is that ω is now a function of the number of grid intervals (i.e., sampling) per wavelength

$$\frac{L}{\Delta x} = \frac{2\pi}{k\Delta x}, \qquad (2.3.4)$$

rather than the wavelength L.

Phase and group velocities are calculated from (2.3.3) by extending (2.2.10) and (2.2.13) to

$$C = Re\left(\frac{\omega\Delta x}{k\Delta x}\right) \qquad (2.3.5a)$$

$$G = Re\left(\frac{\partial \omega}{\partial k \Delta x}\right)\Delta x. \qquad (2.3.5b)$$

They may be interpreted as arising from a numerical scheme where the time dependency can be solved exactly. They thus provide a measurement of inaccuracy solely due to the spatial discretization. However, this does not mean that a subsequent time discretizaion will contribute further errors. It is possible that some cancellation may occur and the fully discretized equations may be more accurate.

Schoenstadt [Sc80] and Williams [Wi81] use a Fourier transform technique to analyse several discretizations of simplified versions of the two dimensional equations (1.2.1). Basically, their analyses consist of comparisons of analytic and numerical phase velocities, group velocities, and amplitude functions. A similar analysis is now performed for the following eight discretizations:

D1. a centred FDM with an unstaggered grid,

D2. a centred FDM with a staggered grid,

D3. a Galerkin FEM with piecewise linear basis functions for both variables and unstaggered elements,

D4. a Galerkin FEM with piecewise linear basis functions for both variables and staggered elements,

D5. a residual least squares FEM with piecewise linear basis functions for both variables and unstaggered elements,

D6. a Galerkin FEM with unstaggered elements, piecewise constant basis functions for one variable and piecewise linear for the other,

D7. a Galerkin FEM with unstaggered elements, piecewise linear basis functions for one variable, piecewise quadratic for the other and

a) Δx = distance between adjacent ***linear variables***,

b) Δx = distance between adjacent ***quadratic variables***,

D8. a Galerkin FEM with unstaggered elements, piecewise quadratic basis functions for both variables and

a) Δx = distance between nodes of the same type, (i.e., between mid-element nodes

or end-element nodes),

b) Δx = distance between adjacent nodes.

The dispersion relationships for D6 and D7 are independent of the particular basis function assignments. D1, D2, D3, and D4 were included in Schoenstadt's investigations. Williams used D3 and D4 to compare with three discretizations of the vorticity-divergence form of the governing equations. This formulation is often preferred when planetary waves, rather than gravity waves, are the more important solution. Such is generally the case with atmospheric models.

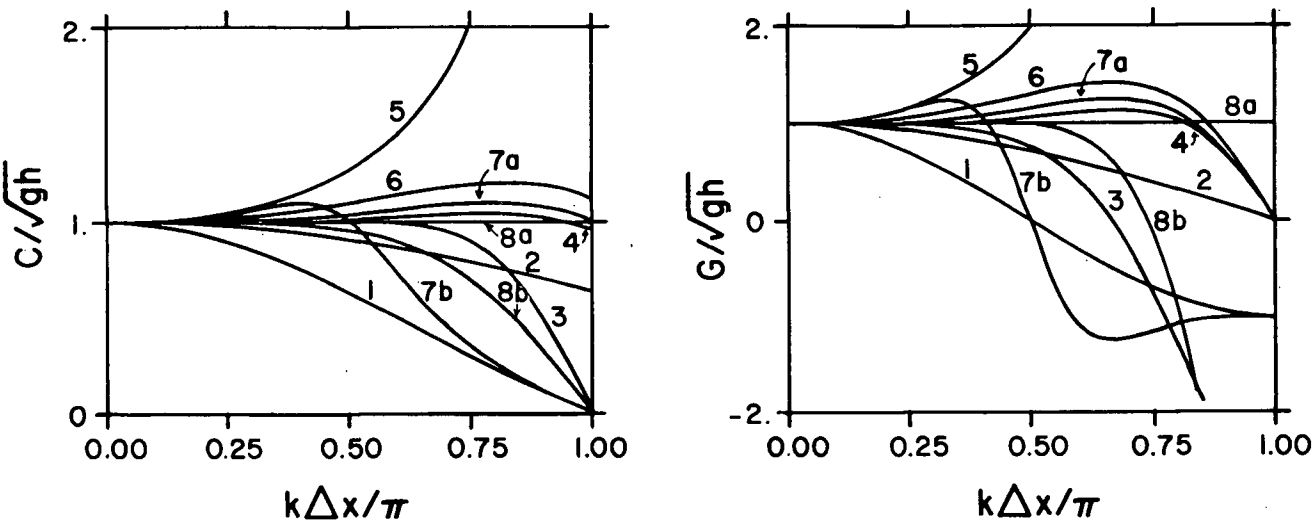

Fig. 2.1. Nondimensional phase and group velocities for several spatial discretizations. Analytic values are identically equal to 1.0 and scarcely distinguishable from those of D8a).

For $\tau = 0$, the spatially discretized equations and their corresponding dispersion relationships are listed in Tables I and II respectively. Fig. 2.1 plots the non-dimensional phase and group velocities versus $k\Delta x/\pi$. Both analytic velocities are identically equal to 1.0 and are shown with a dotted line. All numerical wave amplitudes are identically equal to the analytic amplitude. They have not been shown.

The $(0, \pi]$ range for $k\Delta x$ reflects grid sampling per wavelength. The upper value corresponds to the shortest resolvable wavelength, namely $2\Delta x$, while the lower value represents infinite sampling. Numerical models are usually designed so that desired wavelengths are at least $20\Delta x$ (i.e., $k\Delta x/\pi < 0.1$). Fig. 2.1 shows that most of the selected discretizations

TABLE I

Spatially Discretized Shallow Water Equations

Spatial Discretization	References	Continuity Equation(s) $\frac{\partial}{\partial t}\sum_{i=-1}^{1} a_i z_{j+i} + \frac{h}{\Delta x}\sum_{i=-3/2}^{3/2} b_i u_{j+i} = 0$												Momentum Equation(s) $\frac{\partial}{\partial t}\sum_{i=-1}^{1} c_i u_{j+i} + \frac{g}{\Delta x}\sum_{i=-3/2}^{3/2} d_i z_{j+i} = 0$									
		a_{-1}	$a_{-1/2}$	a_0	$a_{1/2}$	a_1	$b_{-3/2}$	b_{-1}	$b_{-1/2}$	b_0	$b_{1/2}$	b_1	$b_{3/2}$	c_{-1}	$c_{-1/2}$	c_0	$c_{1/2}$	c_1	$d_{-3/2}$	d_{-1}	$d_{-1/2}$	d_0	$d_{1/2}$	d_1	$d_{3/2}$
D1	[Sc80,Vi75]			1				$-\frac{1}{2}$				$\frac{1}{2}$				1				$-\frac{1}{2}$				$\frac{1}{2}$	
D2	[Vi75]			1					-1		1					1					-1		1		
D3	[Sc80,Vi75]	$\frac{1}{6}$		$\frac{2}{3}$		$\frac{1}{6}$		$-\frac{1}{2}$				$\frac{1}{2}$		$\frac{1}{6}$		$\frac{2}{3}$		$\frac{1}{6}$		$-\frac{1}{2}$				$\frac{1}{2}$	
D4	[Vi75]	$\frac{1}{6}$		$\frac{2}{3}$		$\frac{1}{6}$	$-\frac{1}{8}$		$-\frac{5}{8}$		$\frac{5}{8}$		$-\frac{1}{8}$	$\frac{1}{6}$		$\frac{2}{3}$		$\frac{1}{6}$	$-\frac{1}{8}$		$-\frac{5}{8}$		$\frac{5}{8}$		$\frac{1}{8}$
D5		$\frac{-1}{2\Delta x}$				$\frac{1}{2\Delta x}$		$\frac{1}{\Delta x}$		$\frac{-2}{\Delta x}$		$\frac{1}{\Delta x}$		$\frac{-1}{2\Delta x}$				$\frac{1}{2\Delta x}$		$\frac{1}{\Delta x}$		$\frac{-2}{\Delta x}$		$\frac{1}{\Delta x}$	
D6	[Wi81b]			1						-1		1		$\frac{1}{6}$		$\frac{2}{3}$		$\frac{1}{6}$		-1		1			
D7a	[Pi77]	$\frac{1}{6}$		$\frac{2}{3}$		$\frac{1}{6}$		$-\frac{1}{6}$	$-\frac{2}{3}$		$\frac{2}{3}$	$\frac{1}{6}$		$-\frac{1}{10}$	$\frac{1}{5}$	$\frac{4}{5}$	$\frac{1}{5}$	$-\frac{1}{10}$		$-\frac{1}{2}$				$\frac{1}{2}$	
																$\frac{1}{10}$	$\frac{4}{5}$	$\frac{1}{10}$				-1		1	
D8a	[Cu82]	$\frac{-1}{10}$	$\frac{1}{5}$	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{-1}{10}$		$\frac{1}{2}$	-2		2	$\frac{-1}{2}$		$\frac{-1}{10}$	$\frac{1}{5}$	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{-1}{10}$		$\frac{1}{2}$	-2		2	$\frac{-1}{2}$	
				$\frac{1}{10}$	$\frac{4}{5}$	$\frac{1}{10}$				-1		1				$\frac{1}{10}$	$\frac{4}{5}$	$\frac{1}{10}$				-1		1	

Note. $\Delta x = x_{j+1} - x_j = x_{j+1/2} - x_{j-1/2}$ is the distance between end-element nodes and mid-element nodes.

TABLE II

Dispersion Relationships for the Spatial Discretizations of Fig. 2.1

Spatial Discretization	$\omega\Delta x/(gh)^{1/2}$
D1	$\pm \sin(k\Delta x)$
D2	$\pm 2 \sin\left(k\frac{\Delta x}{2}\right)$
D3	$\pm \frac{3\sin(k\Delta x)}{(2+\cos(k\Delta x))}$
D4	$\pm \frac{3}{4}\left(\frac{\sin(\frac{3}{2}k\Delta x)+5\sin(k\Delta x/2)}{2+\cos(k\Delta x)}\right)$
D5	$\pm 2\frac{(1-\cos(k\Delta x))}{\sin(k\Delta x)}$
D6	$\pm\left[\frac{6(1-\cos(k\Delta x)}{2+\cos(k\Delta x)}\right]^{1/2}$
D7a	$0, \pm 2\sin\left(\frac{k\Delta x}{2}\right)\left[\frac{2(4-\cos(k\Delta x))}{(2+\cos(k\Delta x))(3-\cos(k\Delta x))}\right]^{1/2}$
D7b	$0, \pm\sin(k\Delta x)\left[\frac{2(4-\cos(2k\Delta x))}{(2+\cos(2k\Delta x))(3-\cos(2k\Delta x))}\right]^{1/2}$
D8a	$\pm 2\sin\left(\frac{k\Delta x}{2}\right)\left[\frac{(10-\cos^2(k\Delta x/2))^{1/2}\pm 2\cos(k\Delta x/2)}{2-\cos^2(k\Delta x/2)}\right]$
D8b	$\pm\sin(k\Delta x)\left(\frac{(10-\cos^2(k\Delta x))^{1/2}\pm 2\cos(k\Delta x)}{2-\cos^2(k\Delta x)}\right]$

are quite accurate in this range.

The interpretation of Δx necessitates two representations for D7 and D8. In both cases, representation a) is simply the first half of representation b) stretched by a factor of 2. As was seen in Section 1.3, piecewise quadratic approximation requires two types of basis function and the introduction of mid-element nodes. Consequently, waves of length Δx may exist in the approximated $u(t)$ variables. In order to represent these waves in Fig. 2.1, either the upper limit for $k\Delta x$ should be extended to 2π, or Δx should be halved. The latter approach is adopted here.

Ideally, the phase and group velocities of a spatial discretization should be close to

their analytic values. Few of the discretizations shown in Fig. 2.1 are close, particularly for large wavenumbers. Since $2\Delta x$ waves are frequently troublesome in shallow water models (see Section 1.5), their behaviour is important. Fig. 2.1 shows that $2\Delta x$ waves for D1, D3, D7b), and D8b) have zero phase velocity and thus do not propagate. Their corresponding group velocities are negative. Hence the energy associated with these waves is moving, but in the wrong direction. One might therefore see the same generation of spurious waves at an interface with these discretizations as was demonstrated by Trefethen [Tr82]. Furthermore, an inappropriate choice of boundary condition could also cause the instability that he mentions. (This is demonstrated in Section 5.4.) Zero group velocities for discretizations D2, D4, D6, and D7a) indicate that although $2\Delta x$ waves are propagating, the associated energy is not.

A numerical model can not support frequencies larger than the cutoff frequency. This is the largest $|\omega\Delta x/(gh)^{1/2}|$ permitted by the numerical method. Cutoff frequencies exist for D1, D3, D7b), and D8b) since they all attain zero group velocity for waves longer than $2\Delta x$. Precise values for these frequencies can be calculated from the dispersion relationships in Table II. As demonstrated by Trefethen [Tr82], when using one of these discretizations in a problem containing an interface (e.g., due to a mesh refinement or a change of coefficient), it may happen that a wave incident from one side has a frequency which is not sustainable on the other. As shown by Vichnevetsky [Vi80], difficulties may also arise with time-varying boundary conditions which oscillate at frequencies higher than the cutoff.

Graphically it would seem that D8a) is the best spatial discretization. In actual computations, it may not be. Representation D8a) ignores waves shorter than twice the distance between end-element nodes. This is valid provided measures such as artificial viscosity can effectively eliminate these waves. Otherwise, intra-element oscillations can exist and may contaminate the highly accurate longer waves. An additional complication for D8 is that only two of its four dispersion relationships (as given in Table II) have non-dimensional phase and group velocities whose magnitudes tend to 1.0 as $k\Delta x$ tends to zero. (These are the ones shown in Fig. 2.1.) The other two tend to ± 5. and thus are not consistent

with the analytic solution. If waves represented by these spurious curves are generated and sustained in a numerical model, further inaccuracies can be expected. Cullen [Cu82] investigates D8 in more detail.

Provided intra-element oscillations can be avoided, D7 is another promising spatial discretization. Walters and Carey [Wa83] recommend the linear basis functions for $z(x,t)$ and the quadratic functions for $u(x,t)$ since this choice generates fewer spurious modes than vice versa. Consistent with this analysis, they remark that a small amount of dissipation may be necessary in the nonlinear equations to remove $2\Delta x$ waves in the velocity field.

Fig. 2.1 also indicates good accuracy with D4 and D6. In fact D4 is superior to the three vorticity-divergence formulations investigated by Williams [Wi81]. Unfortunately, a convenient triangular element analogue in two dimensions is not apparent, especially for the case of irregular geometry. D6 is also difficult to extend to two dimensions since the piecewise constant variable is discontinuous at the inter-element nodes [Wa83]. However, the *wave equation* FEM of Gray and Lynch [Gr77b,Ly79] has the same dispersion relationship (thus phase and group velocity) as D6 in one dimension (this will be shown in Chapter 3), and has been extended to two dimensions. In some sense it may therefore be viewed as a two dimensional version of D6.

A realistic non-zero value for τ would have little effect on the plots of Fig. 2.1. The analytic non-dimensional phase velocity would become slightly less than 1. for all wavenumbers, and the associated group velocity would become slightly greater than 1. All velocities for the eight spatial discretizations would also exhibit small shifts in varying degrees. All velocities would equal zero for small k, since a wave solution to (2.2.1) can not be supported there. As seen from (2.2.8), this occurs when Ω is imaginary. However a more significant change would occur with D7. Its secondary or spurious dispersion relationship would no longer be zero, thereby permitting the existence of associated spurious waves in the numerical solution.

The preceding analysis illustrates how phase and group velocity accuracy can aid in the selection of a spatial discretization. Because of its restrictive nature (i.e., one dimensional

linearized equations, constant Δx and h, $\tau = 0$.) and the fact that economy of the calculations has been ignored, an analysis of this type should be only one part of the selection process. It must also be stressed that implementation of a time-stepping technique can change the relative accuracy of two spatial discretizations. Analyses of the fully discretized equations should therefore always accompany analyses of spatial discretizations. It will be seen in Sections 2.5, 2.6, and 2.7, however, that many characteristics of the spatially discretized solution, such as non-propagating $2\Delta x$ waves, remain after the introduction of time-stepping.

2.4 A Class of ODE Methods; Linear Two-Step Methods

The system of ODEs corresponding to spatially discretized versions of (2.2.1) can be solved by one of many methods discussed in texts such as Gear [Ge71] or Lambert [La73]. In this study, prospective time-stepping methods are restricted to the broad class of linear two-step methods. (This same class was also used in the stability studies of Beam, Warming, and Yee [Be82] (see Section 5.1), and Trefethen [Tr82b].) Each method in this class is uniquely described by six parameters. Consequently, for a specific spatial discretization it may be possible the to optimize these parameters in order to find the most accurate time-stepping method. This will be the objective in Sections 2.5, 2.6, and 2.7. In this section, linear two-step methods are introduced.

For solving the ODE

$$\frac{\partial y}{\partial t} = f(y), \tag{2.4.1}$$

all two-step methods are characterized by the formula [Ge71,La73]

$$a_2 y^{n+2} + a_1 y^{n+1} + a_0 y^n = \Delta t(b_2 f^{n+2} + b_1 f^{n+1} + b_0 f^n) \tag{2.4.2}$$

where a_2, a_1, a_0, b_2, b_1, b_0 are real numbers. In the sense that both sides of (2.4.2) can be multiplied by any constant and not alter the relationship, this equation requires a normalization. Lambert [La73] suggests $a_2 = 1$, while Gear [Ge71] recommends

$$b_0 + b_1 + b_2 = 1. \tag{2.4.3}$$

The latter convention is adopted here.

For at least second order accuracy (i.e., the truncation error is $O(\Delta t^3)$), only two parameters remain free. Choosing them to be a_2 and b_2, the others are specified as

$$\begin{aligned} a_0 &= a_2 - 1, & b_0 &= \tfrac{1}{2} - a_2 + b_2, \\ a_1 &= 1 - 2a_2, & b_1 &= \tfrac{1}{2} + a_2 - 2b_2. \end{aligned} \tag{2.4.4}$$

These relationships are derived in [Ge71,La73,La71].

Some familiar second order methods with their (a_2, b_2) values are: trapezoid or Crank Nicolson, $(1, \frac{1}{2})$; Gear stiffly stable [Ge71], $(\frac{3}{2}, 1)$; Adams-Bashforth, (1,0); Adams-Moulton, $(1, \frac{5}{12})$; Milne, $(\frac{1}{2}, \frac{1}{6})$; and leapfrog, $(\frac{1}{2}, 0)$. Explicit methods are characterized by $b_2 = 0$. Third order methods include Adams-Moulton and have the additional constraint

$$b_2 = \tfrac{1}{2}a_2 - \tfrac{1}{12}. \tag{2.4.5}$$

Milne's method is fourth order.

A two-step method is stable if and only if the roots of

$$a_2 x^2 + a_1 x + a_0 = 0 \tag{2.4.6}$$

are either inside the unit circle, or simple and on the unit circle. This is known as the *root condition* [Ge71]. Multistep methods that satisfy this condition are called *zero-stable* [La73]. When $a_2 \geq 0.5$, all second order two-step methods are zero-stable.

The stability region of a two-step method consists of the set of all complex values of $\gamma\Delta t$ for which the equation

$$(a_2 - \gamma\Delta t b_2)x^2 + (a_1 - \gamma\Delta t b_1)x + (a_0 - \gamma\Delta t b_0) = 0 \tag{2.4.7}$$

satisfies the root condition. A two-step method is then said to be $A - stable$ [Da63] if its stability region contains all of the left half of the complex $\gamma\Delta t$ plane including the imaginary axis. Beam and Warming [Be79] show that a second order two-step method is $A - stable$ if and only if

$$a_2 \geq \tfrac{1}{2} \tag{2.4.8a}$$

$$b_2 \geq \tfrac{1}{2}a_2. \tag{2.4.8b}$$

2.5 The Galerkin FEM with Linear Basis Functions

In this section, the effects of combining an ODE from the class of second order two-step methods with the particular spatial discretization, D3, are studied. Although Section 2.3 has shown that D3 is not the most accurate discretization, it is commonly used and it does effectively illustrate the analysis. Similar analyses with D7 and the *wave equation* approach of Gray and Lynch [Gr77b,Ly79] will follow in Section 2.7 and Chapter 3 respectively.

Define

$$\tilde{s}_j = \frac{1}{6}(s_{j-1} + 4s_j + s_{j+1}) \tag{2.5.1a}$$

$$\Delta s_j = s_{j+1} - s_{j-1} \tag{2.5.1b}$$

where s can be either z or u. Solving (2.3.1) with a two-step method produces the following fully discretized equations:

$$a_2\tilde{z}_j^{n+2} + a_1\tilde{z}_j^{n+1} + a_0\tilde{z}_j^n + \frac{h\Delta t}{2\Delta x}(b_2\Delta u_j^{n+2} + b_1\Delta u_j^{n+1} + b_0\Delta u_j^n) = 0 \tag{2.5.2a}$$

$$(a_2 + \tau\Delta t b_2)\tilde{u}_j^{n+2} + (a_1 + \tau\Delta t b_1)\tilde{u}_j^{n+1} + (a_0 + \tau\Delta t b_0)\tilde{u}_j^n + \frac{g\Delta t}{2\Delta x}(b_2\Delta z_j^{n+2} + b_1\Delta z_j^{n+1} + b_0\Delta z_j^n) = 0. \tag{2.5.2b}$$

If (2.4.4) is satisfied, these equations are second order accurate in time.

Travelling wave solutions to (2.5.2) have the form

$$\begin{pmatrix} z_j^n \\ u_j^n \end{pmatrix} = \begin{pmatrix} \wp_0 \\ \mu_0 \end{pmatrix} e^{i(jk\Delta x - n\omega\Delta t)}. \tag{2.5.3}$$

With

$$\lambda = e^{-i\omega\Delta t}, \tag{2.5.4}$$

the characteristic equation

$$(a_0 + a_1\lambda + a_2\lambda^2)[a_0 + a_1\lambda + a_2\lambda^2 + \tau\Delta t(b_0 + b_1\lambda + b_2\lambda^2)] + gh\left(\frac{\Delta t}{\Delta x}\right)^2\left(\frac{3\sin k\Delta x}{2 + \cos k\Delta x}\right)^2(b_0 + b_1\lambda + b_2\lambda^2)^2 = 0 \tag{2.5.5}$$

must now be satisfied for nontrivial solutions. Changes in the numerical solution over the interval Δt are now studied through the amplitude and phase of the roots of this polynomial. These roots are also eigenvalues of the amplification matrix resulting from a linear stability analysis [Ri67]. Consequently, these roots shall also be referred to as eigenvalues. Richtmyer and Morton [Ri67] call them *amplification factors.*

The roots of (2.5.5) are

$$\lambda_{1,2} = \frac{-T_1 \pm (T_1^2 - 4T_0T_2)^{1/2}}{2T_2} \tag{2.5.6a}$$

$$\lambda_{3,4} = \frac{-R_1 \pm (R_1^2 - 4R_0R_2)^{1/2}}{2R_2} \tag{2.5.6b}$$

where

$$T_j = a_j + b_j S_+ \qquad j = 0, 2$$

$$R_j = a_j + b_j S_- \qquad j = 0, 2$$

$$S_\pm = \tfrac{1}{2}\tau\Delta t \pm i\left[gh\left(\frac{\Delta t}{\Delta x}\right)^2\left(\frac{3\sin k\Delta x}{2+\cos k\Delta x}\right)^2 - \left(\tfrac{1}{2}\tau\Delta t\right)^2\right]^{1/2}. \tag{2.5.6c}$$

Complex eigenvalues occur in conjugate pairs corresponding to progressive and retrogressive travelling waves. Two progressive waves arise when all four roots have non-zero imaginary parts. In this case, only the principal root represents the desired solution; the other is called *spurious* or *parasitic.* Real valued eigenvalues signify a non-propagating wave and frequently arise for $2\Delta x$ waves (when $k\Delta x = \pi$).

Assume the travelling wave solution has no multiple eigenvalues. Then for some functions $P_j(k\Delta x)$, the component of $z(x,t)$ (or $u(x,t)$) with wavenumber sampling $k\Delta x$ has the following complex valued amplitude at time step n:

$$z_n(k\Delta x) = \sum_{j=1}^{4} P_j(k\Delta x)\left(\lambda_j(k\Delta x)\right)^n. \tag{2.5.7}$$

As n increases this amplitude is dominated by the eigenvalue with the largest modulus. For stability, it is necessary that the dominant eigenvalue have modulus less than or equal to 1.0 for all $k\Delta x$. This is a special case of the von Neumann stability condition for initial value problems

$$|\lambda| \leq 1 + O(\Delta t). \tag{2.5.8}$$

The $O(\Delta t)$ term is usually omitted (e.g., [Me76], [Ri67]) when the exact solution does not grow exponentially. Since $h(x)$ is constant and $\tau \geq 0$, this is the case here.

Each numerical eigenvalue has its own dispersion relationship, and thus its own phase and group velocity. Dominant eigenvalues imply dominant dispersion relationships and dominant velocities. Since the same eigenvalue may not be dominant for all $k\Delta x$, switch points may exist. At these points the dominant dispersion relationship is usually multivalued. Numerical difficulties can be expected at wavenumbers where the parasitic dispersion relationship dominates. If through boundary conditions, initial conditions, or an interface, parasitic waves or wave packets are generated at such wavenumbers, they will eventually overshadow principal waves of the same length.

Associated with each dominant dispersion relationship is a dominant or favoured wavenumber. At this $k\Delta x$ value, the amplitude of the dominant eigenvalue is maximum. A favoured wavenumber therefore denotes the wave which grows most rapidly, or decays most slowly, as time advances. Dissipative schemes such as Lax-Wendroff have amplitudes curves which decrease with increasing wavenumber [Me76]. Small wavenumbers therefore dominate and shorter waves are increasingly damped. Schemes where $k\Delta x = \pi$ is favoured can expect problems with $2\Delta x$ waves.

Fig. 2.2 illustrates the numerical eigenvalues, dispersion curves, and phase and group velocities for three two-step methods. Values for the analytic and spatially discretized solutions, and the discrete numerical solution (i.e., from the eigenvalues of the matrix equation solved at each time step) arising from a ring domain test model with 10 grid points, are also included. Results are parameterized in terms of

$$f_1 = \frac{\tau \Delta x}{(gh)^{1/2}} \tag{2.5.9a}$$

and

$$f_2 = (gh)^{1/2} \frac{\Delta t}{\Delta x}. \tag{2.5.9b}$$

The latter parameter is commonly referred to as the Courant number. Gray and Lynch [Gr77b] use

$$f_1' = \frac{\tau}{(gh)^{1/2} k} \tag{2.5.9c}$$

rather than f_1 for their analyses of various time stepping schemes. For constant depth, constant f_1' implies that τ varies with wavenumber. Consequently, the term τu is not a conventional linear friction. However, f_1 is constant and friction is linear when a constant Δx is assumed. For this reason, f_1 will be used in these investigations.

The eigenvalue spectra diagrams show ranges of the four numerical eigenvalues for $k\Delta x$ in the interval $(0, \pi]$. The unit circle is included as a reference for stability. All eigenvalue paths lie entirely in either the non-negative imaginary half plane and correspond to progressive wave solutions, or in the non-positive imaginary half plane and correspond to retrogressive waves. For these examples, paths of the principal numerical solutions lie almost entirely in either the first or fourth quadrants while the spurious numerical solutions are in the second and third. As $k\Delta x$ increases from zero, the principal progressive numerical eigenvalue moves in a counterclockwise direction from the positive real axis. When $k\Delta x$ is approximately $2\pi/3$, this excursion reverses and returns to the real axis along exactly the same path. Platzman [Pl81] refers to the k value at this turning point as the folding wavenumber, k_f, and discusses the aliasing problems that result from its existence. At the folding wavenumber, the real part of the principal progressive dispersion curve is maximum and the corresponding group velocity is zero. The associated frequency is therefore a cutoff frequency. Although the analytic and principal numerical eigenvalue paths are close when $k < k_f$, it cannot be determined from this diagram if adjacent points in these paths arise for the same $k\Delta x$ value.

The second series of diagrams in Fig. 2.2 permits such a comparison by plotting angular displacement (real part of the dispersion relationship) as a function of $k\Delta x$. Only the progressive wave solutions have been shown. Notice that curves arising from the fully discretized numerical solution are determined to a large extent by those solely due to the spatial discretization. However the principal numerical dispersion curve in the second example (leapfrog method) does illustrate that a subsequent time discretization can improve accuracy for some range of $k\Delta x$. For larger $k\Delta x$ values, the first and third examples demonstrate that the spurious numerical solution can provide a better approximation to

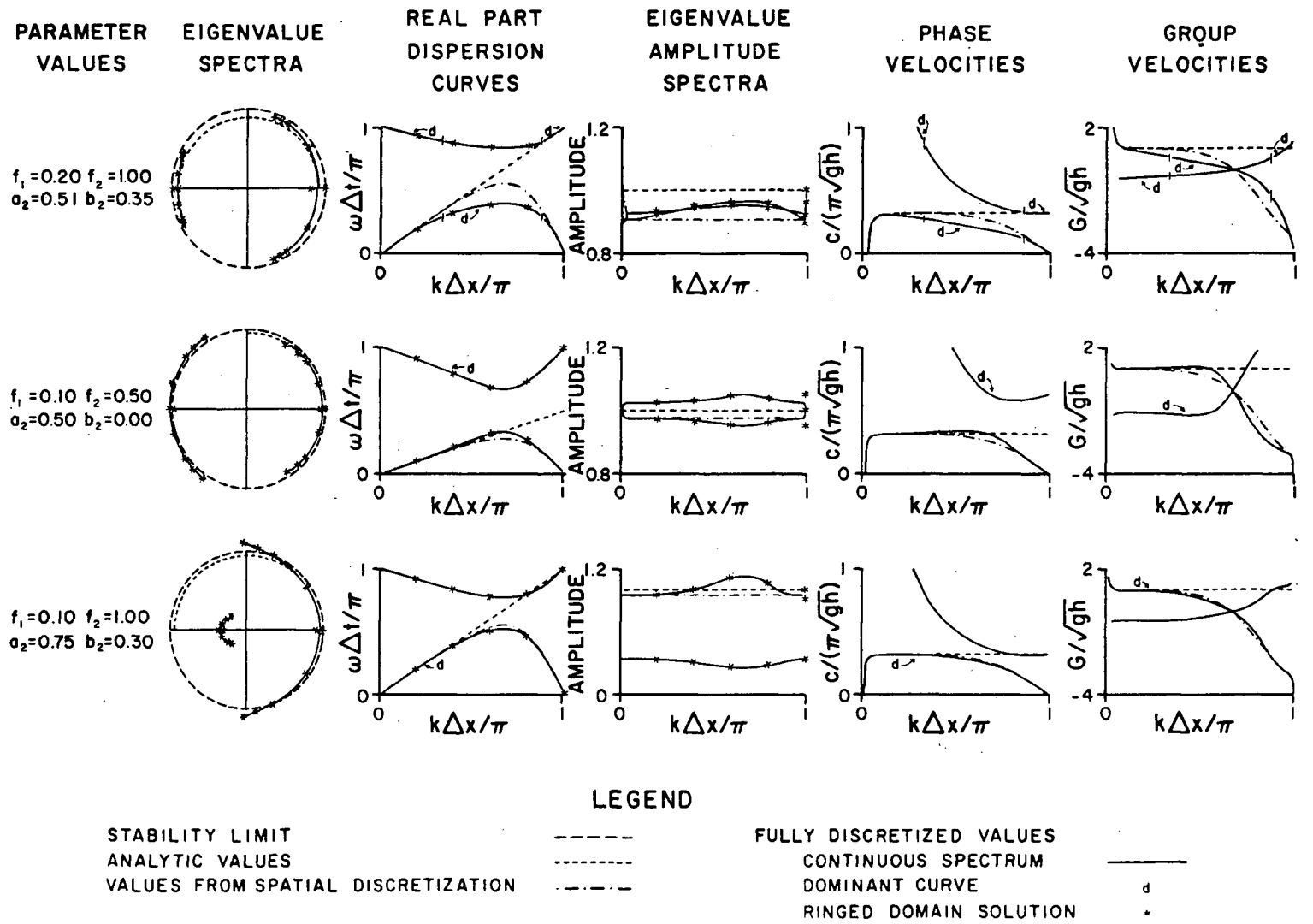

Fig. 2.2. Eigenvalue spectra, dispersion curves, and phase and group velocities for three Galerkin FEMs with linear basis functions.

the analytic dispersion curve than the principal numerical solution.

The third series of diagrams permits determination of instability and the dominant numerical eigenvalue. The first example shows a switch of dominance between the principal and spurious numerical eigenvalues. Specifically, the principal eigenvalue is only dominant for $.325 < k\Delta x/\pi < .880$. In the second example, the spurious eigenvalue is both dominant and unstable, while in the third, the principal eigenvalue is dominant and unstable in the neighbourhood of the folding wavenumber.

In each example the dominant and folding wavenumbers are identical and approximately equal to $2\pi/3$. This is apparent from the eigenvalue spectra plots since the reversal point in each eigenvalue path corresponds to the maximum amplitude.

The fourth and fifth diagrams plot the corresponding non-dimensional phase and group velocities. (They do not have a limiting value of 1.0 at $k\Delta x = 0$ because nonzero friction does not permit a wave solution there.) Phase velocities for the first example are less than the analytic values thereby indicating that the numerical wave solutions travel too slowly. In fact, $2\Delta x$ waves do not travel at all. Group velocities for these cases are also too small and are seen to become negative beyond the folding wavenumber. This indicates energy propagating in the wrong direction. Dominance of the spurious numerical eigenvalue in the second example results in substantially inaccurate phase and group velocities. Switching of the dominant eigenvalue in the first example produces double-valued phase and group velocities at the switch points.

2.6 An Accuracy Analysis of the Galerkin FEM with Linear Basis Functions

The preceding discussion suggests three functions to measure accuracy of the numerical solution; one for each of the amplitude, phase velocity, and group velocity. Their respective definitions are

$$M_A = \left|\frac{\lambda_n}{\lambda_a}\right| \tag{2.6.1a}$$

$$M_C = \frac{C_n - C_a}{C_a} \tag{2.6.1b}$$

$$M_G = \frac{G_n - G_a}{G_a} \qquad (2.6.1c)$$

where λ_n is the dominant progressive numerical eigenvalue, λ_a is the analytic progressive eigenvalue, and C_n, C_a, G_n, G_a are the corresponding phase and group velocities. These error functions are similar to the phase error ratios computed by Warming and Hyett [Wa74].

The velocity accuracy measures are simply relative errors. Negative values denote waves travelling too slowly while zero values are optimal. For example, $-.01$ denotes a numerical velocity which is 1% too slow. The amplitude measure is a ratio denoting the growth (or decay) factor per time step relative to the analytic solution. Values greater than the optimum of 1.0 signify a solution which will decay too slowly or grow too rapidly, whereas values less than 1.0 signify a solution which will decay too rapidly or grow too slowly. After n time steps, the ratio of the numerical amplitude to the analytic will be $(M_A)^n$.

Fig. 2.3 and 2.4 show accuracy measure contours as functions of the second order two-step parameters a_2 and b_2 (designated there by A_2 and B_2). In all plots, a dotted line represents third order methods while asterisks locate the six familiar methods listed in Section 2.4. The stability region is bounded to the left by $a_2 = 0.5$ and from below by the heavy solid line. All methods corresponding to (a_2, b_2) values outside this region have a dominant eigenvalue modulus greater than 1.0 for some $k\Delta x$. They will therefore be unstable.

Fig. 2.3 shows accuracy measure changes as $k\Delta x$ increases and f_1 and f_2 remain fixed at 0.1 and 1.0 respectively. Notice that the most accurate methods may not coincide for all three measures or even lie within the stability region. Thus a method which is most accurate for one $k\Delta x$ value may be unstable for others. Also notice that a method which has more accurate phase velocity may not have more accurate group velocity, and vice versa. For example, with $k\Delta x/\pi = 0.4$ Adams-Bashforth ($(a_2, b_2) = (1, 0)$) has a better phase velocity than Adams-Moulton ($(a_2, b_2) = (1, \frac{5}{12})$); but the latter has a better group velocity. (Since both methods are unstable, this is admittedly a poor example.) High

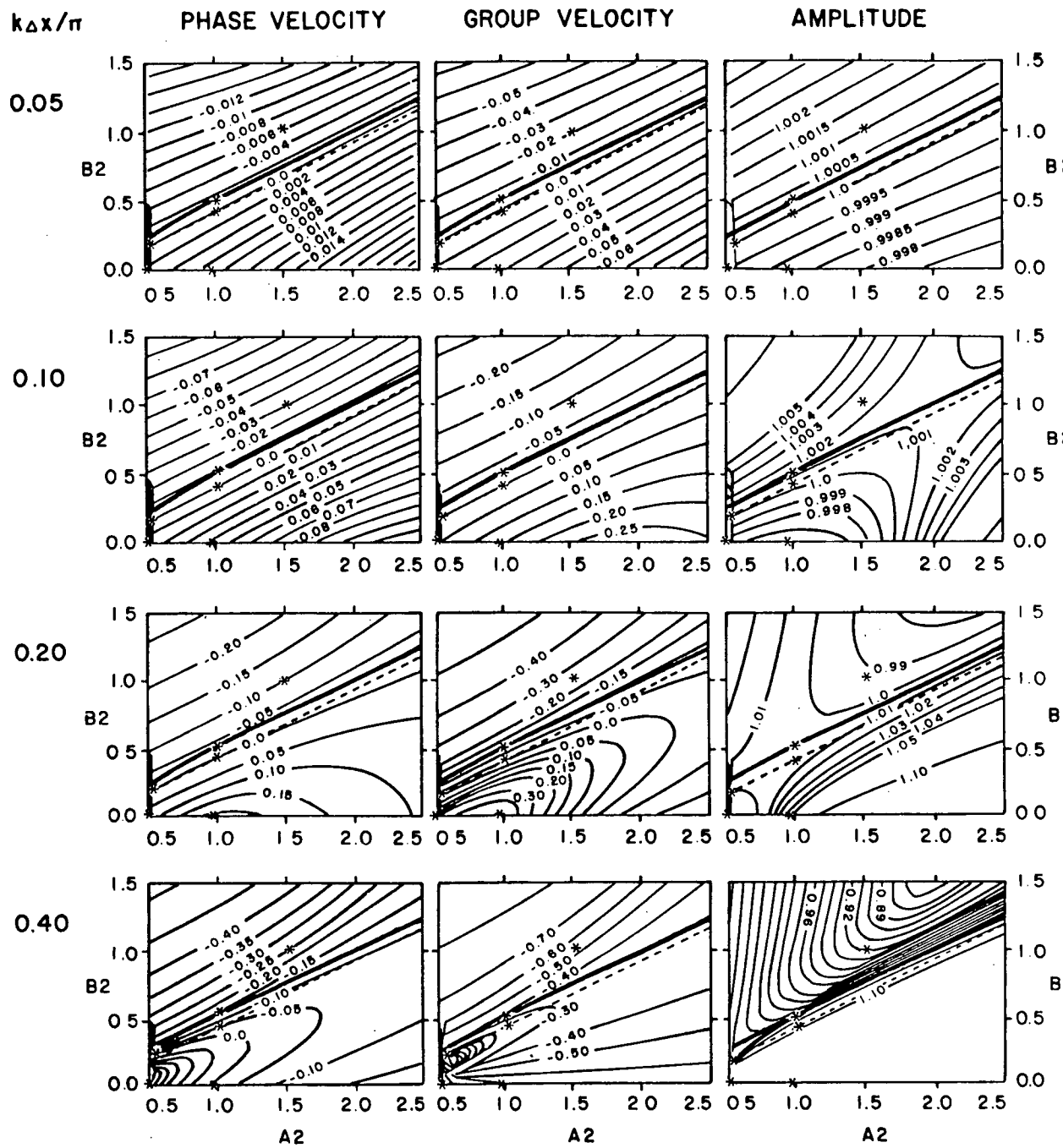

Fig. 2.3. Accuracy measure values for the Galerkin FEM with linear basis functions and $(f_1, f_2) = (0.1, 1.0)$.

accuracy measure values along the lower b_2 axis arise because the parasitic eigenvalue is dominant.

In most numerical models, desired waves have $k\Delta x/\pi < 0.1$ (wavelengths longer than $20\Delta x$). Fig. 2.4 illustrates the accuracy measure changes as f_1 increases with $k\Delta x$ and f_2 fixed at $\pi/10$ and 1.0 respectively. The stability region has the lower boundary $b_2 = 0.5a_2$

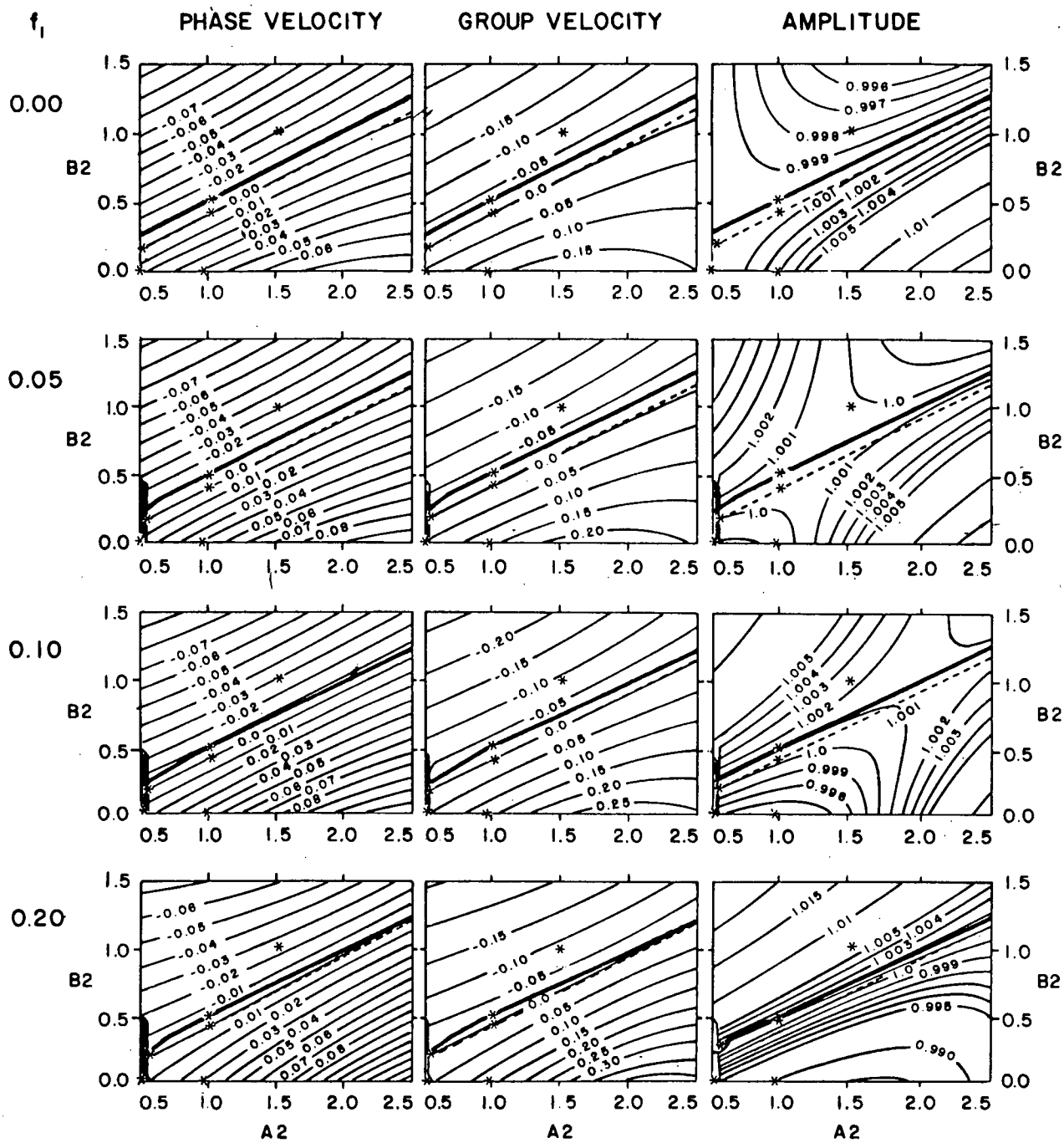

Fig. 2.4. Accuracy measure values for the Galerkin FEM with linear basis functions and $(k\Delta x/\pi, f_2) = (0.1, 1.0)$.

for $f_1 = 0.0$ and becomes less restrictive as f_1 increases. This suggests that for $\tau = 0$, the stability region is defined by (2.4.10), the conditions for A-stability of a second order two-step method. In all cases, the most accurate and stable methods lie on, or very close to the line $b_2 = \frac{1}{2}a_2$. With $f_2 = 0.5$, the same series of plots reveals similar patterns but a less restrictive lower boundary for stability. Optimal accuracy now occurs along the

dotted line or as close to it as stability permits.

It is interesting to notice that all stable schemes in Fig. 2.3 and 2.4 have phase velocities that are slow. This result is not true for all FEMs (it may not even extend to all f_1, f_2, and $k\Delta x$ values with this FEM), as is demonstrated in Fig. 3.1.

Choosing the most accurate two-step method will depend on f_1, f_2, $k\Delta x$, and the relative importance of amplitude and velocity. In most cases, accuracy measure and stability results indicate that all methods along the line $b_2 = \frac{1}{2}a_2$ are very good choices. In fact, the four numerical eigenvalues for all methods in this subset are

$$\lambda_1 = \frac{1 - \frac{1}{2}S_+}{1 + \frac{1}{2}S_+} \tag{2.6.2a}$$

$$\lambda_3 = \frac{1 - \frac{1}{2}S_-}{1 + \frac{1}{2}S_-} \tag{2.6.2b}$$

$$\lambda_2 = \lambda_4 = \frac{a_2 - 1}{a_2} \tag{2.6.2c}$$

where S_+ and S_- are defined in (2.5.6c). λ_1 and λ_3 are principal numerical eigenvalues and are independent of a_2. They are identical for all methods. The other two spurious eigenvalues vary with the two-step method but are constant for all $k\Delta x$. Hence the associated numerical solution will not propagate. Provided λ_1 and λ_2 dominate, all accuracy measures (and numerical solutions after many time steps) for this subset of methods will be identical. However, for some wavenumbers, λ_3 may dominate. The accuracy measures and numerical solution will then vary with a_2. From this perspective, the Crank Nicolson method ($a_2 = 1$) is optimal within the subset. Both its spurious eigenvalues are zero and thus can never dominate. Furthermore, being a one-step method it should also have the most economical storage requirements. However it is implicit and may be expensive with regard to computing time. No second order explicit method exists within the subset.

2.7 A Mixed Interpolation Galerkin FEM

In this section, the analysis of Sections 2.5 and 2.6 is repeated for the Galerkin FEM that approximates $z(x,t)$ with piecewise linear basis functions, and $u(x,t)$ with piecewise quadratic basis functions. Fig. 1.1 shows that there are two types of quadratic basis

function. Even numbered functions such as ψ_{2i} are nonzero over a $2\Delta x$ interval while odd numbered functions such as $\bar{\psi}_{2i+1}$ are nonzero over a Δx interval. This means that discrete velocity variables must be defined at both mid-element nodes and end-element nodes. For constant grid spacing Δx, the variables are located as shown in Fig. 2.5.

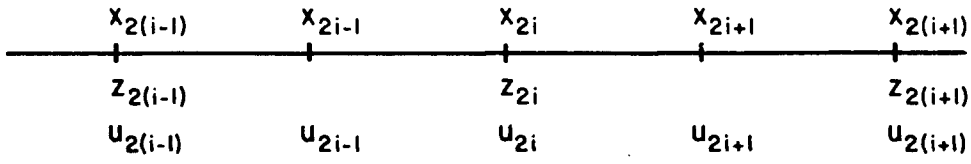

Fig. 2.5. Discrete variables for the mixed interpolation FEM.

The spatially discretized shallow water equations for this mixed interpolation Galerkin FEM are [Pi77]

$$\frac{1}{6}\frac{\partial}{\partial t}(z_{2j-2}+4z_{2j}+z_{2j+2})+\frac{h}{6\Delta x}[u_{2j+2}-u_{2j-2}+4(u_{2j+1}-u_{2j-1})]=0 \quad (2.7.1a)$$

$$\frac{1}{10}\left(\frac{\partial}{\partial t}+\tau\right)(-u_{2j-2}+2u_{2j-1}+8u_{2j}+2u_{2j+1}-u_{2j+2})$$
$$+\frac{g}{2\Delta x}(z_{2j+2}-z_{2j-2})=0 \quad (2.7.1b)$$

$$\frac{1}{10}\left(\frac{\partial}{\partial t}+\tau\right)(u_{2j}+8u_{2j+1}+u_{2j+2})+\frac{g}{\Delta x}(z_{2j+2}-z_{2j})=0. \quad (2.7.1c)$$

When $\tau = 0$, these equations reduce to those listed in Table I. For nontrivial travelling wave solutions of the form

$$\begin{aligned} z_{2j} &= \zeta_0 e^{i(jk\Delta x-\omega t)} \\ u_{2j} &= \mu_0 e^{i(jk\Delta x-\omega t)} \\ u_{2j+1} &= \mu_1 e^{i((j+\frac{1}{2})k\Delta x-\omega t)} \end{aligned} \quad (2.7.2)$$

the following cubic equation must be satisfied:

$$(\omega+i\tau)[\omega^2(2+\cos k\Delta x)(3-\cos k\Delta x)+i\tau\omega(2+\cos k\Delta x)(3-\cos k\Delta x)$$
$$-8\frac{gh}{(\Delta x)^2}\sin^2(\tfrac{1}{2}k\Delta x)(4-\cos k\Delta x)]=0. \quad (2.7.3)$$

When $\tau = 0$, the roots of this equation are the dispersion relationships listed in Table II for D7a).

Assume that the system of ODEs given by (2.7.1) is solved with a linear two-step method. The associated characteristic equation can be calculated directly as in Section 2.5, or indirectly as follows. If the simple ODE given by (2.4.2) has the wave solution

$$\begin{pmatrix} y(t) \\ f(t) \end{pmatrix} = \begin{pmatrix} y_0 \\ f_0 \end{pmatrix} e^{-i\omega t}, \tag{2.7.4}$$

then

$$-i\omega y_0 = f_0. \tag{2.7.5}$$

Assuming the similar wave solution

$$\begin{pmatrix} y_n \\ f_n \end{pmatrix} = \begin{pmatrix} y_0 \\ f_0 \end{pmatrix} e^{-in\omega\Delta t} = \begin{pmatrix} y_0 \\ f_0 \end{pmatrix} \lambda^n \tag{2.7.6}$$

for the general two-step equation (2.4.3) yields

$$y_0(a_2\lambda^2 + a_1\lambda + a_0) = f_0\Delta t(b_2\lambda^2 + b_1\lambda + b_0). \tag{2.7.7}$$

Comparing (2.7.5) and (2.7.7) reveals that the fully discretized characteristic equation can be obtained from the spatially discretized characteristic equation through the substitution

$$-i\omega = \frac{a_2\lambda^2 + a_1\lambda + a_0}{\Delta t(b_2\lambda^2 + b_1\lambda + b_0)}. \tag{2.7.8}$$

A polynomial of order six arises when this approach is applied to (2.7.3). Two of the polynomial roots are given by

$$\lambda_{1,2} = \frac{-(a_1 + \tau\Delta t b_1) \pm [(a_1 + \tau\Delta t b_1)^2 - 4(a_2 + \tau\Delta t b_2)(a_0 + \tau\Delta t b_0)]^{1/2}}{2(a_2 + \tau\Delta t b_2)}. \tag{2.7.9}$$

The remaining four roots have the same form as (2.5.6b), but with $S_\pm$ in (2.5.6c) replaced by

$$S_\pm = \tfrac{1}{2}\tau\Delta t \pm i\left[8gh\left(\frac{\Delta t}{\Delta x}\right)^2 \sin^2(\tfrac{1}{2}k\Delta x)\frac{4 - \cos k\Delta x}{(2 + \cos k\Delta x)(3 - \cos k\Delta x)} - (\tfrac{1}{2}\tau\Delta t)^2\right]^{1/2}. \tag{2.7.10}$$

For Crank-Nicolson time-stepping, these roots reduce to the values quoted by Pinder and Gray [Pi77, page 255]. Four of the six eigenvalues, including λ_1 and λ_2, are spurious. Only the remaining two principal eigenvalues approximate analytic gravity waves.

Because λ_1 and λ_2 are independent of $k\Delta x$, they have zero group velocity and do not propagate energy. For small $\tau\Delta t$ they are also real valued and have zero phase velocity.

Fig. 2.6 shows the eigenvalues, dispersion curves, phase and group velocities for three second order two-step methods together with various (f_1, f_2) parameter values. In all three examples, λ_1 and λ_2 are real valued. This means that their dispersion curve values are either 0 or π, and their amplitude spectra are constant.

For the first example, $\lambda_1 = .904$ and $\lambda_2 = -.268$. The former is dominant when $k\Delta x > .78\pi$. The dominant phase and group velocities are therefore zero in this range.

The second example uses leapfrog time-stepping. It is unstable for the selected values of f_1 and f_2. $\lambda_1 = .951$ and $\lambda_2 = -1.051$. λ_2 is dominant when $k\Delta x < .52\pi$. Outside this range, another spurious eigenvalue is dominant. So even if the leapfrog time-stepping were stable, the numerical solution could be highly contaminated by the spurious roots.

The third example uses Adams-Moulton time-stepping. Instability now arises from the principal numerical root. The dominant phase and group velocities are very accurate for $k\Delta x < .5\pi$. The constant eigenvalues are .951 and -.004 and are always subdominant.

Figures 2.7 and 2.8 plot M_A, M_C, and M_G for the same parameter values as Fig. 2.3 and Fig. 2.4 respectively. For $k\Delta x/\pi = 0.05$ and 0.10, the contours of Fig. 2.7 are virtually identical to those of Fig. 2.3. The only apparent difference is a lower stability boundary that is slightly higher (i.e., a more restrictive stability condition) for the mixed interpolation FEM. When $k\Delta x/\pi = 0.20$, slight differences in the contour lines are evident. For $k\Delta x/\pi = 0.4$ they are more pronounced. Large regions of -1.0 values (i.e., 100% error) also appear in both velocity diagrams when $k\Delta x/\pi = 0.4$. They correspond to the zero phase and group velocities that arise from a dominant, constant, real-valued eigenvalue. The corresponding amplitude measure remains approximately constant at .950 for all a_2 and b_2.

Fig. 2.8 has many of the same features as as Fig. 2.4. For $f_1 = 0$, a large highly inaccurate region results from a constant eigenvalue. However outside this region, and at higher values of f_1, the contour lines are virtually identical. The lower stability limit is

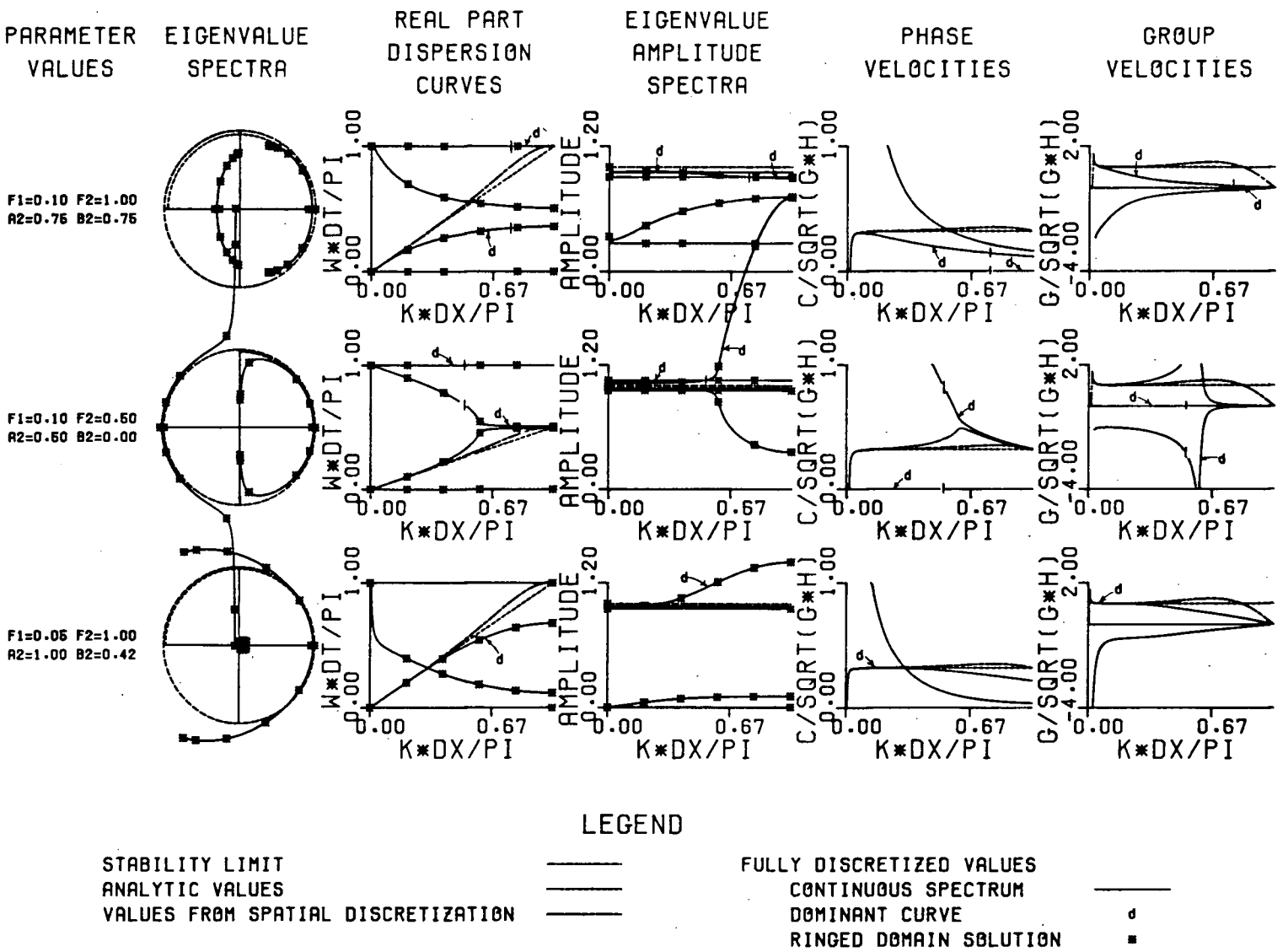

Fig. 2.6. Eigenvalue spectra, dispersion curves, and phase and group velocities for three mixed interpolation Galerkin FEMs.

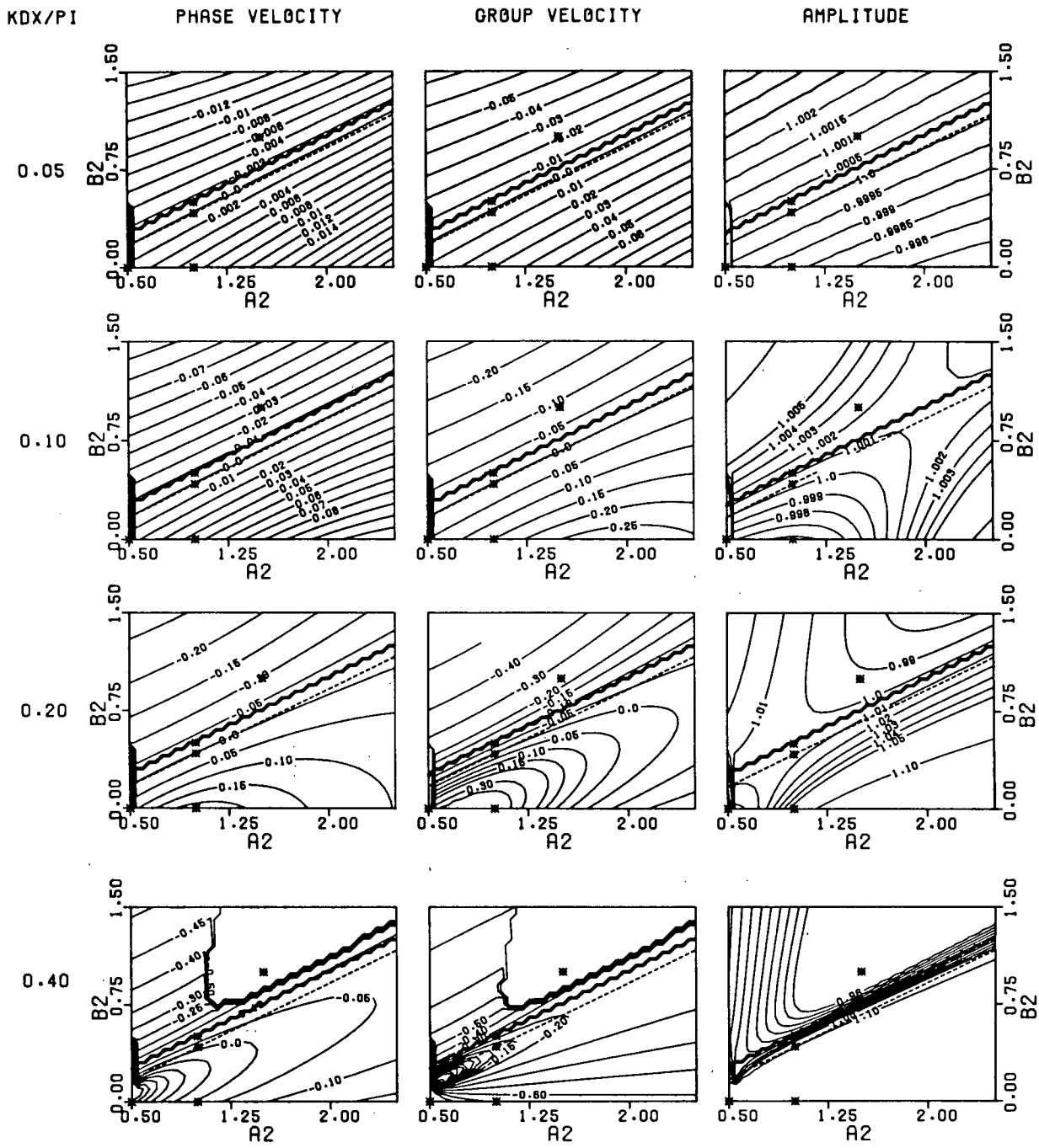

Fig. 2.7. Accuracy measure values for the mixed interpolation Galerkin FEM and $(f_1, f_2) = (0.1, 1.0)$.

seen to become less restrictive as f_1 increases in Fig. 2.4.

The accuracy measure contours for the FEMs with linear and linear/quadratic basis functions indicate that second order two-step methods with $b_2 = \frac{1}{2}a_2$ are most accurate. This result may be explained by re-expressing (2.7.8) as

$$\ln(\lambda) = \frac{a_2\lambda^2 + a_1\lambda + a_0}{b_2\lambda^2 + b_1\lambda + b_0}. \tag{2.7.11}$$

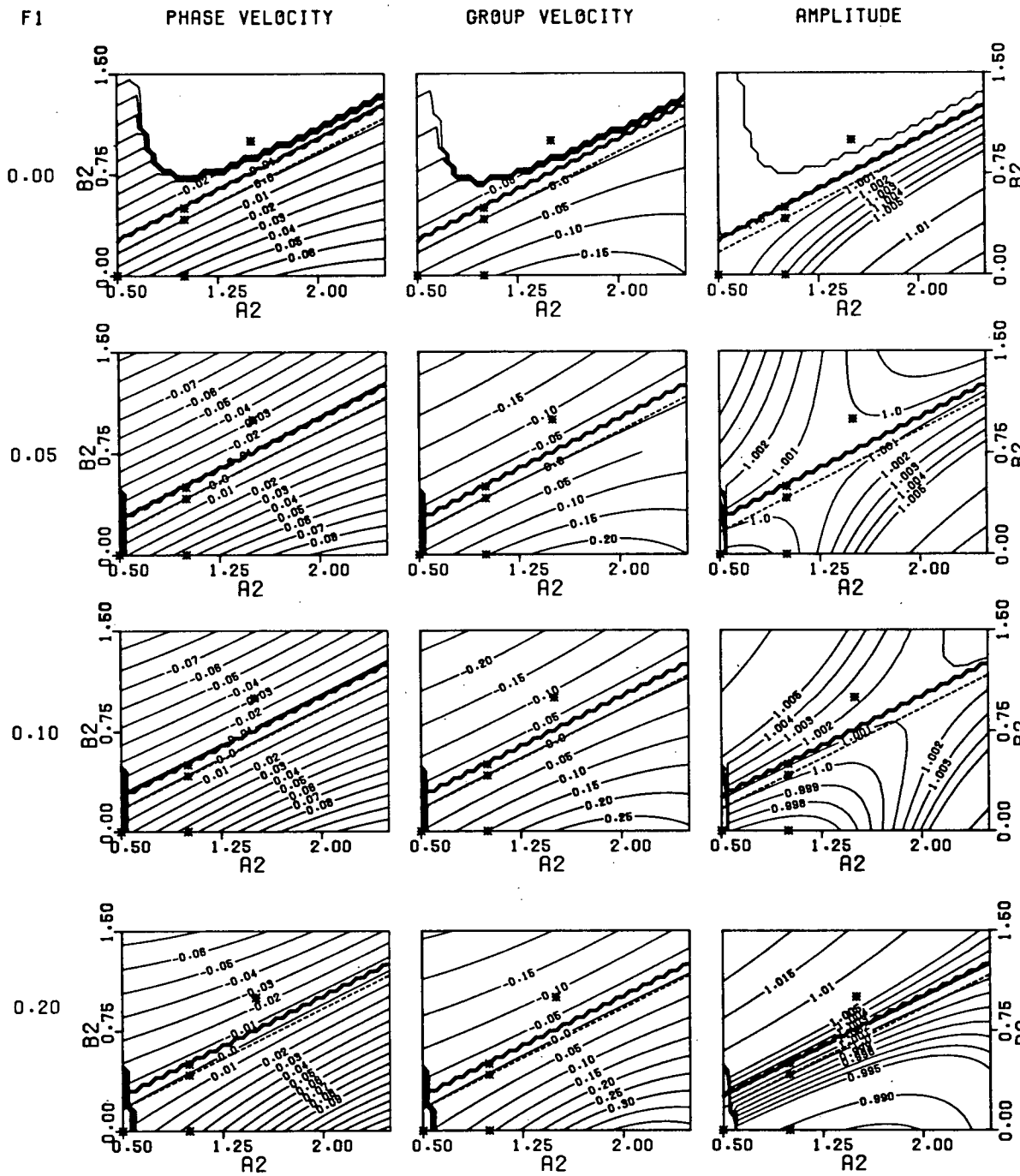

Fig.2.8. Accuracy measure values for the mixed interpolation Galerkin FEM and $(k\Delta x/\pi, f_2) = (0.1, 1.0)$.

If a_2, a_1, a_0, b_2, b_1, b_0 are then chosen to provide a good Padé approximation to $\ln(\lambda)$, travelling wave solutions should be accurately represented by the associated time discretization. Since $\ln(\lambda)$ has a branch point at $\lambda = 0$ and a branch cut along the negative real axis, it is likely that there will be different best approximants in different neighbourhoods of the complex λ plane.

If the two-step method is assumed to be second order accurate, (2.7.11) becomes

$$\ln(\lambda) = \frac{(\lambda-1)[a_2(\lambda-1)+1]}{(\lambda-1)[b_2(\lambda-1)+a_2]+\frac{1}{2}(\lambda+1)}. \tag{2.7.12}$$

When $Re(z) \geq 0$ and $z \neq 0$, $\ln(z)$ has the series expansion [Ab65, page 68]

$$\ln(z) = 2\left[\left(\frac{z-1}{z+1}\right)+\frac{1}{3}\left(\frac{z-1}{z+1}\right)^3+\frac{1}{5}\left(\frac{z-1}{z+1}\right)^5+\cdots\right]. \tag{2.7.13}$$

When $b_2 = \frac{1}{2}a_2$, (2.7.12) reduces to the first term in (2.7.13). Consequently, $b_2 = \frac{1}{2}a_2$ is the best second order Padé approximant to $\ln(z)$ when $Re(z) \geq 0$.

Unfortunately, numerical eigenvalues are not restricted to the nonnegative half plane (e.g., see Fig. 2.2 and Fig. 2.6). However, it is more important to represent long waves accurately. Long waves have small values of $k\Delta x$ and $\omega\Delta t$, and are thus associated with eigenvalues near 1.0. Hence for long waves, the multistep methods suggested by the Padé approximant should be most accurate.

The accuracy of the mixed interpolation FEM, and the FEM with linear basis functions, are quite similar when each uses the same second order two-step method, and $k\Delta x$ is small. This should not be surprising. Fig. 2.1 shows that for long waves, C and G for D3 and D7a) are quite similar. Hence the choice of a spatial discretization has made little difference to the numerical accuracy. However, the accuracy measure contour diagrams show that the choice of a particular time-stepping method can greatly affect the accuracy of the fully discretized equations. So the accuracy of long waves is much more dependent on the choice of the time-stepping method than the choice of either D3 or D7.

Unfortunately, a model can not be restricted to long wavelengths. Short waves can be generated by roundoff errors and may eventually grow to the extent that they severely contaminate the longer waves. Both the linear/linear and linear/quadratic spatial discretizations permit the accumulation of $2\Delta x$ waves, and both have been reported (see Section 1.5) to have difficulties with short wave oscillations. Avoiding the growth of short waves should therefore be a major consideration in choosing a spatial discretization. In this regard, the *wave equation* model proposed by Lynch and Gray [Gr77b,Ly79] holds promise. It will be examined in Chapter 3.

2.8 Verification of the Accuracy Measures

In this section, the previous accuracy measure analysis is compared with an analysis of the truncation errors, and validated with several numerical tests. Depth, Δx, and Δt were constant throughout each test and the additional complication of boundary conditions was avoided by choosing a ring as the test domain. All tests were initial value problems where the propagation characteristics of one or two progressive waves were studied as they travelled around the ring. Numerical solutions were obtained with the FEM that combines D3 with a second order two-step method.

Two series of tests were made. The first was designed for checking only amplitude and phase velocity. It was characterized by initial conditions which were spatially sampled values of a travelling wave (as given by (2.2.9)) with wavelength equal to the ring circumference.

Five test problems were selected, each with f_1, f_2, and $k\Delta x/\pi$ values corresponding to one of the plots in Fig. 2.3, 2.4, or the counterpart to Fig. 2.4 with $f_2 = 0.5$. Wavelength and depth were chosen so that the resultant problem would be realistic for semi-diurnal tides along a one dimensional continental shelf.

Each test problem was run for approximately ten periods and solved with ten different second order two-step methods. If the numerical solution remained stable, a spectral analysis of the $z(x,t)$ and $u(x,t)$ values over the ring was first used to determine if the original travelling wave had dispersed into other wavelengths. As expected for linear equations, this never occurred. The amplitude and phase lag for the wave were then calculated and compared to the analytic result. 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. From these model values, ratios were formed as in (2.6.1) and compared to the accuracy measure values.

The ten second order two-step methods were loosely selected upon the following criteria:

i) representation of most regions in the domain $\frac{1}{2} \leq a_2 \leq \frac{5}{2}$, $0 \leq b_2 \leq \frac{3}{2}$,

ii) inclusion of some well known methods,

iii) inclusion of some expected unstable methods (i.e. those for which $b_2 < \frac{1}{2}a_2$),

iv) inclusion of some methods with small truncation errors.

The chosen methods are listed in Table III and shown in Fig. 2.9.

TABLE III

Second Order Two-Step Methods Used in the Numerical Tests

Parameter Value \ Method Name	Crank–Nicolson	Adams–Moulton					Gear Stiffly Stable			Leapfrog
a_2	1.0	0.75	1.0	2.0	0.6	2.5	1.5	1.0	0.5	0.5
b_2	0.5	0.75	0.417	0.917	0.3	1.25	1.0	1.25	1.0	0.0

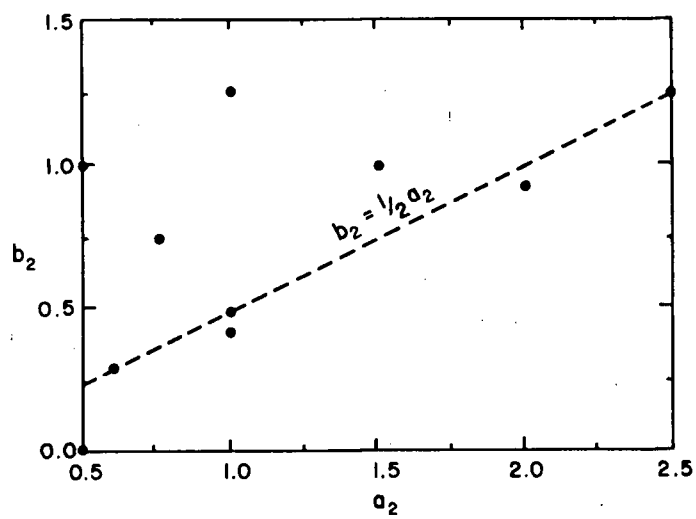

Fig. 2.9. (a_2, b_2) coordinates of the second order two-step methods used in the numerical tests.

The truncation error which arises when the continuity equation is solved numerically

by combining a second order two-step method with D3 is

$$\Delta x\Delta t^3\frac{\partial^2}{\partial t^2}\left(1+\frac{\Delta x^2}{6}\frac{\partial^2}{\partial x^2}\right)\left[\frac{1}{6}\frac{\partial z}{\partial t}+\frac{1}{2}\left(\frac{1}{2}-a_2+2b_2\right)h\frac{\partial u}{\partial x}\right]$$
$$+\Delta x\Delta t^4\left(\frac{2a_2-1}{24}\right)\frac{\partial^3}{\partial t^3}\left(1+\frac{\Delta x^2}{6}\frac{\partial^2}{\partial x^2}\right)\left[\frac{\partial z}{\partial t}+2h\frac{\partial u}{\partial x}\right] \qquad (2.8.1)$$
$$+O(\Delta x^5)O(\Delta t)+O(\Delta t^5)O(\Delta x),$$

where $z=z(x,t)$ and $u=u(x,t)$ are the true solutions to (2.2.1). The first term in (2.8.1) becomes zero for methods whose parameterization satisfies (2.4.5). With Milne's method, the second term also becomes zero. The test methods denoted by $(a_2,b_2)=(1.0,.417)$ and (2.0,.917) have smaller error coefficients than the others. In the subsequent discussion they will be referred to as M3 and M4 respectively.

Results for the five test problems are given in Table IV. Initial conditions at times 0 and Δt were specified exactly. A run was judged unstable when the absolute value of the first elevation point became greater than ten times the initial amplitude, ς_0, of 1.0. Only the Adams-Moulton and leapfrog methods became unstable and only the latter was unstable for all tests. Instability can occur even though $|\lambda|\leq 1.$ for the wavelength of the initial travelling wave. During the numerical computations round-off errors produce signals at all wavelengths. So if $|\lambda|>1.$ for any $k\Delta x$, this signal will grow without bound and eventually dominate the initial wave.

Table IV shows that the dispersion analysis and test model results are very close. In most cases, differences in the amplitude changes and non-dimensional phase velocities occurred in the fifth digit. Consequently, accuracy measures calculated from the test models were virtually the same as those from the dispersion analysis. In fact, only for the second problem and the method $(a_2,b_2)=(0.5,1.0)$ are the discrepancies as large as 1%.

Two sets of analytic solutions are shown in Table IV. In each numerical test, analytic values were calculated at every time step and the resultant time series was analysed in the same manner as the numerical time series. Results are shown in the last row and should be identical to the purely analytic values in the preceding row. Differences between the last two rows are therefore an indication of inaccuracies arising from the least squares analysis.

The relative performance of M3 and M4 varied with each test. With tests 4 and 5 they

TABLE IV

Results for the First Series of Numerical Tests.

		Problem Number and Parameter Values														
		1			2			3			4			5		
		f_1	f_2	$\frac{k\Delta x}{\pi}$	f_1	f_2	$\frac{k\Delta x}{\pi}$	f_1	f_2	$\frac{k\Delta x}{\pi}$	f_1	f_2	$\frac{k\Delta x}{\pi}$	f_1	f_2	$\frac{k\Delta x}{\pi}$
		.10	1.0	.1	.10	1.0	.4	.00	1.0	.1	.05	.5	.1	.20	.5	.1
Two-step method parameters (a_2, b_2)	Source of results	$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$		$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$		$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$		$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$		$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$	
(1.0,0.5)	analysis	.95234	.97999		.96444	.88055		1.0000	.99184		.98765	.99478		.95148	.94677	
	model	.95234	.98000		.96445	.88055		1.0000	.99184		.98765	.99478		.95148	.94677	
(.75,.75)	analysis	.95613	.94836		.94865	.65918		.99925	.95831		.98793	.98591		.95257	.94122	
	model	.95613	.94836		.94906	.65931		.99925	.95831		.98793	.98591		.95257	.94122	
(1.0,.417)	analysis	.95157	.98782		1.02658	.95197		1.0004	.99976		.98760	.99681		.95124	.94808	
	model	unstable			1.02658[a]	.95197		unstable			.98760	.99681		.95124	.94808	
(2.0,.917)	analysis	.95234	.98782		1.01216	.90152		1.0010	.99843		.98765	.99681		.95127	.94837	
	model	.95234	.98782		1.01210[a]	.90158		1.0010[a]	.99843		.98765	.99681		.95127	.94837	
(0.6,0.3)	analysis	.95234	.97999		.96444	.88055		1.0000	.99184		.98765	.99478		.95148	.94677	
	model	.95234	.98000		.96480	.88046		1.0000	.99184		.98765	.99478		.95148	.94677	
(2.5,1.25)	analysis	.95234	.97999		.96444	.88055		1.0000	.99184		.98765	.99478		.95148	.94677	
	model	.95234	.98000		.96440	.88080		1.0000	.99184		.98765	.99478		.95148	.94677	
(1.5,1.0)	analysis	.95346	.95726		.87190	.75200		.99805	.97066		.98773	.98875		.95216	.94241	
	model	.95346	.95726		.87191	.75201		.99805	.97066		.98773	.98875		.95216	.94241	
(1.0,1.25)	analysis	.95810	.91886		.92298	.56281		.99747	.92954		.98809	.97721		.95327	.93506	
	model	.95810	.91886		.92257	.56259		.99747	.92954		.98809	.97721		.95327	.93506	
(0.5,1.0)	analysis	.96039	.92224		.98897	.56060		1.0000	.92857		.98829	.97751		.95365	.93632	
	model	.96039	.92221		.98127	.56505		1.0009	.92830		.98829	.97750		.95365	.93632	
(0.5,0.0)	analysis	1.05397	8.9977		1.96666	1.30461		1.0000	1.01717		1.01274	18.999		1.05184	19.0497	
	model	unstable			unstable			unstable			.99379[a]	.98044		unstable		
Analytic solution	analysis	.95123	.98725		.95123	.99921		1.0000	1.0000		.98758	.99683		.95123	.94799	
	model	.95123	.98725		.95123	.99921		1.0000	1.0000		.98758	.99683		.95123	.94799	

[a] Going unstable but does not satisfy instability criterion.

most accurately represented phase velocity and amplitude decay. With tests 2 and 3, they had accurate velocities but were unstable. With test 1, M4 was most accurate while M3 was unstable. The truncation error analysis therefore predicted the high accuracy, but did not foresee the potential stability problems. This is to be expected.

Fig. 2.10 shows the $z(x,t)$ and $u(x,t)$ profiles around the ring domain for test problem 1 when solved analytically and with the Gear method. The wave is moving leftward and the numerical solution is seen to be too slow (by 3% as calculated from values in Table IV). After 42 time steps, this translates to a phase discrepancy of 21.9° between the numerical

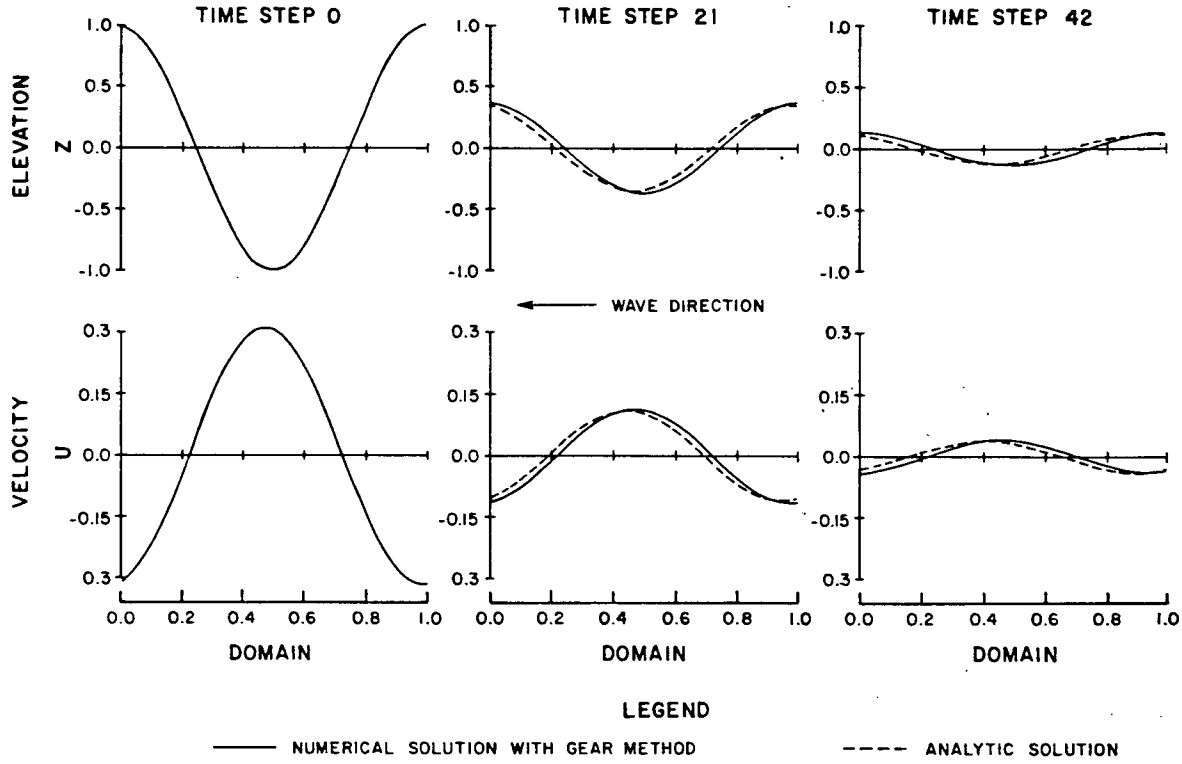

Fig. 2.10. Elevation and velocity profiles for problem 1 in the first series of numerical tests.

and analytic solutions. It is also evident that the numerical amplitude is not decaying as quickly as it should. An error of 0.234% in $|\lambda|$ (from Table IV) in this case compounds to an amplitude error of 10% after 42 time steps.

The second series of tests is similar to the first but permits a check of the group velocity calculations. Two travelling waves of equal amplitude but different wavelength were now initially specified on the ring domain. As time progresses, their combined effect is a short wavelength carrier wave moving inside and at a different speed than a long wavelength envelope (e.g., see Fig. 2.12). Algebraically, this is seen [Br60] by considering two close frequency/wavenumber coordinates, (ω_1, k_1) and (ω_2, k_2), on a dispersion curve as shown in Fig. 2.11.

Defining

$$\begin{aligned} \omega_0 &= \tfrac{1}{2}(\omega_1 + \omega_2), \qquad & k_0 &= \tfrac{1}{2}(k_1 + k_2), \\ \Delta\omega &= \tfrac{1}{2}|\omega_2 - \omega_1|, \qquad & \Delta k &= \tfrac{1}{2}|k_2 - k_1|, \end{aligned} \tag{2.8.2}$$

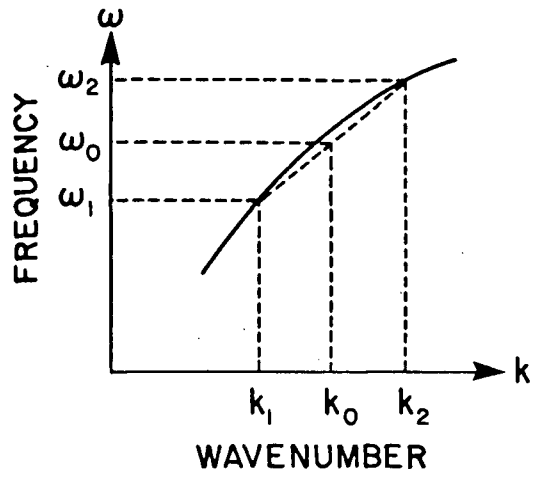

Fig. 2.11. A sample dispersion curve.

the combined effect of two equal amplitude progressive waves at these frequencies is then

$$A\cos(\omega_1 t - k_1 x) + A\cos(\omega_2 t - k_2 x) = 2A\cos(\omega_0 t - k_0 x)\cos(\Delta k x - \Delta\omega t). \qquad (2.8.3)$$

This represents an envelope with wavenumber Δk and a carrier wave with wavenumber k_0. As k_1 and k_2 approach k_0, the speeds of the envelope and carrier waves approximate the group and phase velocity respectively since

$$C(k_0) = \frac{\omega(k_0)}{k_0} = \lim_{k_1,k_2 \to k_0}\left(\frac{\omega_0}{k_0}\right) \qquad (2.8.4a)$$

and

$$G(k_0) = \frac{\partial\omega(k_0)}{\partial k} = \lim_{k_1,k_2 \to k_0}\left(\frac{\Delta\omega}{\Delta k}\right). \qquad (2.8.4b)$$

So if k_1 and k_2 are sufficiently close, speeds of the envelope and carrier waves are approximately $G(k_0)$ and $C(k_0)$ respectively.

Numerical tests to measure the group velocity using the preceding approach have an additional complication. Since the eigenvalue amplitudes for wavenumbers k_1 and k_2 are generally not the same, the two waves do not decay (or grow) at the same rate. So even though they may have equal amplitudes initially, after one step, there is a slight difference. In order that (2.8.3) be a reasonable representation of the two waves, all numerical tests were run for only a few time steps.

In all numerical experiments, wavelengths L_1 and L_2 of the two travelling waves were chosen so that the ring circumference, L, was one lobe of the envelope (as in Fig. 2.12) and L/L_1 and L/L_2 were both integer valued. Consequently, for this series of tests the parameter values f_1, f_2, and $k\Delta x/\pi$ for the six selected test problems could only approximate

those for one of the accuracy measure plots in Fig. 2.3, 2.4, or the counterpart to Fig. 2.4 with $f_2 = 0.5$.

Assuming the $z(x,t)$ and $u(x,t)$ profiles around the domain can be approximated at any time step by

$$A\cos(k_0 x - \phi_1) \cdot B\cos(\Delta k x - \phi_2), \tag{2.8.5}$$

where k_0 and Δk are specified, the parameters A, B, ϕ_1, and ϕ_2 then characterize the wave packet. Specifically, if k_1 and k_2 are sufficiently close, then AB is the amplitude of the envelope and $(\phi_1/k_0 t)$ and $(\phi_2/\Delta k t)$ respectively approximate the phase and group velocity. Values for these parameters were calculated from nonlinear least squares fits to the $z(x,t)$ and $u(x,t)$ profiles. In all cases, velocities and amplitude changes were the same for both variables.

All tests were for only ten time steps with the initial conditions at times 0 and Δt specified exactly. The same ten second order two-step methods were tested in this series as before.

Results for these numerical tests are presented in Table V. Dispersion analysis and test values are not as close as before but due to the several approximations involved, this is expected. Comparisons between tabulated results with the same f_1, f_2, and $k\Delta x/\pi$ values (e.g., test 1 in the first series and test 2 in the second) provide an estimate of the error associated with these approximations. In all cases, increasing the number of grid points in the domain would decrease $k\Delta x$ and reduce this error. In each numerical test, analytic values were calculated at every time step and the resultant time series was analysed in the same manner as the numerical time series. Results are shown in the last row of Table V. Differences between the last two rows are an indication of inaccuracies arising from the nonlinear least squares analysis.

For most tests, discrepancies between the dispersion analysis and test model estimates of the phase velocity, group velocity, and eigenvalue amplitude were less than 1%. The two-step method with the poorest correspondence was $(a_2, b_2) = (0.5, 1.0)$. This was also poorest for the first series of tests and is because the parasitic eigenvalue is only slightly

TABLE V

Results for the Second Series of Numerical Tests

Problem Number and Parameter Values	f_1	f_2	$\frac{k_0 \Delta x}{\pi}$	$\frac{\Delta k \Delta x}{\pi}$
1	.10	1.0	.208	.042
2	.10	1.0	.104	.021
3	.00	1.0	.104	.021
4	.05	.5	.104	.021
5	.20	.5	.104	.021
6	.10	1.0	.367	.033

Two-step method parameters (a_2, b_2)	Source of results	1 $\lvert\lambda\rvert$	1 $\frac{C}{(gh)^{1/2}}$	1 $\frac{G}{(gh)^{1/2}}$	2 $\lvert\lambda\rvert$	2 $\frac{C}{(gh)^{1/2}}$	2 $\frac{G}{(gh)^{1/2}}$	3 $\lvert\lambda\rvert$	3 $\frac{C}{(gh)^{1/2}}$	3 $\frac{G}{(gh)^{1/2}}$	4 $\lvert\lambda\rvert$	4 $\frac{C}{(gh)^{1/2}}$	4 $\frac{G}{(gh)^{1/2}}$	5 $\lvert\lambda\rvert$	5 $\frac{C}{(gh)^{1/2}}$	5 $\frac{G}{(gh)^{1/2}}$	6 $\lvert\lambda\rvert$	6 $\frac{C}{(gh)^{1/2}}$	6 $\frac{G}{(gh)^{1/2}}$
(1.0,0.5)	analysis	.956	.963	.902	.952	.980	.986	1.000	.991	.974	.988	.995	.996	.952	.951	1.044	.963	.898	.710
	model	.956	.959	.901	.952	.979	.987	1.000	.990	.973	.988	.994	.996	.952	.948	1.047	.963	.896	.710
(.75,.75)	analysis	.961	.851	.632	.956	.946	.883	.999	.955	.873	.988	.985	.967	.953	.944	1.018	.952	.690	.347
	model	.959	.844	.640	.956	.941	.884	.999	.951	.874	.988	.984	.967	.953	.941	1.020	.946	.694	.365
(1.0,.417)	analysis	.959	.996	.991	.952	.989	1.013	1.000	1.000	.999	.988	.997	1.003	.951	.952	1.051	1.009	.966	.825
	model	.960	.995	.989	.952	.988	1.014	1.000	1.000	.999	.988	.997	1.003	.951	.950	1.053	1.011	.963	.823
(2.0,.917)	analysis	.965	.983	.935	.953	.989	1.009	1.001	.998	.992	.988	.997	1.002	.951	.953	1.051	1.003	.920	.725
	model	.967	.979	.934	.953	.988	1.010	1.001	.998	.992	.988	.997	1.003	.951	.950	1.054	1.007	.916	.714
(0.6,0.3)	analysis	.956	.963	.902	.952	.980	.986	1.000	.991	.974	.988	.995	.996	.952	.951	1.044	.963	.898	.710
	model	.957	.960	.902	.952	.979	.988	1.000	.990	.974	.988	.995	.996	.952	.948	1.047	.968	.896	.703
(2.5,1.25)	analysis	.956	.963	.902	.952	.980	.986	1.000	.991	.974	.988	.995	.996	.952	.951	1.044	.963	.898	.710
	model	.957	.959	.896	.953	.980	.988	1.000	.990	.974	.988	.995	.997	.952	.948	1.048	.965	.891	.695
(1.5,1.0)	analysis	.942	.892	.743	.953	.956	.916	.998	.968	.912	.988	.988	.977	.952	.946	1.024	.887	.776	.510
	model	.942	.884	.739	.953	.953	.917	.998	.966	.911	.988	.987	.977	.952	.943	1.027	.887	.771	.503
(1.0,1.25)	analysis	.955	.777	.499	.958	.914	.800	.997	.924	.795	.988	.976	.940	.954	.937	.990	.929	.594	.239
	model	.952	.765	.497	.958	.907	.798	.997	.918	.794	.988	.973	.939	.954	.933	.992	.918	.590	.240
(0.5,1.0)	analysis	.976	.778	.493	.961	.918	.805	1.000	.923	.791	.988	.976	.941	.954	.939	.995	.988	.592	.235
	model	.973	.773	.529	.959	.909	.797	.998	.917	.776	.988	.976	.946	.954	.935	.998	.972	.613	.288
(0.5,0.0)	analysis	.936[a]	1.082	1.307	.949[a]	1.005	1.064	1.000[a]	1.019	1.058	.987[a]	1.001	1.013	.951[a]	.955	1.059	1.694	1.442	-.311
	model	.910	1.073	1.228	.949	1.007	1.077	1.000	1.021	1.059	.987	1.001	1.014	.951	.952	1.059	unstable		
Analytic solution	analysis	.951	.997	1.003	.951	.988	1.012	1.000	1.000	1.000	.988	.997	1.003	.951	.952	1.050	.951	.999	1.001
	model	.951	.997	1.003	.951	.988	1.012	1.000	1.000	1.000	.988	.997	1.003	.951	.950	1.053	.951	.999	1.001

[a] Calculated from the sub-dominant eigenvalue.

smaller than the principal eigenvalue. Many time steps are therefore required before the energy assigned to the parasitic solution by the initial conditions becomes insignificant. In fact, were it not for round-off errors and initial conditions which are, in varying degrees, inconsistent with each numerical method, the results from the first series of tests would be exactly the same as those predicted by the principal numerical eigenvalue.

Throughout the second series of tests, M3 and M4 most accurately represented the phase and group velocity. In fact, only for tests 1 and 6 were their amplitude decay factors not the most accurate. These two tests have the highest $k\Delta x$ values thereby suggesting that difficulties with these methods arise with shorter wavelengths. Plots similar to those of Fig. 2.2 confirm this. The $|\lambda|$ values for test 6 indicate future instability. Although those for test 1 suggest stability, Fig. 2.3 indicates magnitudes greater than 1. at other wavelengths. Hence eventual instability can be expected here also.

Due to the shortness of the tests, instability (judged as before) occurred only once. Had the runs been longer, dominant eigenvalues with magnitudes greater than 1.0 would have caused other numerical solutions to become unstable.

Fig. 2.12 illustrates the results of solving test problem 6 with the Crank Nicolson method. Both travelling waves are again moving leftward and decaying at the rate of 4% per time step. The phase velocity is 19% larger than the group velocity causing the carrier wave to move leftward inside the envelope. Analytic values are also shown and after 10 time steps have the following features relative to the numerical solution:

i) an envelope amplitude which is about 11% smaller,

ii) a carrier wave which is about 61° further advanced,

iii) an envelope which is about 16° further advanced.

Even though the numerical group velocity error is larger than the numerical phase velocity (29% vs. 10% from Table IV), the envelope has less phase error after 10 time steps because its frequency is smaller by a factor of 11.

The results of both sets of numerical tests validate the accuracy measure calculations of Section 2.6. Only for the method $(a_2, b_2) = (.5, 1.0)$ were there notable discrepancies

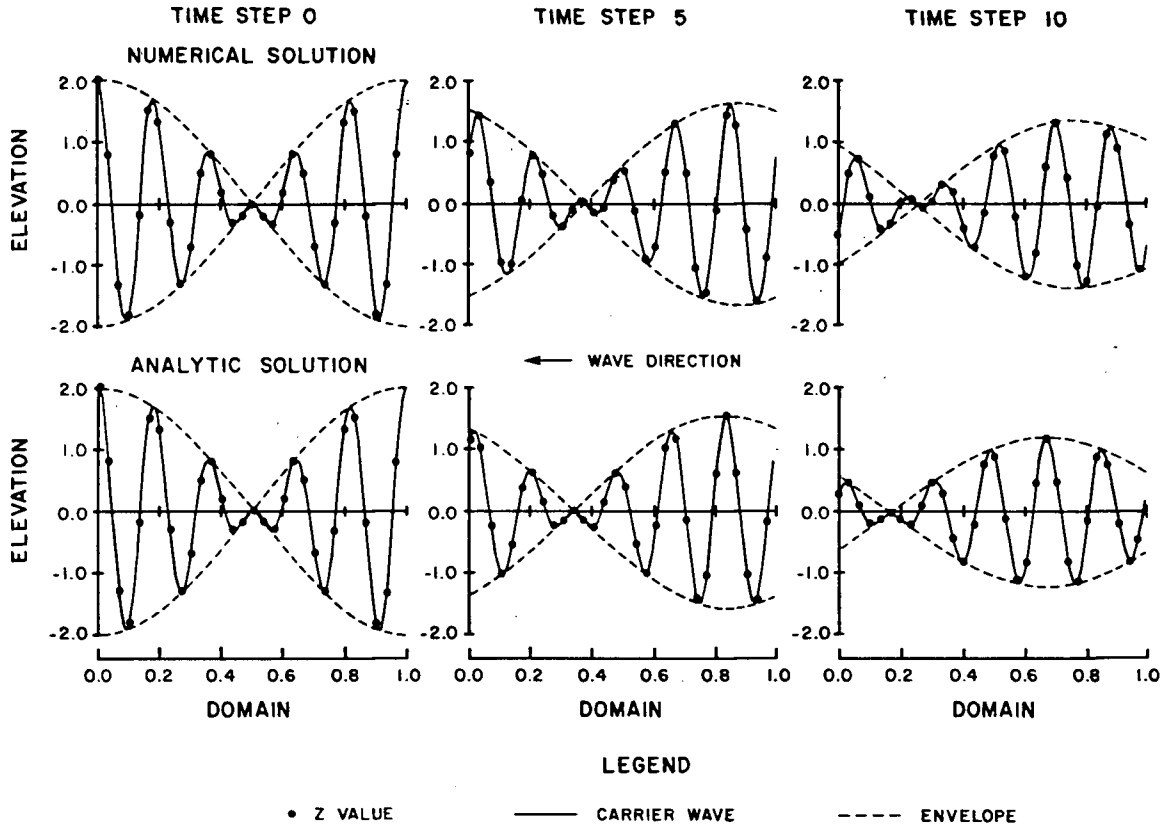

Fig. 2.12. Numerical and analytic elevation profiles for problem 6 in the second series of numerical tests.

between the test results and the accuracy measure calculations. These can be attributed to the fact that the spurious and principal numerical eigenvalues had virtually the same magnitude. Hence, over the test period, neither one dominated the other.

The performance of methods M3 and M4 confirms the high accuracy predicted by their truncation errors. An investigation of their absolute stability regions could be expected to predict the instability. The relatively good performance of methods in the subset $b_2 = \frac{1}{2}a_2$ is also substantiated by (2.8.1). They all have the same error constant. In fact, for each constant c, all methods related by

$$b_2 = \tfrac{1}{2}a_2 + c \tag{2.8.6}$$

have the same truncation error. This explains the general tendency toward contour lines of this slope in Fig. 2.3 and 2.4, and further validates the analysis of Section 2.6.

2.9 Summary and Conclusions

Here is a summary of some highlights of this chapter.

In Section 2.3, several spatial discretizations were examined for the accuracy of their phase and group velocities. Each of the four most accurate were shown to have drawbacks which could affect their performance or implementation in two dimensions.

In Section 2.5, the class of second order two-step methods was combined with the particular spatial discretization, a Galerkin FEM with piecewise linear basis functions. The concepts of dominant dispersion relationship, dominant phase and group velocity, and dominant or favoured wavenumber were defined and illustrated. It was shown that the same dispersion relationship may not be dominant for all wavenumbers, and the dominant dispersion relationship may be multivalued at some points.

In Section 2.6, three accuracy measure functions were defined to facilitate the search for an optimally accurate two-step method. It was shown that the most accurate methods for wave amplitude, phase velocity, and group velocity may not coincide. In particular, it was demonstrated that the best method for phase velocity may not be best for group velocity, and vice versa. Furthermore, a method which most accurately represents either velocity may be unstable. In general, the choice of an optimally accurate method depends on f_1, f_2, $k\Delta x$, and the relative importance of amplitude, phase velocity, and group velocity.

In Section 2.7 the accuracy analysis was repeated for the Galerkin FEM that approximates z with piecewise linear basis functions, and u with piecewise quadratics. The higher order basis functions were seen to result in six numerical solutions, four of which were spurious. However, the accuracy measure values were found to be very similar to those for the Galerkin FEM with linear basis functions. The relatively high accuracy of two-step methods satisfying $b_2 = \frac{1}{2}a_2$ was also explained in terms of Padé approximations to $\ln(\lambda)$.

In Section 2.8, numerical tests validated the phase velocity, group velocity, and amplitude decay factors which were calculated in Section 2.6. Only in cases where the spurious and principal eigenvalues had approximately the same magnitude were there significant discrepancies between the analysis and test results. Truncation errors were also calculated

and correctly predicted the most accurate methods, when they remained stable.

For a Galerkin FEM with piecewise linear basis functions and a Galerkin FEM with linear and quadratic basis functions, the most accurate and stable two-step methods are characterized by $b_2 = \frac{1}{2}a_2$. Crank Nicolson ($a_2 = 1$) is the best among these since it has no spurious eigenvalues. However it is implicit and may be expensive with respect to computing time. Crank Nicolson may not be the best time-stepping for all spatial discretizations. Due to second derivatives in their continuity equation, a variation of the linear two-step methods introduced in Section 2.4 is required for the Gray and Lynch *wave equation* method. As will be seen in Chapter 3, an accuracy measure analysis of this approach shows that the Crank Nicolson analogue is not the most accurate. Furthermore, in this case, the most accurate methods are independent of wavenumber.

Again it must be emphasized that in the preceding analysis, accuracy was the only consideration for determining a good method. In two dimensional problems this is no longer a sufficient criterion. Storage requirements and computational costs are now at least as important and may necessitate the use of a method which is less accurate but more economical.

Travelling wave solutions of the form (2.2.3) do not exist when the depth in (2.2.1a) is assumed non-constant. With a forcing frequency ω, solutions can now be expected to have the form

$$\begin{pmatrix} z(x,t) \\ u(x,t) \end{pmatrix} = \begin{pmatrix} \varsigma_0(x) \\ \mu_0(x) \end{pmatrix} e^{-i\omega t} \tag{2.9.1}$$

where $\varsigma_0(x)$ and $\mu_0(x)$ are complex functions representing the spatial amplitude and phase variations. Their precise nature will depend on $h(x)$. For example, in the absence of friction and with a linear depth, Lamb [La32] shows that

$$|\varsigma_0(x)| = cJ_0(2(\kappa x)^{1/2}), \tag{2.9.2a}$$

where

$$h(x) = h_0 x \tag{2.9.2b}$$

$$\kappa = \omega^2/gh_0 \tag{2.9.2c}$$

and c is some constant. Lynch and Gray [Ly78b] extend this result to the case $h(x) = h_0 x^n$ for integer n, and include linear friction as in (2.2.1b).

In general, the depth dependency of the solution will be such that waves of constant frequency will have their wavelength decrease and their amplitude increase as they enter shallow water. Phase and group velocity will also become depth dependent. This same behaviour can be expected in a numerical model, although it may not be accurately represented. Unfortunately, the model will not differentiate between spurious and principal waves; all will become shorter and grow. Although it may not be the case analytically, it is possible that with particular numerical schemes and depth variations, shorter waves will grow more quickly. This could be disastrous, for if the short waves are spurious, they may eventually contaminate the numerical solution.

For some depth variations, it is possible to forecast the rapid growth of short waves with an analysis similar to that of Section 2.5. Since amplitude is now a function of both space and time, spatial growth curves (with $k\Delta x$ along the abscissa) are required in addition to the temporal growth curves of Fig. 2.2. In fact, it may be necessary to produce these curves for several depth characteristics (e.g., ratios of depth gradient to depth). Numerical schemes which favour high wavenumbers could then be expected to exhibit rapid growth of short waves and should be avoided.

In the absence of nonlinear terms, short waves may be generated numerically by boundary conditions, an interface, round-off errors, or arise naturally such as through a transition from deep to shallow water. Intuitively, this last source can be controlled by maintaining the same sampling rate per wavelength everywhere in the model. This requires a constant $k\Delta x$ for each wave as it moves thoughout the model domain. Therefore any transition from deep to shallow water would not correspond to a rightward shift on a spatial amplitude growth curve which has $k\Delta x$ as the abscissa and which may favour short wavenumbers. Using the dispersion relationship for constant depth (2.2.8), a first approximation to uniform sampling is attained by choosing Δx proportional to $(h(x))^{1/2}$. This choice has further

appeal. Stability conditions when they arise are frequently in the form

$$\Delta t \leq c\Delta x/(h(x))^{1/2} \tag{2.9.3}$$

for some constant c. Therefore a constant value for $\Delta x/(h(x))^{1/2}$ implies that deep regions of the model where there may be little variation in the numerical solution, are not dictating the largest possible time step.

However choosing Δx proportional to $(h(x))^{1/2}$ will not affect the generation of short waves due to round-off errors, boundary conditions, or an interface. It may only control their subsequent wavenumber transitions. If an amplitude growth curve shows that these waves will grow faster than the desired longer waves, numerical difficulties can be expected.

3. THE 'WAVE EQUATION' MODEL

3.1 Introduction

Recently Gray and Lynch [Gr77b,Ly78,Ly79] introduced a FEM 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 PDE. The revised system of equations is then solved with a Galerkin FEM, piecewise linear basis functions, and centered 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 chapter, the one dimensional, linearized version of both the *wave equation* method (WEM), and the *lumped wave equation* method (LWEM) are studied using the dispersion analysis developed in Chapter 2. Section 3.2 calculates the spatially discretized equations for both the WEM and LWEM. It also shows that the principal dispersion relationships for each discretization is identical to one of those listed in Table II. Section 3.3 introduces the time-stepping methods proposed by Gray and Lynch, and specifies the numerical eigenvalues and stability restrictions for both fully discretized schemes. Section 3.4 shows that these time-stepping methods are a subset of the second order two-step methods associated with the *wave equation* ODEs. Section 3.5 applies these generalized two-step methods to both the lumped and unlumped spatial discretizations, and determines the particular methods with the most accurate wave propagation, and wave amplitude growth (or decay) characteristics. These results are confirmed with asymptotic analyses in Section 3.6, and numerical tests in Section 3.7. Section 3.8 summarizes and briefly discusses the results.

3.2 An Analysis of the Spatial Discretizations

Gray and Lynch [Ly79] calculate their *wave equation* by differentiating (1.2.1a) with respect to time. Substitutions from the momentum equations (1.2.1b) and (1.2.1c) are then used to eliminate the velocity time derivatives. The one dimensional, linearized, constant depth *wave equation* is

$$\frac{\partial^2 z}{\partial t^2} + \tau \frac{\partial z}{\partial t} - gh \frac{\partial^2 z}{\partial x^2} = 0. \qquad (3.2.1a)$$

This PDE is solved in combination with the momentum equation

$$\frac{\partial u}{\partial t} + \tau u + g \frac{\partial z}{\partial x} = 0. \qquad (3.2.1b)$$

When nontrivial travelling wave solutions of the form (2.2.3) are assumed, the characteristic equation for (3.2.1) is the product of (2.2.6) and $(-i\omega + \tau)$. Replacing (2.2.1a) with (3.2.1a) has therefore produced an additional dispersion relationship whose associated solution is a stationary wave that decays in time when $\tau > 0$. This solution will arise from components within the initial conditions that do not satisfy the continuity equation (2.2.1a).

Gray and Lynch solve (3.2.1) with a Galerkin FEM, piecewise linear basis functions, and centered time-stepping. Applying the Galerkin condition to the $(\partial^2 z/\partial x^2)$ term necessitates an integration by parts. This is called a weak form of the Galerkin condition [Ly79]. When depth and Δx are assumed constant, and z_j, u_j are the time dependent variable values at any node j away from boundaries, the spatially discretized system of ODEs for the WEM are

$$\left(\frac{\partial^2}{\partial t^2} + \tau \frac{\partial}{\partial t}\right)\left(\tfrac{1}{6} z_{j-1} + \tfrac{2}{3} z_j + \tfrac{1}{6} z_{j+1}\right) - \frac{gh}{\Delta x^2}(z_{j+1} - 2z_j + z_{j-1}) = 0 \qquad (3.2.2a)$$

$$\left(\frac{\partial}{\partial t} + \tau\right)\left(\tfrac{1}{6} u_{j-1} + \tfrac{2}{3} u_j + \tfrac{1}{6} u_{j+1}\right) + \frac{g}{2\Delta x}(z_{j+1} - z_{j-1}) = 0. \qquad (3.2.2b)$$

Assuming the travelling wave solutions (2.3.2), the associated dispersion relationships are

$$\omega = -i\tau \qquad (3.2.3a)$$

and

$$\omega = -i\tfrac{1}{2}\tau \pm \left[\frac{6gh}{\Delta x^2}\left(\frac{1 - \cos k\Delta x}{2 + \cos k\Delta x}\right) - (\tfrac{1}{2}\tau)^2\right]^{1/2}. \qquad (3.2.3b)$$

The latter expression is identical to the dispersion relationship for the mixed interpolation approach recently described by Williams and Zienkiewicz [Wi81b]. They solve (2.2.1) with a Galerkin FEM, piecewise linear basis functions for approximating $u(x,t)$, and piecewise constant functions for $z(x,t)$. In Section 2.3, this FEM is denoted as D6.

The 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 $z(x,t)$ 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 (3.2.3b) is shown in Fig. 2.1 for the case $\tau = 0$. For the one dimensional equations, this discretization 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 inter-element nodes [Wa83]. 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.

In the practical application of FEMs, numerical quadrature must be used to evaluate the coefficients in equations such as (3.2.2). Traditionally [Ly79], the method of choice has been Gaussian quadrature. It gives the greatest accuracy for a fixed number of integration points [Zi77]. However, Gray and van Genuchten [Gr78] have pointed out that other types of quadrature formulas, when tailored to specific element types, can provide economical alternatives to Gaussian quadrature. In particular, any element-quadrature combination for which the integration points exactly coincide with the nodes, has the effect of lumping the WEM [Ly79]. If such a quadrature is used instead of the exact integration which produced (3.2.2), the coefficient matrices associated with the $(\partial/\partial t)$ and $(\partial^2/\partial t^2)$ terms become diagonal and the resultant spatial discretization becomes the LWEM.

The system of ODEs for the LWEM is

$$\left(\frac{\partial^2}{\partial t^2}+\tau\frac{\partial}{\partial t}\right)z_j-\frac{gh}{\Delta x^2}(z_{j+1}-2z_j+z_{j-1})=0 \tag{3.2.4a}$$

$$\left(\frac{\partial}{\partial t}+\tau\right)u_j+\frac{g}{2\Delta x}(z_{j+1}-z_{j-1})=0. \tag{3.2.4b}$$

The associated dispersion relationships are

$$\omega=-i\tau \tag{3.2.5a}$$

$$\omega=-i\tfrac{1}{2}\tau\pm\left[\frac{2gh}{\Delta x^2}(1-\cos k\Delta x)-(\tfrac{1}{2}\tau)^2\right]^{1/2}$$

$$=-i\tfrac{1}{2}\tau\pm\left[\frac{4gh}{\Delta x^2}\sin^2(\tfrac{1}{2}k\Delta x)-(\tfrac{1}{2}\tau)^2\right]^{1/2}. \tag{3.2.5b}$$

The latter expression is identical to the dispersion relationship for a centered FDM with spatial staggering of the $z(x,t)$ and $u(x,t)$ variables. In Section 2.3, this FDM is denoted as D2. When $\tau=0$, the relative accuracy of its phase and group velocities is shown in Fig. 2.1. Notice that phase velocities for the LWEM are too slow, whereas those for the WEM are too fast.

3.3 Numerical Eigenvalues for the WEM and LWEM

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

$$\frac{\partial}{\partial t}z_j(n\Delta t)\simeq\frac{z_j^{n+1}-z_j^{n-1}}{2\Delta t} \tag{3.3.1a}$$

$$\frac{\partial^2}{\partial t^2}z_j(n\Delta t)\simeq\frac{z_j^{n+1}-2z_j^n+z_j^{n-1}}{\Delta t^2} \tag{3.3.1b}$$

$$z_j(n\Delta t)\simeq\tfrac{1}{2}\theta(z_j^{n+1}+z_j^{n-1})+(1-\theta)z_j^n. \tag{3.3.1c}$$

In order to relax the stability constraints, the friction term in the momentum equation is treated as in (3.3.1c), but with the separate weighting parameter α.

Applying these approximations to (3.2.2) and assuming the non-trivial travelling wave solutions (2.5.3) requires satisfaction of one of the following two quadratics:

$$\lambda^2(1+\alpha\tau\Delta t)+2\tau\Delta t(1-\alpha)\lambda-(1-\alpha\tau\Delta t)=0, \tag{3.3.2a}$$

$$\text{or} \quad \lambda^2[1+\tfrac{1}{2}(\theta E^2+\tau\Delta t)]+\lambda[-2+(1-\theta)E^2]+1+\tfrac{1}{2}(\theta E^2-\tau\Delta t)=0, \tag{3.3.2b}$$

$$\text{where} \quad E^2=6gh\left(\frac{\Delta t}{\Delta x}\right)^2\left(\frac{1-\cos k\Delta x}{2+\cos k\Delta x}\right). \tag{3.3.2c}$$

Constant depth, constant Δx and Δt are also assumed. λ is defined in (2.5.4). The product of these quadratics is the characteristic polynomial for the WEM.

The roots of (3.3.2) are

$$\lambda_{1,2}=\frac{-\tau\Delta t(1-\alpha)\pm[1+(\tau\Delta t)^2(1-2\alpha)]^{1/2}}{1+\alpha\tau\Delta t} \tag{3.3.3a}$$

$$\lambda_{3,4}=\frac{1-\frac{1}{2}(1-\theta)E^2\pm i[E^2+\frac{1}{4}(2\theta-1)E^4-(\frac{1}{2}\tau\Delta t)^2]^{1/2}}{1+\frac{1}{2}(\theta E^2+\tau\Delta t)}. \tag{3.3.3b}$$

λ_1 and λ_2 are parasitic and arise from the spatially discretized solution (3.2.3a). They are independent of wavenumber 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 non-zero, they represent progressive and retrogressive waves.

With constant depth and $\tau \geq 0$, a necessary condition for the stability of the WEM is $|\lambda| \leq 1$ for all eigenvalues (3.3.3). This condition translates to the following restrictions on θ and α:

i) for the parasitic roots,

$$\alpha \geq \tfrac{1}{2}; \tag{3.3.4a}$$

ii) for the propagating principal roots,

$$\theta \geq \frac{-\Delta x^2}{6gh\Delta t^2}; \tag{3.3.4b}$$

iii) for the non-propagating principal roots,

$$\theta \geq \frac{1}{2}\left(1-\frac{\Delta x^2}{3gh\Delta t^2}\right). \tag{3.3.4c}$$

With non-zero friction, (3.3.4b) can be made less restrictive. However, this is not essential since the constraint imposed by (3.3.4c) dominates. Conditions i) and iii) are given in [Ly79].

An analysis of the LWEM yields similar results. With the following substitution for (3.3.2c)

$$E^2 = 2gh\left(\frac{\Delta t}{\Delta x}\right)^2(1-\cos k\Delta x), \tag{3.3.5}$$

equations (3.3.2) and (3.3.3) again specify the characteristic polynomial and its roots. Stability of the parasitic root is again dictated by (3.3.4a) while the counterparts to (3.3.4b) and (3.3.4c) are

$$\theta \geq \frac{-\Delta x^2}{2gh\Delta t^2} \tag{3.3.6a}$$

$$\text{and} \qquad \theta \geq \frac{1}{2}\left(1-\frac{\Delta x^2}{gh\Delta t^2}\right), \tag{3.3.6b}$$

respectively. As with the WEM, condition (3.3.6b) overrides (3.3.6a).

3.4 Two-Step Methods for Solving the ODEs

A simple ODE corresponding to (3.2.2a) is

$$\frac{\partial^2 y}{\partial t^2} + \frac{\partial y}{\partial t} = f(y) \tag{3.4.1}$$

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

$$\begin{aligned} c_2y^{n+2} + c_1y^{n+1} + c_0y^n + \Delta t(a_2y^{n+2} + a_1y^{n+1} + a_0y^n) \\ = \Delta t^2(b_2f^{n+2} + b_1f^{n+1} + b_0f^n). \end{aligned} \tag{3.4.2}$$

When each term is expanded in a Taylor series about y^n or f^n, (3.4.2) becomes

$$\sum_{j=0}^{\infty} h_j\Delta t^j = 0 \tag{3.4.3a}$$

where

$$h_0 = (c_2 + c_1 + c_0)y \tag{3.4.3b}$$

$$h_1 = (2c_2 + c_1)y' + (a_2 + a_1 + a_0)y \tag{3.4.3c}$$

$$h_2 = (2c_2 + \tfrac{1}{2}c_1)y'' + (2a_2 + a_1)y' - (b_2 + b_1 + b_0)(y'' + y') \tag{3.4.3d}$$

$$h_3 = (\tfrac{4}{3}c_2 + \tfrac{1}{6}c_1)y''' + (2a_2 + \tfrac{1}{2}a_1)y'' - (2b_2 + b_1)(y''' + y'') \tag{3.4.3e}$$

$$h_4 = (\tfrac{2}{3}c_2 + \tfrac{1}{24}c_1)y'''' + (\tfrac{4}{3}a_2 + \tfrac{1}{6}a_1)y''' - (2b_2 + \tfrac{1}{2}b_1)(y'''' + y''') \qquad (3.4.3f)$$

and primed values denote true derivatives of y^n. For j^{th} order accuracy (i.e., the truncation error is $O(\Delta t^{j+1})$) of the two-step method, it is necessary that

$$h_0 = h_1 = h_2 = \cdots = h_j = 0. \qquad (3.4.4)$$

Requiring second order accuracy and assuming Gear's normalization (2.4.4) leaves only three coefficients to be specified freely. Choosing them to be a_2, b_2, and b_1, the others are

$$\begin{aligned} c_2 &= c_0 = 1, \quad & a_0 &= a_2 - 1, \quad & b_0 &= 1 - b_1 - b_2, \\ c_1 &= -2, & a_1 &= 1 - 2a_2. \end{aligned} \qquad (3.4.5)$$

Third order methods have the additional constraints

$$a_2 = \tfrac{1}{2}, \qquad b_0 = b_2, \qquad (3.4.6)$$

while fourth order accuracy is not possible since (3.4.5) and (3.4.6) are inconsistent with $h_4 = 0$.

For solving the simple ODE

$$\frac{\partial y}{\partial t} = f(y) + g(y) \qquad (3.4.7)$$

with the two-step method

$$\begin{aligned} &a_2 y^{n+2} + a_1 y^{n+1} + a_0 y^n \\ &\qquad = \Delta t(b_2 f^{n+2} + b_1 f^{n+1} + b_0 f^n + d_2 g^{n+2} + d_1 g^{n+1} + d_0 g^n), \end{aligned} \qquad (3.4.8)$$

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

$$\begin{aligned} a_0 &= a_2 - 1, \quad & b_0 &= \tfrac{1}{2} - a_2 + b_2, \quad & d_0 &= \tfrac{1}{2} - a_2 + d_2, \\ a_1 &= 1 - 2a_2, \quad & b_1 &= \tfrac{1}{2} + a_2 - 2b_2, \quad & d_1 &= \tfrac{1}{2} + a_2 - 2d_2. \end{aligned} \qquad (3.4.9)$$

In this case, third order methods also require

$$b_2 = d_2 = \tfrac{1}{2}a_2 - \tfrac{1}{12} \qquad (3.4.10)$$

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

The ODEs in (3.2.2) 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:

$$\begin{aligned} c_2 &= c_0 = 1, \quad & a_0 &= a_2 - 1, \quad & b_0 &= \tfrac{1}{2} - a_2 + b_2, \quad & d_0 &= \tfrac{1}{2} - a_2 + d_2, \\ c_1 &= -2, \quad & a_1 &= 1 - 2a_2, \quad & b_1 &= \tfrac{1}{2} + a_2 - 2b_2, \quad & d_1 &= \tfrac{1}{2} + a_2 - 2d_2. \end{aligned} \tag{3.4.11}$$

The particular case

$$a_2 = \tfrac{1}{2}, \qquad b_2 = \tfrac{1}{2}\theta, \qquad d_2 = \tfrac{1}{2}\alpha, \tag{3.4.12}$$

makes (3.4.2) third order accurate and is precisely the subset of time-stepping methods proposed by Lynch and Gray. Fourth order accuracy for (3.4.8) and third order accuracy for (3.4.2) 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 (3.3.4a) indicates that it will be unstable.

3.5 A Dispersion Analysis

In Chapter 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 (3.4.11). The methods proposed by Lynch and Gray are a subset of this class.

Define

$$\tilde{s}_j = \tfrac{1}{6}(s_{j-1} + 4s_j + s_{j+1}), \tag{3.5.1a}$$

$$\hat{s}_j = s_{j+1} - 2s_j + s_{j-1}, \tag{3.5.1b}$$

$$\Delta s_j = s_{j+1} - s_{j-1}, \tag{3.5.1c}$$

where s can be either z or u. Application of a second order two-step method to solve (3.2.2) then produces the following system of equations:

$$\tilde{z}_j^{n+2} - 2\tilde{z}_j^{n+1} + \tilde{z}_j^{n} + \tau\Delta t(a_2\tilde{z}_j^{n+2} + a_1\tilde{z}_j^{n+1} + a_0\tilde{z}_j^{n})$$

$$= gh\left(\frac{\Delta t}{\Delta x}\right)^2 (b_2\hat{z}_j^{n+2} + b_1\hat{z}_j^{n+1} + b_0\hat{z}_j^n), \qquad (3.5.2a)$$

$$a_2\tilde{u}_j^{n+2} + a_1\tilde{u}_j^{n+1} + a_0\tilde{u}_j^n = -\tau\Delta t(d_2\tilde{u}_j^{n+2} + d_1\tilde{u}_j^{n+1} + d_0\tilde{u}_j^n) + \frac{g\Delta t}{2\Delta x}(b_2\Delta z_j^{n+2} + b_1\Delta z_j^{n+1} + b_0\Delta z_j^n), \qquad (3.5.2b)$$

where the restrictions imposed by (3.4.11) are assumed but have not been included.

For non-trivial travelling wave solutions to (3.5.2) one of the following two quadratics must be satisfied:

$$a_2\lambda^2 + a_1\lambda + a_0 + \tau\Delta t(d_2\lambda^2 + d_1\lambda + d_0) = 0 \qquad (3.5.3a)$$

$$\lambda^2 - 2\lambda + 1 + \tau\Delta t(a_2\lambda^2 + a_1\lambda + a_0) + E^2(b_2\lambda^2 + b_1\lambda + b_0) = 0. \qquad (3.5.3b)$$

For the WEM, E^2 and λ are defined by (3.3.2c) and (2.5.4) respectively. For the LWEM, E^2 is defined by (3.3.5).

The root $\lambda = 1$ can be troublesome for it represents an undamped non-propagating wave. If energy is transferred to a wavenumber 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., [Wa83]). However it can be avoided in this case. From (3.4.11) and (3.5.3), 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 (3.5.3a) or (3.5.3b) only for certain values of (a_2, b_2, d_2). Therefore with non-zero friction and a judicious choice of these parameters, the generalized *wave equation* method given by (3.5.2) can avoid the troublesome accumulation of $2\Delta x$ waves.

Fig. 3.1 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 and f_2, as they were in Chapter 2.

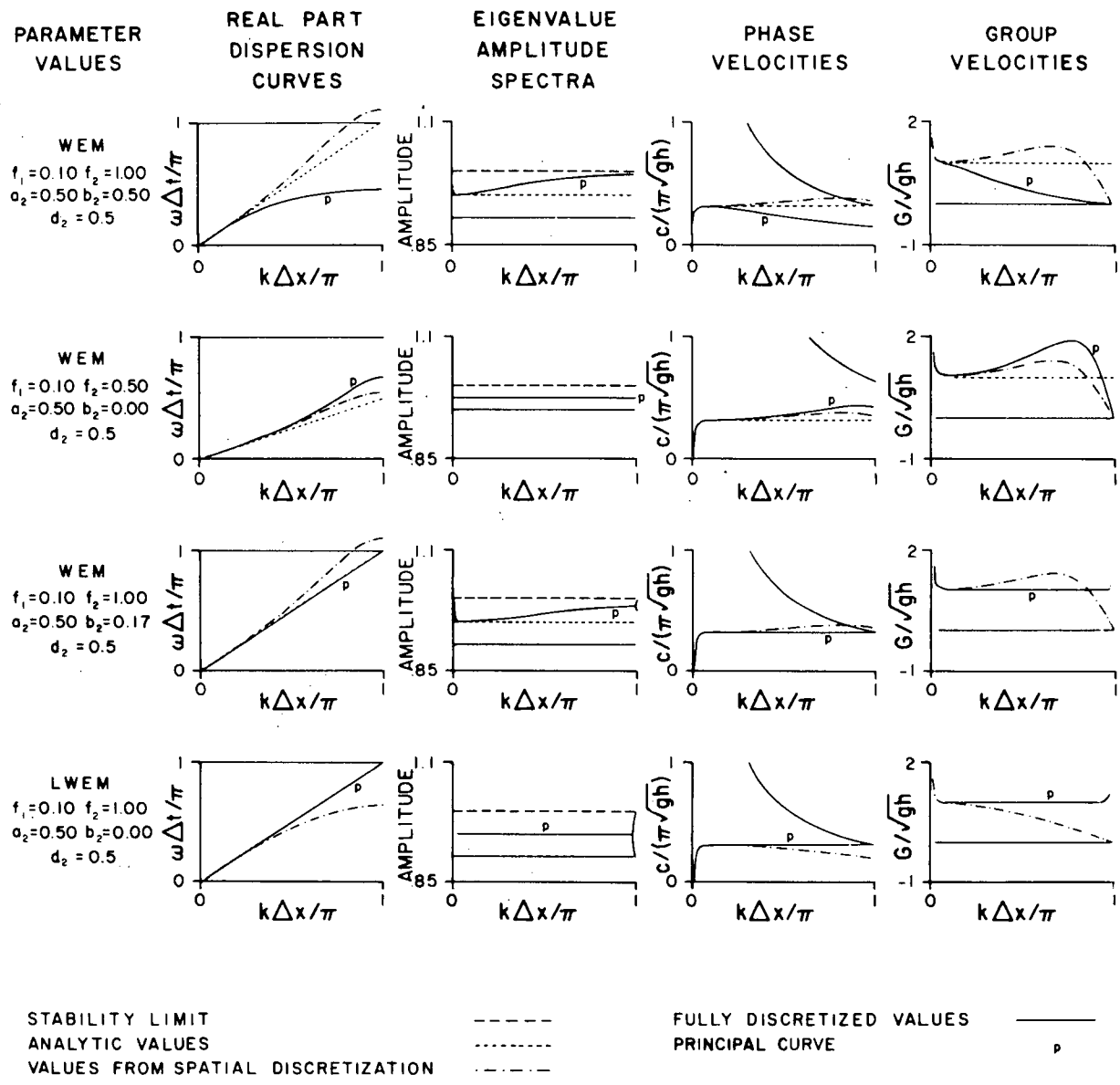

Fig. 3.1. Dispersion curves, eigenvalue amplitudes, and phase and group velocities for four *wave equation* models.

All four methods are from the subset proposed by Lynch and Gray (i.e., they satisfy (3.4.11) and (3.4.12)). The first three methods have not been lumped while the fourth has. Stability of the parasitic eigenvalue is insured 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 [Gr77b]. The third method produces fourth order accuracy for (3.4.8) and second order for (3.4.2). Its principal dispersion curve and phase and group velocities are seen to closely approximate the analytic values. 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 $\frac{1}{6}$, (3.4.2) would become third 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. 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.

Fig. 3.2 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 (3.5.3a) 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 $a_2 = \frac{1}{2}$. 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 $a_2 = \frac{1}{2}$. (Recall from Chapter 2 that 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 $k\Delta x/\pi$

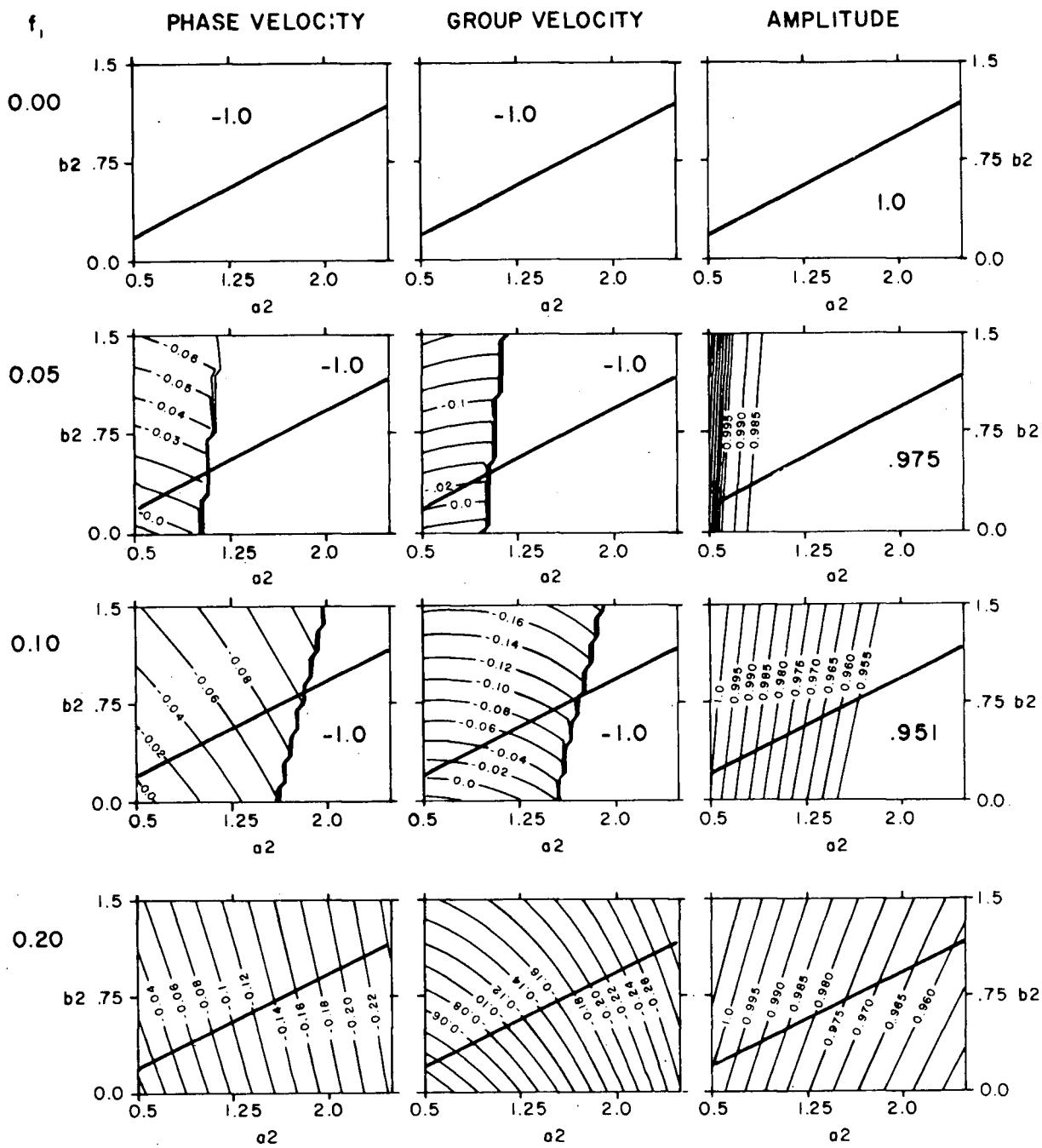

Fig. 3.2. Accuracy measure values for the WEM as functions of (a_2, b_2) for $f_2 = 1.0$, $d_2 = 0.5$, and $k\Delta x/\pi = 0.1$.

values 0.05, 0.2, and 0.4.

In light of the results in Section 3.4, greater accuracy with $a_2 = \frac{1}{2}$ is not surprising. It substantiates the desirability of third order accuracy for (3.4.2). 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 (3.4.12).

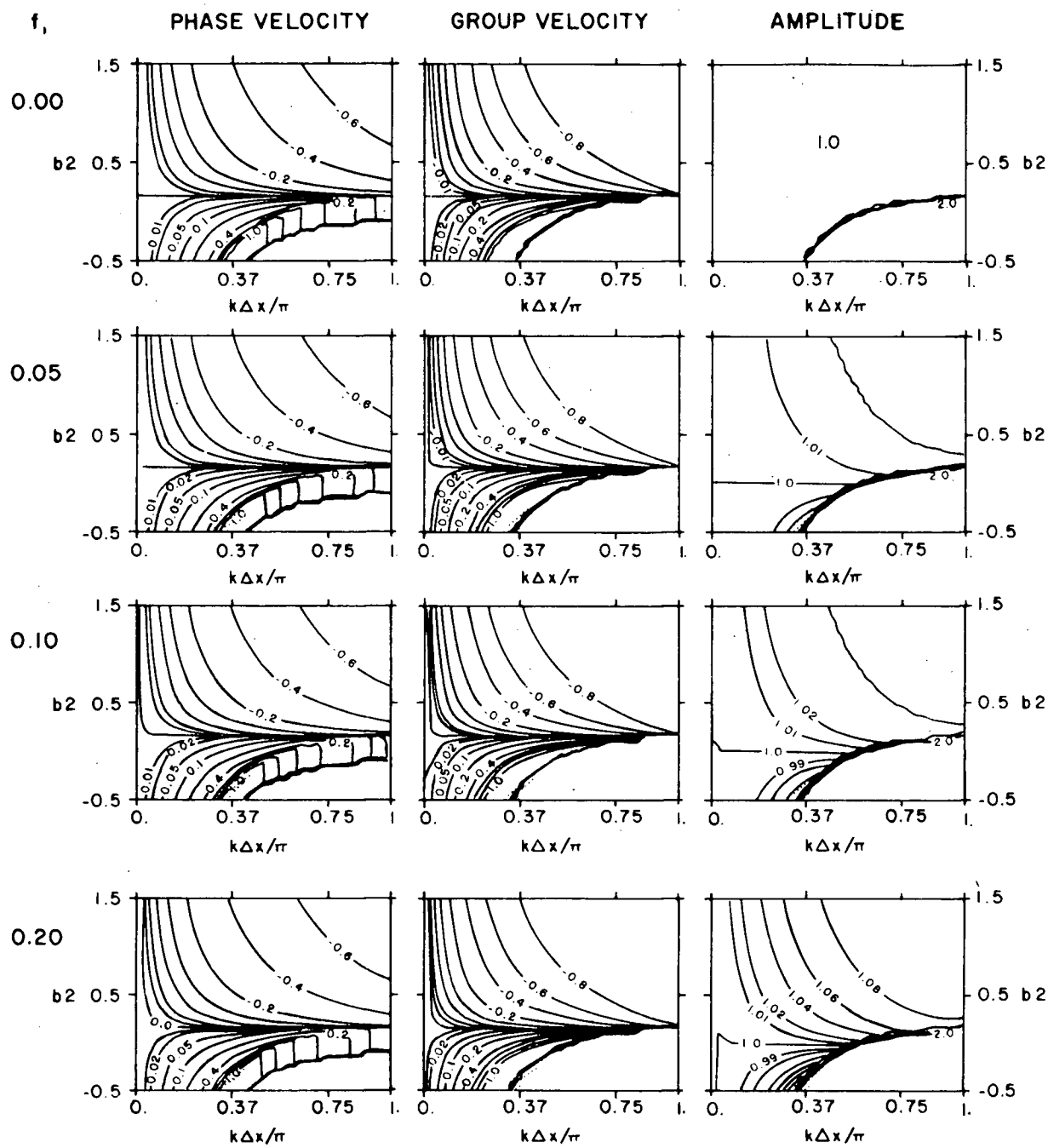

Fig. 3.3. Accuracy measure values for the WEM as functions of $(b_2, k\Delta x)$ for $a_2 = d_2 = 0.5$ and $f_2 = 1.0$.

Fig. 3.3 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 non-zero 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. 3.3. The first is that except for very small wavenumbers, 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 wavenumbers, $b_2 = 0$ produces optimal accuracy of wave amplitudes. It is also independent of $k\Delta x$ and f_1 , although for $f_1 = 0.0$, it does occur over a large region.

Fig. 3.4 is a similar plot with $f_2 = \frac{1}{2}$. 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 $b_2 = 0$.

Fig. 3.5 shows the accuracy measure contours for the LWEM with the same parameter values as Fig. 3.3. Fig. 3.5 and Fig. 3.3 are remarkably similar. The amplitude measures are virtually identical 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 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 3.7.

Combining the restriction $a_2 = \frac{1}{2}$ with (3.5.3) and (3.4.11) has the following implications for $2\Delta x$ waves. $\lambda = 1$ does not satisfy (3.5.3b) and only satisfies (3.5.3a) when $\tau = 0$. $\lambda = -1$ satisfies (3.5.3a) when either $d_2 = \frac{1}{4}$ or $\tau\Delta t = 0$, and satisfies (3.5.3b) when

$$b_2 = \tfrac{1}{4}(1 - (1/3f_2^2)) \qquad \text{for the WEM,} \tag{3.5.4a}$$

$$b_2 = \tfrac{1}{4}(1 - (1/f_2^2)) \qquad \text{for the LWEM.} \tag{3.5.4b}$$

Therefore with $d_2 > \frac{1}{4}$ ($\theta > \frac{1}{2}$) and $\tau \neq 0$, $2\Delta x$ waves should accumulate only when the

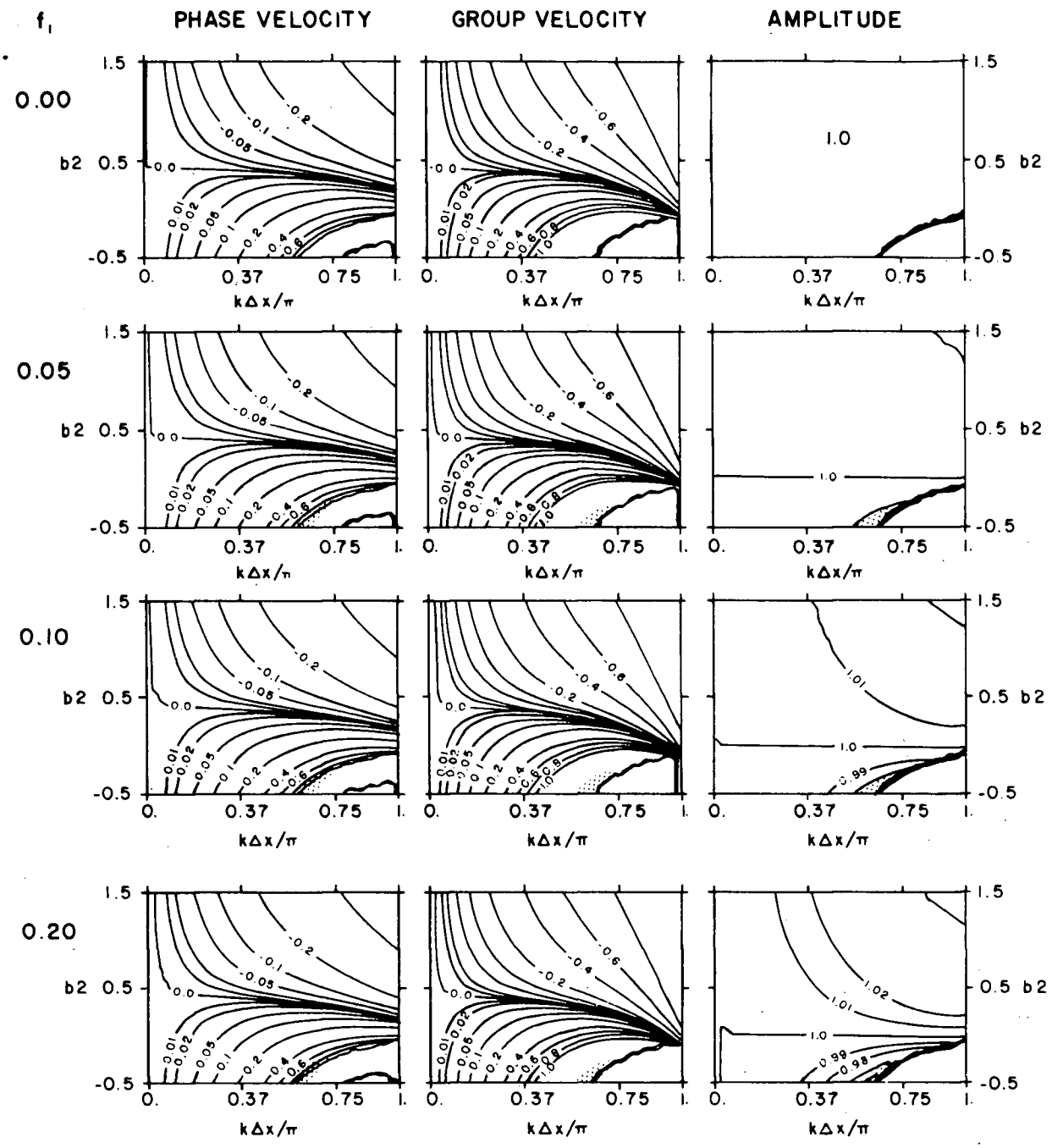

Fig. 3.4. Accuracy measure values for the WEM as functions of $(b_2, k\Delta x)$ for $a_2 = d_2 = 0.5$ and $f_2 = 0.5$.

time-stepping method parameter b_2, and the Courant number, f_2, are related by (3.5.4). Since $d_2 \geq \frac{1}{4}$ is required for stability, and bottom friction is usually included in a model, these first two restrictions are normally satisfied.

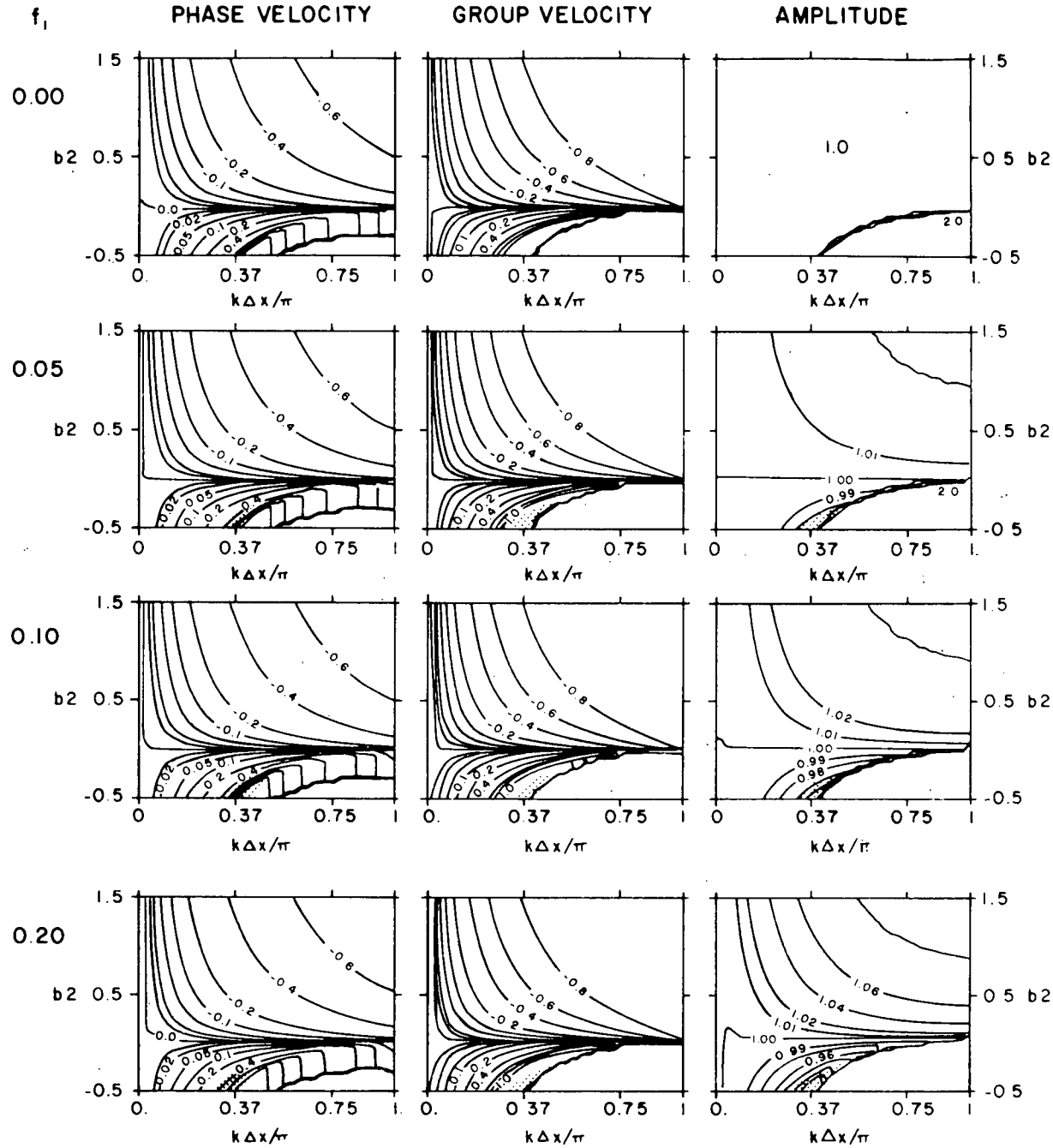

Fig. 3.5. Accuracy measure values for the LWEM as functions of $(b_2, k\Delta x)$ for $a_2 = d_2 = 0.5$ and $f_2 = 1.0$.

3.6 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., $k\Delta x/\pi \leq 0.1$), 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.

From the analytic dispersion relationship (2.2.8) it follows that

$$Re(\omega\Delta t)^2 = f_2^2(k\Delta x)^2 - (\tfrac{1}{2}\tau\Delta t)^2. \tag{3.6.1}$$

Assuming a non-zero imaginary part for the complex root

$$\lambda = re^{i\phi} \tag{3.6.2}$$

of the general quadratic

$$a\lambda^2 + b\lambda + c = 0 \tag{3.6.3}$$

implies

$$\phi = \arcsin \beta^{1/2} \tag{3.6.4a}$$

$$\text{where} \qquad \beta = 1 - (b^2/4ac). \tag{3.6.4b}$$

Provided $\beta < 1$, the associated power series expansion [Ab65] is

$$\phi = \beta^{1/2} + \tfrac{1}{6}\beta^{3/2} + \tfrac{3}{40}\beta^{5/2} + O(\beta^{7/2}). \tag{3.6.5}$$

Therefore

$$\phi^2 = \beta + \tfrac{1}{3}\beta^2 + \tfrac{8}{45}\beta^3 + O(\beta^4). \tag{3.6.6}$$

Applying these results to the quadratic for the principal eigenvalues arising from the WEM, (3.3.2b), yields

$$\beta = \frac{E^2 - E^4(\frac{1}{4} - \frac{1}{2}\theta) - (\frac{1}{2}\tau\Delta t)^2}{(1 + \frac{1}{2}\theta E^2)^2 - (\frac{1}{2}\tau\Delta t)^2}. \tag{3.6.7}$$

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

$$E^2 = f_2^2\xi^2\left(1 + \tfrac{1}{12}\xi^2 + \tfrac{1}{360}\xi^4\right) + O(\xi^8) \tag{3.6.8}$$

and a similar expansion for E^4 is

$$E^4 = f_2^4\xi^4(1 + \tfrac{1}{6}\xi^2) + O(\xi^8). \tag{3.6.9}$$

Substituting these expansions into (3.6.7) yields

$$\beta = f_2^2\xi^2 + f_2^4\xi^4[-\tfrac{1}{2}\theta + (1/12f_2^2) - \tfrac{1}{4}] + O(\xi^6) - (\tfrac{1}{2}\tau\Delta t)^2(1 + O(\xi^2) + O((\tau\Delta t)^2)) \tag{3.6.10}$$

and substituting (3.6.10) into (3.6.6) produces

$$\begin{aligned}\phi^2 =& f_2^2\xi^2 + f_2^4\xi^4[-\tfrac{1}{2}\theta + \tfrac{1}{12}(1 + (1/f_2^2))] + O(\xi^6) \\ &- (\tfrac{1}{2}\tau\Delta t)^2(1 + O(\xi^2) + O((\tau\Delta t)^2)).\end{aligned} \tag{3.6.11}$$

But in this case, $\phi = -\omega_r\Delta t$ where ω_r is the frequency arising from the principal numerical eigenvalue. Matching (3.6.11) with (3.6.1), it then follows that ω_r will be a good approximation to $Re(\omega)$ when

$$\theta = \frac{1}{6}\left(1 + \frac{1}{f_2^2}\right). \tag{3.6.12}$$

For the f_2 values 1 and $\frac{1}{2}$, (3.6.12) 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 Fig. 3.3 and 3.4.

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

$$E^2 = f_2^2\xi^2\left(1 - \tfrac{1}{12}\xi^2 + \tfrac{1}{360}\xi^4\right) + O(\xi^8) \tag{3.6.13}$$

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

$$\theta = \frac{1}{6}\left(1 - \frac{1}{f_2^2}\right). \tag{3.6.14}$$

Denoting the optimal parameter values of (3.6.12) and (3.6.14) by θ^*, both the lumped and unlumped versions of (3.6.11) can be re-expressed as

$$\begin{aligned}\phi^2 = f_2^2\xi^2 + \tfrac{1}{2}f_2^4\xi^4(\theta^* - \theta) + f_2^6\xi^6\left[\frac{1}{4}(\theta^* - \theta)^2 + \frac{1}{240}\left(1 - \frac{1}{f_2^4}\right)\right] \\ + O(\xi^8) + O((\tau\Delta t)^2).\end{aligned} \tag{3.6.15}$$

This explains the similar contour patterns in Fig. 3.3 and 3.5. 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 scheme 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

$$|\lambda| = e^{-\frac{1}{2}\tau\Delta t} \tag{3.6.16}$$

and its counterpart for a propagating principal numerical eigenvalue is

$$|\lambda| = \left(\frac{1+\frac{1}{2}\theta E^2 - \frac{1}{2}\tau\Delta t}{1+\frac{1}{2}\theta E^2 + \frac{1}{2}\tau\Delta t}\right)^{1/2}. \tag{3.6.17}$$

The time-stepping parameter value

$$\theta = 0 \tag{3.6.18}$$

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 (3.5.17) 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 Fig. 3.3, 3.4, and 3.5.

Fig. 3.6 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. $(\theta, f_2) = (\frac{1}{3}, 1)$ 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.

3.7 Numerical Tests

The analysis of Section 3.5 is now confirmed with numerical tests similar to the first

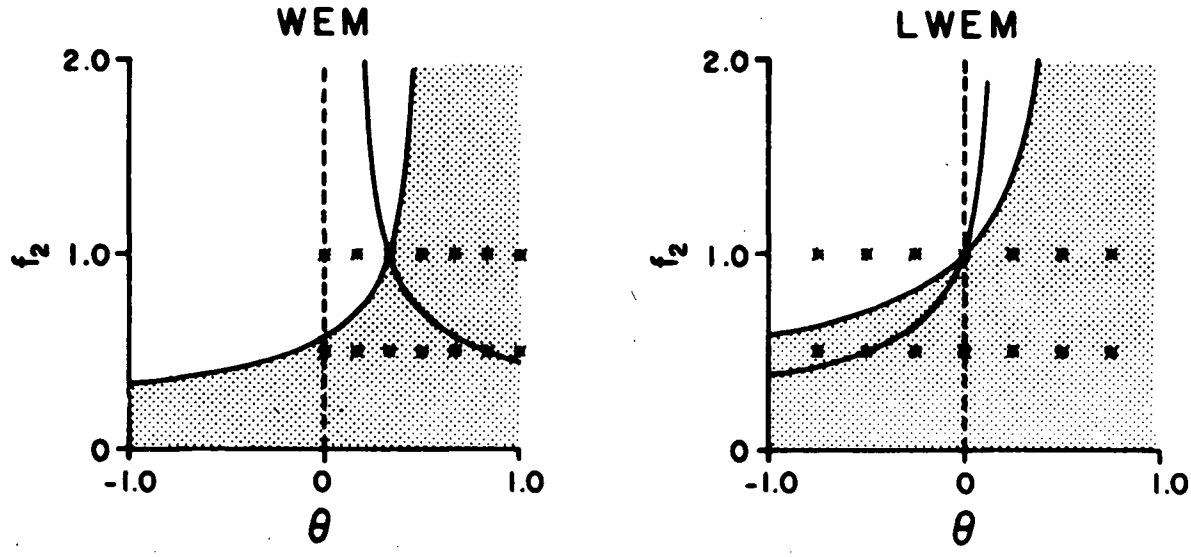

Fig. 3.6. Stability regions and lines of optimal accuracy for the LWEM and WEM. Asterisks denote methods used in the test problems, shaded areas denote stability, and the most accurate methods for wave propagation and wave amplitude decay are shown with solid and dotted lines lines respectively.

series reported in Section 2.8. 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 Section 2.8.

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. 3.3 or 3.4. Wavelength and depth were chosen so that the resultant problem would be realistic for semi-diurnal tides along a one dimensional continental shelf.

Each test problem was run for approximately ten 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. (θ, f_2) pairs for the test problems are shown with asterisks in Fig. 3.6. 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 VI. A run was judged unstable when the absolute value of the first elevation point became greater than ten times the initial amplitude. All unstable methods are predicted by (3.3.4). Methods which produce the most accurate representations of wave amplitude decay and phase velocity are designated. For all tests, they confirm the predictions in Fig. 3.3 and 3.4.

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

These same six problems were also solved with the LWEM. b_2 values for the time-stepping methods were now chosen as $-.375$, $-.25$, $-.125$, 0., .125, .25, and .375. They are illustrated in Fig. 3.6. As predicted by (3.3.6b), 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, 5 and 6, $b_2 = 0$ was most accurate for amplitude while $b_2 = -.25$ was most accurate for phase velocity. These results validate (3.6.18) and (3.6.14). As with the results in Table VI, the maximum discrepancy between the analysis and model results was less than 1%.

3.8 Summary and Conclusions

The preceding analysis has determined the following features of the one dimensional *wave equation* FEM:

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

TABLE VI

Numerical Test Results for the WEM

		Problem number and parameter values																	
		1			2			3			4			5			6		
		f_1	f_2	$\frac{k\Delta x}{\pi}$	f_1	f_2	$\frac{k\Delta x}{\pi}$	f_1	f_2	$\frac{k\Delta x}{\pi}$	f_1	f_2	$\frac{k\Delta x}{\pi}$	f_1	f_2	$\frac{k\Delta x}{\pi}$	f_1	f_2	$\frac{k\Delta x}{\pi}$
		.10	1.0	.4	.10	1.0	.1	.00	1.0	.2	.05	.5	.4	.05	.5	.1	.20	.5	.2
Two-step method parameter: b_2 =	Source of results	$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$		$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$		$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$		$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$		$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$		$\lvert\lambda\rvert$	$\frac{C}{(gh)^{1/2}}$	
0.0	analysis	.95119[a]	1.16840		.95119[a]	.99610		1.00000[a]	1.03464		.98758[a]	1.08719		.98758[a]	1.00203		.95119[a]	1.00882	
	model	unstable			unstable			unstable			.98681	1.08811		.98756	1.00205		.95118	1.00876	
0.08333	analysis	.95741	1.07407		.95158	.99205		1.00000[a]	1.01688		.98802	1.06663		.98760	1.00100		.95159	1.00462	
	model	unstable			unstable			unstable			.98771	1.06613		.98759	1.00101		.95159	1.00463	
0.16667	analysis	.96223	.99981[a]		.95197	.98806[a]		1.00000[a]	1.00000[a]		.98844	1.04719		.98763	.99997		.95199	1.00048	
	model	.96223	.99981		.95197	.98810		1.00000	1.00000		.98832	1.04754		.98762	.99997		.95199	1.00054	
0.25	analysis	.96607	.93929		.95235	.98411		1.00000[a]	.98394		.98882	1.02878		.98765	.99894		.95238	.99638	
	model	.96638	.93961		.95237	.98415		1.00015	.98397		.98849	1.02870		.98765	.99895		.95238	.99646	
0.33333	analysis	.96920	.88870		.95272	.98022		1.00000[a]	.96862		.98919	1.01131		.98768	.99792		.95276	.99233	
	model	.96921	.88872		.95276	.98028		1.00013	.96917		.98916	1.01119		.98768	.99793		.95276	.99242	
0.41667	analysis	.97180	.84555		.95309	.97636		1.00000[a]	.95400		.98952	.99470[a]		.98770	.99690[a]		.95313	.98834[a]	
	model	.97009	.84379		.95314	.97649		.99982	.95437		.98955	.99470		.98770	.99691		.95315	.98842	
0.5	analysis	.97399	.80818		.95345	.97256		1.00000[a]	.94003		.98984	.97889		.98773	.99588		.95351	.98439	
	model	.96840	.80720		.95350	.97276		1.00018	.94004		.99002	.97913		.98773	.99590		.95353	.98448	
Analytic solution	analysis	.95123	.99921		.95123	.98725		1.00000	1.00000		.98758	.99980		.98758	.99683		.95123	.98725	
	model	.95123	.99921		.95123	.98725		1.00000	1.00000		.98758	.99980		.98758	.99683		.95123	.98725	

[a] Most accurate value.

ii) a similarity of the LWEM to the FDM with spatial staggering of the variables;

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

iv) 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;

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

In particular the analysis indicates that an explicit ($b_2 = \theta = 0$) 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 [Ly79], 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 with $f_2 = 1$ also has a major disadvantage; it may have problems with $2\Delta x$ waves. This is evident from (3.5.4). With $f_2 = 1.0$ and the optimal values given by (3.6.12) or (3.6.14), $\lambda = -1$ 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. 3.1 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. 3.1 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

$$\Delta x = ch^{1/2} \tag{3.8.1}$$

for some constant c. Intuitively, this is not an unreasonable strategy. Constant frequency (e.g., tidal) waves have their wavenumbers 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 (2.2.8), a first approximation to uniform sampling is attained through (3.8.1). Such a choice also implies that the stability constraints (3.3.4c) and (3.3.6b) 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 wavenumbers, 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 the magnitudes depend on $\tau\Delta t$. As demonstrated in Fig. 3.3, 3.4, and 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

$$d_2 = \tfrac{1}{2}\alpha = \tfrac{1}{4}(1 + 1/(\tau\Delta t)^2) \tag{3.8.2}$$

and have the value

$$|\lambda| = \left|\frac{1-\tau\Delta t}{1+\tau\Delta t}\right|. \tag{3.8.3}$$

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 crucial. Provided $\tau\Delta t < 1$, $d_2 = \frac{1}{2}$ is a reasonable compromise. When $\theta \geq 0$, this choice guarantees a smaller parasitic eigenvalue for both the WEM and LWEM. And the same dominance is insured for negative θ provided

$$\theta \geq \frac{-1}{12 f_2^2} \tag{3.8.4a}$$

and

$$\theta \geq \frac{-1}{4 f_2^2} \tag{3.8.4b}$$

for the unlumped and lumped approaches respectively. In fact, Fig. 3.3, 3.4, and 3.5 suggest that these conditions may be overly restrictive.

In summary, the best wave equation method is the LWEM with $\theta = 0$ and (in most cases) $\alpha = 1$. The spatial discretization should be chosen so that everywhere in the model domain, f_2 equals, or is slightly less than 1.

4. TWO DIMENSIONAL DISPERSION ANALYSES

4.1 Introduction

The preceding two chapters analyzed finite element (and finite difference) solutions to the one dimensional shallow water equations. This chapter extends the analysis to two dimensions, where the advantages and disadvantages of FEMs are more apparent. Since FEMs permit grids of variable size, shape, and orientation, they are usually able to provide a better approximation of the spatial domain than FDMs. Specifically, better coastline fits are possible at model boundaries and grid size can be reduced in regions where the solution is expected to require greater resolution. However most FEMs are not cost competitive with explicit FDMs. Their initialization costs and bookkeeping are more extensive, and more computations are usually required at each time step. For many applications this extra cost outweighs the advantages.

Some FEMs are able to significantly reduce their computations by *lumping* the matrix involved in the equation to be solved at each time step. Lumping refers to the procedure of replacing a matrix row with a new row whose diagonal entry is the sum of all entries in the old row, and whose other entries are all zero. Although lumping is sometimes applied with little justification, it can result from the numerical quadrature that is used to calculate matrix entries [Gr78,Zi77]. When a lumped matrix is combined with explicit time-stepping, the matrix equation at each time step is diagonal and trivially solved. Generally, lumping also reduces accuracy [Mu82,St73]. However in Chapter 3, it was shown that the explicit LWEM need not be less accurate than the WEM. In this chapter, the same result is shown to extend to two dimensions for a particular configuration of triangular elements.

In this chapter, three FEMs for solving the two dimensional shallow water equations are compared with a traditional explicit FDM. The comparison is based on accuracy and

cost. Accuracy is measured by comparing numerical and analytic plane wave solutions, as was done in Chapters 2 and 3. Cost is measured as the number of computations per unit of real time and per unit of model area. It ignores the model initialization. Using these measures, two of the FEMs are found to be cost competitive, and as accurate as the chosen explicit FDM.

This chapter is divided into nine sections. Section 4.2 specifies the two dimensional linearized shallow water equations and their plane wave solutions. It also redefines the concepts of phase and group velocity for two dimensions.

Section 4.3 investigates the Richardson-Sielecki [He69,He76] finite difference scheme. It is a popular and successful explicit technique whose dispersion relationship has been previously calculated [Me76,He81].

Section 4.4 studies the Galerkin FEM with piecewise linear basis functions and Crank-Nicolson time-stepping. The analysis is restricted to two combinations of six triangular elements. Since accuracy is dependent on the shape and configuration of the elements, this examination is meant to be illustrative rather than comprehensive. Nevertheless, one of the configurations is found to be more accurate and may well be optimal.

Section 4.5 studies Thacker's *irregular grid finite-difference* technique [Th77,Th78b]. For the chosen element configuarations, it is simply a lumped version of the FEM in Section 4.4.

Section 4.6 studies the mixed interpolation FEM with piecewise linear basis functions for approximating elevation, and piecewise quadratic functions for the velocity components.

Section 4.7 studies the WEM and LWEM. It extends many of the results in Chapter 3.

Section 4.8 assesses the cost and accuracy of the Richardson-Sielecki, Thacker, and lumped *wave equation* methods.

Finally Section 4.9 summarizes and briefly discusses the results.

4.2 Analytic Results

The two dimensional linearized shallow water equations are

$$\frac{\partial z}{\partial t} + \frac{\partial (hu)}{\partial x} + \frac{\partial (hv)}{\partial y} = 0 \tag{4.2.1a}$$

$$\frac{\partial u}{\partial t} + g\frac{\partial z}{\partial x} - fv + \tau u = 0 \tag{4.2.1b}$$

$$\frac{\partial v}{\partial t} + g\frac{\partial z}{\partial y} + fu + \tau v = 0 \tag{4.2.1c}$$

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

$u(x, y, t)$ = x component of the velocity,

$v(x, y, t)$ = y component of the velocity,

$h(x, y)$ = mean sea depth,

g = gravity,

$f(x, y)$ = Coriolis coefficent,

τ = linear bottom friction coefficient.

Assuming constant values for the depth and Coriolis coefficient, plane wave solutions of the form

$$\begin{pmatrix} z(x,y,t) \\ u(x,y,t) \\ v(x,y,t) \end{pmatrix} = \begin{pmatrix} \varsigma_0 \\ \mu_0 \\ \nu_0 \end{pmatrix} e^{i(k_1 x + k_2 y - \omega t)} \tag{4.2.2}$$

can be found for (4.2.1). ω is frequency and

$$\mathbf{k} = (k_1, k_2) \tag{4.2.3a}$$

are the (x, y) components of wavenumber. The wavelength is now defined as

$$L = 2\pi/k \tag{4.2.3b}$$

where

$$k = (k_1^2 + k_2^2)^{1/2}. \tag{4.2.3c}$$

For nontrivial solutions, the following cubic characteristic equation

$$\omega^3 + 2i\tau\omega^2 - \omega(\tau^2 + f^2 + ghk^2) - i\tau ghk^2 = 0 \tag{4.2.4}$$

must be satisfied. Dispersion relationships are obtained from its roots.

Two cases are possible; either all three roots are purely imaginary, or one is purely imaginary and the other two, when multiplied by i, are complex conjugates. For the latter case, the roots are [Se65]

$$\omega_1 = i[A + B - \tfrac{2}{3}\tau] \tag{4.2.5a}$$

$$\omega_2 = -\tfrac{1}{2}(A - B)3^{1/2} - i[\tfrac{1}{2}(A + B) + \tfrac{2}{3}\tau] \tag{4.2.5b}$$

$$\omega_3 = \tfrac{1}{2}(A - B)3^{1/2} - i[\tfrac{1}{2}(A + B) + \tfrac{2}{3}\tau] \tag{4.2.5c}$$

where

$$A = (-\tfrac{1}{2}b + d)^{1/3} \tag{4.2.5d}$$

$$B = (-\tfrac{1}{2}b - d)^{1/3} \tag{4.2.5e}$$

$$d^2 = \tfrac{1}{4}b^2 + \tfrac{1}{27}a^3 \tag{4.2.5f}$$

$$a = f^2 + ghk^2 - \tfrac{1}{3}\tau^2 \tag{4.2.5a}$$

$$b = \tfrac{1}{27}\tau[9(ghk^2 - 2f^2) - 2\tau^2]. \tag{4.2.5h}$$

This case arises when

$$d^2 > 0. \tag{4.2.5i}$$

The first root corresponds to a steady current that will decay in time when $\tau > 0$. The other two roots correspond to gravity waves (or inertial waves when $\tau = k = 0$) that travel at the same speed in opposite directions, and have the same rate of amplitude decay (or growth).

The second case arises when

$$d^2 \leq 0 \tag{4.2.5j}$$

and produces three nonpropagating, decaying waves. It only occurs for relatively large τ.

When $\tau = 0$, all wave amplitudes are constant in time. The dispersion relationships now become [Le78]

$$\omega_1 = 0 \tag{4.2.6a}$$

$$\omega_{2,3} = \pm (f^2 + ghk^2)^{1/2}. \tag{4.2.6b}$$

In two dimensions, phase and group velocity are defined as [Le78]

$$\mathbf{C} = Re(\omega)\mathbf{k}/k^2 \tag{4.2.7a}$$

$$\mathbf{G} = Re\left(\frac{\partial\omega}{\partial k_1}, \frac{\partial\omega}{\partial k_2}\right). \tag{4.2.7b}$$

For (4.2.6), the nontrivial velocities are

$$\mathbf{C} = \pm (f^2 + ghk^2)^{1/2}\mathbf{k}/k^2 \tag{4.2.8a}$$

$$\mathbf{G} = \pm ghk(f^2 + ghk^2)^{-1/2}. \tag{4.2.8b}$$

Waves whose propagation speed $\mathbf{C}$ varies with the wavelength are said to be *dispersive*. If $\mathbf{C}$ is independent of direction, these waves are also said to be *isotropic* [Li78]. The waves described by (4.2.8) are isotropic, and nondispersive only when $f = 0$.

4.3 The Richardson-Sielecki FDM

A FDM which has been used successfully in many tide and storm surge problems [He69,He76,Cr76] is the Richardson-Sielecki (henceforth RS) scheme. It involves calculating variables on a Richardson grid [Ri22,Pl63] (also known as Arakawa's lattice C grid [Me76]) using a particular method of handling the Coriolis terms introduced by Sielecki [Si68]. Assuming a constant depth and Coriolis coefficient, its difference equations for solving (4.2.1) are

$$\frac{z_{rs}^{n+1} - z_{rs}^{n}}{\Delta t} + h\left(\frac{u_{r+1,s}^{n+\frac{1}{2}} - u_{rs}^{n+\frac{1}{2}}}{\Delta x} + \frac{v_{r,s+1}^{n+\frac{1}{2}} - v_{rs}^{n+\frac{1}{2}}}{\Delta y}\right) = 0 \tag{4.3.1a}$$

$$\frac{u_{rs}^{n+\frac{3}{2}} - u_{rs}^{n+\frac{1}{2}}}{\Delta t} + g\left(\frac{z_{rs}^{n+1} - z_{r-1,s}^{n+1}}{\Delta x}\right)$$

$$-\frac{f}{4}\left(v_{r-1,s}^{n+\frac{1}{2}} + v_{rs}^{n+\frac{1}{2}} + v_{r-1,s+1}^{n+\frac{1}{2}} + v_{r,s+1}^{n+\frac{1}{2}}\right) + \tau\left(\theta u_{rs}^{n+\frac{3}{2}} + (1-\theta)u_{rs}^{n+\frac{1}{2}}\right) = 0 \tag{4.3.1b}$$

$$\frac{v_{rs}^{n+\frac{3}{2}} - v_{rs}^{n+\frac{1}{2}}}{\Delta t} + g\left(\frac{z_{rs}^{n+1} - z_{r,s-1}^{n+1}}{\Delta y}\right)$$

$$+\frac{f}{4}\left(u_{rs}^{n+\frac{3}{2}}+u_{r,s-1}^{n+\frac{3}{2}}+u_{r+1,s}^{n+\frac{3}{2}}+u_{r+1,s-1}^{n+\frac{3}{2}}\right)+\tau\left(\theta v_{rs}^{n+\frac{3}{2}}+(1-\theta)v_{rs}^{n+\frac{1}{2}}\right)=0. \quad (4.3.1c)$$

θ is a frictional weighting parameter and Δx, Δy, and Δt are the space and time step sizes. The elevation and velocity components are seen to be staggered by a half time step. The spatial placement of the variables is also staggered, as shown in Fig. 4.1. The scheme is explicit. When restricted to one dimension, the RS scheme has the spatial discretization D2 studied in Section 2.3.

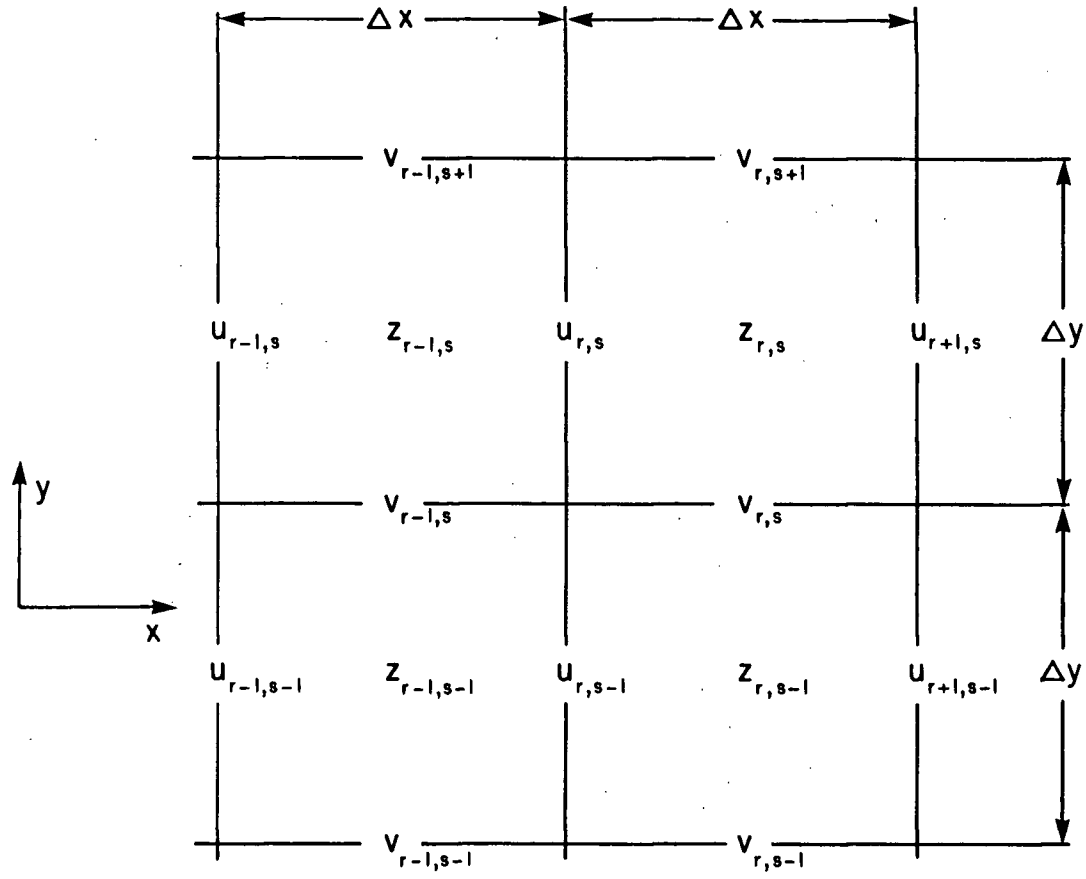

Fig. 4.1. Spatially discretized variables in the RS or lattice C grid

The dispersion relationship for (4.3.1) can be found by assuming plane wave solutions of the form

$$z_{rs}^{n}=\zeta_0 e^{i(rk_1\Delta x+sk_2\Delta y-n\omega\Delta t)} \quad (4.3.2a)$$

$$u_{rs}^{n+\frac{1}{2}}=\mu_0 e^{i[(r-\frac{1}{2})k_1\Delta x+sk_2\Delta y-(n+\frac{1}{2})\omega\Delta t]} \quad (4.3.2b)$$

$$v_{rs}^{n+\frac{1}{2}}=\nu_0 e^{i[rk_1\Delta x+(s-\frac{1}{2})k_2\Delta y-(n+\frac{1}{2})\omega\Delta t]}. \quad (4.3.2c)$$

A nontrivial solution requires

$$(\lambda-1)\left\{[\lambda-1+\tau\Delta t(\theta\lambda+(1-\theta))]^2+\lambda[f\Delta t\cos(\tfrac{1}{2}k_1\Delta x)\cos(\tfrac{1}{2}k_2\Delta y)]^2\right\}$$
$$+4\lambda gh\Delta t^2[\lambda-1+\tau\Delta t(\theta\lambda+(1-\theta))]\left(\frac{\sin^2(\frac{1}{2}k_1\Delta x)}{\Delta x^2}+\frac{\sin^2(\frac{1}{2}k_2\Delta y)}{\Delta y^2}\right) \tag{4.3.3}$$
$$-\lambda(\lambda-1)f\Delta tgh\left(\frac{\Delta t^2}{\Delta x\Delta y}\right)\sin k_1\Delta x\sin k_2\Delta y=0,$$

where λ is again defined by (2.5.4).

For specific values of f, h, Δx, Δy, Δt, and τ, the roots of (4.3.3) are functions of wavenumber. For $\tau = 0$, these roots and the resultant dispersion relationship can be expressed algebraically [He81]. For nonzero τ, the results can be found numerically. In particular, with $\Delta y = \Delta x$, $\omega\Delta t$ can be expressed in terms of the wavenumber sampling coordinates $(k_1\Delta x, k_2\Delta x)$ and the three parameters

$$f_1=\frac{\tau\Delta x}{(gh)^{1/2}} \tag{4.3.4a}$$

$$f_2=(gh)^{1/2}\frac{\Delta t}{\Delta x} \tag{4.3.4b}$$

$$f_3=\frac{f\Delta x}{(gh)^{1/2}}. \tag{4.3.4c}$$

f_1 and f_2 were also defined in (2.5.9), while f_3 is a nondimensional inverse of the radius of deformation parameter used in [Me76].

Nondimensional phase and group velocities for the RS scheme are calculated from the roots of (4.3.3) as

$$\frac{\mathbf{C}}{(gh)^{1/2}}=\frac{Re(\omega\Delta t)}{f_2}\frac{(k_1\Delta x,k_2\Delta x)}{(k\Delta x)^2} \tag{4.3.5a}$$

$$\frac{\mathbf{G}}{(gh)^{1/2}}=-Im\left(\left(\frac{\partial\lambda}{\partial k_1\Delta x},\frac{\partial\lambda}{\partial k_2\Delta x}\right)/(\lambda f_2)\right). \tag{4.3.5b}$$

An exhaustive comparison of the RS and analytic solutions will not be attempted here. Two roots of (4.3.3) are associated with gravity waves. They will be studied in some detail. The third root will be considered only for its stability and its potential contamination of the gravity wave solution. Whereas $k_1\Delta x$ and $k_2\Delta x$ will vary over their complete domain $(-\pi,\pi]$ only a small portion of the (f_1, f_2, f_3, θ) parameter space will

be examined. Subsequent figures for the RS scheme and the FEMs will be shown for $(f_1, f_3) = (.05, .10)$. These are typical values for shallow water models at mid-latitudes. f_2 and time-stepping parameters such as θ will have order unity ($O(1)$) or less, and will generally be chosen for high accuracy of the gravity wave solutions.

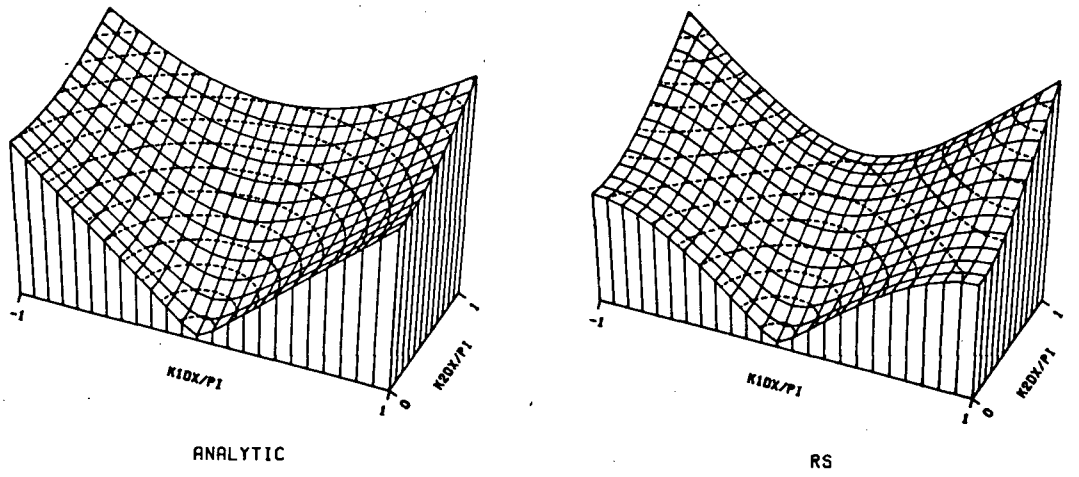

Fig. 4.2. Analytic and RS dispersion surfaces ($|\omega|\Delta x/(\pi(gh)^{1/2})$) for the parameter values $f_1 = .05$, $f_2 = .7071$, and $f_3 = .10$. $\theta = 0.5$ and $\Delta x = \Delta y$ for the RS scheme. Dotted line contours are in increments of .10.

The RS dispersion surface for $\Delta y = \Delta x$ and (f_1, f_2, f_3, θ) =(.05,.7071,.10,.5) is shown in Fig. 4.2. From (4.3.3) it is seen that $(k_1\Delta x, k_2\Delta x)$ and $-(k_1\Delta x, k_2\Delta x)$ produce the same values. (This will be referred to as symmetry through the origin.) Hence only positive $k_2\Delta x$ need be displayed. The *progressive wave* (positive ω) surface has been shown. A corresponding *retrogressive* surface (negative ω) exists and is simply the mirror image about the $(k_1\Delta x, k_2\Delta x)$ plane of the progressive surface. $f_2 = 2^{-1/2}$ is the maximum permitted for stability when $\Delta y = \Delta x$ [He81]. It is also the most accurate value for wave propagation in the $x = \pm y$ direction when $f = \tau = 0$.

The analytic dispersion surface has been included in Fig. 4.2 for comparison. As seen from (4.2.4), it is symmetric about both planes $k_1 = 0$ and $k_2 = 0$, and through the origin. Both the analytic and RS surfaces have a maximum values of approximately $2^{1/2}$ at $(k_1\Delta x/\pi, k_2\Delta x/\pi) = (\pm 1, 1)$. If f were equal zero, the analytic dispersion surface would be a cone with straight sides. With nonzero f, these sides develop a slight curvature.

Mesinger and Arakawa [Me76] show $|\omega|/f$ contours for the spatially discretized RS

scheme (the lattice C grid) with $f_3 = 0.5$. It has the same basic characteristics as Fig. 4.2. Notice that for small wavenumbers and $k_1 \simeq k_2$, RS surface values closely approximate the analytic.

Fig. 4.3 displays the accuracy of the RS scheme. It plots the two accuracy measure functions

$$M_A = \left| \frac{\lambda_n}{\lambda_a} \right| \tag{4.3.6a}$$

$$M_C = \frac{|\mathbf{C}_n| - |\mathbf{C}_a|}{|\mathbf{C}_a|} \tag{4.3.6b}$$

where λ_n is the principal progressive numerical eigenvalue, $\lambda_a$is the analytic progressive eigenvalue, and $\mathbf{C}_n$, $\mathbf{C}_a$ are the corresponding phase velocities. (4.3.6a) is identical to (2.6.1a) when the principal eigenvalue is dominant. (4.3.6b) is a two dimensional extension of (2.6.1b). Normalized group velocity vectors for both the analytic and RS solutions are also shown in Fig. 4.3.

M_C is the relative error in phase velocity magnitude. Since it is calculated as a function of k, it also equals the relative error in frequency . Negative values denote waves travelling too slowly while zero values are optimal. For example, $-.01$ denotes a numerical wave speed which is 1% too slow. The amplitude measure, M_A, is a ratio denoting the growth (or decay) factor per time step relative to the analytic solution. Values greater than the optimum of 1. signify a solution which decays too slowly or grows too rapidly. After n time steps, the ratio of the numerical amplitude to the analytic is $(M_A)^n$.

Wave amplitudes are seen to be accurately represented by the RS scheme. However waves travelling to the north-east will be slightly too large while those to the north-west will be slightly too small. This effect is solely due to the asymmetric treatment of the Coriolis terms in (4.3.1). Specifically, when $f_3 = 0$, both M_A and M_C become symmetric about $k_1 = 0$. Henry [He81] discusses alternative treatments of the Coriolis term which improve accuracy.

Comparing (4.3.5a) with (4.2.7a) it is evident that all phase velocity directions are correct. (This may not be true when $\Delta x \neq \Delta y$.) There is however some error in $|\mathbf{C}|$. As with M_A, there is an asymmetry about $k_1 = 0$ which disappears when $f_3 = 0$. Since the

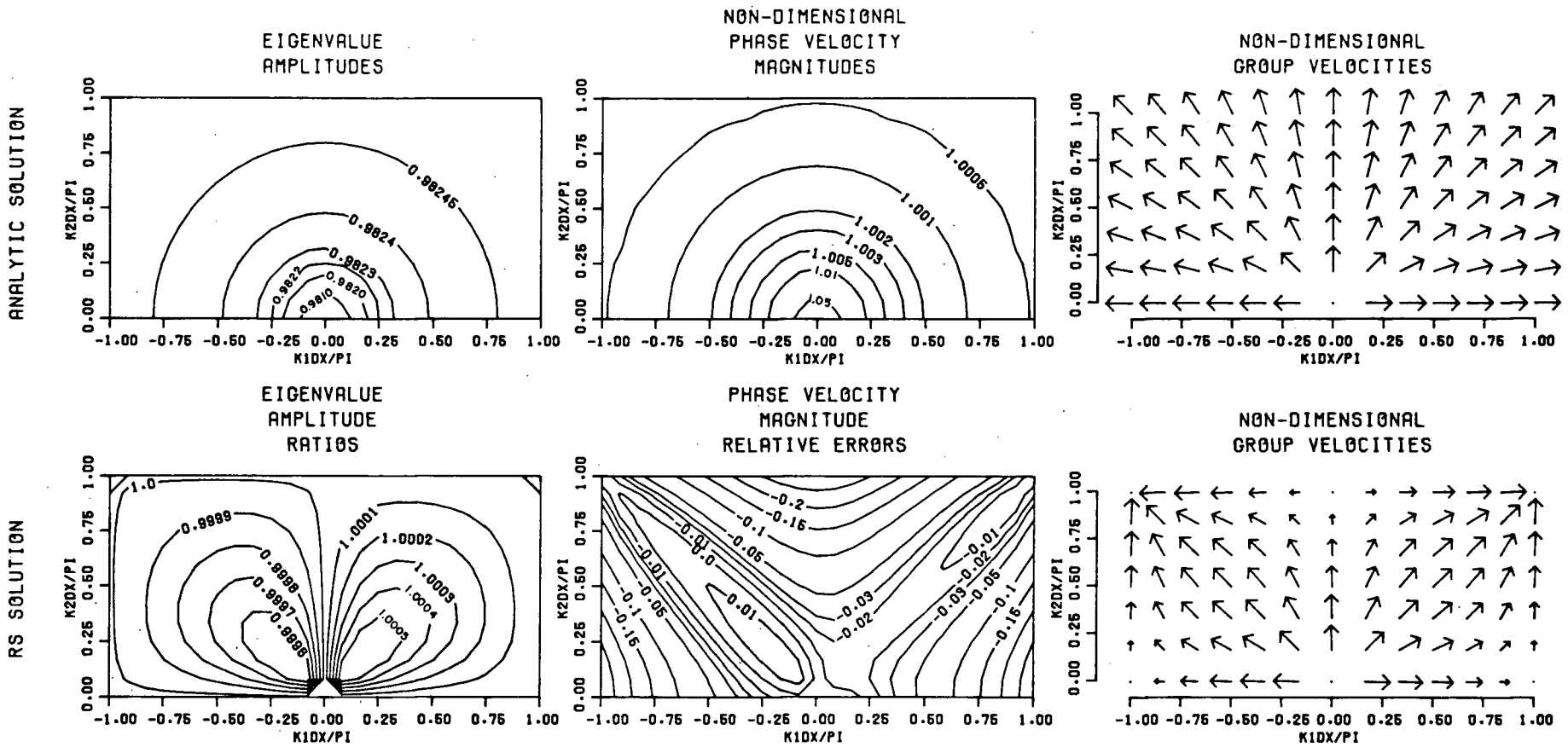

Fig. 4.3. Analytic solution, RS accuracy measures, and $\mathbf{G}/(gh)^{1/2}$ for the parameter values of Fig. 4.2.

M_C contours are not concentric circles about $(k_1\Delta x, k_2\Delta x) = (0,0)$, the numerical waves are anisotropic. They remain so when $f_3 = 0$.

The RS group velocity vectors display errors in both magnitude and direction. Analytic and RS values have not been shown for the case $(k_1\Delta x, k_2\Delta x) = (0,0)$, which represents infinite sampling per wavelength. Consistency of the numerical solution however indicates that for this limit, the RS values approach the analytic. Except for one dimensional motion (i.e., $k_1\Delta x = 0$ or $k_2\Delta x = 0$) directions err toward the diagonal. This suggests that wave energy, which travels at the group velocity, tends to favour this direction. The one dimensional $2\Delta x$ waves denoted by $(k_1\Delta x, k_2\Delta x) = (\pi, 0)$, $(-\pi, 0)$, $(0, \pi)$ are seen to have zero group velocity. This is to be expected since they correspond to saddle points in the dispersion surface. Zero group velocity has also been calculated for the diagonal waves of length $2^{1/2}\Delta x$ which are associated with $(k_1\Delta x, k_2\Delta x) = (\pi, \pi)$ or $(-\pi, \pi)$. When $\tau = 0$, calculations based on the dispersion relationship in [He81] show this to occur when $f_2 < 2^{-1/2}$.

4.4 The Galerkin FEM with Piecewise Linear Basis Functions

This section extends the one dimensional analysis of the Galerkin FEM with piecewise linear basis functions (henceforth GLFEM) to two dimensions. Two simple configurations of triangle elements are assumed. As might be expected, the particular shape and configuration of the triangles affect the accuracy of a FEM implementation. Platzman [Pl81] examines two triangular meshes in his study of FEM tidal models, while Mullen and Belytschko [Mu82] study the effects of four meshes on spatial discretizations of the wave equation. Both investigations assume linear basis functions.

The two configurations studied here are meant to be illustrative rather than comprehensive. They are shown in Fig. 4.4. Both consist of six equal triangles with the three variables $z(x, y, t)$, $u(x, y, t)$, and $v(x, y, t)$ defined at each vertex or node. The first mesh involves right triangles. A mirror image of the particular case $\Delta x = \Delta y$ is examined in [Mu82].

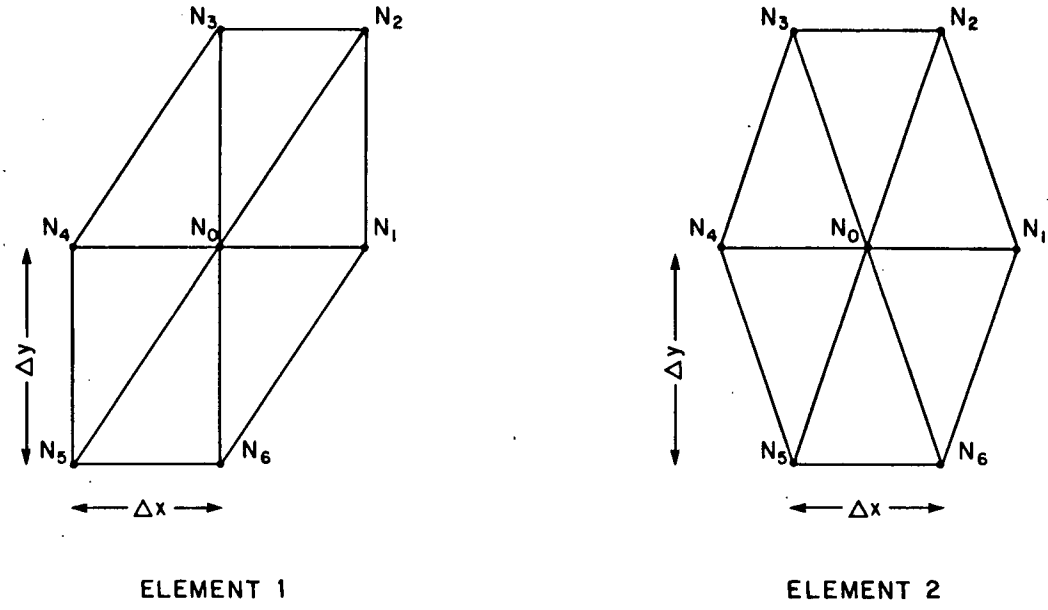

Fig. 4.4. Triangular element configurations for the FEM analyses.

The second mesh consists of isoceles triangles and is considered in [Pl81]. The special case of equilateral triangles is studied in [Mu82]. Because of its symmetry, one would intuitively expect equilateral triangles to be more accurate. Indeed, Mullen and Belytschko conclude that for their problem, this arrangement almost removes the directional dependence of phase velocity. Numerical experiments [Hi82] have also demonstrated that equilateral triangles are more accurate than right triangles.

Imposing the Galerkin condition with the basis function corresponding to node N_0 in element 1, the spatially discretized versions of (4.2.1) become

$$\begin{aligned}
&\frac{\partial}{\partial t}\left[\tfrac{1}{2}z_0+\tfrac{1}{12}(z_1+z_2+z_3+z_4+z_5+z_6)\right]\\
&\quad+h\left[\frac{2}{3}\left(\frac{u_1-u_4}{2\Delta x}\right)+\frac{1}{6}\left(\frac{u_2-u_3}{\Delta x}\right)+\frac{1}{6}\left(\frac{u_6-u_5}{\Delta x}\right)\right]\\
&\qquad+h\left[\frac{2}{3}\left(\frac{v_3-v_6}{2\Delta y}\right)+\frac{1}{6}\left(\frac{v_2-v_1}{\Delta y}\right)+\frac{1}{6}\left(\frac{v_4-v_5}{\Delta y}\right)\right]=0 \qquad (4.2.1a)\\
&\left(\frac{\partial}{\partial t}+\tau\right)\left[\tfrac{1}{2}u_0+\tfrac{1}{12}(u_1+u_2+u_3+u_4+u_5+u_6)\right]\\
&\quad+g\left[\frac{2}{3}\left(\frac{z_1-z_4}{2\Delta x}\right)+\frac{1}{6}\left(\frac{z_2-z_3}{\Delta x}\right)+\frac{1}{6}\left(\frac{z_6-z_5}{\Delta x}\right)\right]\\
&\qquad-f\left[\tfrac{1}{2}v_0+\tfrac{1}{12}(v_1+v_2+v_3+v_4+v_5+v_6)\right]=0 \qquad (4.4.1b)
\end{aligned}$$

$$\left(\frac{\partial}{\partial t}+\tau\right)\left[\tfrac{1}{2}v_0+\tfrac{1}{12}(v_1+v_2+v_3+v_4+v_5+v_6)\right]$$
$$+g\left[\frac{2}{3}\left(\frac{z_3-z_6}{2\Delta y}\right)+\frac{1}{6}\left(\frac{z_2-z_1}{\Delta y}\right)+\frac{1}{6}\left(\frac{z_4-z_5}{\Delta y}\right)\right]$$
$$+f\left[\tfrac{1}{2}u_0+\tfrac{1}{12}(u_1+u_2+u_3+u_4+u_5+u_6)\right]=0. \qquad (4.4.1c)$$

The algebra required to derive these ODEs is facilitated by the *triangular area coordinates* described by Pinder and Gray [Pi77].

The analogous result for element 2 is

$$\frac{\partial}{\partial t}\left[\tfrac{1}{2}z_0+\tfrac{1}{12}(z_1+z_2+z_3+z_4+z_5+z_6)\right]$$
$$+h\left[\frac{2}{3}\left(\frac{u_1-u_4}{2\Delta x}\right)+\frac{1}{6}\left(\frac{u_2-u_3}{\Delta x}\right)+\frac{1}{6}\left(\frac{u_6-u_5}{\Delta x}\right)\right]$$
$$+h\left[\frac{1}{2}\left(\frac{v_2-v_6}{2\Delta y}\right)+\frac{1}{2}\left(\frac{v_3-v_5}{2\Delta y}\right)\right]=0 \qquad (4.4.2a)$$

$$\left(\frac{\partial}{\partial t}+\tau\right)\left[\tfrac{1}{2}u_0+\tfrac{1}{12}(u_1+u_2+u_3+u_4+u_5+u_6)\right]$$
$$+g\left[\frac{2}{3}\left(\frac{z_1-z_4}{2\Delta x}\right)+\frac{1}{6}\left(\frac{z_2-z_3}{\Delta x}\right)+\frac{1}{6}\left(\frac{z_6-z_5}{\Delta x}\right)\right]$$
$$-f\left[\tfrac{1}{2}v_0+\tfrac{1}{12}(v_1+v_2+v_3+v_4+v_5+v_6)\right]=0 \qquad (4.4.2b)$$

$$\left(\frac{\partial}{\partial t}+\tau\right)\left[\tfrac{1}{2}v_0+\tfrac{1}{12}(v_1+v_2+v_3+v_4+v_5+v_6)\right]$$
$$+g\left[\frac{1}{2}\left(\frac{z_2-z_6}{2\Delta y}\right)+\frac{1}{2}\left(\frac{z_3-z_5}{2\Delta y}\right)\right]$$
$$+f\left[\tfrac{1}{2}u_0+\tfrac{1}{12}(u_1+u_2+u_3+u_4+u_5+u_6)\right]=0. \qquad (4.4.2c)$$

Assuming plane wave solutions of the form

$$\begin{pmatrix} z(r\Delta x, s\Delta y, t) \\ u(r\Delta x, s\Delta y, t) \\ v(r\Delta x, s\Delta y, t) \end{pmatrix} = \begin{pmatrix} \varsigma_0 \\ \mu_0 \\ \nu_0 \end{pmatrix} e^{i(rk_1\Delta x+sk_2\Delta y-\omega t)}, \qquad (4.4.3)$$

dispersion relationships solely due to these spatial discretizations can be found. They are calculated from the cubic polynomials which result when requiring nonzero values for ς_0, μ_0, and ν_0. For element 1, the cubic is

$$\omega^3A^2+2i\omega^2A^2\tau+\omega\left[-A^2(f^2+\tau^2)-gh\left(\frac{G_x^2}{\Delta x^2}+\frac{G_y^2}{\Delta y^2}\right)\right]$$
$$-i\tau gh\left(\frac{G_x^2}{\Delta x^2}+\frac{G_y^2}{\Delta y^2}\right)=0 \qquad (4.4.4)$$

where

$$A = \tfrac{1}{2} + \tfrac{1}{6}[\cos k_1\Delta x + \cos k_2\Delta y + \cos(k_1\Delta x + k_2\Delta y)] \tag{4.4.5a}$$

$$G_x = \tfrac{2}{3}\sin k_1\Delta x - \tfrac{1}{3}\sin k_2\Delta y + \tfrac{1}{3}\sin(k_1\Delta x + k_2\Delta y) \tag{4.4.5b}$$

$$G_y = \tfrac{2}{3}\sin k_2\Delta y - \tfrac{1}{3}\sin k_1\Delta x + \tfrac{1}{3}\sin(k_1\Delta x + k_2\Delta y). \tag{4.4.5c}$$

Walters and Carey [Wa83] obtain this result for the particular case $f = \tau = 0$ and $\Delta x = \Delta y$. With $f = \tau = 0$ and either k_1 or k_2 equal to zero, the nontrivial dispersion relationships arising from (4.4.4) and (4.4.5) reduce to the one dimensional result (2.3.3).

For element 2, the cubic polynomial is unchanged but

$$A = \tfrac{1}{2} + \tfrac{1}{6}\cos k_1\Delta x + \tfrac{1}{3}\cos(\tfrac{1}{2}k_1\Delta x)\cos k_2\Delta y \tag{4.4.6a}$$

$$G_x = \tfrac{2}{3}[\sin k_1\Delta x + \sin(\tfrac{1}{2}k_1\Delta x)\cos k_2\Delta y] \tag{4.4.6b}$$

$$G_y = \sin k_2\Delta y \cos(\tfrac{1}{2}k_1\Delta x). \tag{4.4.6c}$$

Again, with $f = \tau = 0$ and $k_1 = 0$, the nontrivial dispersion relationships simplify to the one dimensional result (2.3.3). With equilateral triangles, this simplification also occurs when $k_1 = 3^{1/2}k_2$.

Phase velocities, group velocities, and wave amplitude decay factors can be calculated from the roots of (4.4.4). However, as was seen in Chapters 2 and 3, their accuracy does not always indicate the accuracy of the fully discretized numerical solution. In some cases, a subsequent time discretization may partially cancel the errors arising from the spatial discretization, thereby making the fully discretized equations more accurate. It is therefore best to continue the analysis by introducing a particular time-stepping method for solving the system of ODEs given by (4.4.1) or (4.4.2).

The one dimensional analysis in Chapter 2 suggests that Crank Nicolson (CN) is the best time-stepping method to use in combination with a Galerkin FEM and piecewise linear basis functions. Although CN time-stepping may not be most accurate for all two dimensional wave directions, it will be more efficient than other implicit two-step methods and it should avoid problems with spurious numerical solutions. In fact, combining any

other linear two-step method with this spatial discretization leads to six numerical dispersion relationships; three of which are spurious and not present with the CN method. For these reasons, the subsequent analysis will assume CN time-stepping. However, the same approach can also be followed with any other two-step method.

Dispersion relationships for the fully discretized equations are calculated by assuming the plane wave solutions

$$\begin{pmatrix} z(r\Delta x, s\Delta y, n\Delta t) \\ u(r\Delta x, s\Delta y, n\Delta t) \\ v(r\Delta x, s\Delta y, n\Delta t) \end{pmatrix} = \begin{pmatrix} \varsigma_0 \\ \mu_0 \\ \nu_0 \end{pmatrix} e^{i(rk_1\Delta x + sk_2\Delta y - n\omega\Delta t)}. \tag{4.4.7}$$

Assuming nontrivial solutions then leads to a characteristic equation whose roots are the numerical eigenvalues λ. With CN time-stepping, these values may also be calculated as

$$\lambda = \frac{1 - \frac{1}{2}i\omega_0\Delta t}{1 + \frac{1}{2}i\omega_0\Delta t} \tag{4.4.8}$$

where ω_0 is a root of (4.4.4). This result follows from (2.7.8). Dispersion surfaces, phase and group velocities, and wave amplitude decay factors can now be calculated from the λs.

In order to compare the relative accuracy of elements 1 and 2 in Fig. 4.4, the total area of the elements should be considered. Since both elements have the same storage requirements for the nodal variables, equal area implies equal storage costs for a model of pre-specified spatial dimensions. Accuracy can then be compared on an equal-cost basis. In particular, consider $\Delta x = \Delta y$ in element 1, and equilateral triangles with sides of length d in element 2. Equal area then requires

$$d = (4/3)^{1/4}\Delta x. \tag{4.4.9}$$

In order that the accuracy of these two FEMs might also be compared to the RS results, the parameters of (4.3.4) should be re-defined for triangular elements. Specifically

$$f_1 = \tau\left(\frac{2A_r}{gh}\right)^{1/2} \tag{4.4.10a}$$

$$f_2 = \left(\frac{gh}{2A_r}\right)^{1/2}\Delta t \tag{4.4.10b}$$

$$f_3 = f\left(\frac{2A_r}{gh}\right)^{1/2} \tag{4.4.10c}$$

where A_r is the area of each triangle in the respective element. With $\Delta x = \Delta y$ in element 1, (4.4.10) and (4.3.4) become equal.

With equilateral triangles, it is convenient to re-define G_y in (4.4.6c). In particular, setting

$$G_y = [2 \sin k_2 \Delta y \cos(\tfrac{1}{2} k_1 \Delta x)]/3^{1/2} \tag{4.4.11}$$

with $\Delta x = d$ and $\Delta y = 3^{1/2}d/2$ permits the replacement of $G_y/\Delta y$ with G_y/d in (4.4.4).

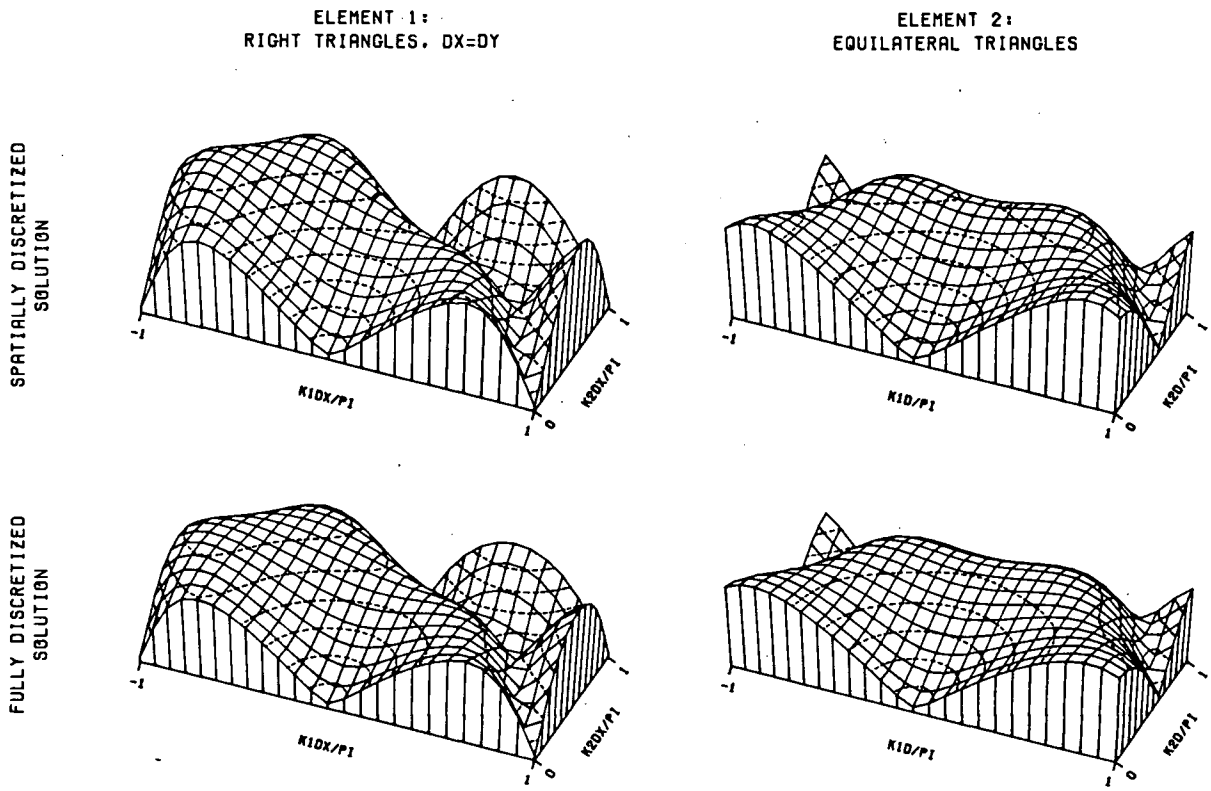

Fig. 4.5. Dispersion surfaces ($|\omega|\Delta x/(\pi(gh)^{1/2})$) for the GLFEM with CN time-stepping. Parameter values are $f_1 = .05$, $f_2 = .7071$, $f_3 = .10$. Dotted line contours are in increments of .10.

Fig. 4.5 shows dispersion surfaces for the same parameter values as in Fig. 4.2, namely (f_1, f_2, f_3) =(.05,.7071,.10). It also has the same scale and is viewed from the same perspective. Surfaces are shown for both elements of Fig. 4.4 and both the spatially discretized and fully discretized equations. Again, only positive $k_2\Delta x$ need be displayed

since (4.4.4) is symmetric through the origin for element 1, and both through the origin and about the $k_2 d$ axis for element 2.

Fig. 4.5 has several notable points. The first is that there is little difference between the spatially discretized and fully discretized surfaces. This implies that virtually all the inaccuracy of the fully discretized equations is due to the spatial discretization. Hence CN has scarcely affected the accuracy. This may not be true for all time-stepping schemes.

Each of the dispersion surfaces in Fig. 4.5 is symmetric. Considering the symmetries in the elements themselves, these are to be expected. The surface for element 1 is symmetric about the planes $k_2 = k_1$ and $k_2 = -k_1$. The element 2 surface is symmetric about $k_2 = k_1 \tan\phi$, where $\phi = 30°, 60°, 90°, 120°$, or $150°$.

The most striking feature of both surfaces is their poor accuracy for higher values of wavenumber sampling. Accuracy is reasonable for small wavenumbers (i.e., $k\Delta x/\pi < 0.1$) but it deteriorates as $k\Delta x$ increases. This is consistent with the one dimensional analysis of Chapter 2. Particularly disturbing are the frequency valleys. For the case $\tau = 0$, the nontrivial roots of (4.4.4) are

$$\omega = \pm\left[f^2 + \frac{gh}{A^2}\left(\frac{G_x^2}{\Delta x^2} + \frac{G_y^2}{\Delta y^2}\right)\right]^{1/2}. \tag{4.4.12}$$

Assuming $\Delta x = \Delta y$, minimal values of $|\omega|$ occur when

$$G_x^2 = G_y^2 = 0. \tag{4.4.13}$$

In particular, for element 1 they occur at the following seven values of $(k_1\Delta x/\pi, k_2\Delta x/\pi)$: (0,0), (1,1), (-1,1), (0,1), (1,0), (-1,0), $(\frac{2}{3}, \frac{2}{3})$. The latter corresponds to a diagonal wave of length $(4.5)^{1/2}\Delta x$, while the fourth, fifth, and sixth minimal values are associated with one dimensional $2\Delta x$ waves. The second and third minima correspond to diagonal waves of length $2^{1/2}\Delta x$. (In a two dimensional grid, plane waves shorter than $2\Delta x$ are possible.) All these waves have the inertial frequency f. In the particular case when $f = 0$, $\omega = |C| = 0$ at these seven points. Hence the progressive and retrogressive surfaces touch. These seven minimal values for $|\omega|$ are perturbed slightly by the nonzero values of f in Fig. 4.5.

For element 2, the surface minima occur only for the $(k_1 d/\pi, k_2 d/\pi)$ values of (0,0), $(1, 3^{-1/2})$, and $(-1, 3^{-1/2})$. The latter two correspond to waves of length $3^{1/2}$. A similar minimum also exists at $(0, 2(3)^{-1/2})$ but is not shown.

Comparing accuracy on the basis of equal area now means that $k_1 \Delta x \neq k_1 d$. That is, even though waves may have the same lengths on the element 1 and element 2 meshes, their sampling rates per wavelength will differ. The shortest one dimensional wave supported by element 1 is $2\Delta x$ while for element 2 it is the slightly longer value of $2d$. In order to permit accuracy comparisons on the basis of wavelength, the $k_1 d$ and $k_2 d$ axes should be scaled. This is done in Fig. 4.6.

Fig. 4.6 shows M_A, M_C, and $\mathbf{G}/(gh)^{1/2}$ values associated with the fully discretized dispersion surfaces of Fig. 4.5. Both elements configurations produce wave amplitudes which are too large and phase velocities which are too small. However for both M_A and M_C, when $k_1 \Delta x/\pi < 0.5$ the element 2 contour levels are further away from $(k_1, k_2) = (0, 0)$ than those of element 1. This implies that element 2 is more accurate for longer waves.

Comparing Fig. 4.6 with Fig. 4.3, one cannot conclude that either numerical scheme is consistently more accurate. Generally the RS scheme is more accurate, however there are some regions near $(k_1 \Delta x, k_2 \Delta x) = (0, 0)$ where the GLFEM schemes are better for both wave amplitude and phase velocity.

The accuracy of the GLFEM numerical group velocity deteriorates significantly as the wavenumber increases. Errors exist in both magnitude and direction. In fact, for some short waves, $\mathbf{G}$ is not only much too large but also in virtually the opposite direction from what it should be. Group and phase velocities which are not co-directional signify energy propagating in a different direction than the wave crests. Although this should not occur for shallow water waves, it is clearly seen to do so for both the RS and GLFEM schemes.

4.5 Thacker's Irregular Grid FDM

Thacker [Th77,Th78,Th78b] has recently presented a technique for defining FDMs over irregular grids of triangular elements. The underlying concept is that in the vicinity of a triangle, the partial derivatives of a function can be approximated by the slopes of a

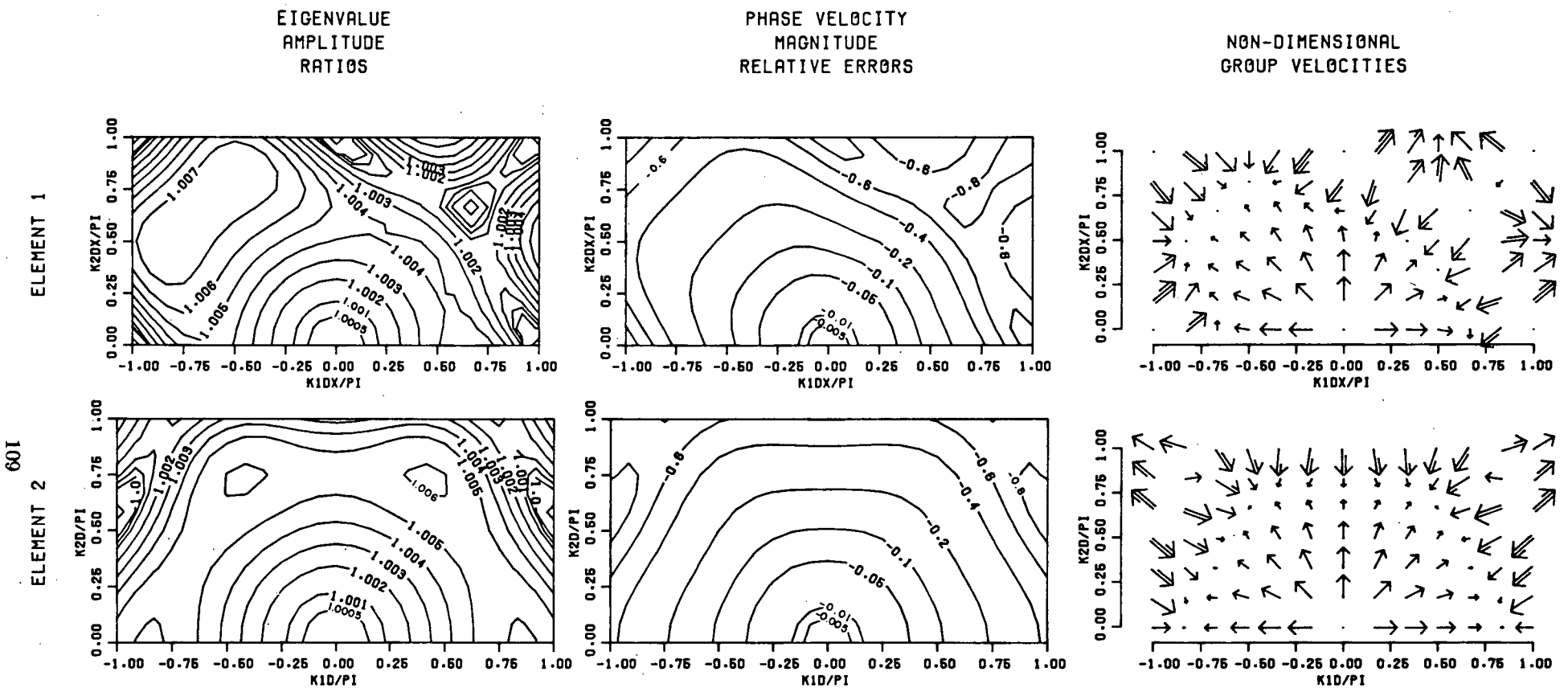

Fig. 4.6. M_A, M_C, and $\mathbf{G}/(gh)^{1/2}$ for the GLFEM with CN time-stepping and the parameter values of Fig. 4.5. Each full shaft of multi-shafted vectors denotes 1 unit (i.e., $|\mathbf{G}| = (gh)^{1/2}$).

plane determined by the values of the function at the vertices. At a vertex, the partial derivatives are then approximated by a weighted average of the approximations in each of the triangles which contain that vertex. For equal area triangles such as those of Fig. 4.4, the resultant spatial derivative approximations are equivalent to those for the GLFEM. In fact, for elements 1 and 2 respectively, the spatially discretized equations are simply found by replacing all terms of the form $\frac{1}{2}z_0 + \frac{1}{12}(z_1 + z_2 + z_3 + z_4 + z_5 + z_6)$ in (4.4.1) and (4.4.2) with z_0. As Wang [Wa78] points out, this means that the spatial discretization for Thacker's method is simply a lumped mass matrix version of the Galerkin FEM with piecewise linear basis functions.

Thacker employs an explicit leapfrog time-stepping similar to the RS scheme of Section 4.3. For solving (4.2.1), his fully discretized equations may be generalized to

$$\frac{z_j^{n+1} - z_j^n}{\Delta t} + h\left(\overline{\frac{\partial u}{\partial x}}\right)_j^{n+1/2} + h\left(\overline{\frac{\partial v}{\partial y}}\right)_j^{n+1/2} = 0 \qquad (4.5.1a)$$

$$\frac{u_j^{n+1/2} - u_j^{n-1/2}}{\Delta t} + g\left(\overline{\frac{\partial z}{\partial x}}\right)_j^{n} - f[\theta v_j^{n+1/2} + (1-\theta)v_j^{n-1/2}] + \tau[\theta u_j^{n+1/2} + (1-\theta)u_j^{n-1/2}] = 0 \qquad (4.5.1b)$$

$$\frac{v_j^{n+1/2} - v_j^{n-1/2}}{\Delta t} + g\left(\overline{\frac{\partial z}{\partial y}}\right)_j^{n} + f[\theta u_j^{n+1/2} + (1-\theta)u_j^{n-1/2}] + \tau[\theta v_j^{n+1/2} + (1-\theta)v_j^{n-1/2}] = 0 \qquad (4.5.1c)$$

where

$$\left(\overline{\frac{\partial}{\partial x}}\right) \quad \text{and} \quad \left(\overline{\frac{\partial}{\partial y}}\right)$$

denote the spatial derivative approximations. For elements 1 and 2, these approximations are identical to those in (4.4.1) and (4.4.2). The particular scheme discussed by Thacker has $\tau = 0$ and $\theta = 1$.

Assuming a uniform grid of equilateral triangles and $f = 0$, Thacker [Th78b] calculates dispersion relationships for his scheme. These results can be extended to include the element 1 grid with $\Delta x = \Delta y$, and to allow for nonzero friction and Coriolis. For both grids, the spatially discretized relationships are simply found by setting $A = 1$ in (4.4.4).

Assuming plane wave solutions, the characteristic equation arising from (4.5.1) is

$$(\lambda-1)\{[\lambda(1+\theta\tau\Delta t)-1+(1-\theta)\tau\Delta t]^2+(f\Delta t)^2(\lambda\theta+(1-\theta))^2\}$$
$$+gh\Delta t^2\lambda[\lambda(1+\theta\tau\Delta t)-1+(1-\theta)\tau\Delta t]\left(\frac{G_x^2}{\Delta x^2}+\frac{G_y^2}{\Delta y^2}\right)=0 \tag{4.5.2}$$

with G_x, G_y defined as in (4.4.5), or (4.4.6) and (4.4.11). With $\tau = 0$, the respective dispersion relationships for elements 1 and 2 are

$$\cos\omega\Delta t = \frac{1-(f\Delta t)^2\theta(1-\theta)-\frac{1}{2}gh(\Delta t/\Delta x)^2G_{xy}}{1+(\theta f\Delta t)^2} \tag{4.5.3a}$$

$$\cos\omega\Delta t = \frac{1-(f\Delta t)^2\theta(1-\theta)-\frac{1}{2}gh(\Delta t/d)^2G_{xy}}{1+(\theta f\Delta t)^2} \tag{4.5.3b}$$

$$\text{where} \qquad G_{xy} = G_x^2+G_y^2. \tag{4.5.3c}$$

With $f = 0$, (4.5.3b) simplifies to Thacker's relationship.

For comparison, when Thacker's time-stepping method is combined with the GLFEM, the dispersion relationships are still expressed by (4.5.3a) and (4.5.3b) but have

$$G_{xy} = \frac{G_x^2+G_y^2}{A^2}. \tag{4.5.4}$$

Necessary conditions for stability can be determined from (4.5.3) by requiring

$$-1 \le \cos\omega\Delta t \le 1. \tag{4.5.5}$$

For equilateral triangles, Thacker obtains the following conditions for his scheme and the FEM with similar time-stepping

$$(gh)^{1/2}\frac{\Delta t}{d} = f_2' \le 1.70437 \tag{4.5.6a}$$

$$\text{and} \qquad f_2' \le 0.90288. \tag{4.5.6b}$$

These conditions assume $f = 0$ and $\theta = 1$. For element 1 with $\Delta x = \Delta y$, the analogous conditions are

$$(gh)^{1/2}\frac{\Delta t}{\Delta x} = f_2 \le 1.4142 \tag{4.5.7a}$$

$$\text{and} \qquad f_2 \le 0.79830. \tag{4.5.7b}$$

On the basis of equal area, element 2 has less restrictive stability constraints than element 1. Furthermore, for both grids Thacker's scheme is less restrictive than its GLFEM counterpart. This implies that Thacker's scheme can use a larger time step. Increased stability with lumping is also noted by Strang and Fix [St73].

Thacker claims that his scheme is most accurate with the maximum possible time step. This can be verified with an asymptotic analysis.

With $\tau = 0$, the analytic dispersion relationship for the configuration of equilateral triangles is

$$\omega\Delta t = \pm\left[(f\Delta t)^2 + (f_2')^2(\xi^2 + \eta^2)\right]^{1/2} \tag{4.5.8}$$

where $\xi = k_1 d$ and $\eta = k_2 d$. When f_2' is $O(1)$ and f_3 and kd are is small, Taylor expansions then give

$$\cos\omega\Delta t \simeq 1 - \tfrac{1}{2}[(f\Delta t)^2 + (f_2')^2(\xi^2 + \eta^2)] + \tfrac{1}{24}[(f\Delta t)^2 + (f_2')^2(\xi^2 + \eta^2)]^2 \tag{4.5.9}$$

to powers of order 4.

With equilateral triangles and small values of ξ and η

$$G_x^2 + G_y^2 \simeq \xi^2 + \eta^2 - \tfrac{1}{4}(\xi^2 + \eta^2)^2. \tag{4.5.10}$$

Substitution into (4.5.3b), and matching terms with (4.5.9) then shows that

$$f_2' = 3^{1/2} \tag{4.5.11a}$$

and

$$\theta = \tfrac{1}{2} \tag{4.5.11b}$$

produce the best approximation of the analytic dispersion relationship. It also shows that accuracy increases as f_2' approaches $3^{1/2}$ from below. This verifies Thacker's remarks. Notice that the optimal value of f_2' is slightly larger than the stability limit given in (4.5.6a).

For right triangles with $\Delta x = \Delta y$, a similar analysis is less conclusive. Accuracy of the numerical method is now dependent on wave direction as well as f_2. In particular, the optimal f_2 value is

$$f_2 = 2\left(1 + \frac{\xi\eta}{\xi^2 + \eta^2}\right)^{1/2} \tag{4.5.12}$$

where $\xi = k_1 \Delta x$ and $\eta = k_2 \Delta x$. This implies a minimal optimal value of $2^{1/2}$ when $\xi = -\eta$. As seen from (4.5.7) this is also the maximum stable value. $\theta = \frac{1}{2}$ still provides the most accurate representation of the Coriolis terms.

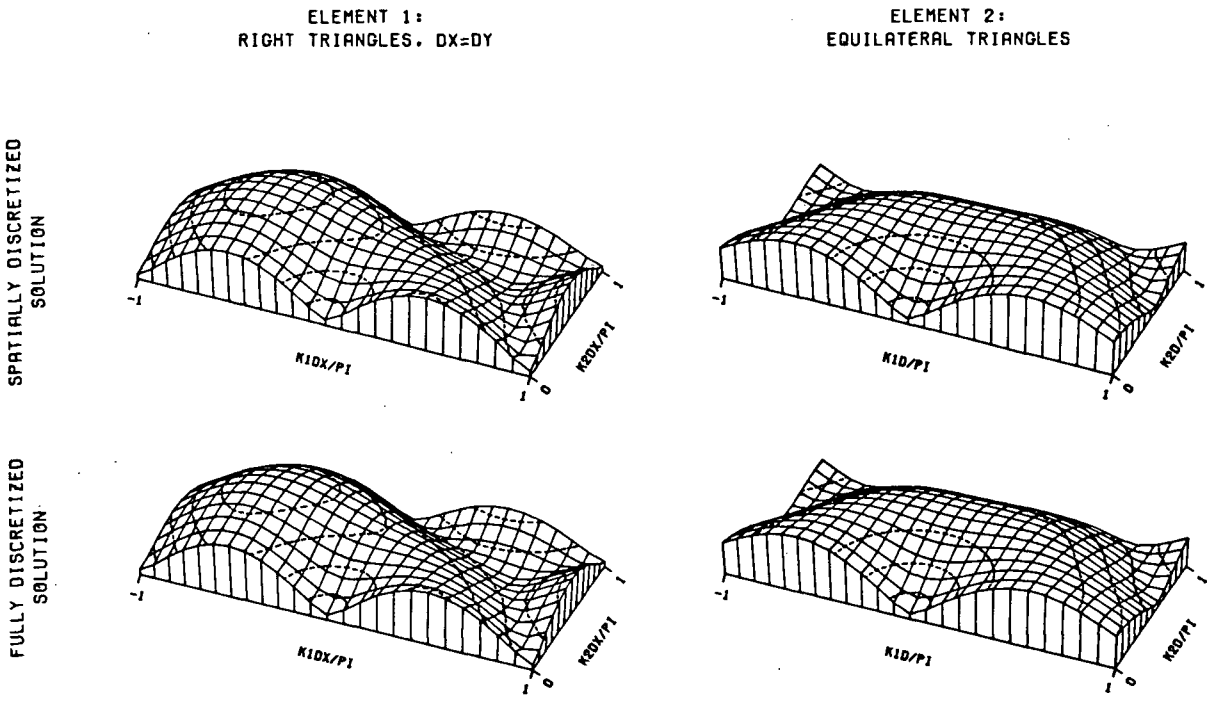

Fig. 4.7. Dispersion surfaces ($|\omega|\Delta x/(\pi(gh)^{1/2})$) for Thacker's method. Parameter values are $f_1 = .05$, $f_2 = .7071$, $f_3 = .10$, $\theta = 1$. Dotted line contours are in increments of .10.

Fig. 4.7 shows the dispersion surfaces for the spatially and fully discretized versions of Thacker's scheme. It has the same parameter values, scale, and perspective as Fig. 4.2 and Fig. 4.5. The spatially discretized surfaces are simply lumped versions of those in Fig. 4.5. They have the same characteristic shape but, as seen from the dotted line contour levels, have smaller values. Again, the time-stepping method has little effect on the fully discretized surface values. The symmetries and location of the surface minima are the same for Fig. 4.7 as for Fig. 4.5. This implies that as with the GLFEM, problems with short waves can also be expected with Thacker's scheme.

Fig. 4.8 displays the two accuracy measure functions and group velocity vectors corresponding to the surfaces of Fig. 4.7. As with the GLFEM, the element 2 mesh is generally more accurate than the element 1. Fig. 4.9 is similar but has the near optimal parameter values $\theta = \frac{1}{2}$, and $f_2 = 1.40$, $f_2 = 1.826$ ($f_2' = 1.70$) for the right triangle and equilateral

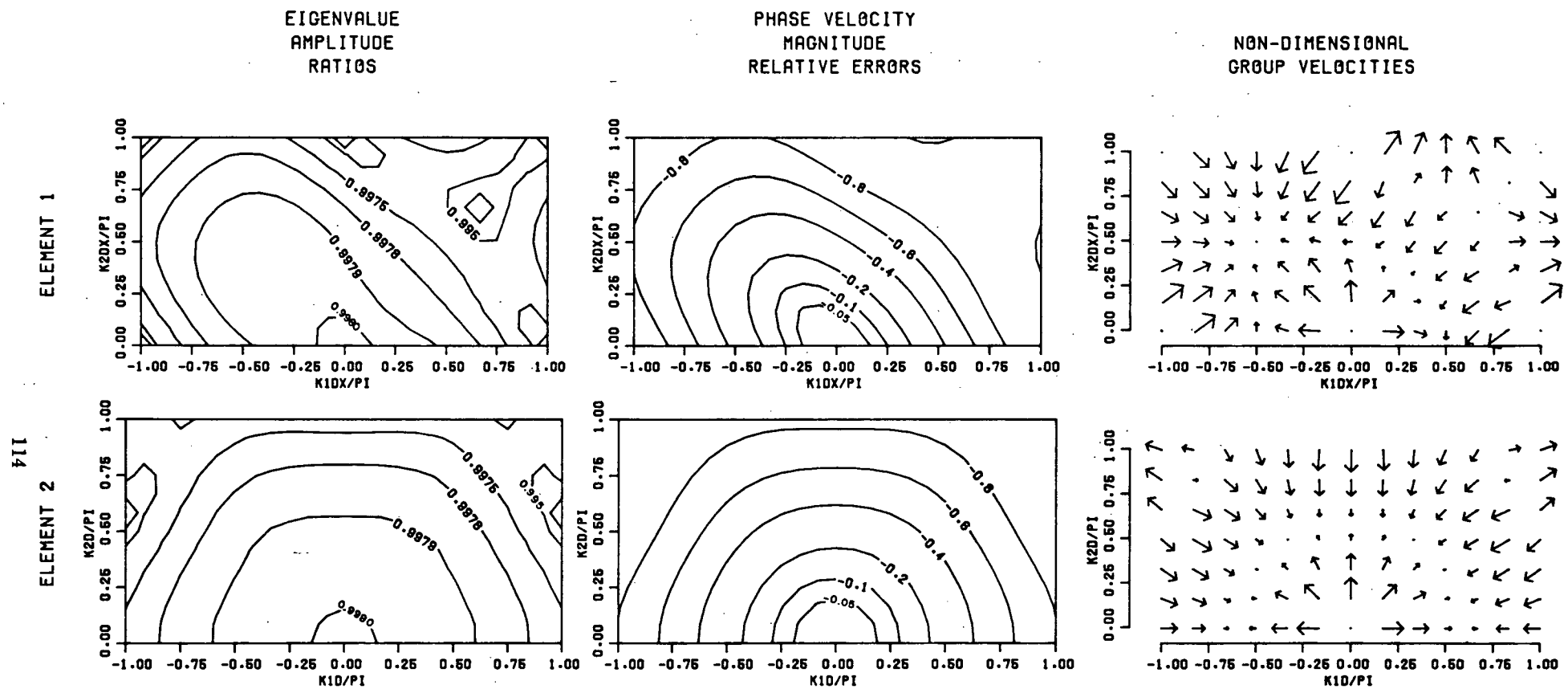

Fig. 4.8. M_A, M_C, and $\mathbf{G}/(gh)^{1/2}$ for Thacker's method with the parameter values of Fig. 4.7. Each full shaft of multi-shafted vectors denotes 1 unit (i.e., $|\mathbf{G}| = (gh)^{1/2}$).

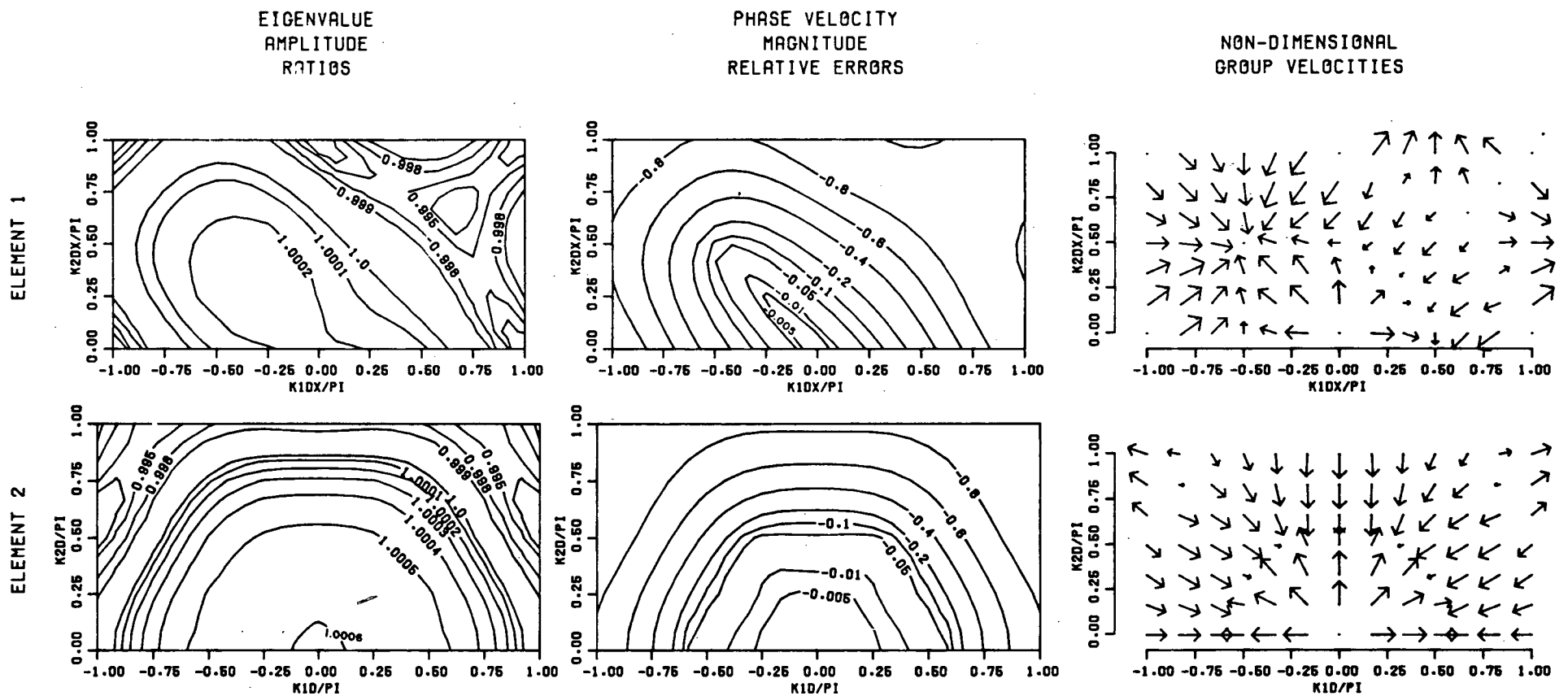

Fig. 4.9. M_A, M_C, and $\mathbf{G}/(gh)^{1/2}$ for Thacker's method with $f_1 = .05$, $f_3 = .10$, and $\theta = 0.5$. $f_2 = 1.40$ for element 1 and $f_2' = 1.70$ for element 2. Each full shaft of multi-shafted vectors denotes 1 unit (i.e., $|\mathbf{G}| = (gh)^{1/2}$).

triangle cases respectively. As theory predicts, Fig. 4.9 does display more accurate phase velocity magnitude and group velocity for small wavenumbers. The amplitude ratios are also more accurate, though they are too large in Fig. 4.9 and too small in Fig. 4.8. Notice that for small increasing wavenumbers, Fig. 4.9 shows an improvement in amplitude accuracy.

A comparison of Fig. 4.8 and Fig. 4.6 demonstrates that mass lumping can cause an accuracy loss. However, to some extent, the different time-stepping methods for the two techniques has influenced the accuracy measure values. Replacing the CN time-stepping used with the GLFEM of Fig. 4.6 with Thacker's time-stepping, actually improves, for the same parameter values, the element 2 accuracy. However, Thacker's scheme remains less accurate for both elements.

A comparison of Fig. 4.9 with Fig. 4.6 is also revealing. It illustates that an optimal Thacker scheme can be more accurate for long waves than the GLFEM. Considering the much smaller costs of running Thacker's scheme, this is a significant result. However, the most accurate Thacker scheme is not more accurate than the most accurate GLFEM, since the latter improves as f_2 decreases. But a less accurate implementation of Thacker's scheme can always be made as accurate a GLFEM by increasing its spatial resolution. Thacker claims that even with the associated cost of this refinement, his scheme will be cheaper because of its explicit nature. Further cost and accuracy comparisons with Thacker's method are given in Section 4.8.

4.6 A Mixed Interpolation FEM

In this section, the mixed interpolation FEM (henceforth GMFEM) studied in Section 2.7 is extended to two dimensions. Piecewise linear basis functions are used for the spatial approximations to $z(x, y, t)$ while piecewise quadratic functions are used for both $u(x, y, t)$ and $v(x, y, t)$. Walters and Cheng [Wa79,Wa80] have successfuly used this FEM for their tidal models of San Francisco Bay.

As in one dimension, two types of quadratic basis function are required. In addition to basis functions associated with the corner nodes illustrated in Fig. 4.4, there are also basis

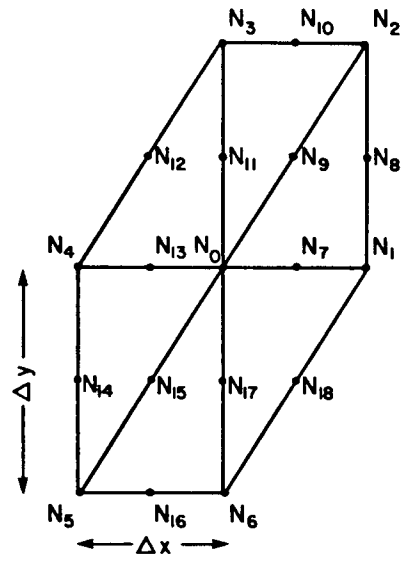

Fig. 4.10. Nodes for the mixed interpolation FEM.

functions associated with midpoint nodes. All midpoint and corner nodes for element 1 are shown in Fig. 4.10. Approximations to $z(x,y,t)$, $u(x,y,t)$, and $v(x,y,t)$ now have the form

$$\hat{z}(x,y,t) = \sum_{i=1}^{N_c} \varsigma_i(t)\phi_i(x,y) \tag{4.6.1a}$$

$$\hat{u}(x,y,t) = \sum_{i=1}^{N_c} \mu_i(t)\psi_i(x,y) + \sum_{j=1}^{N_m} \bar{\mu}_j(t)\bar{\psi}_j(x,y) \tag{4.6.1b}$$

$$\hat{v}(x,y,t) = \sum_{i=1}^{N_c} \nu_i(t)\psi_i(x,y) + \sum_{j=1}^{N_m} \bar{\nu}_j(t)\bar{\psi}_j(x,y) \tag{4.6.1c}$$

where N_c and N_m are the number of corner and midpoint nodes respectively, ϕ_i is the piecewise linear basis function associated with corner node i, ψ_i is the piecewise quadratic basis function associated with corner node i, $\bar{\psi}_j$ is the piecewise quadratic basis function associated with midpoint node j, and ς_i, μ_i, ν_i, $\bar{\mu}_j$, $\bar{\nu}_j$ are the time dependent coefficients for these basis functions. Residual continuity and momentum equations are formed by substituting these approximations into (4.2.1).

In calculating a numerical dispersion relationship, it is assumed that the solution in each region of the domain is obtained in an identical manner to the solution in every other region [Pi77]. This assumption means that the calculation of plane wave solutions can be

restricted to a small representative region of the domain. Inspection of Fig. 4.10 reveals that all eighteen nodes can be grouped into four classes: corner nodes, and midpoint nodes along vertical, horizontal, and diagonal sides. Three types of midpoint node are required because each has a different arrangement of other nodes around it. This means that each will then have different approximations to the $(\partial/\partial x)$ and $(\partial/\partial y)$ terms in (4.2.1).

Assume that N_0, N_7, N_9, and N_{11} are representative nodes of each of the four types. (All other nodes can be obtained through one of the following basic shifts: $(\pm\Delta x, 0)$, $(0, \pm\Delta y)$, $\pm(\Delta x, \Delta y)$.) It is sufficient to consider only the Galerkin conditions that arise from the basis functions associated with these four nodes. With reference to Section 1.3, these Galerkin conditions are formed by using

i) the linear basis function associated with N_0 as the weight function for the continuity equation residual,

ii) the quadratic basis function associated with N_0 as the weight function for each of the two momentum equations residuals,

iii) the quadratic basis functions associated with N_7, N_9, and N_{11} as the weight functions for each of the two momentum equation residuals.

The spatially discretized equations that arise when following this procedure are:

$$\begin{aligned}
&\frac{\partial}{\partial t}[\tfrac{1}{2}z_0 + \tfrac{1}{12}(z_1 + z_2 + z_3 + z_4 + z_5 + z_6)] \\
&+\frac{h}{3\Delta x}[u_7 - u_{13} + \tfrac{1}{2}(u_8 - u_{11} + u_9 - u_{12} + u_{18} - u_{15} + u_{17} - u_{14})] \\
&\quad +\frac{h}{3\Delta y}[v_{11} - v_{17} + \tfrac{1}{2}(v_9 - v_{18} + v_{10} - v_7 + v_{12} - v_{15} + v_{13} - v_{16})] = 0
\end{aligned} \tag{4.6.2a}$$

$$\begin{aligned}
&(\frac{\partial}{\partial t} + \tau)[u_0 - \tfrac{1}{18}(u_1 + u_2 + u_3 + u_4 + u_5 + u_6) - \tfrac{1}{9}(u_8 + u_{10} + u_{12} + u_{14} + u_{16} + u_{18})] \\
&\qquad -f[v_0 - \tfrac{1}{18}(v_1 + v_2 + v_3 + v_4 + v_5 + v_6) \\
&\qquad\qquad -\tfrac{1}{9}(v_8 + v_{10} + v_{12} + v_{14} + v_{16} + v_{18})] = 0
\end{aligned} \tag{4.6.2b}$$

$$\begin{aligned}
&(\frac{\partial}{\partial t} + \tau)[v_0 - \tfrac{1}{18}(v_1 + v_2 + v_3 + v_4 + v_5 + v_6) - \tfrac{1}{9}(v_8 + v_{10} + v_{12} + v_{14} + v_{16} + v_{18})] \\
&\qquad f[u_0 - \tfrac{1}{18}(u_1 + u_2 + u_3 + u_4 + u_5 + u_6) \\
&\qquad\qquad -\tfrac{1}{9}(u_8 + u_{10} + u_{12} + u_{14} + u_{16} + u_{18})] = 0
\end{aligned} \tag{4.6.2c}$$

$$\frac{\partial}{\partial t}+\tau)[-\tfrac{1}{45}(u_2+u_6)+\tfrac{16}{45}u_7+\tfrac{4}{45}(u_8+u_9+u_{17}+u_{18})]+\frac{2g}{3\Delta x}(z_1-z_0)$$
$$-f[-\tfrac{1}{45}(v_2+v_6)+\tfrac{16}{45}v_7+\tfrac{4}{45}(v_8+v_9+v_{17}+v_{18})]=0 \qquad (4.6.2d)$$

$$(\frac{\partial}{\partial t}+\tau)[-\tfrac{1}{45}(v_2+v_6)+\tfrac{16}{45}v_7+\tfrac{4}{45}(v_8+v_9+v_{17}+v_{18})]+\frac{g}{3\Delta y}(z_2-z_1+z_0-z_6)$$
$$+f[-\tfrac{1}{45}(u_2+u_6)+\tfrac{16}{45}u_7+\tfrac{4}{45}(u_8+u_9+u_{17}+u_{18})]=0 \qquad (4.6.2e)$$

$$(\frac{\partial}{\partial t}+\tau)[-\tfrac{1}{45}(u_3+u_1)+\tfrac{16}{45}u_9+\tfrac{4}{45}(u_7+u_8+u_{10}+u_{11})]+\frac{g}{3\Delta x}(z_1-z_0+z_2-z_3)$$
$$-f[-\tfrac{1}{45}(v_3+v_1)+\tfrac{16}{45}v_9+\tfrac{4}{45}(v_7+v_8+v_{10}+v_{11})]=0 \qquad (4.6.2f)$$

$$(\frac{\partial}{\partial t}+\tau)[-\tfrac{1}{45}(v_3+v_1)+\tfrac{16}{45}v_9+\tfrac{4}{45}(v_7+v_8+v_{10}+v_{11})+\frac{g}{3\Delta y}(z_2-z_1+z_3-z_0)$$
$$+f[-\tfrac{1}{45}(u_3+u_1)+\tfrac{16}{45}u_9+\tfrac{4}{45}(u_7+u_8+u_{10}+u_{11})]=0 \qquad (4.6.2g)$$

$$(\frac{\partial}{\partial t}+\tau)[-\tfrac{1}{45}(u_2+u_4)+\tfrac{16}{45}u_{11}+\tfrac{4}{45}(u_9+u_{10}+u_{12}+u_{13})]+\frac{g}{3\Delta x}(z_2-z_3+z_0-z_4)$$
$$-f[-\tfrac{1}{45}(v_2+v_4)+\tfrac{16}{45}v_{11}+\tfrac{4}{45}(v_9+v_{10}+v_{12}+v_{13})]=0 \qquad (4.6.2h)$$

$$(\frac{\partial}{\partial t}+\tau)[-\tfrac{1}{45}(v_2+v_4)+\tfrac{16}{45}v_{11}+\tfrac{4}{45}(v_9+v_{10}+v_{12}+v_{13})]+\frac{2g}{3\Delta y}(z_3-z_0)$$
$$+f[-\tfrac{1}{45}(u_2+u_4)+\tfrac{16}{45}u_{11}+\tfrac{4}{45}(u_9+u_{10}+u_{12}+u_{13})]=0. \qquad (4.6.2i)$$

All these equations were calculated using ***triangular area coordinates*** described and illustrated in Pinder and Gray [Pi77, pg. 96-101]. Notice that the momentum equations arising from the basis function associated with N_0 have no $(\partial z/\partial x)$ or $(\partial z/\partial y)$ approximations. At first glance, this is somewhat disconcerting. However upon reflection, it suggests that these equations serve the purpose of linking the four types of velocity, rather than approximating the momentum equations *per se*. Since all nine equations must be solved simultaneously, approximations to the $(\partial z/\partial x)$ and $(\partial z/\partial y)$ terms in (4.6.2b) and (4.6.2c) actually arise through coupling with the other equations.

Plane wave solutions to (4.6.2) have the same form as (4.4.3) but the complex constants u_0 and v_0 must each be replaced with a sum of four constants, one for each type of velocity approximant. In order to have nontrivial plane wave solutions to (4.6.2), the determinant of a 9 by 9 matrix must now be zero. This matrix is

$$\begin{pmatrix} -i\omega(\frac{1}{2}+\frac{1}{6}A) & \frac{2hi}{3\Delta x}\mathbf{P}_1^T & \frac{2hi}{3\Delta y}\mathbf{P}_2^T \\ \frac{2gi}{\Delta x}\mathbf{P}_1 & (-i\omega+\tau)M & -fM \\ \frac{2gi}{\Delta y}\mathbf{P}_2 & fM & (-i\omega+\tau)M \end{pmatrix} \qquad (4.6.3a)$$

where

$$M = \frac{1}{15}\begin{pmatrix} \frac{9}{2}(1-\frac{1}{9}A) & -c_6 & -c_4 & -c_5 \\ -c_6 & 8 & 4c_2 & 4c_3 \\ -c_4 & 4c_2 & 8 & 4c_1 \\ -c_5 & 4c_3 & 4c_1 & 8 \end{pmatrix} \tag{4.6.3b}$$

$$\mathbf{P}_1 = s_1\begin{pmatrix} 0 \\ 1 \\ c_2 \\ c_3 \end{pmatrix} \qquad \mathbf{P}_2 = s_2\begin{pmatrix} 0 \\ c_3 \\ c_1 \\ 1 \end{pmatrix} \tag{4.6.3c}$$

and

$$\begin{aligned} s_1 &= \sin(\tfrac{1}{2}k_1\Delta x) & s_2 &= \sin(\tfrac{1}{2}k_2\Delta y) \\ c_1 &= \cos(\tfrac{1}{2}k_1\Delta x) & c_2 &= \cos(\tfrac{1}{2}k_2\Delta y) \\ c_3 &= \cos(\tfrac{1}{2}k_1\Delta x + \tfrac{1}{2}k_2\Delta y) & c_4 &= \cos(\tfrac{1}{2}k_1\Delta x - \tfrac{1}{2}k_2\Delta y) \\ c_5 &= \cos(k_1\Delta x + \tfrac{1}{2}k_2\Delta y) & c_6 &= \cos(\tfrac{1}{2}k_1\Delta x + k_2\Delta y) \end{aligned} \tag{4.6.3d}$$

$$A = \cos k_1\Delta x + \cos k_2\Delta y + \cos(k_1\Delta x + k_2\Delta y).$$

One check of these algebraic calculations is to confirm that the matrix reduces to the one dimensional result whose determinant is (2.7.3). This can be done by projecting the configuration of triangles in Fig. 4.10 onto the x-axis. Applying this projection to (4.6.3) requires setting $f = k_2\Delta y = 0$, and dropping the momentum equations and the matrix columns associated with the velocity component $v(x, y, t)$. As $\Delta y \to 0$, N_9 coalesces with N_7, and N_{11} coalesces with N_0. The four $u(x, y, t)$ approximants therefore reduce to two, as they should for one dimension. Adding the momentum equations associated with the coalescing nodes now produces a 3 by 3 matrix that is identical to the one dimensional matrix in Section 2.7.

Despite the structure of the 9 by 9 matrix, manual calculation of its determinant is a huge undertaking, prone to many errors. In order to avoid these difficulties, the algebraic manipulation package REDUCE was used instead. This package was able to calculate the determinant, but the resultant expression was so long and complicated that it was essentially useless. Even with the simplification $f = 0$, the algebraic expression for the determinant required 600 lines of computer output!

Were the result not so complicated, the next step would be to re-arrange the determinant expression into a polynomial of order 9 in ω. Then with the aid of a numerical

routine for finding polynomial roots (e.g., from the NAG or IMSL libraries), nine values of ω could be calculated for specific values of $(k_1\Delta x, k_2\Delta x)$. Of these nine values, two would be principal roots and the other seven would be spurious. Assuming Crank Nicolson time-stepping, numerical eigenvalues for the fully discretized equations could then be calculated with (4.4.8). From these λs, frequencies, phase velocities, and amplitude decay factors, could be calculated as they were for the GLFEM. Group velocity calculations, however, would require differentiation of the polynomial and could be quite messy, even with REDUCE. Approximations to **G** through the use of difference approximations to the partial derivatives would be much easier.

An analysis of the GMFEM with the equilateral triangles of element 2 would encounter the same difficulties. Three distinct midpoint nodes would again be required for velocities in the 0°, 60°, and 120° directions, and the determinant of a 9 by 9 matrix would also be required to calculate the spatially discretized dispersion relationships.

In short, the GMFEM can be analysed in a similar manner to the GLFEM. However, much more work is involved. It was felt that the effort would not justify the results, so no further calculations were made.

4.7 The WEM and LWEM

This section extends the results of Chapter 3 to the two dimensional triangular elements of Fig. 4.4.

The linearized, two dimensional, constant depth version of the continuity equation solved by Lynch and Gray [Ly79] is

$$\frac{\partial^2 z}{\partial t^2} + \tau\frac{\partial z}{\partial t} - gh\left(\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2}\right) + hf\left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right) = 0. \tag{4.7.1}$$

It is solved in combination with the momentum equations (4.2.1b) and (4.2.1c). The characteristic equation for this system of PDEs is simply the product of (4.2.4) and $(-i\omega + \tau)$. Consequently, replacing (4.2.1a) with (4.7.1) produces an additional dispersion relationship whose associated solution is a stationary wave that decays in time when $\tau > 0$. This same effect was noted in one dimension.

Since Lynch and Gray employ piecewise linear approximating functions, imposing the Galerkin conditions to the Laplacian term in (4.7.1) necessitates an integration by parts. For the basis function associated with node N_0 in element 1, the spatially discretized Galerkin equation arising from (4.7.1) is

$$\begin{aligned}\left(\frac{\partial^2}{\partial t^2}+\tau\frac{\partial}{\partial t}\right)&[\tfrac{1}{2}z_0+\tfrac{1}{12}(z_1+z_2+z_3+z_4+z_5+z_6)]\\ &+hf\left[\frac{2}{3}\left(\frac{v_1-v_4}{2\Delta x}\right)+\frac{1}{6}\left(\frac{v_2-v_3}{\Delta x}\right)+\frac{1}{6}\left(\frac{v_6-v_5}{\Delta x}\right)\right]\\ &-hf\left[\frac{2}{3}\left(\frac{u_3-u_6}{2\Delta y}\right)+\frac{1}{6}\left(\frac{u_2-u_1}{\Delta y}\right)+\frac{1}{6}\left(\frac{u_4-u_5}{\Delta y}\right)\right]\\ &-gh\left[\frac{z_1-2z_0+z_4}{\Delta x^2}+\frac{z_3-2z_0+z_6}{\Delta y^2}\right]=0.\end{aligned}\tag{4.7.2}$$

The analogous result for element 2 is

$$\begin{aligned}\left(\frac{\partial^2}{\partial t^2}+\tau\frac{\partial}{\partial t}\right)&[\tfrac{1}{2}z_0+\tfrac{1}{12}(z_1+z_2+z_3+z_4+z_5+z_6)]\\ &+hf\left[\frac{2}{3}\left(\frac{v_1-v_4}{2\Delta x}\right)+\frac{1}{6}\left(\frac{v_2-v_3}{\Delta x}\right)+\frac{1}{6}\left(\frac{v_6-v_5}{\Delta x}\right)\right]\\ &-hf\left[\frac{1}{2}\left(\frac{u_2-u_6}{2\Delta y}\right)+\frac{1}{2}\left(\frac{u_3-u_5}{2\Delta y}\right)\right]-gh\left(\frac{z_1-2z_0+z_4}{\Delta x^2}\right)\\ &-\frac{gh}{\Delta y^2}[\tfrac{1}{4}(z_2-z_1-z_0+z_6)+\tfrac{1}{4}(z_3-z_0-z_4+z_5)+\tfrac{1}{2}(\tfrac{1}{2}(z_2+z_3+z_6+z_5)-2z_0)]=0.\end{aligned}\tag{4.7.3}$$

The associated spatially discretized momentum equations are given by (4.4.1b), (4.4.1c), and (4.4.2b), (4.4.2c) respectively.

Assuming plane wave solutions of the form (4.4.3), dispersion relationships can be found for these spatial discretizations. In both cases they are derived from the roots of the polynomial

$$\begin{aligned}\omega^4+3i\tau\omega^3-\omega^2(3\tau^2+f^2+2ghB/A)-i\tau\omega(\tau^2+f^2+4ghB/A)\\ +gh\left[2(\tau^2+f^2)\frac{B}{A}-\frac{f^2}{A^2}\left(\frac{G_x^2}{\Delta x^2}+\frac{G_y^2}{\Delta y^2}\right)\right]=0.\end{aligned}\tag{4.7.4}$$

For element 1, A, G_x, and G_y are defined by (4.4.5) while

$$B=\frac{1-\cos k_1\Delta x}{\Delta x^2}+\frac{1-\cos k_2\Delta y}{\Delta y^2}.\tag{4.7.5a}$$

For element 2, A, G_x, and G_y are defined by (4.4.6) while

$$B = \frac{1-\cos k_1\Delta x}{\Delta x^2} + \frac{\frac{1}{4}(3+\cos k_1\Delta x) - \cos(\frac{1}{2}k_1\Delta x)\cos k_2\Delta y}{\Delta y^2}. \qquad (4.7.5b)$$

The fully discretized *wave equations* are obtained by applying the time-stepping approximations (3.3.1) to the spatially discretized system of ODEs. Again θ is the weighting parameter for the gravity terms, and α, is the weighting parameter for the friction terms in the momentum equations. All Coriolis terms are evaluated directly at time level n. Consequently, the fully discretized continuity equation does not involve velocities at time level $n+1$, although the momentum equations do require elevations at that level. This permits a computational time saving since the continuity and momentum equations may now be solved sequentially rather than simultaneously.

The two dimensional dispersion relationships for the fully discretized equations are calculated by assuming plane wave solutions of the form (4.4.7). With G_x, G_y, A, and B defined appropriately for elements 1 and 2, the characteristic equation is

$$[\Upsilon_1 + \tfrac{1}{2}\tau\Delta t\Upsilon_2 + 2gh(\Delta t)^2\Upsilon_\theta B/A][\Upsilon_2^2 + 4\tau\Delta t\Upsilon_2\Upsilon_\alpha + 4(\Delta t)^2(\tau^2\Upsilon_\alpha^2 + f^2\lambda^2)] - \frac{4gh\lambda^2 f^2\Upsilon_\theta(\Delta t)^4}{A^2}\left(\frac{G_x^2}{\Delta x^2} + \frac{G_y^2}{\Delta y^2}\right) = 0 \qquad (4.7.6a)$$

where

$$\Upsilon_1 = (\lambda - 1)^2 \qquad (4.7.6b)$$

$$\Upsilon_2 = \lambda^2 - 1 \qquad (4.7.6c)$$

$$\Upsilon_\theta = \tfrac{1}{2}\theta(\lambda^2+1) + (1-\theta)\lambda \qquad (4.7.6d)$$

$$\Upsilon_\alpha = \tfrac{1}{2}\alpha(\lambda^2+1) + (1-\alpha)\lambda. \qquad (4.7.6e)$$

Two roots of (4.7.6) approximate the gravity wave solutions (4.2.5b) and (4.2.5c). They are the principal roots. The remaining four roots are either approximations to the stationary modes $\omega = -i\tau$ and (4.2.5a), or spurious roots.

A linear stability analysis of the WEM is difficult when f is nonzero. However, necessary stability conditions with realistic nonzero values of f should only be perturbations of

the conditions derived by assuming $f = 0$. Therefore, the restrictions obtained by assuming $f = 0$ should be close to those required with nonzero Coriolis. Numerical computations confirm this.

When $f = 0$, (4.7.6) reduces to

$$Q_1 Q_2^2 = 0 \tag{4.7.7a}$$

$$\text{where} \quad Q_1 = \Upsilon_1 + \tfrac{1}{2}\tau\Delta t\Upsilon_2 + 2gh(\Delta t)^2\Upsilon_\theta B/A \tag{4.7.7b}$$

$$\text{and} \quad Q_2 = \Upsilon_2 + 2\tau\Delta t\Upsilon_\alpha. \tag{4.7.7c}$$

This result is similar to the one dimensional characteristic equation , $Q_1 Q_2 = 0$, specified by (3.3.2). Specifically, Q_2 is identical to (3.3.2a), and Q_1 can be expressed as (3.3.2b) when E^2 is defined as

$$E^2 = 2gh(\Delta t)^2 B/A. \tag{4.7.8}$$

This similarity implies that the one dimensional stability analysis can be followed here.

The roots of Q_2 are parasitic. In one dimension, they are stable when [Ly79]

$$\alpha \geq \tfrac{1}{2}. \tag{4.7.9}$$

In two dimensions, each of the parasitic roots has multiplicity 2, thereby requiring the more restrictive condition $\alpha > \frac{1}{2}$. As discussed in Section 3.8, $\alpha = 1$ is a good choice since it generally ensures that the spurious root magnitudes are less than those of the principal roots.

The propagating principal roots of Q_1 are stable when

$$\theta \geq -\frac{2}{E^2}(1 + \tfrac{1}{2}\tau\Delta t) \tag{4.7.10a}$$

for all E^2 and $\tau\Delta t$. The nonpropagating principal roots are stable when

$$\theta \geq \frac{1}{2}\left(1 - \frac{4}{E^2}\right) \tag{4.7.10b}$$

for all E^2. Assuming positive $\tau\Delta t$, the second condition is more restrictive. For element 1 with $\Delta x = \Delta y$, it reduces to

$$\theta \geq \frac{1}{2}\left(1 - \frac{1}{6.4641 f_2^2}\right) \tag{4.7.11a}$$

while for the equilateral triangles of element 2, it becomes

$$\theta \geq \frac{1}{2}\left(1 - \frac{1}{5.5783(f_2')^2}\right). \tag{4.7.11b}$$

With $k_1 = 0$, E^2 is identical to its one dimensional counterpart. Hence all the roots and the stability constraints reduce to those in (3.3.3) and (3.3.4). Setting $k_1 = 0$ is equivalent to projecting both elements of Fig. 4.4 onto the y-axis. The six nodes coalesce to three nodes which are uniformly separated by Δy. Consequently, it is not surprising that the two dimensional characteristic equation is closely related to its one dimensional counterpart.

In one dimension it was possible to choose a value of θ which produces the most accurate WEM dispersion relationship. This is also possible in two dimensions. Assuming $f = \tau = 0$, the principal numerical eigenvalues are the roots of the quadratic

$$(\lambda - 1)^2 + 2gh(\Delta t)^2[\tfrac{1}{2}\theta(\lambda^2 + 1) + (1 - \theta)]B/A = 0. \tag{4.7.12}$$

With element 1 and $\Delta x = \Delta y$, substitution for λ leads to the dispersion relationship

$$\cos\omega\Delta t = \frac{1 - (1 - \theta)D}{1 + \theta D} \tag{4.7.13a}$$

$$\text{where} \qquad D = \frac{f_2^2(2 - \cos\xi - \cos\eta)}{\frac{1}{2} + \frac{1}{6}(\cos\xi + \cos\eta + \cos(\xi + \eta))} \tag{4.7.13b}$$

and as before, $\xi = k_1\Delta x$, $\eta = k_2\Delta x$. This dispersion relationship can be compared with the analytic result for small ξ and η.

An asymptotic expansion for D is

$$D \simeq f_2^2[\tfrac{1}{2}(\xi^2 + \eta^2) + \tfrac{1}{12}(\xi^2 + \eta^2)^2 + \tfrac{1}{12}\xi\eta(\xi^2 + \eta^2) - \tfrac{1}{24}(\xi^4 + \eta^4)]. \tag{4.7.14}$$

Assuming $|\theta D| < 1$, substitution in (4.7.13a) then gives

$$\begin{aligned}\cos\omega\Delta t \simeq 1 - \tfrac{1}{2}f_2^2(\xi^2 + \eta^2) + (\xi^2 + \eta^2)^2(\tfrac{1}{4}\theta f_2^4 - \tfrac{1}{12}f_2^2) \\ - \tfrac{1}{12}f_2^2[\xi\eta(\xi^2 + \eta^2) - \tfrac{1}{2}(\xi^4 + \eta^4)].\end{aligned} \tag{4.7.15}$$

Matching this expansion with the corresponding analytic result

$$\cos\omega\Delta t \simeq 1 - \tfrac{1}{2}f_2^2(\xi^2 + \eta^2) + \tfrac{1}{24}f_2^4(\xi^2 + \eta^2)^2 \tag{4.7.16}$$

does not yield one value of θ which is best for all wave directions. In particular with $\eta = s\xi$, matching terms in (4.7.15) and (4.7.16) requires

$$\theta = \frac{1}{6} + \frac{1}{3f_2^2} + \frac{1}{6f_2^2}\left[\frac{2s}{1+s^2} - \frac{1+s^4}{(1+s^2)^2}\right]. \tag{4.7.17}$$

With $s = 0$ or $s = \infty$, (4.7.17) reduces to the one dimensional result (3.6.12). This value produces the most accurate representation of wave propagation along either axis of element 1 in Fig. 4.4. However for waves propagating along the diagonal (i.e., $s = 1$), the optimal time-stepping parameter is

$$\theta = \frac{1}{6} + \frac{5}{12f_2^2}. \tag{4.7.18}$$

With $f_2 = 2^{-1/2}$, these two optimal values are appreciably different, namely $\frac{1}{2}$ and 1.

Accuracy which varies with wave direction is clearly undesirable. It implies that grid orientation can affect the model accuracy and that by simply changing direction, a wave may be less accurately represented. Fortunately, with the preceding simplifying assumptions this directional dependence can be avoided with equilateral triangles.

With $\xi = k_1 d$, $\eta = k_2 d$, and $rd = \Delta y$, the dispersion relationship for element 2 is again given by (4.7.13a), but

$$D = (f_2')^2\left(\frac{1 - \cos\xi + (\frac{3}{4} + \frac{1}{4}\cos\xi - \cos(\frac{1}{2}\xi)\cos(r\eta))/r^2}{\frac{1}{2} + \frac{1}{6}\cos\xi + \frac{1}{3}\cos(\frac{1}{2}\xi)\cos(r\eta)}\right). \tag{4.7.19}$$

Its asymptotic expansion for small ξ and η is

$$D \simeq (f_2')^2\left[\frac{1}{2}(\xi^2+\eta^2) + \xi^4\left(\frac{1}{48} + \frac{1}{128r^2}\right) + \frac{1}{24}r^2\eta^4 + \frac{1}{12}r^2\xi^2\eta^2\right]. \tag{4.7.20}$$

The associated dispersion relationship expansion is

$$\begin{aligned}\cos\omega\Delta t \simeq 1 &- \tfrac{1}{2}(f_2')^2(\xi^2+\eta^2) + \tfrac{1}{4}\theta(f_2')^4(\xi^2+\eta^2)^2 \\ &-(f_2')^2\left[\xi^4\left(\frac{1}{48} + \frac{1}{128r^2}\right) + \frac{1}{24}r^2\eta^4 + \frac{1}{12}r^2\eta^2\xi^2\right].\end{aligned} \tag{4.7.21}$$

For equilateral triangles, $r = 3^{1/2}/2$ and (4.7.21) becomes

$$\cos\omega\Delta t \simeq 1 - \tfrac{1}{2}(f_2')^2(\xi^2+\eta^2) + (f_2')^4(\xi^2+\eta^2)^2\left(\frac{\theta}{4} - \frac{1}{32(f_2')^2}\right). \tag{4.7.22}$$

Matching with the analytic expansion then yields

$$\theta = \frac{1}{6} + \frac{1}{8(f_2')^2} \tag{4.7.23}$$

as the time-stepping parameter value which most accurately approximates analytic wave propagation. Unlike (4.7.17), it is not directionally dependent. Furthermore, with the substitution $\Delta y = (3^{1/2}/2)d$, (4.7.23) is identical to the one dimensional result (3.6.12). Again this substitution is equivalent to projecting element 2 onto the k_2 axis.

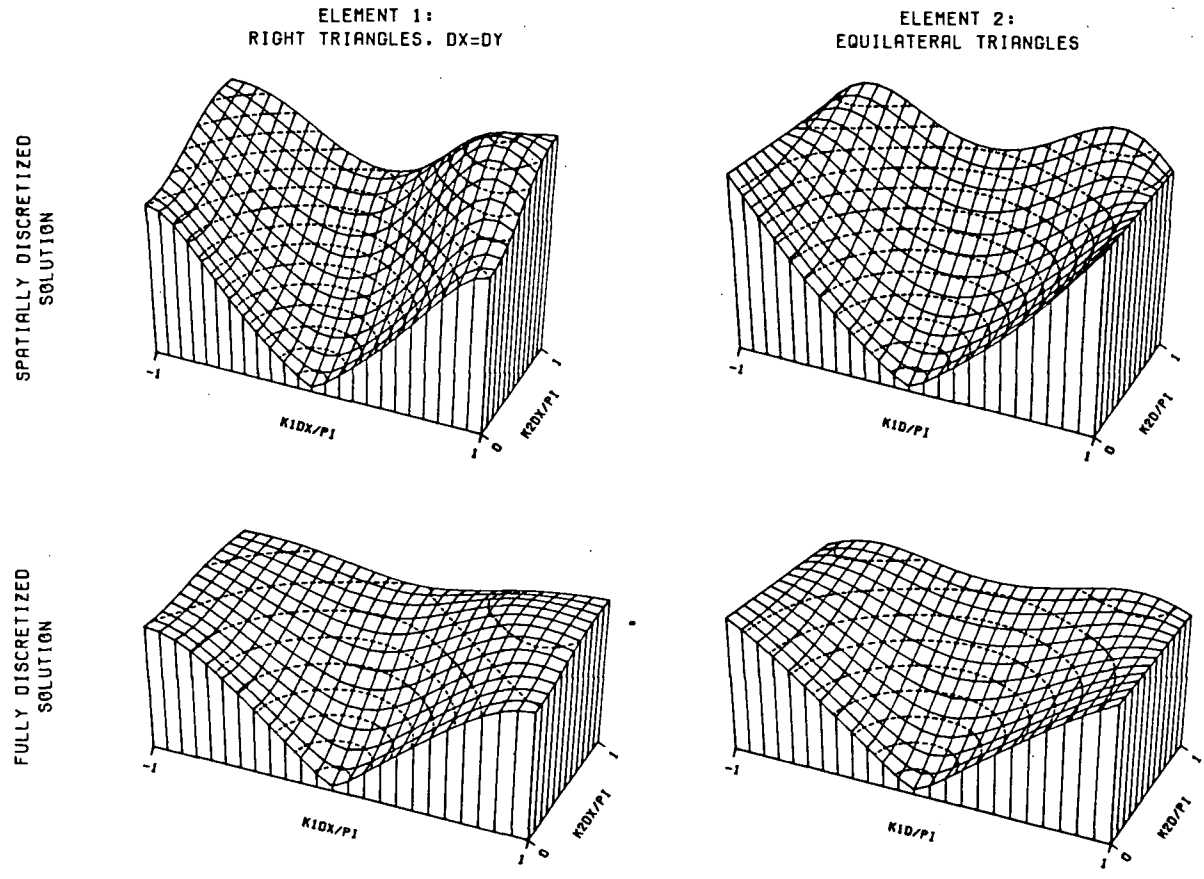

Fig. 4.11. Dispersion surfaces ($|\omega|\Delta x/(\pi(gh)^{1/2})$) for the WEM. Parameter values are $f_1 = .05$, $f_2 = .7071$, $f_3 = .10$, and $\alpha = 1$. $\theta = 0.5$ for element 1 and $\theta = .45534$ for element 2. Dotted line contours are in increments of .10.

Fig. 4.11 shows the principal dispersion surfaces for the spatially and fully discretized versions of the WEM. It has the same parameter values, scale, and perspective as Fig. 4.2, 4.5, and 4.7. The time-stepping parameters for the fully discretized equations are $\theta = 0.5$ for element 1 and $\theta = .45534$ for element 2. The former is optimal for one dimensional wave propagation along the ξ or η axis, while the latter is optimal for all directions. Both

values satisfy the principal eigenvalue stability conditions (4.7.11). For both elements, choosing $\alpha = 1$ ensures stability of the parasitic eigenvalues.

Unlike the GLFEM and Thacker's scheme, these surfaces do not have local minima at large wavenumbers. This implies that short waves do not have the small inertial phase velocities discussed in Sections 4.4 and 4.5. Provided the parasitic waves do not contaminate the numerical solution, the WEM (with these f_1, f_2, f_3 values) should therefore not have the same short wave problems as the GLFEM and Thacker's method.

The M_A and M_C contours and group velocity vectors corresponding to the dispersion surface plots of Fig. 4.11 are shown in Fig. 4.12. High phase velocity accuracy along the axes of element 1 is evident but seems to occur at the expense of accuracy in other directions. For virtually all wavenumbers, element 2 more accurately approximates wave propagation than element 1. And for small wavenumbers, its wave amplitude approximations are also slightly more accurate.

The numerical quadrature employed by Lynch and Gray [Ly79] has the effect of lumping their equations. As with Thacker's scheme, this lumping causes all terms of the form $\frac{1}{2}z_0 + \frac{1}{12}(z_1 + z_2 + z_3 + z_4 + z_5 + z_6)$ in the spatially discretized equations to be replaced by z_0. The associated dispersion equation (4.7.4) then requires the re-definition $A = 1$. The fully discretized lumped equations, and their associated characteristic equation (4.7.6) require these same substitutions. When $\theta = 0$, the LWEM is explicit.

Necessary stability restrictions for the LWEM can also be found when $f = 0$. The parasitic eigenvalues are identical to those for the WEM and are thus governed by the same stability conditions. Similarly, with $A = 1$ substituted in (4.7.8), the principal eigenvalues are stable when (4.7.10b) is satisfied. These conditions reduce to

$$\theta \geq \frac{1}{2}\left(1 - \frac{1}{2f_2^2}\right) \tag{4.7.24a}$$

for $\Delta x = \Delta y$ and element 1, and

$$\theta \geq \frac{1}{2}\left(1 - \frac{1}{1.47218(f_2')^2}\right) \tag{4.7.24b}$$

for the equilateral triangles of element 2.

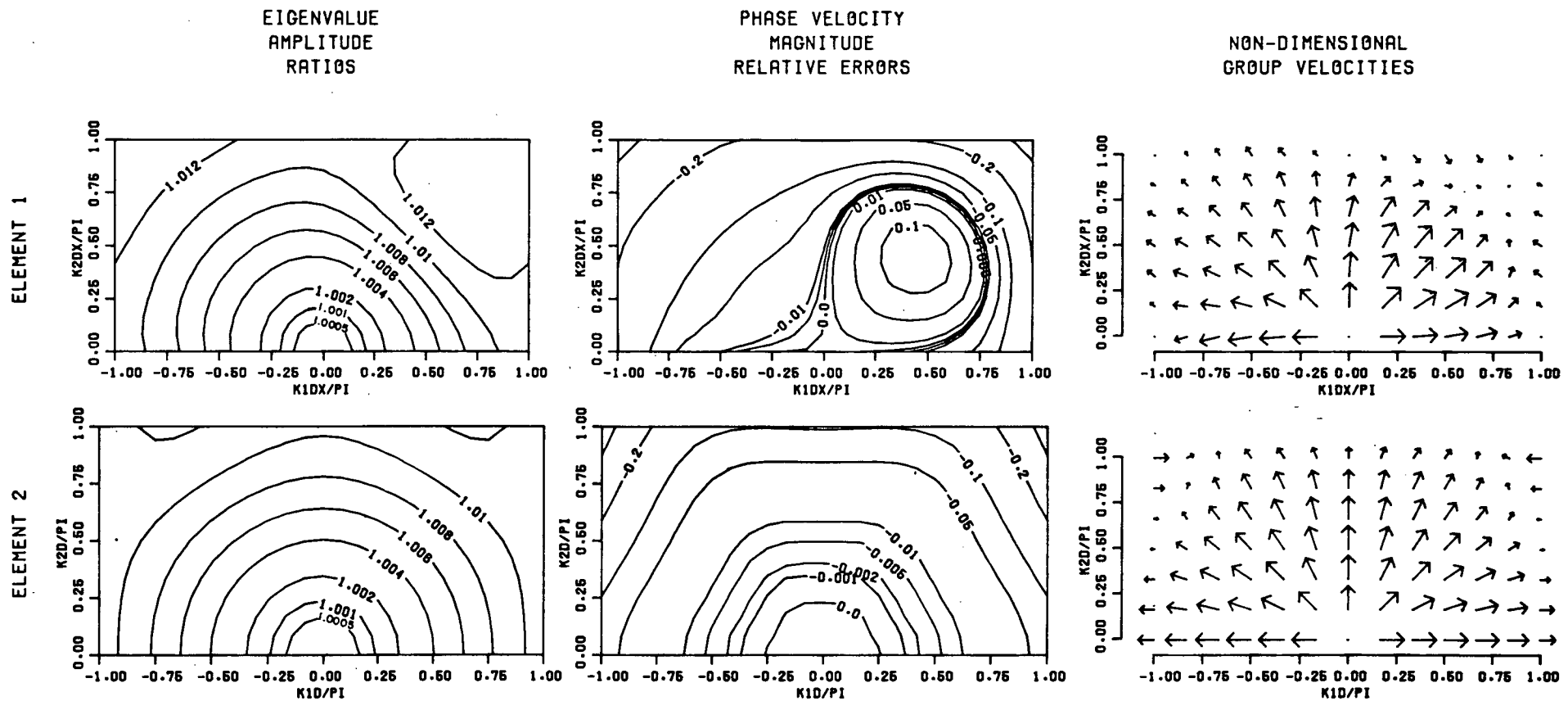

Fig. 4.12. M_A, M_C, and $\mathbf{G}/(gh)^{1/2}$ for the WEM with the parameter values of Fig. 4.11. Each full shaft of multi-shafted vectors denotes 1 unit (i.e., $|\mathbf{G}| = (gh)^{1/2}$).

The two dimensional LWEM also has values of θ that are most accurate for wave propagation. Assume $f = \tau = 0$. With element 1 and $\Delta x = \Delta y$, the LWEM dispersion relationship is again given by (4.7.13a), but

$$D = f_2^2(2 - \cos\xi - \cos\eta). \tag{4.7.25}$$

The asymptotic dispersion relationship for small ξ and η then becomes

$$\cos\omega\Delta t \simeq 1 - \tfrac{1}{2}f_2^2(\xi^2 + \eta^2) + \tfrac{1}{24}f_2^2(\xi^4 + \eta^4) + \tfrac{1}{4}\theta f_2^4(\xi^2 + \eta^2)^2. \tag{4.7.26}$$

Matching with (4.7.16) and setting $\eta = s\xi$ then requires

$$\theta = \frac{1}{6}\left[1 - \frac{1 + s^4}{f_2^2(1 + s^2)^2}\right], \tag{4.7.27}$$

which again varies with wave direction. As with the WEM, $s = 0$ or $s = \infty$ produces the one dimensional result (3.6.14).

With $\xi = k_1 d$, $\eta = k_2 d$, and $rd = \Delta y$, the LWEM dispersion relationship for element 2 is also given by (4.7.13a) but

$$D = (f_2')^2[1 - \cos\xi + (\tfrac{3}{4} + \tfrac{1}{4}\cos\xi - \cos(\tfrac{1}{2}\xi)\cos(r\eta))/r^2]. \tag{4.7.28}$$

The asymptotic dispersion relationship for small ξ and η becomes

$$\begin{aligned}\cos\omega\Delta t \simeq 1 + (f_2')^2\left[-\frac{1}{2}(\xi^2 + \eta^2) - \xi^4\left(\frac{1}{128r^2} - \frac{1}{24}\right) + \frac{1}{24}r^2\eta^4 + \frac{1}{16}\xi^2\eta^2\right] \\ + \frac{\theta}{4}(\xi^2 + \eta^2)^2(f_2')^4.\end{aligned} \tag{4.7.29}$$

For equilateral triangles, (4.7.26) becomes

$$\cos\omega\Delta t \simeq 1 - \tfrac{1}{2}(f_2')^2(\xi^2 + \eta^2) + (f_2')^4(\xi^2 + \eta^2)^2\left[\frac{\theta}{4} + \frac{1}{32(f_2')^2}\right]. \tag{4.7.30}$$

Matching with the analytic expansion then yields

$$\theta = \frac{1}{6} - \frac{1}{8(f_2')^2}. \tag{4.7.31}$$

Again, this optimal value is not directionally dependent. And with the substitution $\Delta y = (3^{1/2}/2)d$, it is identical to the one dimensional result (3.6.14).

In one dimension, $\theta = 0$ was seen to produce the most accurate approximation of gravity wave amplitudes for both the WEM and LWEM. The same is true in two dimensions when $f = 0$ and $\tau > 0$. In fact, it is true for both element 1 when $\Delta x = \Delta y$, and the equilateral triangles of element 2. Furthermore, when $f = \tau = 0$, all stable values of θ produce exact amplitudes.

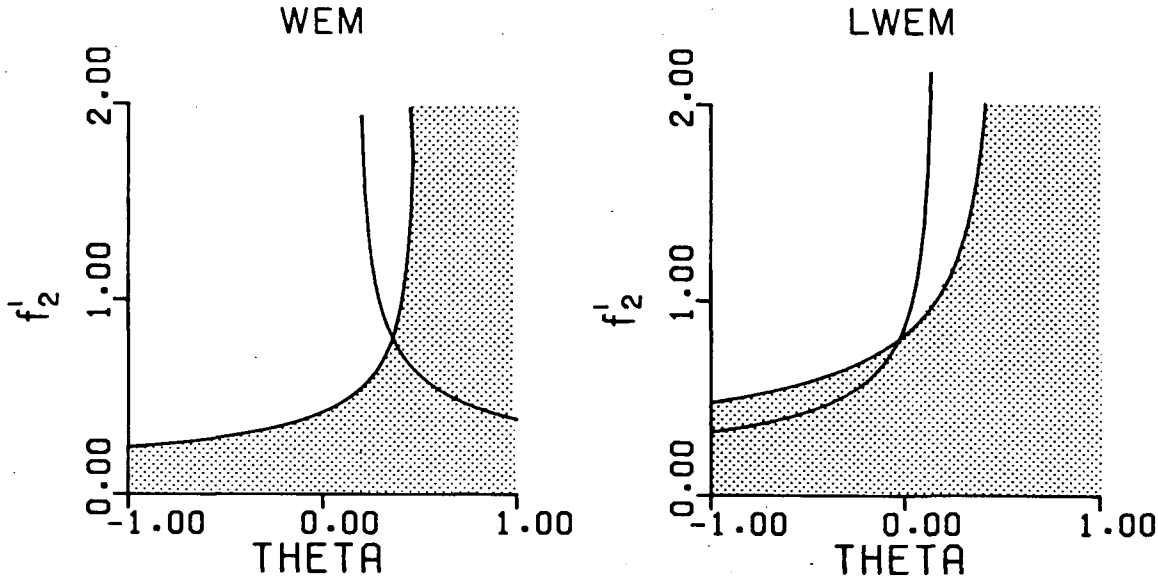

Fig. 4.13. Stability and accuracy for the WEM and LWEM over equilateral triangles. Shaded regions denote stability. Solid lines designate the most accurate values of θ and f_2' for wave propagation.

Fig. 4.13 illustrates the stability regions and the most accurate time-stepping parameter values of θ for the WEM and LWEM over a configuration of equilateral triangles. Values for f_2' and θ should be chosen so that the resultant numerical method is stable. The particular choice may be a compromise between accuracy and time step size. Large values of Δt (or f_2') result in less computation cost but may be less accurate. Computationally, the explicit LWEM ($\theta = 0$) should be most economical. Unfortunately, the associated f_2' value which yields optimal accuracy ($f_2' = .866025$) is outside the stability region. $f_2' = .824175$ is the most accurate and stable choice.

Fig. 4.14 shows the principal dispersion surfaces for the spatially and fully discretized versions of the LWEM. It has the same parameter values, scale, and perspective as Fig. 4.2, 4.5, 4.7, and 4.11. The time-stepping parameters for the fully discretized equations

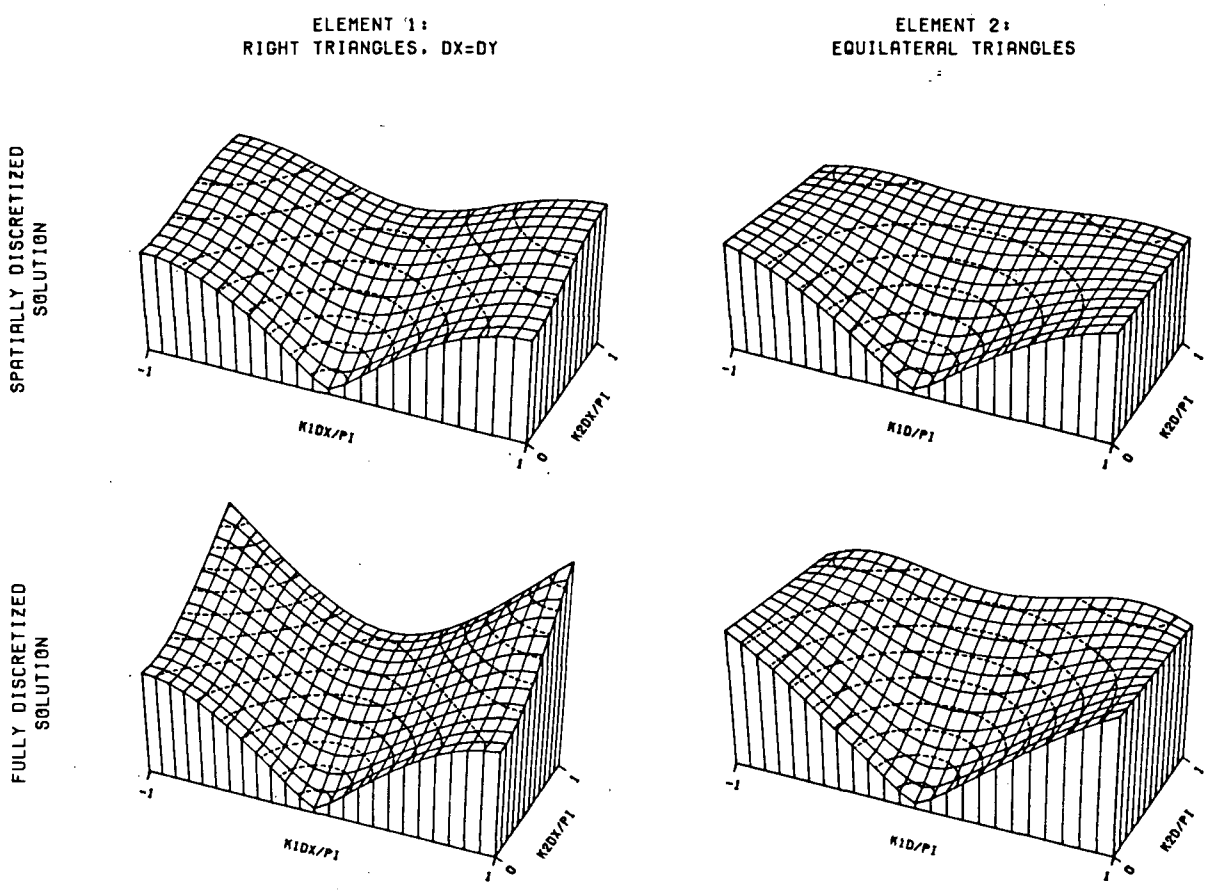

Fig. 4.14. Dispersion surfaces ($|\omega|\Delta x/(\pi(gh)^{1/2})$) for the LWEM. Parameter values are $f_1 = .05$, $f_2 = .7071$, $f_3 = .10$, and $\alpha = 1$. $\theta = 0$ for element 1 and $\theta = -.122$ for element 2. Dotted line contours are in increments of .10.

are $\theta = 0$ for element 1 and $\theta = -.122$ for element 2. The former is optimal for wave propagation along the directions $\eta = \pm\xi$, and as seen from (4.7.24a), lies just within the stability limit. (Choosing $\theta = -\frac{1}{6}$, the optimal value for wave propagation along the ξ or η axis, would be unstable.) The value for element 2 is optimal for all wave directions and satisfies the stability constraint (4.7.24b).

Comparing the spatially discretized surfaces in Fig. 4.11 and Fig. 4.13, it is evident that lumping has reduced the ω values. However the chosen time-stepping methods are seen to lower the WEM values and raise the LWEM values so that the fully discretized surfaces are more similar.

Fig. 4.15 shows the M_A and M_C contours and the group velocity vectors associated with the dispersion surfaces of Fig. 4.14. Accurate wave propagation along the lines $\eta = \pm\xi$ for element 1 is evident, but appears to be at the expense of accuracy in other

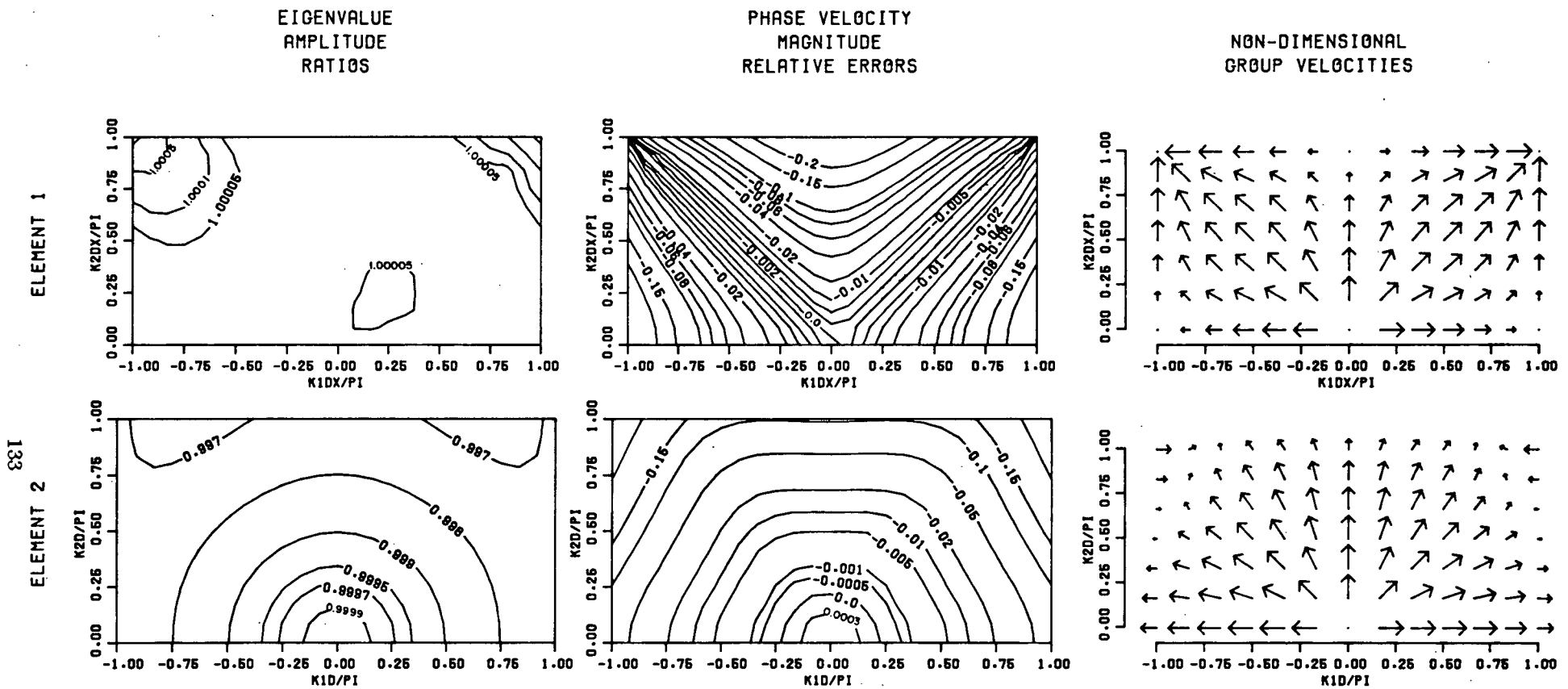

Fig. 4.15. M_A, M_C, and $\mathbf{G}/(gh)^{1/2}$ for the LWEM with the parameter values of Fig. 4.14. Each full shaft of multi-shafted vectors denotes 1 unit (i.e., $|\mathbf{G}| = (gh)^{1/2}$).

directions. For small wavenumbers, element 2 displays the same accuracy in all directions and is generally more accurate than element 1. Wave amplitude accuracy also seems to be independent of direction for element 2. However it is slightly less accurate than the amplitudes associated with element 1.

The M_C values for element 2 are virtually identical in Fig. 4.12 and Fig. 4.15. In fact, with $f_1 = f_3 = 0$, they would be equal. Denoting the optimal parameter values of (4.7.23) and (4.7.31) by θ^*, both (4.7.22) and (4.7.30) when expanded to terms of order 6 can be expressed as

$$\begin{aligned}\cos\omega\Delta t \simeq & 1 - \tfrac{1}{2}(f_2')^2(\xi^2+\eta^2) + \tfrac{1}{4}(f_2')^4(\xi^2+\eta^2)^2[\tfrac{1}{6} + \theta - \theta^*] \\ & - \tfrac{1}{8}(f_2')^2(\xi^2+\eta^2)^3[(f_2')^4(\theta-\theta^*+\tfrac{1}{6})^2 - \tfrac{1}{64}] \\ & - \tfrac{1}{256}(f_2')^2[\tfrac{11}{45}\xi^6 + \tfrac{1}{5}\eta^6 + \xi^2\eta^4 + \tfrac{1}{3}\xi^4\eta^2].\end{aligned} \tag{4.7.32}$$

This implies that around their respective θ^* values, both the WEM and LWEM have the same accuracy deterioration for $\cos\omega\Delta t$. As was found in one dimension, the best time-stepping method for the lumped scheme produces the same propagation accuracy (to order 8 in $(k_1 d, k_2 d)$) as the best time-stepping method for the unlumped method. Notice that the associated M_A values indicate amplitudes which are too small for the LWEM and too large for the WEM.

The fully discretized LWEM principal dispersion surface shown in Fig. 4.14 for element 1 is remarkably similar to the RS dispersion surface in Fig. 4.2. In fact, when $f = \tau = 0$ and $\theta = 0$ for the LWEM, not only are the principal dispersion surfaces identical, but the principal characteristic equations are also identical, even when $\Delta x \neq \Delta y$. This is seen as follows. Set

$$Q_* = (\lambda - 1)^2 + 4\lambda gh(\Delta t)^2\left(\frac{\sin^2(\frac{1}{2}k_1\Delta x)}{(\Delta x)^2} + \frac{\sin^2(\frac{1}{2}k_2\Delta y)}{(\Delta y)^2}\right). \tag{4.7.33}$$

Then with $f = \tau = 0$, the RS characteristic equation becomes (from (4.3.3))

$$(\lambda - 1)Q_* = 0, \tag{4.7.34a}$$

and the LWEM characteristic equation with $\theta = 0$ becomes (from (4.7.7) and (4.7.5a) with $A = 1$)

$$(\lambda^2 - 1)^2 Q_* = 0. \tag{4.7.34b}$$

This implies that the two principal numerical solutions not only propagate identically (i.e., have the same dispersion relationship), but their wave amplitudes also decay or grow at the same rate. (The LWEM does however have three spurious modes which may contaminate the numerical solution.) Since the accuracy measure and asymptotic analyses indicate that the LWEM is more accurate for wave propagation when combined with equilateral rather than right triangles, it seems that the LWEM can be more accurate than the RS scheme. Moreover, since both schemes are explicit, they should be comparable economically. These points are examined further in the next section.

4.8 Comparisons of Accuracy and Economy

Most FEMs are more expensive than explicit FDMs. This is the case with the GLFEM and the RS scheme. This disadvantage is primarily due to the nondiagonal matrix equation which must be solved at each time step. In Sections 4.5 and 4.7, it was seen that with explicit time-stepping, both Thacker's scheme and the LWEM produce diagonal matrices. Hence they should be much cheaper than the GLFEM. In this section, cost and accuracy comparisons are given for the RS scheme over a square grid, and Thacker's method and the LWEM over a configuration of equilateral triangles.

Unfortunately both Thacker's scheme and the explicit LWEM are unstable at the f_2' values which produce their best wave propagation accuracy. Furthermore, their most accurate and stable f_2' values are significantly different. For identical configurations of equilateral triangles, the associated Δt values therefore differ. This means that for an arbitrary frequency ω, dispersion relationships such as those expressed in the form $\cos \omega \Delta t$ can not be used to compare accuracy.

Nondimensional phase and group velocities are independent of Δt (assuming f_2' is specified) and thus provide a better basis for comparison. Asymptotic expansions for small $(k_1 d, k_2 d)$ can be obtained from (4.5.3) and (4.7.22). The phase velocities when $f = \tau = 0$ are

$$|\mathrm{C}|/(gh)^{1/2} \simeq 1 + (kd)^2(\tfrac{1}{24}(f_2')^2 - \tfrac{1}{8}) \qquad (4.8.1a)$$

$$|\mathbf{C}|/(gh)^{1/2} \simeq 1 + (kd)^2(\tfrac{1}{24}(f_2')^2 - \tfrac{1}{32}) \qquad (4.8.1b)$$

for Thacker's scheme and the LWEM respectively. k is defined by (4.2.3c). The associated respective group velocities are

$$\mathbf{G}/(gh)^{1/2} \simeq \left[k^{-1} + 3d^2k^3(\tfrac{1}{24}(f_2')^2 - \tfrac{1}{8})\right]\mathbf{k} \qquad (4.8.2a)$$

$$\mathbf{G}/(gh)^{1/2} \simeq \left[k^{-1} + 3d^2k^3(\tfrac{1}{24}(f_2')^2 - \tfrac{1}{32})\right]\mathbf{k}. \qquad (4.8.2b)$$

Notice that for high wave resolution, both phase velocities are isotropic and both group velocities have no directional error.

Assuming the optimal stable values for f_2', specifically $f_2' = 1.70437$ for Thacker and $f_2' = 0.824175$ for the LWEM, (4.8.1) becomes

$$|\mathbf{C}|/(gh)^{1/2} \simeq 1 - .00396345(kd)^2 \qquad (4.8.3a)$$

$$|\mathbf{C}|/(gh)^{1/2} \simeq 1 - .00294732(kd)^2. \qquad (4.8.3b)$$

Since the corresponding analytic values are 1.0, the second term in each case is the phase velocity error. Both errors in the group velocity magnitude are larger by a factor of three.

(4.8.3) indicates that for identical configurations of equilateral triangles, the best explicit LWEM is more accurate than the best Thacker scheme. However Thacker's scheme is cheaper since it uses a much larger time step. By reducing both d and Δt with Thacker's scheme, it is possible to attain the LWEM accuracy and retain the cost advantage.

If the same accuracy is assumed for both methods, Thacker's Δt becomes larger by the factor 1.78329. However his smaller d requires 1.34476 more nodes per unit area, and thus more calculations over one time step. The net result is that Thacker's scheme can have the same wave propagation accuracy for small wavenumbers as the LWEM, yet require only .75409 the number of calculations per unit area and unit of time.

Despite this cost advantage, Thacker's method may not be preferable to the LWEM. Boundary conditions often introduce short waves into a numerical model. Their accumulation can contaminate the desired longer wave solutions. Problems of this type have been reported with the GLFEM (see Section 1.5). Since both Thacker's scheme and the GLFEM

do not represent short waves accurately, similar problems may also arise with Thacker's scheme. They should not exist with the LWEM.

With $\Delta x = \Delta y$ and $f = \tau = 0$, the RS dispersion relationship is

$$\sin^2(\tfrac{1}{2}\omega\Delta t) = f_2^2[\sin^2(\tfrac{1}{2}k_1\Delta x) + \sin^2(\tfrac{1}{2}k_2\Delta x)]. \tag{4.8.4}$$

For small values of $(k_1\Delta x, k_2\Delta x)$, the asymptotic expansion for the associated nondimensional phase velocity magnitude is

$$|C|/(gh)^{1/2} \simeq 1 + \frac{1}{24}\left((f_2^2 - 1)(k\Delta x)^2 + \frac{2(k_1\Delta x)^2(k_2\Delta x)^2}{(k\Delta x)^2}\right). \tag{4.8.5}$$

Since it is anisotropic, comparisons with (4.8.1) are not straightforward.

Let us compare the RS scheme with the LWEM. One grid square of the RS has three unique variables and area Δx^2. One triangular element of the LWEM has area $(3^{1/2}/4)d^2$ and has the equivalent of 3/2 variables, since each node shares its variables with five other triangles. For a comparison based on equal density of the variables, set $\Delta x = .930605d$. Assume the optimal f_2 value (from (4.8.5)) when $k_1 = k_2$, namely $f_2 = 2^{-1/2}$. Then, if $\Delta t'$ is the optimal time step for the LWEM, $\Delta t = .79842\Delta t'$ is the optimal time step for the RS. The LWEM is therefore more economical. Its relative accuracy depends on the wave direction. When $k_1 = k_2$, the RS more accurately approximates wave speed. However when $k_1 = 0$ or $k_2 = 0$, the LWEM is more accurate.

4.9 Summary and Conclusions

The preceding analysis has demonstrated that FEMs can be cost competitive and as accurate as explicit FDMs. In particular, Thacker's scheme and the explicit LWEM were found to be cheaper and generally more accurate than the RS finite difference method.

Of the two configurations of triangular elements, the preceding analysis indicates that equilateral triangles are the better choice. Their phase and group velocities are independent of direction, and more accurate for long waves. Numerical tests [Hi82] substantiate this result. In fact, because equilateral triangles seem to produce isotropic waves when the wave resolution is high, they may be the optimal triangular discretization.

Optimal accuracy for Thacker's scheme, the WEM, and the LWEM depends on the parameter f_2'. As discussed in Section 3.8, it is both possible and reasonable to keep this parameter approximately constant throughout a model. Consequently, an ideal triangular discretization should employ equilateral triangles whose side length is proportional to $h^{1/2}$. In practice, this strategy may be difficult to implement.

Specific results from the preceding analysis are now summarized by section. The RS scheme studied in Section 4.3 was found to be quite accurate for small wavenumbers, and for waves travelling at 45° to the grid axes. However the numerical phase velocity was seen to be anisotropic. Asymmetric treatment of the Coriolis terms was also seen to affect the accuracy.

The GLFEM studied in Section 4.4 displayed accuracy comparable to the RS for small wavenumbers but became very inaccurate at larger wavenumbers. The numerical dispersion surface was seen to have peaks and valleys, implying waves with zero group velocity. Some short waves were calculated to have small inertial speeds while others had group velocities whose directions were incorrect by almost 180° . The configuration of equilateral triangles was found to be more accurate at small wavenumbers, than the grid of right triangles.

Thacker's scheme, studied in Section 4.5, was found to have the same short wave problems as the GLFEM. Stability conditions were calculated for both elements and the f_2 value which most accurately approximates wave propagation was also calculated. Phase velocities were isotropic with the equilateral grid.

Section 4.6 attempted a two dimensional analysis of the Galerkin FEM with piecewise linear basis functions for $z(x, y, t)$ and piecewise quadratics for $u(x, y, t)$ and $v(x, y, t)$. The spatially discretized version of this scheme has nine numerical solutions, only two of which approximate gravity waves. In order to calculate the numerical dispersion relationships, the determinant of a 9 by 9 matrix had to be calculated. This was done with the computer routine REDUCE, but the resultant expression was too complicated and long to warrant further analysis.

Section 4.7 extended many of the results from Chapter 3 to two dimensions. Stability

conditions when $f = 0$ were determined for both the WEM and the LWEM. An asymptotic analysis for small wavenumbers was also used to determine the most accurate time-stepping method for each scheme. Accuracy was again seen to be directionally dependent with element 1, but independent for the equilateral triangles of element 2. It was also shown that with an appropriate time-stepping method and a grid of equilateral triangles, wave propagation accuracy can be preserved in going from the WEM to the LWEM. It was also shown that with $f = \tau = 0$, the explicit LWEM when applied to the right triangles of element 1 has the same principal characteristic equation as the RS finite difference scheme.

Section 4.8 found that for small wavenumbers, the most accurate version of Thacker's scheme can more cheaply attain the same accuracy as the most accurate version of the LWEM. However it is less accurate and may experience difficulties with short waves. The RS scheme was seen to be more expensive per unit of real time than the WLEM. However its accuracy is directionally dependent. For some directions it more accurately models wave propagation, while for others it is less accurate.

5. A DISPERSION ANALYSIS WITH BOUNDARY CONDITIONS

5.1 Introduction

The dispersion analyses in previous chapters assumed a periodic domain. This meant that the accuracy of various numerical methods could be studied without the additional complexities introduced by boundary conditions. However, boundary conditions are required for most oceanographic models of the shallow water equations, so it is important to study their effects on the numerical solution.

Boundary conditions for a hyperbolic problem can affect both accuracy and stability. They 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 changing wave amplitudes, and indirectly, by generating undesirable short waves which contaminate the solution. Consequently, some of the accurate and stable methods studied in previous chapters may be less attractive when combined with inappropriate boundary conditions. In this chapter, it is shown that dispersion analyses can be extended to study both the accuracy and stability of initial boundary value (IBV) problems.

Most stability theory for finite difference models of hyperbolic IBV problems is based on a classic yet complex paper by Gustafsson, Kreiss, and Sundström [Gu72], henceforth GKS. Their normal mode analysis for stability [Gu72, Definition 3.3] involves substitutions similar to those in our previous dispersion analyses, and checks for nontrivial solutions associated with eigenvalues whose magnitudes are not less than unity. Trefethen [Tr83] has recently shown that the GKS criterion has physical interpretation in terms of group velocity. This interpretation does not provide an alternative to the algebraic GKS stability test, which is often difficult to carry out, but it does clarify the algebra. 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. With a dispersion analysis that is extended to include boundary conditions, it should therefore be possible to investigate some aspects of GKS stability.

The accuracy of boundary conditions is often determined by examining truncation errors. Gustafsson [Gu75] has shown that boundary and initial approximations may be one order of accuracy lower than the interior approximations without decreasing the overall accuracy. Sköllermo [Sk75,Sk79] extends 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 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 studies boundary condition accuracy indirectly. In particular, she measures the number of meshpoints per wavelength that are needed to compute each Fourier component of the solution to some pre-assigned relative accuracy. In keeping with the general theme of this thesis, it is shown in this chapter that accuracy can also be examined through physical concepts such as wave amplitude profiles and reflection coefficents.

The analysis technique will be demonstrated for the one dimensional linearized shallow water equations with constant depth. Such a problem has two boundaries, one at either end of a channel. Periodic forcing will be assumed at one of the boundaries while the other will be either closed or radiating. Boundaries such as these are common in tidal models. An accurate *steady state* solution over the channel is the desired result.

At this point, it is important to differentiate between the terms *steady state* and *stable*. A numerical method which for a fixed step size gives solutions which converge to a steady state may not be GKS stable. Consider the numerical scheme

$$u^{n+1} = Qu^n \tag{5.1.1}$$

for some matrix operator Q. For a fixed spatial discretization, this scheme will converge to a steady state if all eigenvalues of Q are strictly inside the unit circle. However this does not imply that the scheme is GKS stable, since stability is defined in terms of a

norm of Q rather than the maximum eigenvalue modulus. Indeed, Gustafsson [Gu82] gives an example where this case occurs. Given a GKS stable method of the form (5.1.1), he then presents sufficient conditions which ensure that all the eigenvalues of Q lie strictly inside the unit circle. In that way, both convergence to a steady state and stability are guaranteed.

Beam, Warming, and Yee [Be82] (henceforth BWY) adopt a similar approach. They define the concept of *P-stability* as follows:

> A difference scheme for an inital-boundary value problem is said to be *P-stable* if it is GKS-stable and all eigenvalues (corresponding to nontrivial eigenvectors) of the resolvent equations for a finite number of spatial mesh intervals have modulus less than or equal to unity.

The resolvent equations are obtained by substituting $u_j^n = z^n v^j$ into the difference equations for the domain interior and for the homogeneous boundary conditions. This same substitution is made in the GKS normal mode analysis. It is clearly similar to the substitutions used in Chapters 2 and 3 when forming the dispersion relationship.

Gustafsson [Gu 82] notes that since all eigenvalues with modulus equal to unity are permitted, P-stability does not exclude all growing modes. (This is discussed further in Section 5.3.) Hence a P-stable method may not converge to a steady state.

BWY generalize some of the specific boundary conditions that Gustafsson and Oliger [Gu80] show to be stable. Other aspects of the BWY paper link it closely to Chapter 2. BWY present necessary conditions for the P-stability of specific difference methods when combined with space and space-time extrapolation boundary conditions. Space and space-time extrapolation of order $q-1$ are respectively defined as

$$(F-1)^q u_j^n = 0 \tag{5.1.2a}$$

$$(FE^{-1}-1)^q u_j^n = 0. \tag{5.1.2b}$$

F is the spatial shift operator

$$F u_j^n = u_{j+1}^n, \tag{5.1.2c}$$

and E is the temporal shift operator

$$Eu_j^n = u_j^{n+1}. \tag{5.1.2d}$$

In particular, BWY solve the problem

$$\frac{\partial u}{\partial t} - c\frac{\partial u}{\partial x} = 0 \qquad 0 \le x \le L \tag{5.1.3a}$$

$$u(L,t) = g(t) \qquad t \ge 0 \tag{5.1.3b}$$

$$u(x,0) = f(x) \tag{5.1.3c}$$

with a centred three-point spatial difference approximation to $\partial u/\partial x$, and linear multistep time stepping (as in Chapter 2). For example, their two step difference method would be

$$\begin{aligned} &a_2u_j^{n+2} + a_1u_j^{n+1} + a_0u_j^n \\ &\qquad = \frac{c\Delta t}{2\Delta x}\left[b_2(u_{j+1}^{n+2} - u_{j-1}^{n+2}) + b_1(u_{j+1}^{n+1} - u_{j-1}^{n+1}) + b_0(u_{j+1}^n - u_{j-1}^n)\right] \end{aligned} \tag{5.1.4}$$

for real numbers a_2, a_1, a_0, b_2, b_1, b_0. Except for the spatial spreading of the $\partial u/\partial t$ approximation introduced by a GFEM with piecewise linear basis functions, (5.1.4) is identical to the general difference formulas studied in Chapter 2. In fact, (5.1.4) describes the general time stepping formulas that would arise from lumping the GFEM.

BWY give necessary and sufficient conditions for a linear two-step method to be A-stable [Da63] (see Section 2.4). In particular, a second order linear two-step method is A-stable when, in addition to the conditions (2.4.5),

$$b_2 \ge \tfrac{1}{2}a_2 \tag{5.1.5a}$$

$$a_2 \ge \tfrac{1}{2}. \tag{5.1.5b}$$

It is therefore not surprising that these conditions define (fairly closely) the stability regions seen in the accuracy measure figures of Chapter 2.

The major BWY results are sufficient conditions for P-stability. Specifically, they prove that if the linear multistep method is A-stable, then their difference formula (e.g., (5.1.4)), when combined with space extrapolation condition at the left boundary and a specified right boundary, is P-stable. They also prove a similar result for strongly A-stable methods

and space-time extrapolation. The close relationship between the BWY difference method and the GFEMs in Chapter 2 suggest that we may be able to extend their stability results to our methods. This will not be investigated in this thesis, but certainly warrants further attention.

The five subsequent sections within Chapter 5 have the following contents. Section 5.2 defines the mathematical problem and confirms that the chosen boundary conditions are well-posed. For $\tau = 0$, our radiating boundary conditions are also shown to be equivalent to the absorption conditions of Engquist and Majda [En77]. Section 5.3 develops the analysis for the Richardson-Sielecki FDM and studies the stability and relative accuracy of several numerical boundary conditions. Section 5.4 performs a GKS stability analysis for one set of boundary conditions in combination with the GFEM which uses piecewise linear basis functions and Crank-Nicolson time stepping. An example of unstable boundary conditions of the Trefethen type is also given. Section 5.5 studies the stability and relative accuracy of five sets of boundary conditions with that same GFEM. Finally, Section 5.6 summarizes the results.

5.2 Boundary Conditions for the Shallow Water Equations

Let us begin with a mathematical definition of the problem and a confirmation that our boundary conditions are well-posed. Assuming constant depth and linear friction, the governing equations (2.2.1) can be expressed in matrix form as

$$\frac{\partial \mathbf{w}}{\partial t} = A\frac{\partial \mathbf{w}}{\partial x} + B\mathbf{w} \tag{5.2.1a}$$

where

$$\mathbf{w} = \begin{pmatrix} z \\ u \end{pmatrix} \tag{5.2.1b}$$

and

$$A = \begin{pmatrix} 0 & -h \\ -g & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 0 \\ 0 & -\tau \end{pmatrix}. \tag{5.2.1c}$$

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

$$\mathbf{w}(x, 0) = \mathbf{f}(x). \tag{5.2.1d}$$

for some function **f**.

In order to determine proper boundary conditions for this hyperbolic problem, we follow Chapter 15 in Kreiss and Oliger [Kr73]. The eigenvalues of A are $\pm(gh)^{1/2}$. Since the number of boundary conditions at $x = 0$ must equal the number of negative eigenvalues, and the number of conditions at $x = L$ must correspond to the number of positive eigenvalues, our problem requires one condition at each boundary. The precise form of these conditions is expressable in terms of the characteristic variables.

The characteristic variables for (5.2.1) are defined as

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = T^{-1}\mathbf{w} = \begin{pmatrix} \left(\frac{g}{4h}\right)^{1/4} z + \left(\frac{h}{4g}\right)^{1/4} u \\ -\left(\frac{g}{4h}\right)^{1/4} z + \left(\frac{h}{4g}\right)^{1/4} u \end{pmatrix} \tag{5.2.2}$$

where the nonsingular matrix

$$T = \begin{pmatrix} \left(\frac{h}{4g}\right)^{1/4} & -\left(\frac{h}{4g}\right)^{1/4} \\ \left(\frac{g}{4h}\right)^{1/4} & \left(\frac{g}{4h}\right)^{1/4} \end{pmatrix} \tag{5.2.3}$$

transforms A to a real diagonal matrix. Specifically

$$T^{-1}AT = (gh)^{1/2} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{5.2.4}$$

The characteristic variable v_1 is associated with the negative eigenvalue $-(gh)^{1/2}$ and must be specified at the left boundary. It is an incoming variable there. The characteristic variable v_2 is associated with the positive eigenvalue $(gh)^{1/2}$ and must be specified at the right boundary where it is the incoming variable. The precise form of these boundary conditions is [Kr73]

$$v_1(0,t) = S_1 v_2(0,t) + g_1(t) \tag{5.2.5a}$$

$$v_2(L,t) = S_2 v_1(L,t) + g_2(t) \tag{5.2.5b}$$

where, in this case, S_1 and S_2 are scalars. Our boundary conditions will be well-posed if they can be expressed in this form.

Assume that the left boundary is either closed or radiating. Closed boundaries are usually expressed as

$$u(0,t) = 0. \tag{5.2.6}$$

With $S_1 = -1$ and $g_1(t) = 0$, this condition is seen to conform to (5.2.5a).

Radiating boundaries for the shallow water equations are often [He76] represented as

$$u(0,t) = -\left(\frac{g}{h}\right)^{1/2} z(0,t). \tag{5.2.7}$$

This is the expected relationship (when $\tau = 0$) between elevation and velocity for a leftward travelling wave. At the left boundary, (5.2.7) should therefore transmit leftward waves without any reflection. Setting $S_1 = g_1(t) = 0$, (5.2.7) is also seen to conform to (5.2.5a).

Re-writing (5.2.7) as

$$v_1(0,t) = 0, \tag{5.2.8}$$

it is apparent that our radiating condition ensures no reflection at the left boundary by setting the ingoing characteristic variable to zero. Gustafsson and Kreiss [Gu79] note that this is the underlying principle behind the absorbing boundary conditions of Engquist and Majda [En77]. In particular, when $\tau = 0$, (5.2.1) can be re-expressed in terms of the characteristic variables as

$$\frac{\partial \mathbf{v}}{\partial x} = \mathcal{V}\mathbf{v} \tag{5.2.9a}$$

where

$$\mathcal{V} = (gh)^{-1/2} \begin{pmatrix} -\frac{\partial}{\partial t} & 0 \\ 0 & \frac{\partial}{\partial t} \end{pmatrix} \tag{5.2.9b}$$

is a pseudo-differential operator. From Section 2 in [En77], it then follows that the *perfectly absorbing boundary condition* for our problem is given by (5.2.7).

At the right boundary we wish to impose a driving condition or a combination driving-radiating condition. The combined condition is designed to generate leftward waves and radiate rightward waves. We would also like the option of expressing such conditions in terms of either z or u. Pure driving conditions for z, and u are attained by setting $S_2 = 1$ and $S_2 = -1$ respectively. The driving-radiating condition

$$u(L,t) = \left(\frac{g}{h}\right)^{1/2} z(L,t) + g_2(t) \tag{5.2.10}$$

is possible with $S_2 = 0$.

Therefore, all our boundary conditions are well-posed.

5.3 The Richardson-Sielecki Scheme

In the previous section it was seen that only one boundary condition is required at each end of the one dimensional channel. Many numerical methods for solving the shallow water equations have their elevation and velocity variables located at the same spatial point. This means that one additional condition is required at each boundary in order to completely specify the numerical problem. Numerical methods such as the Richardson-Sielecki scheme (henceforth RS) stagger z and u spatially. Only one variable is then located at each boundary, and the need for extra boundary conditions is avoided. We therefore introduce the boundary condition analysis with an application to the RS scheme. Section 5.4 will examine a more complicated unstaggered scheme.

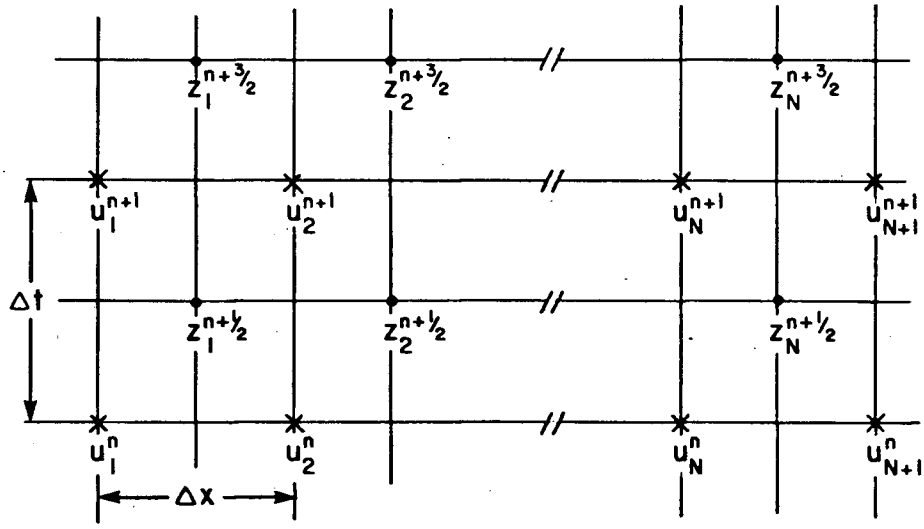

Fig. 5.1 One dimensional RS grid.

The one dimensional RS grid is illustrated in Fig. 5.1. Notice that the z and u variables are staggered in both space and time. Velocity variables are assumed at both boundaries. In particular, it is assumed that the left boundary is either open or closed, while the right boundary is forced.

In the domain interior, the finite difference equations are

$$z_j^{n+1/2} = z_j^{n-1/2} - \frac{h\Delta t}{\Delta x}(u_{j+1}^n - u_j^n) \tag{5.3.1a}$$

$$(1 + \tfrac{1}{2}\tau\Delta t)u_j^{n+1} = (1 - \tfrac{1}{2}\tau\Delta t)u_j^n - \frac{g\Delta t}{\Delta x}(z_j^{n+1/2} - z_{j-1}^{n+1/2}). \tag{5.3.1b}$$

Initial conditions are assumed to be

$$u_j^0 = g_1(j) \tag{5.3.1c}$$

$$z_j^{1/2} = g_2(j) \tag{5.3.1d}$$

for some functions g_1 and g_2.

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

$$u_1^n = 0 \tag{5.3.2a}$$

$$u_{N+1}^n = f(n) \tag{5.3.2b}$$

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 (5.2.7). The most common implementation [He76] is

$$u_1^{n+1} = -\left(\frac{g}{h}\right)^{1/2} z_1^{n+1/2} \tag{5.3.3a}$$

whereby zeroth order space-time extrapolation (5.1.2b)

$$z_{1/2}^{n+1} = z_1^{n+1/2} \tag{5.3.3b}$$

is used to calculate the z value coincident with u_1^{n+1} .

A GKS stability analysis of the RS scheme is quite complicated due to the spatial and temporal staggering of the variables. It will not be attempted here. However the scheme has been widely used and its stability restrictions are well-known. In particular, it is Cauchy stable [He81] when

$$(gh)^{1/2}\frac{\Delta t}{\Delta x} \leq 1. \tag{5.3.4}$$

We now develop the dispersion analysis. Define the vector $\mathbf{X}$ as

$$\mathbf{X}^{n+1} = (u_1^{n+1}, z_1^{n+1/2}, u_2^{n+1}, z_2^{n+1/2}, \ldots z_N^{n+1/2}, u_{N+1}^{n+1}). \tag{5.3.5}$$

Then equations (5.3.1) and (5.3.2) can be expressed in matrix form as

$$\mathbf{X}^{n+1} = A\mathbf{X}^n + \mathbf{X}_D f(n+1). \tag{5.3.6a}$$

A is the finite difference matrix operator and

$$\mathbf{X}_D = (0, 0, 0, \ldots, 1) \tag{5.3.6b}$$

is the vector which locates the driving conditions. Assume a driving condition

$$f(n) = Re[ae^{i(n\omega\Delta t - \phi)}] \tag{5.3.6c}$$

for some frequency, amplitude, and phase, ω, a, and ϕ respectively.

Repeated substitution into (5.3.6a) gives

$$\mathbf{X}^{n+1} = A^{n+1}\mathbf{X}^0 + Re\left[ae^{i[(n+1)\omega\Delta t - \phi]}\left(\sum_{\ell=0}^{n} e^{i(\ell-n)\omega\Delta t}A^{n-\ell}\right)\mathbf{X}_D\right] \tag{5.3.7}$$

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

$$\left(\sum_{\ell=0}^{n} e^{i(\ell-n)\omega\Delta t}A^{n-\ell}\right)B = I - (e^{-i\omega\Delta t}A)^{n+1} \tag{5.3.8a}$$

where

$$B = [I - e^{-i\omega\Delta t}A]. \tag{5.3.8b}$$

So provided B is invertible

$$\sum_{\ell=0}^{n} e^{i(\ell-n)\omega\Delta t}A^{n-\ell} = [I - (e^{-i\omega\Delta t}A)^{n+1}]B^{-1}. \tag{5.3.9}$$

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

$$B\mathbf{x} = \mathbf{0}. \tag{5.3.10}$$

This implies

$$A\mathbf{x} = e^{i\omega\Delta t}\mathbf{x} \tag{5.3.11a}$$

and

$$\lambda(\omega\Delta t) = e^{i\omega\Delta t} \tag{5.3.11b}$$

is an eigenvalue of the matrix A. Therefore, provided λ is not an eigenvalue of A, B is invertible.

When the driving frequency and time step are chosen so that $\lambda(\omega\Delta t)$ is not an eigenvalue of A, (5.3.7) can be re-written as

$$\mathbf{X}^{n+1} = A^{n+1}\mathbf{X}^0 + Re\left\{ae^{i[(n+1)\omega\Delta t-\phi]}\left(I-(e^{-i\omega\Delta t}A)^{n+1}\right)\mathbf{Y}\right\} \tag{5.3.12}$$

$$\text{where} \qquad \mathbf{Y} = B^{-1}\mathbf{X}_D. \tag{5.3.13}$$

$\mathbf{X}^{n+1}$ converges to a steady state when A^{n+1} converges. But [Pu76]

$$\lim_{n\to\infty} A^n = L \text{ (a definable matrix) iff}$$

i) $|\lambda| \leq 1$ for all eigenvalues of A,

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

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

Furthermore [Pu76],

$$\lim_{n\to\infty} A^n = 0$$

$$\text{iff } |\lambda| < 1 \text{ for all eigenvalues of } A.$$

The steady state solution for $\mathbf{X}^{n+1}$ is complicated when A has at least one eigenvalue equal to unity. We therefore adopt the Gustafsson [Gu82] approach and assume that all eigenvalues are strictly inside the unit circle. This implies

$$\mathbf{X}^{n+1} = Re[ae^{i[(n+1)\omega\Delta t-\phi]}\mathbf{Y}]. \tag{5.3.14}$$

$\mathbf{Y}$ then contains the spatial profile of the steady state solution.

The general form of $\mathbf{Y}$ can be found by extending the dispersion analyses of the previous chapters. Assume that solutions to (5.3.1), (5.3.2) have the separable form

$$z_j^{n+1/2} = \varsigma_0\lambda^{n+1/2}\kappa^j \tag{5.3.15a}$$

$$u_j^{n+1} = \mu_0\lambda^{n+1}\kappa^{j-1/2} \tag{5.3.15}$$

for complex numbers λ and κ. (This same substitution is made in the normal mode analysis [Gu72,Be82,Tr83] to form the resolvent equations.) A nontrivial solution ($\varsigma_0 \neq 0$ or $\mu_0 \neq 0$) for (5.3.1) then requires

$$\kappa^2 - 2\kappa\left\{1+\tfrac{1}{2}[\lambda-2+1/\lambda+\tfrac{1}{2}\tau\Delta t(\lambda-1/\lambda)]/gh(\Delta t/\Delta x)^2\right\}+1=0. \tag{5.3.16}$$

For a specific value of

$$\lambda = e^{i\omega\Delta t}, \tag{5.3.17}$$

there are two values of κ, namely κ_1 and κ_2. Since their product is 1.0

$$\kappa_1 = 1/\kappa_2. \tag{5.3.18a}$$

Set

$$\kappa = \kappa_2 = re^{ik\Delta x} \tag{5.3.18b}$$

for some values of r and $k\Delta x$. The general numerical solution for z is then

$$z_j^{n+1/2} = e^{i(n+1/2)\omega\Delta t}[\varsigma_1 r^{-j} e^{-ijk\Delta x} + \varsigma_2 r^j e^{ijk\Delta x}] \tag{5.3.19}$$

for some complex coefficients ς_1 and ς_2. A similar solution will exist for u_j^{n+1}. The precise values of ς_1 and ς_2 are determined by the boundary conditions.

Notice that the first term in (5.3.19) is a rightward wave with a spatially variant amplitude profile. In particular, if $r > 1$ the wave amplitude decreases as the wave moves rightward. The second term in (5.3.19) is a leftward wave. Again, if $r > 1$ the wave amplitude decreases with propagation.

(5.3.19) can be generalized to allow for the case $|\lambda| \neq 1$ and the case when κ_1 and κ_2 coalesce. The general numerical solution for $z_j^{n+1/2}$ is then identical to the form given by Trefethen [Tr83, equation (2.7)] when he is investigating resolvent solutions with $|z| > 1$.

Comparing (5.3.19) with (5.3.14), it appears that the component of $\mathbf{Y}$ representing z_j should have the form

$$Y_{jz} = \varsigma_1 r^{-j} e^{-ijk\Delta x} + \varsigma_2 r^j e^{ijk\Delta x} \tag{5.3.20}$$

for specific values of ς_1 and ς_2. In fact ς_1 and ς_2 should be constant for all values of j.

Algebraic solutions for ς_1 and ς_2 will be messy. However they can be found numerically without much difficulty. Assume that N and

$$f_1 = \frac{\tau\Delta x}{(gh)^{1/2}} \tag{5.3.21a}$$

$$f_2 = (gh)^{1/2}\frac{\Delta t}{\Delta x} \tag{5.3.21b}$$

are constant, and check that all eigenvalues of A are inside the unit circle. This ensures that (5.3.14) is valid and that $\lambda(\omega\Delta t)$ is not an eigenvalue of A for any (real) values of $\omega\Delta t$. The steady state numerical solution for the driving frequency $\omega\Delta t$ is then found through the following steps:

i) using (5.3.16), solve for κ;

ii) re-arrange (5.3.13) to

$$B\mathbf{Y} = \mathbf{X}_D \tag{5.3.22}$$

and solve for $\mathbf{Y}$;

iii) do a least squares fit on the z (and u) components within $\mathbf{Y}$ (i.e., the vector components located in the same positions as the z components in $\mathbf{X}$) to find the complex coefficients ς_1 and ς_2 in (5.3.20).

Reflection coefficients for each boundary can be found from ς_1 and ς_2. Specifically, if $j = j_L$ at the left boundary, then the ratio of the rightward to the leftward wave,

$$R_L = \left(\frac{\varsigma_1}{\varsigma_2}\right) r^{-2j_L} e^{-2ij_L k\Delta x} \tag{5.3.23}$$

is a measure of the reflection characteristics of the boundary. It will be referred to as the reflection coefficient for the left boundary. An accurate radiating boundary should have small $|R_L|$, while an accurate closed boundary should have $|R_L| = 1$.

At the right boundary, where $j = j_R$, the driving component of the leftward wave must be removed before calculating the reflection coefficient. If the driving component is

$$z_D^{n+1/2} = d_0 e^{i[(n+1/2)\omega\Delta t - \phi]} \tag{5.3.24}$$

then

$$\mathrm{R}_R = \frac{(\varsigma_2 r^{j_R} e^{ij_R k\Delta x} - d_0 e^{-i\phi})}{\varsigma_1 r^{-j_R} e^{-ij_R k\Delta x}}. \tag{5.3.25}$$

This ratio will be referred to as the reflection coefficient for the right boundary.

The analysis technique is now illustrated for several boundary conditions and parameter values. The first problem of study has parameter values $(f_1, f_2, N) = (0., .95, 10)$ and boundary conditions (5.3.3a), and (5.3.2b) with (5.3.6c). The analysis results are shown in Fig. 5.2.

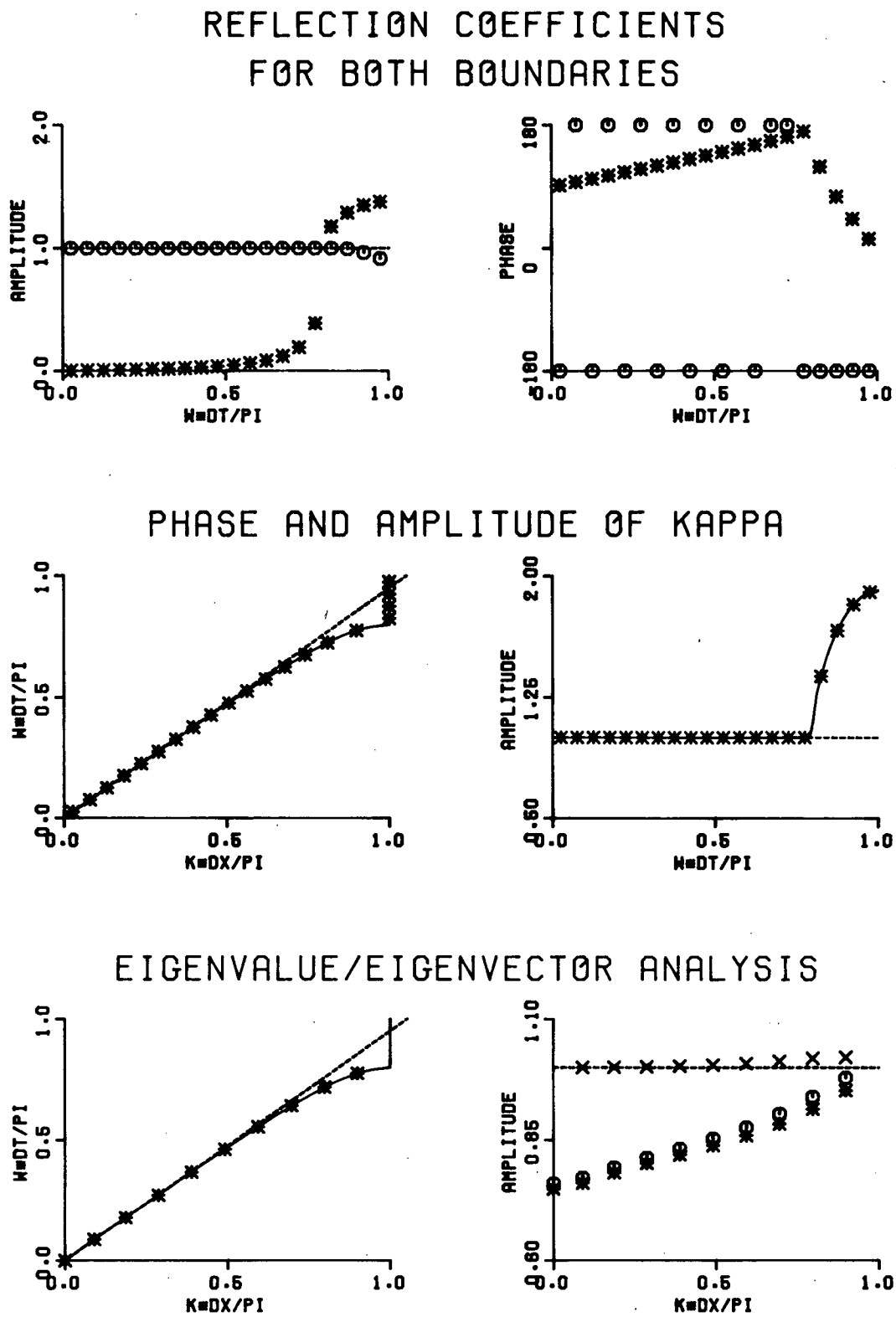

Fig. 5.2 Dispersion analysis of RS scheme with $f_1 = 0.$, $f_2 = .95$, $N = 10$ and boundary conditions (5.3.3a), and (5.3.2b) with (5.3.6c). See text for symbol definitions.

The two lowest diagrams display the relationships between the eigenvalues and eigenvectors of the matrix A. Such information describes the behaviour of transient and random

signals in the numerical model. Transient signals are generated by the initial conditions, while random signals are generated by roundoff errors interacting with the boundary conditions. Provided all $|\lambda| < 1$, neither of these signals will affect the steady state solution.

The diagram on the lower left is a dispersion relationship. The dotted line is the analytic relationship while the solid line is the numerical dispersion relationship for a ring domain. Asterisks plot $\omega\Delta t$ versus $k\Delta x$ (i.e., the arguments of λ versus those for κ). Only positive values of $\omega\Delta t$ and $k\Delta x$ are shown. Notice that all the asterisks lie along the solid line. This implies that the phase and group velocities of all transient and random signals are the same as they would be for some ring domain.

The diagram on the lower right plots $|\lambda|$ and $|\kappa|$ versus $k\Delta x$. The $|\lambda|$ values are shown as circles while the $|\kappa|$ values are asterisks. All the eigenvalue amplitudes are strictly less than unity so a steady state solution does exist. All $|\kappa|$ values are also strictly less than unity. This means that the spatial amplitude profile for these leftward (positive $k\Delta x$) waves increases to the left. Both $|\lambda|$ and $|\kappa|$ increase monotonically with $k\Delta x$. However this does not necessarily mean that short wavelengths are favoured by the numerical scheme, as was the case in Chapter 2. In order to determine which waves grow most quickly or decay least quickly, it is now necessary to introduce the concept of *Lagrangian amplitudes*. These values are designated by crosses and are calculated as follows.

In one time step, Δt, each wave travels

$$C\Delta t = \frac{\omega\Delta t}{k} \tag{5.3.28a}$$

where C is the phase velocity. As a fraction of the spatial grid size Δx, this distance is

$$d = C\frac{\Delta t}{\Delta x}. \tag{5.3.28b}$$

So, in one time step, the amplitude change for each rightward wave is

$$A_R = |\lambda||1/\kappa|^d. \tag{5.3.28c}$$

This is a Lagrangian amplitude change since our perspective is moving with the wave. Lagrangian amplitudes for leftward waves are defined as

$$A_L = |\lambda||\kappa|^{-d} \tag{5.3.28d}$$

and are thus identical to the Lagrangian amplitudes for rightward waves.

The concept of Lagrangian amplitudes was not required in Chapter 2 because, on a periodic domain, wave amplitude growth (or decay) is not spatially dependent. In particular, when $|\kappa| = 1$, $A_L = |\lambda|$. So the amplitude changes per time step that were calculated in Chapter 2 are a special case of the Lagrangian amplitudes studied here.

Notice that short waves have Lagrangian amplitudes which are slightly larger than unity. Although this means that these waves grow as they propagate, it does not necessarily mean that they cause instability. Since all waves are absorbed (in varying degrees) at the left boundary, wave growth is counteracted there. In fact, net wave growth must be bounded since from any perspective that is fixed spatially, wave amplitudes do not increase in time.

The middle row of diagrams shows model response to various driving frequencies. Again the left diagram is a dispersion curve with the same solid and dotted lines as the diagram below it. Asterisks now denote the numerical values obtained when twenty equally spaced values of $\omega\Delta t$ are assumed in (5.3.17) and substituted into (5.3.16). Again these asterisks lie along the numerical dispersion curve for a ring domain. Notice that when $\omega\Delta t > 2.5$, $k\Delta x = \pi$. This means that driving frequencies larger than the cutoff value generate $2\Delta x$ waves. Such waves are similar in origin to the $4\Delta x$ waves that Vichnevetsky [Vi80] predicts for an unstaggered finite difference grid.

The diagram on the right plots $|\kappa|$ versus $\omega\Delta t$. Notice that beyond the cutoff frequency, the amplitude of κ increases dramatically. In particular, $|\kappa|$ is a negative real with magnitude greater than unity. Since $k\Delta x = \pm\pi$, the direction in which $2\Delta x$ waves are travelling cannot be determined. So we can no longer assume that κ_2 is associated with the leftward wave and κ_1 with the rightward. The amplitude profile due to κ_2 decreases as one moves inward from the right boundary, while the profile due to κ_1 increases. However the complex coefficient for κ_2 , namely ς_2, is much larger than the coefficient for κ_1, so the resultant amplitude profile does decrease toward the left. Vichnevetsky [Vi80] also observes this phenomenon. He comments that 'the amplitude decays in space at a rate

which increases monotonically with the excess of frequency above the cutoff'.

The top two diagrams show the amplitude and phase of the reflection coefficients for both boundaries. These coefficents were calculated for the u variable. Values for z have identical amplitudes and phases which differ by 180°. Reflection coefficient amplitudes and phases for the right boundary are shown as circles. Amplitudes equal 1.0 when $\omega\Delta t$ is less than the cutoff frequency. This is to be expected since the homogeneous condition, $u_{N+1} = 0$, simulates a closed boundary. Beyond the cutoff frequency, the amplitudes decrease monotonically. The associated phases are $\pm 180°$.

Reflection coefficient amplitudes and phases for the left boundary are shown as asterisks and illustrate the accuracy of condition (5.3.3a). 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. (5.3.3a) is therefore an accurate open boundary condition for long waves. Reflection coefficients also increase beyond 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. However in this instance, test model runs confirm stability. The reflection coefficient phases for the left boundary indicate the relative phase relationship of the reflected wave to the incoming wave. They are seen to exhibit a smooth transition with $\omega\Delta t$.

Fig. 5.3 illustrates the analysis results for a second test problem whose only difference from the first problem is that $f_1 = 0.05$. Although the $\omega\Delta t$ versus $k\Delta x$ plot of eigenvalues and eigenvectors may seem unchanged, the asterisks have shifted along the solid-line dispersion curve. Values of $|\kappa|$ and $|\lambda|$ differ noticeably from Fig. 5.2, particularly for small values of $k\Delta x$. Although both the λ and κ amplitudes have minima at $k\Delta x \simeq .3\pi$, the Lagrangian amplitude curve has retained a similar shape. It has however been shifted downward so that all values are less than 1.0.

Phases and amplitudes of κ differ slightly from those in Fig. 5.2. Although it is not

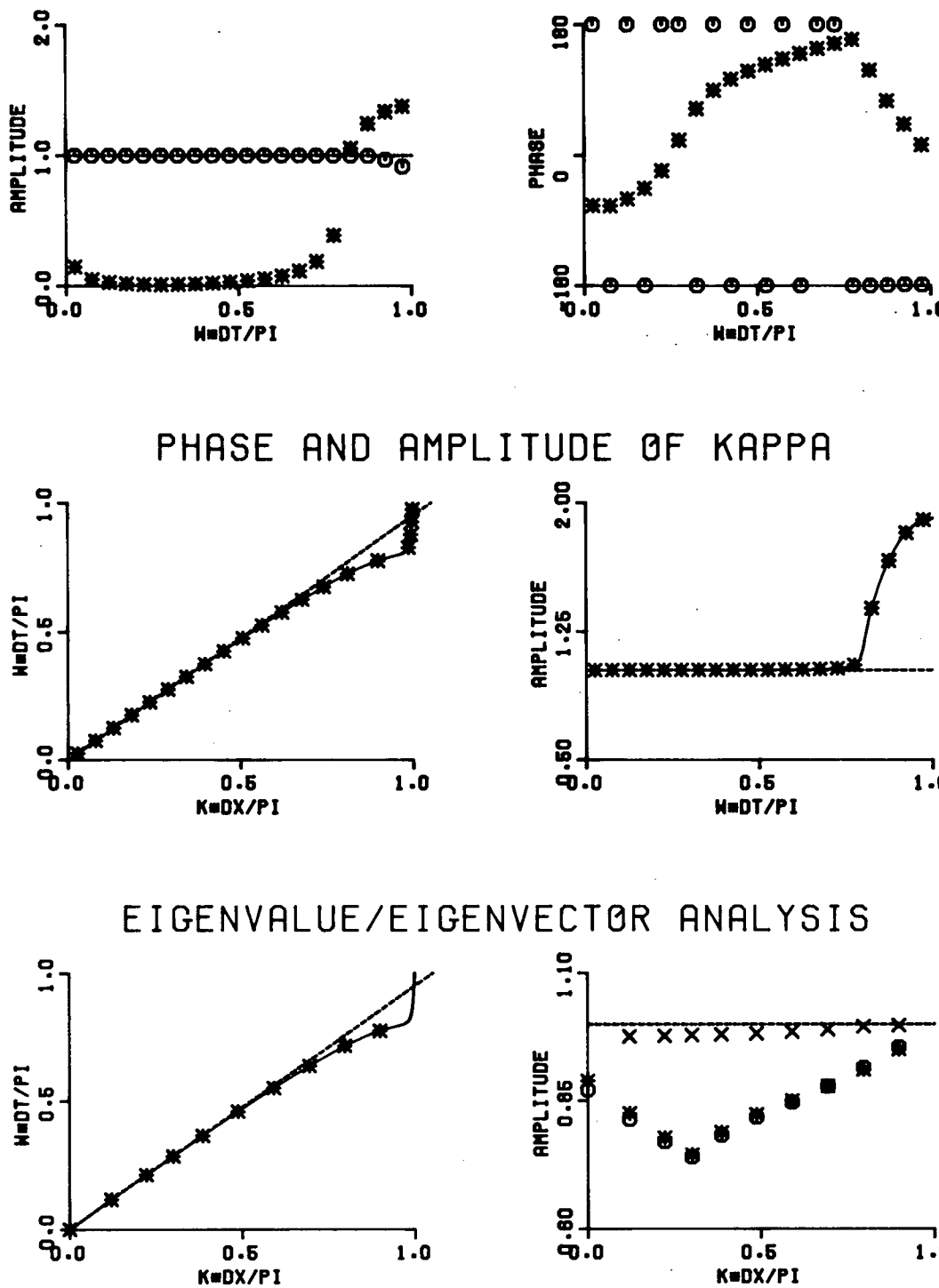

Fig. 5.3 Dispersion analysis of RS scheme with $f_1 = .05$, $f_2 = .95$, $N = 10$ and boundary conditions (5.3.3a), and (5.3.2b) with (5.3.6c).

clear from the diagrams, $\omega\Delta t$ values greater than the cutoff now give rise to $k\Delta x$ values which are slightly less than π . The $|\kappa|$ values are now slightly greater than 1.0, denoting

a spatial wave amplitude decrease. This is an expected consequence of positive friction.

The reflection coefficient diagrams do display a noteworthy change. Specifically, the left boundary condition is no longer most effective with long waves. The reflection coefficient amplitudes have a minimum at about $\omega\Delta t = 0.3\pi$ and seem to increase symmetrically on either side of this value. The reflection coefficients for the right boundary are unchanged from those of Fig. 5.2. This implies that friction does not affect the accuracy of the closed boundary condition $u_{N+1}^{n+1} = 0$.

In order to assess the accuracy of boundary condition (5.3.3a), three other implementations have also been analysed. Fig. 5.4 shows the first of these analyses. First order space-time extrapolation is now used to approximate the absorption condition of Engquist and Majda. The left boundary condition is

$$u_1^{n+1} = -2\left(\frac{g}{h}\right)^{1/2} z_1^{n+1/2} - u_2^n. \tag{5.3.29}$$

All other parameter values and boundary conditions are the same as for Fig. 5.3.

The eigenvalue-eigenvector plots now illustrate an important point. Specifically, it is not necessary that all modes of the boundary value problem satisfy the dispersion relationship for a ring domain. As can be seen from the lower left diagram, the mode associated with $k\Delta x/\pi = .793$ does not lie on the solid line. This means that the phase and group velocity of transient or random signals at this wavenumber have been changed by the boundary conditions. The associated amplitudes $|\lambda|$,$|\kappa|$, and $|A_L|$ also stand out for this eigensolution. However this anomalous mode should not affect the numerical results. All eigenvalue amplitudes are less than unity so all transient and random signals should decay.

Plots of the phase and amplitude of κ are virtually identical to those of Fig 5.3. In particular, all the driving frequencies produce waves that propagate as predicted by the ring domain dispersion relationship.

Reflection coefficient amplitudes for the left boundary are less than those for Fig. 5.3 when $\omega\Delta t$ is in the interval $[1.2, 2.6]$ and slightly higher when $\omega\Delta t < 1.2$. So for the parameter values $(f_1, f_2, N) = (.05, .95, 1.0)$, boundary condition (5.3.29) is not consistently more accurate than (5.3.3a).

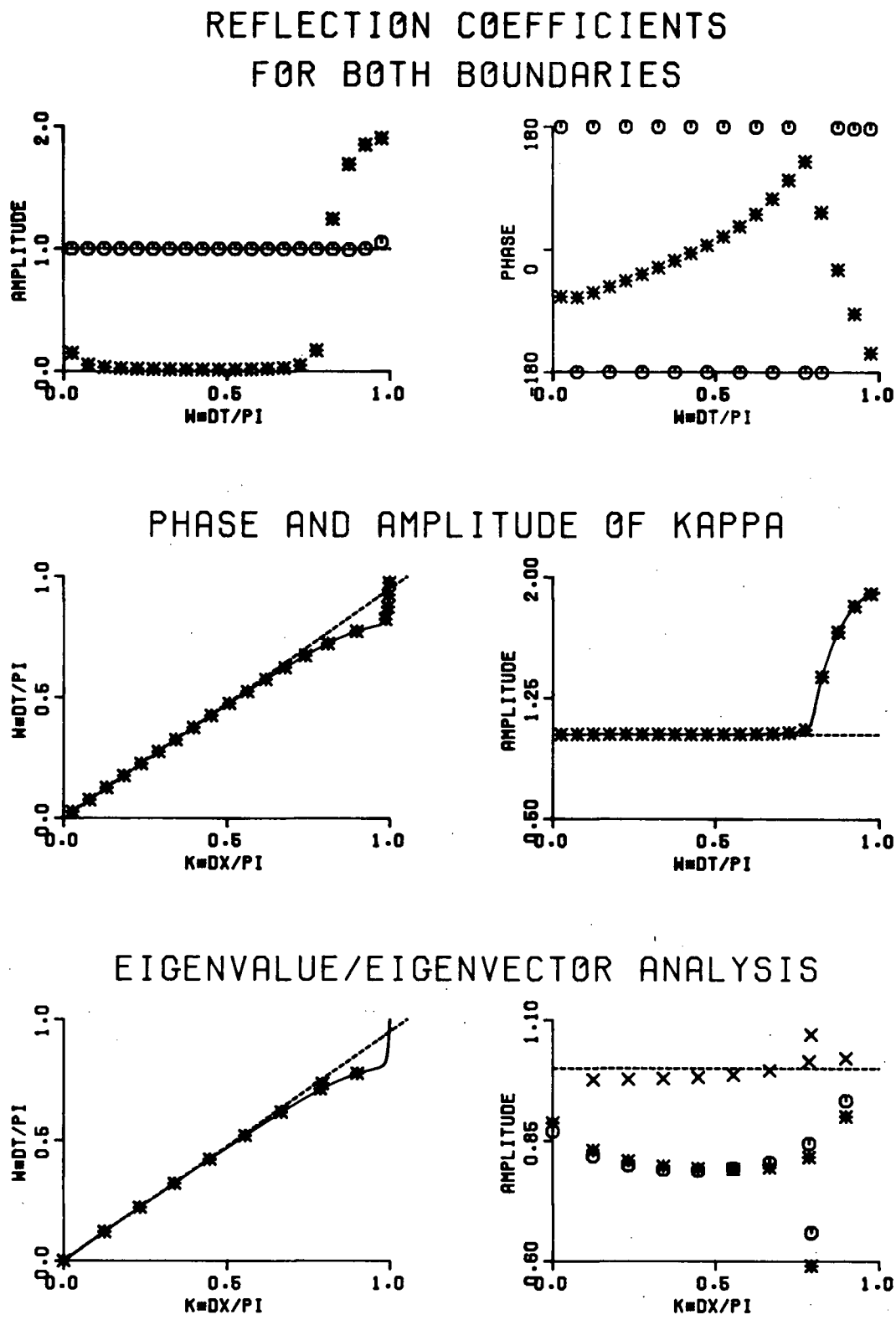

Fig. 5.4 Dispersion analysis of RS scheme with $f_1 = .05$, $f_2 = .95$, $N = 10$ and boundary conditions (5.3.29), and (5.3.2b) with (5.3.6c).

However with zero friction ($f_1 = 0$), (5.3.29) is more accurate. Provided $\omega\Delta t$ is less than the cutoff frequency, the reflection coefficent amplitudes are smaller, by at least a

factor of 2.5 than those in Fig. 5.2. The anomolous mode also disappears when $f_1 = 0$. Specifically, the eigenvalue-eigenvector plots now show that all $(\omega\Delta t, k\Delta x)$ pairs satisfy the numerical dispersion relationship for a ring domain.

Fig. 5.5 shows the analysis results for a third implementation of the absorbing left boundary. Linear extrapolation in time followed by linear extrapolation in space is used to obtain a z value coincident with u_1^{n+1}. The resultant boundary condition is

$$u_1^{n+1} = -\left(\frac{g}{h}\right)^{1/2}\left\{\tfrac{3}{2}[\tfrac{3}{2}z_1^{n+1/2} - \tfrac{1}{2}z_1^{n-1/2}] - \tfrac{1}{2}[\tfrac{3}{2}z_2^{n+1/2} - \tfrac{1}{2}z_2^{n-1/2}]\right\} \qquad (5.3.30)$$

Parameter values and other boundary conditions are identical to those of Fig. 5.3.

The eigenvalue-eigenvector analysis of this problem shows that it is unstable. $\lambda = -2.5142$, $\kappa = -.314048$ satisfy both (5.3.16) and the boundary conditions. This eigensolution is a $2\Delta x$ (and $2\Delta t$) wave whose magnitude increases by 2.5142 each time step. Due to the scale of the amplitude diagram, $|\lambda|$ and $|\kappa|$ for this point seem to be part of the reflection coefficient phase diagram.

Our analysis of the model response to a driving frequency must be modified when there are eigenvalues outside the unit circle. In particular, (5.3.12) no longer converges to (5.3.14). With zero initial conditions, it is seen from (5.3.12) that the numerical solution consists of two components, one of which converges to (5.3.14). The other component causes instability. The two top diagrams of Fig. 5.5 show reflection coefficients of the convergent component. Notice that the left boundary amplitudes are considerably larger than those in Fig. 5.3 and Fig. 5.4. In any event, reflections at the boundary are inconsequential due to the model instability. The problem remains unstable when $\tau = 0$.

A fourth implementation of the absorbing boundary condition combines linear extrapolation with phase velocity. Its development is illustrated in Fig. 5.6. Assume that a numerical wave travels at the phase speed $C^*(< 1)$. Then in the time $\frac{1}{2}\Delta t$, the elevation value $z_*^{n+1/2}$ will have travelled $\frac{1}{2}C^*\Delta t$ and will be coincident with u_1^{n+1}. Setting

$$\Delta z = \tfrac{1}{2}\Delta x - \tfrac{1}{2}C^*\Delta t, \qquad (5.3.31a)$$

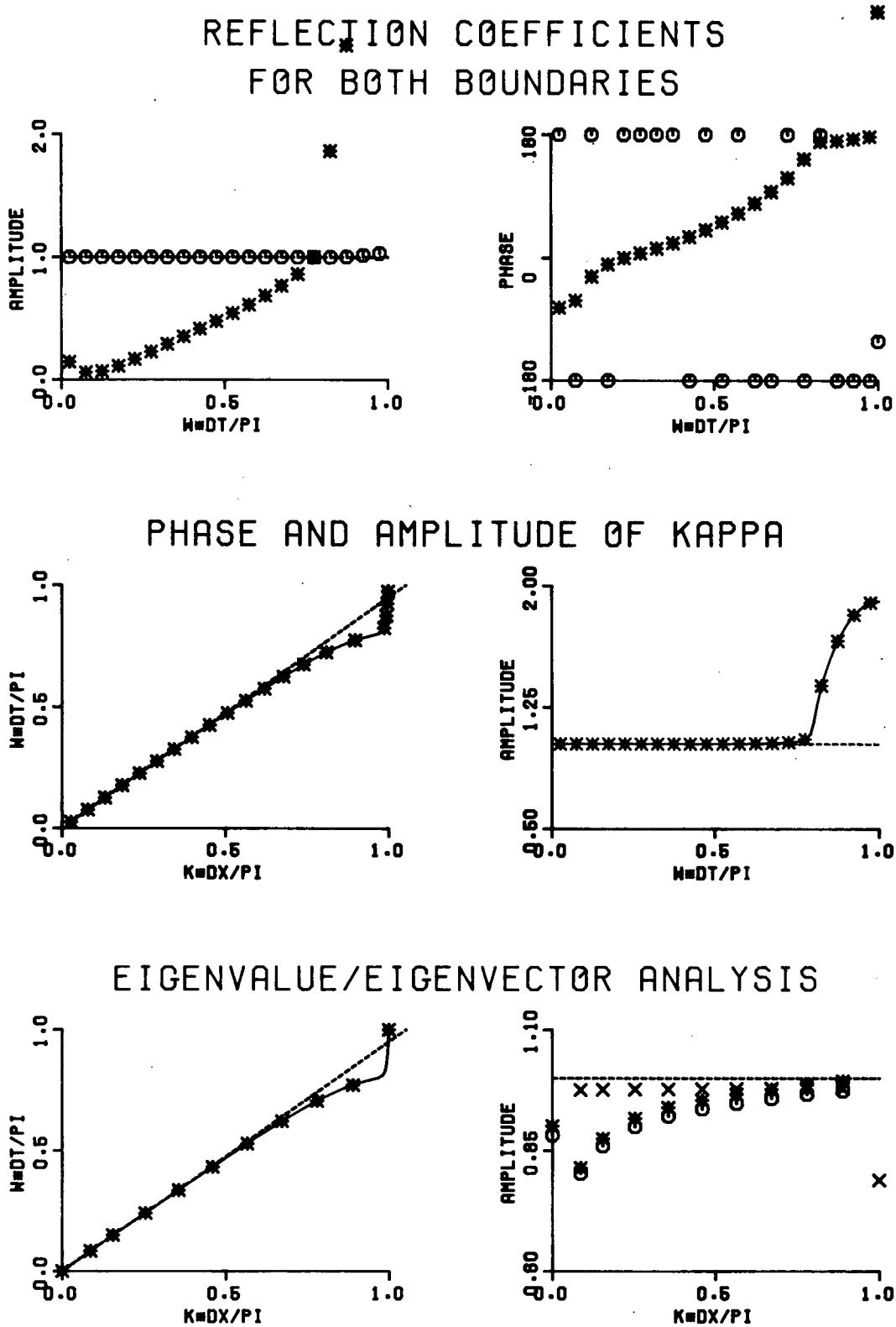

Fig. 5.5 Dispersion analysis of RS scheme with $f_1 = .05$, $f_2 = .95$, $N = 10$ and boundary conditions (5.3.30), and (5.3.2b) with (5.3.6c).

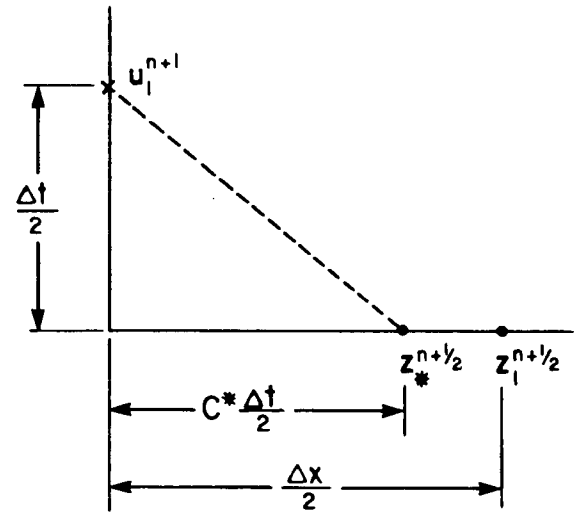

Fig. 5.6 Derivation of boundary condition (5.3.31).

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

$$z_*^{n+1/2} = \left(1 + \frac{\Delta z}{\Delta x}\right) z_1^{n+1/2} - \left(\frac{\Delta z}{\Delta x}\right) z_2^{n+1/2}. \tag{5.3.31b}$$

The associated radiation condition is therefore

$$u_1^{n+1} = -\left(\frac{g}{h}\right)^{1/2} z_*^{n+1/2}. \tag{5.3.31c}$$

Fig. 5.7 shows the analysis results with this boundary condition and $C^* = (gh)^{1/2}$. Parameter values and other boundary conditions are identical to those of Fig. 5.3. In fact, all plots are quite similar to those in Fig. 5.3. Notable exceptions are one anomolous eigenmode which has magnitude considerably less than the others, yet seems to lie on the dispersion curve for a ring domain. Compared to the values of Fig. 5.3, the reflection coefficient amplitudes here are slightly lower when $1.64 < \omega\Delta t < 2.28$. Hence (5.3.31c) is not a significant improvement on (5.3.3a) when $f_1 = .05$. However when $\omega\Delta t$ is less than the cutoff frequency and $\tau = 0$, (5.3.31c) is slightly more accurate than (5.3.3a), but not as accurate as (5.3.29).

The relative accuracy of the preceding four absorbing boundary conditions is illustrated in Fig. 5.8 for $(f_1, f_2, N) = (0., .95, 1.0)$. Reflection coefficient amplitudes are plotted against $\omega\Delta t$. Listed in terms of decreasing accuracy, these methods and their symbol representations in Fig. 5.8 are:

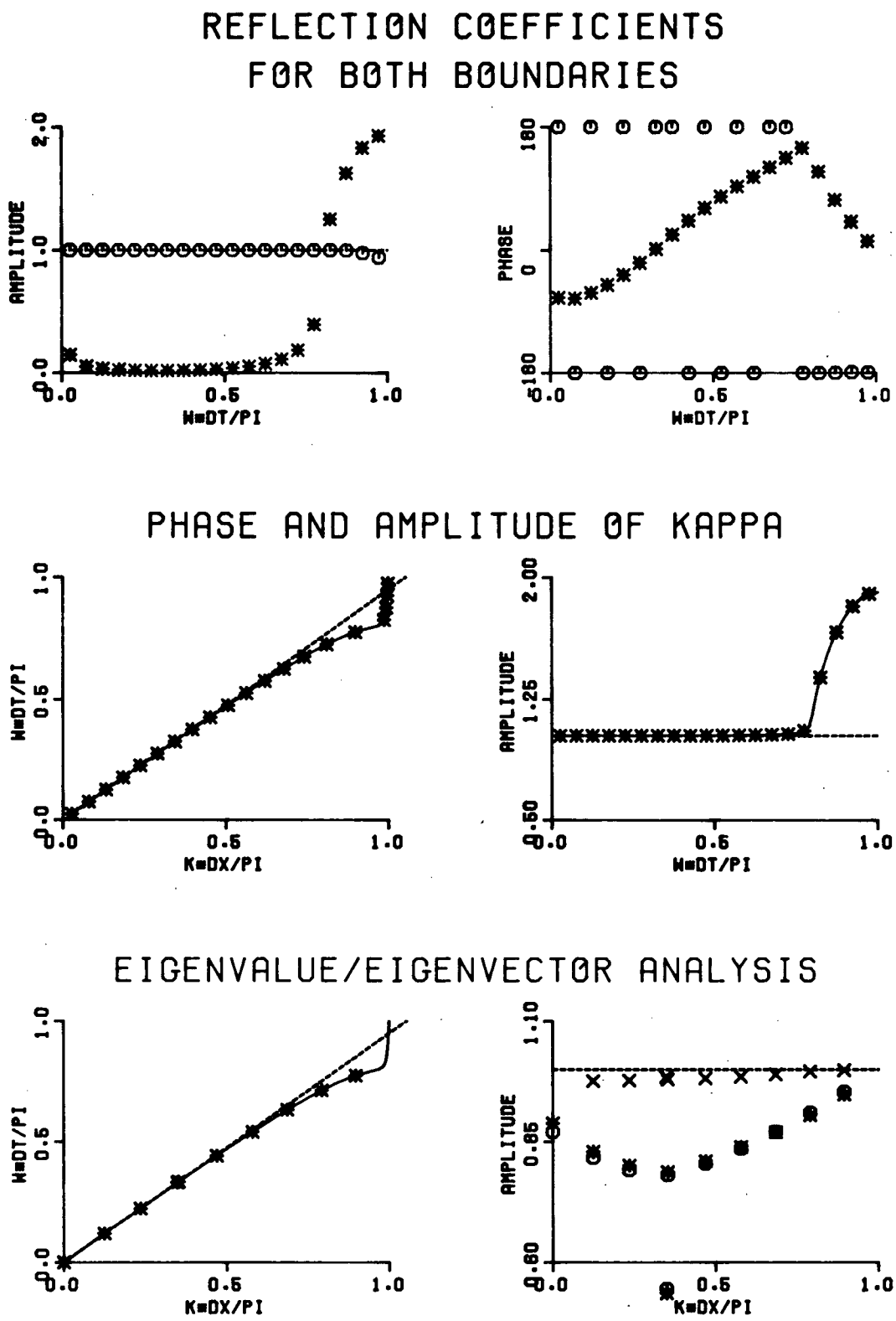

Fig. 5.7 Dispersion analysis of RS scheme with $f_1 = .05$, $f_2 = .95$, $N = 10$ and boundary conditions (5.3.31), and (5.3.2b) with (5.3.6c).

i) first order space-time extrapolation (5.3.29), $*$;

ii) linear spatial extrapolation with phase velocity (5.3.31c), $\times$;

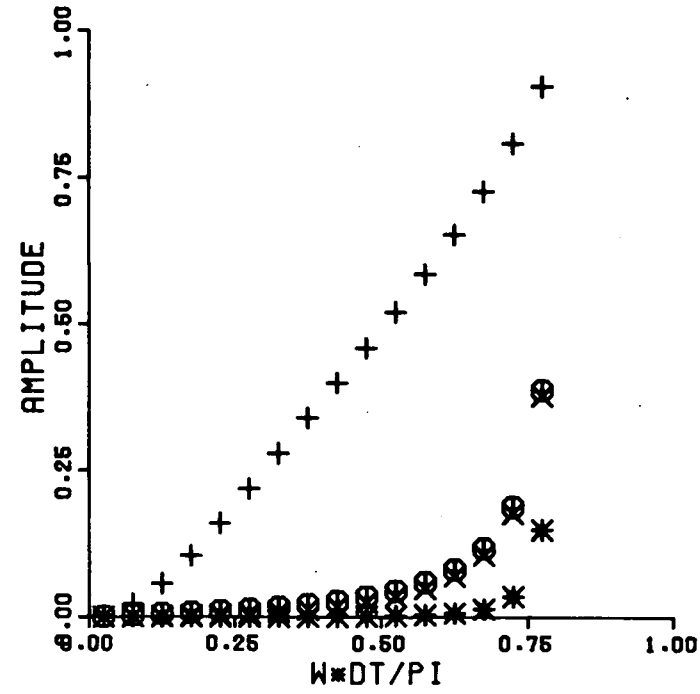

Fig. 5.8 Relative accuracy of radiating boundary conditions for the RS scheme with $f_1 = 0$, $f_2 = .95$, $N = 10$. See text for symbol definitions.

iii) zeroth order space-time extrapolation (5.3.3a), $\oplus$;

iv) linear time extrapolation followed by linear spatial extrapolation (5.3.30), $+$.

The last of these methods is unstable.

In Chapter 2 it was seen that when $\tau > 0$, the velocity and elevation of a travelling wave are no longer in phase. This means that a boundary condition which specifies a scalar relationship between z and u (such as the Engquist and Majda condition) cannot be expected to effectively absorb waves at a boundary. This conclusion is confirmed by the four previous implementations of (5.2.7). With $f_1 = .05$, none of them accurately absorbed long waves.

Is it possible to generalize the absorbing boundary condition for the case $\tau > 0$? Let us follow the Engquist and Majda approach. For a perfectly absorbing left boundary, we want both $z_1^{n+1/2}$ and u_1^{n+1} to consist of only leftward waves. That is, we do not want a reflected right wave. Consistent with (5.3.15), assume

$$z_1^{n+1/2} = \varsigma_L \lambda^{n+1/2} \kappa \tag{5.3.32a}$$

$$u_1^{n+1} = \mu_L \lambda^{n+1} \kappa^{1/2} \tag{5.3.32b}$$

where λ is given by (5.3.17) and $\omega\Delta t > 0$ is a driving frequency. Then

$$u_1^{n+1} = \left(\frac{\mu_L}{\varsigma_L}\right)\left(\frac{\lambda}{\kappa}\right)^{1/2} z_1^{n+1/2}. \tag{5.3.33}$$

Substituting (5.3.32) into (5.3.1a) implies

$$\left(\frac{\mu_L}{\varsigma_L}\right) = -\left(\frac{g}{h}\right)^{1/2}\left(\frac{\lambda^{1/2} - \lambda^{-1/2}}{f_2(\kappa^{1/2} - \kappa^{-1/2})}\right). \tag{5.3.34}$$

Consequently

$$u_1^{n+1} = -z_1^{n+1/2}\left(\frac{g}{h}\right)^{1/2}\left[\frac{\lambda - 1}{f_2(\kappa - 1)}\right] \tag{5.3.35}$$

is perfectly radiating boundary condition for the driving frequency $\omega\Delta t$. (It is interesting to note that (5.3.35) is also obtained when (5.3.32) is substituted into the continuity equation (5.3.1a) for $z_1^{n+3/2}$.)

Unfortunately, there are two difficulties with boundary condition (5.3.35). The first is a dependency on $\omega\Delta t$. This means that the condition will not be perfect for all driving frequencies. The second difficulty is that the condition is complex valued. (This is to be expected since complex scalars permit the phase changes that are required.) Since the finite difference equations are real valued, this boundary condition cannot be applied directly. Two alternatives exist: either we approximate (5.3.35) with a real valued condition; or we solve all the finite difference equations in complex arithmetic. The latter involves substantially more computations and storage, so the former is usually adopted.

(5.3.3a) is a real valued approximation of (5.3.35). It approximates $(\lambda - 1)/(f_2(\kappa - 1))$ with 1. As seen by the minima in the previous reflection coefficient amplitude curves, this is a good estimate when $f_1 = .05$ and $\omega\Delta t = .3\pi$. However better estimates for small $\omega\Delta t$ should exist and would be expected to improve long wave absorption.

A variation of (5.3.31) might be expected to improve boundary absorption when $\tau > 0$. Analytically the velocity leads the elevation by the angle (see (2.2.9c))

$$\theta = \arctan(\tfrac{1}{2}\tau\Delta t, \omega\Delta t). \tag{5.3.36}$$

So in equation (5.3.31c) we actually want the elevation value which is θ behind $z_*^{n+1/2}$, rather than $z_*^{n+1/2}$ itself. In terms of grid intervals, this lag is $\ell = \theta/k\Delta x$. So the elevation

to use is z_{j_*} where $j_* = 1-(\Delta z/\Delta x)+\ell$. Linear interpolation between neighbouring discrete z values could then be used to estimate $z_{j_*}^{n+1/2}$.

Notice that this condition is also dependent on $\omega\Delta t$ and $\tau\Delta t$. In fact, a sample trial with this approach suggests that the accuracy improvement is quite localized. When $(f_1, f_2, \omega\Delta t) = (.05, .95, .08)$, the preceding argument indicates that

$$u_1^{n+1} = -\left(\frac{g}{h}\right)^{1/2} z_4^{n+1/2} \tag{5.3.37}$$

should be more accurate than (5.3.3a) and (5.3.31c). An analysis confirms that the reflection coefficient amplitude is about 40% lower. However this improvement has a high price. At the other nineteen discrete $\omega\Delta t$ values in the same analysis, the coefficients are considerably higher. And, there are now several eigenvalues with modulus greater than unity. So the accuracy improvement for particular values of $\omega\Delta t$ and $\tau\Delta t$ has not only reduced accuracy elsewhere, it has also made the method unstable.

Throughout these investigations we have assumed a pure driving condition at the right boundary. As was seen, this condition acts as a closed boundary for rightward waves which have been generated by an imperfect absorbing left boundary. This fact was confirmed when the left absorbing boundary was replaced with the closed condition (5.3.2a). With parameter values and a driving right boundary as in Fig. 5.2, both boundaries had the same reflection coefficients.

If we wish to minimize the reflection of waves at the right boundary, we can prescribe a condition which not only specifies an inward wave but also attempts to radiate rightward waves. The following condition used by Flather [Fl76] does this.

Decompose $z_N^{n+1/2}$ and $u_{N+1}^{n+1/2}$ into their leftward and rightward components.

$$z_N^{n+1/2} = z_R^{n+1/2} + z_L^{n+1/2} \tag{5.3.38a}$$

$$u_{N+1}^{n+1} = u_R^{n+1} + u_L^{n+1}. \tag{5.3.38b}$$

Then a radiating condition for the right boundary which is analogous to (5.3.3a) is

$$(u_{N+1}^{n+1} - u_L^{n+1}) = \left(\frac{g}{h}\right)^{1/2} (z_N^{n+1/2} - z_L^{n+1/2}). \tag{5.3.39}$$

(Of course, conditions consistent with the other three absorbing boundary implementations could also be used.) u_L^{n+1} and $z_L^{n+1/2}$ are the velocity and elevation for the leftward specified wave. When possible, Flather specifies these individually. However for a travelling wave, they should be related.

Can we find a perfect driving-radiating condition? For a particular value of λ, assume that the leftward wave components are

$$z_L^{n+1/2} = \varsigma_L \lambda^{n+1/2} \kappa^N \tag{5.3.40a}$$

$$u_L^{n+1} = \mu_L \lambda^{n+1} \kappa^{N+1/2} \tag{5.3.40b}$$

for some coefficients z_L, u_L. (5.3.34) then implies

$$z_L^{n+1/2} = -f_2 \left(\frac{h}{g}\right)^{1/2} \left(\frac{1-\kappa^{-1}}{\lambda - 1}\right) u_L^{n+1}. \tag{5.3.41}$$

The associated rightward wave components

$$z_R^{n+1/2} = \varsigma_R \lambda^{n+1/2} \kappa^{-N} \tag{5.3.42a}$$

$$u_R^{n+1} = \mu_R \lambda^{n+1} \kappa^{-(N+1/2)} \tag{5.3.42b}$$

yield

$$u_R^{n+1} = \left(\frac{g}{h}\right)^{1/2} \left[\frac{\lambda - 1}{f_2(\kappa - 1)}\right] z_R^{n+1/2}. \tag{5.3.43}$$

Substituting (5.3.38) into (5.3.43), and using (5.3.41) yields

$$u_{N+1}^{n+1} = u_L^{n+1} + \left(\frac{g}{h}\right)^{1/2} \left[\frac{\lambda - 1}{f_2(\kappa - 1)}\right] \left\{ z_N^{n+1/2} + f_2 \left(\frac{h}{g}\right)^{1/2} \left(\frac{1-\kappa^{-1}}{\lambda - 1}\right) u_L^{n+1} \right\} \tag{5.3.44}$$

as the perfect driving-radiating boundary condition. Approximating $(\lambda - 1)/(f_2(\kappa - 1))$ by unity (this approximation reduces (5.3.35) to (5.3.3a)), this condition can be expressed in real variables as

$$u_{N+1}^{n+1} = \left(\frac{g}{h}\right)^{1/2} z_N^{n+1/2} + Re\left\{ u_L^{n+1} \left[1 + f_2 \left(\frac{1-\kappa^{-1}}{\lambda - 1}\right) \right] \right\}. \tag{5.3.45}$$

The complex multiplier for u_L^{n+1} simply alters the amplitude and phase of this specified function. It does not mean that we require complex variable calculations.

The first test problem (illustrated in Fig. 5.2) was rerun with the purely specified boundary condition replaced by (5.3.45). Not surprisingly, the reflection coefficients for both boundaries were now identical.

Brief investigations with other driving-radiating boundary conditions suggest that the absorption properties of the right boundary are highly sensitive to the consistency of u_L^{n+1} and $z_L^{n+1/2}$. In fact, with the reasonable approximation

$$z_L^{n+3/2} = -\left(\frac{h}{g}\right)^{1/2} u_L^{n+1} \tag{5.3.46}$$

substituted in (5.3.39), it was found that the reflection coefficient amplitudes were close to unity. Hence this particular driving-radiating condition is not a significant improvement over a purely specified condition.

The effects of changing parameters f_2 and N were also briefly investigated. With the same boundary conditions and f_1, N values as in Fig. 5.2, raising f_2 to its limiting stable value of 1.0 made the left boundary condition more accurate. The numerical dispersion relationships also became more accurate. Again using Fig. 5.2 as a reference, N was increased to 16 with the other parameters and boundary conditions left unchanged. All the new eigenvalues and eigenvectors were different, but the top four diagrams were unchanged. This is to be expected. The relationship between λ and $\omega\Delta t$ as given by (5.3.17) is independent of N. The value of N determines only a specific point on the reflection coefficient curves, not the shape of the curves themselves. Increasing N while keeping f_2 constant causes a smaller value of $\omega\Delta t$. So for the same driving frequency ω, a larger N results in a leftward shift on both reflection coefficient curves.

The preceding analysis results were partially confirmed with numerical model tests. Boundary conditions (5.3.3a), (5.3.29), (5.3.30) and (5.3.31) were tested with the driving frequency $\omega\Delta t = .70685835$. This is the fifth discrete forcing frequency shown in all the analysis plots. Parameter values for the model runs were identical to those in the analysis, and the model computations were done in double precision. Each run lasted for 270 time steps 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 (5.3.19).

In order to determine if the coefficients were converging to the values predicted by the analysis, four least squares fits were made over the successive time step ranges [135,167], [168,201], [202,235], [236,270]. As predicted by the analysis, condition (5.3.30) caused model instability. In all other cases, residuals decreased with each successive fit, and the fitted values seemed to be converging. The coefficient values z_1, z_2 obtained from the fourth fit were identical to at least 4 digits, with the analysis results.

5.4 A GKS Stability Analysis

In this section we perform a GKS stability analysis of a Galerkin FEM (GFEM) with selected boundary conditions. Since the GKS analysis assumes that matrix A in (5.2.1a) is Hermitian, we must re-express the governing equations (5.2.1) in terms of characteristic variables. Assuming $\tau = 0$, they are

$$\frac{\partial}{\partial t}\begin{pmatrix} v \\ w \end{pmatrix} = (gh)^{1/2}\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} v \\ w \end{pmatrix}. \tag{5.4.1}$$

The characteristic variables (v, w) are equal to (v_1, v_2) in (5.2.2).

The selected GFEM has piecewise linear basis functions and Crank-Nicolson (CN) time stepping. Its difference equations are

$$\begin{aligned} &\tfrac{1}{6}[(v_{j-1}^{n+1} - v_{j-1}^{n}) + 4(v_j^{n+1} - v_j^{n}) + (v_{j+1}^{n+1} - v_{j+1}^{n})] \\ &\qquad = -\tfrac{1}{4}(gh)^{1/2}(\Delta t/\Delta x)[v_{j+1}^{n} - v_{j-1}^{n} + v_{j+1}^{n+1} - v_{j-1}^{n+1}] \qquad (5.4.2a) \\ &\tfrac{1}{6}[(w_{j-1}^{n+1} - w_{j-1}^{n}) + 4(w_j^{n+1} - w_j^{n}) + (w_{j+1}^{n+1} - w_{j+1}^{n})] \\ &\qquad = \tfrac{1}{4}(gh)^{1/2}(\Delta t/\Delta x)[w_{j+1}^{n} - w_{j-1}^{n} + w_{j+1}^{n+1} - w_{j-1}^{n+1}]. \qquad (5.2.4b) \end{aligned}$$

Unlike the RS scheme,notice that there is no staggering of the variables in either space or time.

It was seen in Section 5.2 that well-posed boundary conditions for (5.4.1) require specifying v at the left boundary and w at the right boundary. Assume the well-posed conditions

$$v_0^n = 0 \qquad (5.4.3a)$$

$$w_N^n = f(n) \tag{5.4.3b}$$

for some specified function f. These represent an absorbing left boundary, and a combined specified-absorbing right boundary. The homogeneous right boundary condition, $w_N^n = 0$, is actually used in place of (5.4.3b) in the stability analysis.

In order to solve (5.4.2), additional boundary conditions are required for v_N and w_0. These conditions are obtained through constant spatial extrapolation; that is, (5.1.2a) with $q = 1$. The complete set of boundary conditions is then

$$v_0^n = 0 \tag{5.4.4a}$$

$$v_N^n = v_{N-1}^n \tag{5.4.4b}$$

$$w_N^n = 0 \tag{5.4.4c}$$

$$w_0^n = w_1^n. \tag{5.4.4d}$$

Stability of the two-boundary problem is guaranteed if each of the associated quarter-plane problems with one boundary is stable [Gu72, Theorem 5.4]. In addition to satisfying one boundary condition, the solution of these quarter-plane problems must be spatially bounded. For a right quarter-plane problem this means

$$\sum_{j=0}^{\infty} v_j^2 \Delta x < \infty. \tag{5.4.5}$$

Since our boundary conditions do not couple the two characteristic variables, (5.4.2a) and (5.4.2b) can be studied separately. This means that for the stability of (5.4.2) with (5.4.4), it is sufficient that each of the following four quarter-plane problems be stable:

i) (5.4.2a) with (5.4.4a),

ii) (5.4.2a) with (5.4.4b),

iii) (5.4.2b) with (5.4.4c),

iv) (5.4.2b) with (5.4.4d).

Traditionally, quarter-plane problems are analysed with their boundaries on the left. Those with boundaries on the right can be transformed to an equivalent left boundary

problem (e.g., [Gu82]). In our case, the left quarter-plane problems ii) and iii) are respectively equivalent to the right quarter-plane problems iv) and i). Consequently, we need only confirm the stability of problems i) and iv) to have stability of all four quarter-plane problems.

We begin with a GKS stability analysis of problem i) using Trefethen [Tr83] as a guide. Assuming *resolvent* solutions of the form

$$v_j^n = \alpha_0 \lambda^n \kappa^j, \tag{5.4.6}$$

(5.4.2a) becomes

$$(\lambda - 1)(1 + 4\kappa + \kappa^2) + \tfrac{3}{2} f_2 (\lambda + 1)(\kappa^2 - 1) = 0. \tag{5.4.7}$$

This is the *resolvent equation*. For a particular value of λ, (5.4.7) has two roots. The general numerical solution is therefore

$$v_j^n = \lambda^n (\alpha_1 \kappa_1^j + \alpha_2 \kappa_2^j). \tag{5.4.8}$$

Boundary condition (5.4.4a) then requires that

$$\alpha_1 + \alpha_2 = 0. \tag{5.4.9}$$

The first step in a GKS stability analysis is to confirm that there are no nontrivial solutions of the form (5.4.8) with $|\kappa| \leq 1$ when $|\lambda| > 1$. Nontrivial solutions with $|\kappa| < 1$ are called *eigensolutions* and cause an instability of the Godunov-Ryabenkii (GR) type.

The spatial stencil for (5.4.2a) extends one point to both the left and right of v_j. Trefethen's proposition [Tr83, page 206] (which is based on [Gu72, Lemma 5.2]), then implies that any solution with $|\lambda| > 1$ has exactly one κ value with modulus greater than unity, and one with modulus less than unity. Assume $|\kappa_2| > 1$ and $|\kappa_1| < 1$. Then $\alpha_2 = 0$, otherwise (5.4.5) is not satisfied. (5.4.9) implies that $\alpha_1 = 0$ also. So only the trivial solution can exist when $|\lambda| > 1$.

The GR check is only necessary for stability. In order to obtain a condition that is also sufficient (or nearly so [Tr83]), we must also investigate the case $|\lambda| = 1$.

The second step in a GKS stability analysis is to look for nontrivial solutions with $|\kappa| \leq 1$ when $|\lambda| = 1$. When $|\lambda| = 1$, the roots of (5.4.7) satisfy

$$\kappa_1 \kappa_2 = \frac{1 + \frac{3}{2} i f_2 \cot(\frac{1}{2}\omega \Delta t)}{1 - \frac{3}{2} i f_2 \cot(\frac{1}{2}\omega \Delta t)}. \tag{5.4.10}$$

Since this product has magnitude unity, two cases are possible:

$$\text{either} \quad \text{i)} \quad |\kappa_1| < 1 \text{ and } |\kappa_2| > 1,$$

$$\text{or} \quad \text{ii)} \quad |\kappa_1| = |\kappa_2| = 1.$$

The dispersion relationship for (5.4.2a) is obtained when

$$\lambda = e^{i\omega \Delta t} \tag{5.4.11a}$$

$$\kappa = e^{-ik\Delta x} \tag{5.4.11b}$$

are substituted into (5.4.7). In particular,κ_1 and κ_2 are associated with wavenumbers k_1 and k_2. When $|\omega \Delta t| < \omega_c$ (the cutoff frequency), both $k_1$and k_2 are real valued and case ii) applies. The resultant dispersion curve is shown in Fig. 5.9. When $|\omega \Delta t| > \omega_c$, both k_1 and k_2 are complex valued and case i) applies. Following the same argument as with $|\lambda| > 1$, (5.4.5) and (5.4.9) imply that only a trivial solution can exist for case i).

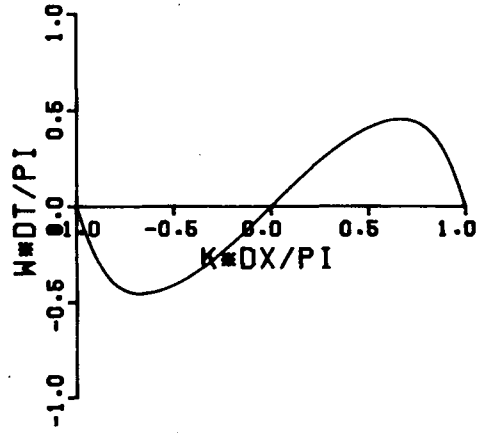

Fig. 5.9 Dispersion relationship for (5.4.2a).

Case ii) arises when $(\omega \Delta t, k\Delta x)$ lie on the dispersion curve shown in Fig. 5.9. Nontrivial solutions satisfying (5.4.8) now exist and have the form

$$v_j^n = \alpha_1 e^{in\omega \Delta t}(e^{-ijk_1 \Delta x} - e^{-ijk_2 \Delta x}). \tag{5.4.12}$$

In particular, notice that such solutions consist of two wavenumbers when $|\omega\Delta t| < \omega_c$.

The existence of nontrivial solutions when $|\lambda| = 1$ necessitates a further test. Specifically, we must perturb λ to λ', where $|\lambda'| > 1$, and observe the behaviour of all the κ values in the nontrivial solution. In our case, assume that κ_1 and κ_2 are respectively perturbed to κ_1' and κ_2'. The quarter-plane problem is then unstable iff both $|\kappa_1'| < 1$ and $|\kappa_2'| < 1$. Nontrivial solutions which perturb in this manner, and which have $|\lambda| = 1$ and $|\kappa| = 1$ for at least one κ, are called *generalized eigensolutions.*

From the GKS point of view, the perturbation determines whether certain resolvent solutions for $|\lambda| > 1$ extend continuously to $|\lambda| = 1$. Since roundoff errors can perturb λ values,it is important to determine if such perturbations could cause instability.

In our case, Trefethen's proposition predicts the behaviour of κ_1 and κ_2 when λ is perturbed. Since $|\lambda'| > 1$, we must have $|\kappa_1'| < 1$ and $|\kappa_2'| > 1$. (The perturbation behaviour of λ and κ was confirmed numerically by substituting (5.4.11a) and $\lambda' = 1.001\lambda$ into (5.4.7) for 100 values of $\omega\Delta t$ in the range $(-\pi, \pi]$.) The nontrivial solutions (5.4.12) are therefore stable.

Since quarter-plane problem i) admits no eigensolutions or generalized eigensolutions with $|\lambda| \geq 1$, it is GKS stable.

Trefethen interprets the perturbation condition in terms of group velocity. However, he only presents a necessary condition for stability [Tr83, Theorem 1]. His theory is developed for the hyperbolic equation

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \tag{5.4.13}$$

over a right quarter-plane. It is therefore directly applicable to our characteristic variable w. In particular, he proves that with $|\lambda| = |\kappa_1| = |\kappa_2| = 1$, if the group velocities associated with $k_1\Delta x$ and $k_2\Delta x$ are non-negative and one is strictly positive, then the difference scheme is unstable. Intuitively this makes sense since the analytic solution to (5.4.13) is a leftward wave with negative phase (and group) velocity. Positive numerical group velocity then corresponds to spontaneous radiation of energy at the left boundary into the problem domain. Conceivably, this will cause growth and instability of the numerical solution.

It would seem that Trefethen's theory still applies when the governing equation for the right quarter-plane problem is

$$\frac{\partial u}{\partial t} = -\frac{\partial u}{\partial x}. \tag{5.4.14}$$

Presumably this is because we do not want any energy radiating into the domain from a homogeneous boundary condition, regardless of whether the characteristic variable is incoming or outgoing there. This theory extension is confirmed by the numerical perturbations performed when analysing the stability of (5.4.2a) with (5.4.4a). It was always the κ value associated with the smaller $|k\Delta x|$ which was perturbed inside the unit circle. This is the κ which might cause GKS instability. Fig. 5.9 shows that the group velocities (=dispersion curve slopes) for the smaller $|k\Delta x|$ values are positive.

In any event, Fig. 5.9 shows that the group velocities associated with $k_1\Delta x$ and $k_2\Delta x$ have opposite signs when $|\omega\Delta t| < \omega_c$, and equal zero when $\omega\Delta t = \pm\omega_c$. So Trefethen's theory cannot be applied. His theorem would indicate instability if only one κ were present in the nontrivial solution (5.4.12), and the associated group velocity were positive.

We now proceed to a GKS stability analysis of quarter-plane problem iv). The resolvent equation for (5.4.2b) is

$$(\lambda - 1)(1 + 4\kappa + \kappa^2) - \tfrac{3}{2} f_2(\lambda + 1)(\kappa^2 - 1) = 0. \tag{5.4.15}$$

The general numerical solution is

$$w_j^n = \lambda^n(\alpha_1\kappa_1^j + \alpha_2\kappa_2^j), \tag{5.4.16}$$

and boundary condition (5.4.4d) implies

$$\alpha_1(1 - \kappa_1) + \alpha_2(1 - \kappa_2) = 0. \tag{5.4.17}$$

We first look for nontrivial solutions with $|\kappa| \leq 1$ when $|\lambda| > 1$. Trefethen's proposition again implies that $|\kappa_1| < 1$ and $|\kappa_2| > 1$ when $|\lambda| > 1$. In order to satisfy the w_j analogue to (5.4.5), we must have $\alpha_2 = 0$. This implies either $\alpha_1 = 0$ or $\kappa = 1$. Since $|\kappa_1| < 1$, $\alpha_1 = 0$. Therefore only trivial solutions exist when $|\lambda| > 1$.

We next look for nontrivial solutions with $|\kappa| \leq 1$ when $|\lambda| = 1$. When $|\lambda| = 1$, the roots of (5.4.15) satisfy

$$\kappa_1 \kappa_2 = \frac{1 - \frac{3}{2} i f_2 \cot(\frac{1}{2}\omega \Delta t)}{1 + \frac{3}{2} i f_2 \cot(\frac{1}{2}\omega \Delta t)}. \tag{5.4.18}$$

This product also has magnitude unity, so the same two cases arise:

$$\text{either} \quad \text{i)} \quad |\kappa_1| < 1 \text{ and } |\kappa_2| > 1,$$

$$\text{or} \quad \text{ii)} \quad |\kappa_1| = |\kappa_2| = 1.$$

The dispersion relationship for (5.4.2b) is calculated by substituting (5.4.11) into (5.4.15). It is shown in Fig. 5.10. As before, the first case arises when $|\omega \Delta t| > \omega_c$. Applying the same argument as we did for $|\lambda| > 1$, it follows that $\alpha_1 = \alpha_2 = 0$. So only trivial solutions occur when $|\lambda| = 1$ and $|\kappa_1| < 1$, $|\kappa_2| > 1$.

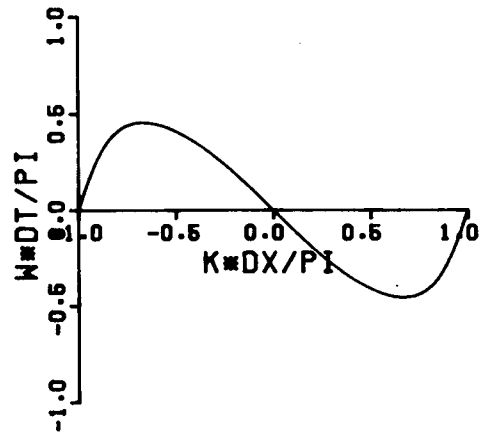

Fig. 5.10 Dispersion relationship for (5.4.2b).

The second case arises when $(\omega \Delta t, k \Delta x)$ lie on the dispersion curve. Nontrivial solutions now exist and have the form

$$w_j^n = \alpha_1 e^{in\omega \Delta t} \left(e^{-ijk_1 \Delta x} - \left[\frac{1 - e^{-ijk_1 \Delta x}}{1 - e^{-ijk_2 \Delta x}} \right] e^{-ijk_2 \Delta x} \right). \tag{5.4.19a}$$

Notice that these solutions need not contain two wavenumbers. $(\lambda, \kappa_1, \kappa_2) = (1, 1, -1)$ satisfies both (5.4.15) and (5.4.17) when $\alpha_2 = 0$. Consequently, the constant nontrivial solution

$$w_j^n = \alpha_1 \tag{5.4.19b}$$

may exist.

We now perform a perturbation analysis. When the nontrivial solution involves both κ_1 and κ_2, Trefethen's proposition can be applied. Specifically, since $|\lambda'| > 1$, $|\kappa_1'| < 1$ and $|\kappa_2'| > 1$. So nontrivial solutions of the form (5.4.19a) which contain two wavenumbers cannot cause instability.

The nontrivial solution (5.4.19b) requires special consideration. Set $\lambda' = 1 + \epsilon$, where the complex valued ϵ is such that $|\lambda'| > 1$. Then $\kappa = 1$ is perturbed to

$$\kappa' = \frac{2\epsilon + \left[3\epsilon^2 + \frac{9}{4}f_2^2(2+\epsilon)^2\right]^{1/2}}{\frac{3}{2}f_2(2+\epsilon) - \epsilon}. \tag{5.4.20}$$

Assuming $f_2 >> \epsilon$, it can be shown that $|\kappa'| > 1$. This was also confirmed numerically by setting $\epsilon = .01e^{i\theta}$ for 100 values of θ such that $|\lambda'| > 1$. In all instances, $\kappa = 1$ was perturbed outside the unit circle. So (5.4.19b) is not a generalized eigensolution. This result is consistent with Trefethen's theory since the group velocity at $k\Delta x = 0 \quad (\kappa = 1)$ is negative.

The quarter-plane problem iv) has no eigensolutions or generalized eigensolutions when $|\lambda| \geq 1$, and is therefore GKS stable. Consequently, the system of equations (5.4.2) with boundary conditions (5.4.4) is GKS stable.

It would be interesting to find boundary conditions that cause an instability of the Trefethen type. Consider the characteristic variable w. When $|\lambda| = 1$ and $|\omega\Delta t| < \omega_c$, it is seen from Fig. 5.10 that $2\Delta x$ waves have positive group velocity. Trefethen's theory then implies that any boundary conditions which support these waves will be unstable.

One such pair of conditions is

$$w_N^n + w_{N-1}^n = 0 \tag{5.4.21a}$$

$$w_0^n = w_2^n. \tag{5.4.21b}$$

The former condition overspecifies the right boundary (it is not well-posed), while the latter is a form of constant extrapolation. Assuming the general solution (5.4.16), (5.4.21a) implies

$$\alpha_1\kappa_1^{N-1}(1+\kappa_1) + \alpha_2\kappa_2^{N-1}(1+\kappa_2) = 0, \tag{5.4.22a}$$

while (5.4.21b) implies

$$\alpha_1(\kappa_1^2 - 1) + \alpha_2(\kappa_2^2 - 1) = 0. \tag{5.4.22b}$$

$2\Delta x$ waves arise when $\lambda = 1$ and $\kappa = -1$ in the characteristic equation (5.4.15). $\kappa = 1$ is also a solution when $\lambda = 1$. Set $\kappa_1 = 1$ and $\kappa_2 = -1$. Then (5.4.22b) is satisfied for all α_1 and α_2, whereas (5.4.22a) requires $\alpha_1 = 0$.

$$w_j^n = \alpha_2(-1)^j \tag{5.4.23}$$

is therefore a nontrivial solution to (5.4.15) and (5.4.21). Since its group velocity is positive, the model will be unstable.

It can be shown that no other eigensolutions or generalized eigensolutions (nontrivial solutions with $|\lambda| \geq 1$) are supported by boundary conditions (5.4.21). Therefore, model instability arises solely from the generalized eigensolution (5.4.23). This instability has been confirmed with a test model which accelerated the accumulation of rounding errors by rounding all newly calculated w_j^n values to three decimal places. A similar trick was also used by Gustafsson [Gu82].

5.5 A Galerkin Finite Element Method

This section examines the accuracy of several boundary conditions for the GFEM which combines piecewise linear basis functions with Crank-Nicolson time stepping. Although this GFEM will be applied to the primitive equation variables z and u, the simple transformation (5.1.4) can be used to re-express results in terms of the characteristic variables. As in Section 5.3, the test problem is a one dimensional channel with an absorbing left boundary, and a combination driving-absorbing right boundary. A closed left boundary will not be investigated here, but the same analysis techniques could certainly be applied.

In the domain interior, the GFEM difference equations are

$$\begin{aligned}\tfrac{1}{6}[(z_{j-1}^{n+1} - z_{j-1}^n) + 4(z_j^{n+1} - z_j^n) + (z_{j+1}^{n+1} - z_{j+1}^n)] \\ + \left(\frac{h\Delta t}{4\Delta x}\right)[u_{j+1}^{n+1} - u_{j-1}^{n+1} + u_{j+1}^n - u_{j-1}^n] = 0\end{aligned} \tag{5.5.1a}$$

$$\frac{1}{6}[(u_{j-1}^{n+1}-u_{j-1}^{n})+4(u_{j}^{n+1}-u_{j}^{n})+(u_{j+1}^{n+1}-u_{j+1}^{n})]$$
$$+\left(\frac{g\Delta t}{4\Delta x}\right)[z_{j+1}^{n+1}-z_{j-1}^{n+1}+z_{j+1}^{n}-z_{j-1}^{n}]$$
$$+\frac{1}{12}\tau\Delta t[u_{j-1}^{n+1}+u_{j-1}^{n}+4(u_{j}^{n+1}+u_{j}^{n})+u_{j+1}^{n+1}+u_{j+1}^{n}]=0. \qquad (5.5.1b)$$

Initial conditions are assumed to be

$$u_j^0 = g_1(j) \qquad (5.5.2a)$$

$$z_j^0 = g_2(j) \qquad (5.5.2b)$$

for some functions g_1 and g_2, while the boundary conditions are

$$u_1^n = -\left(\frac{g}{h}\right)^{1/2} z_1^n \qquad (5.5.3a)$$

$$u_N^n = \left(\frac{g}{h}\right)^{1/2} z_N^n + f(n) \qquad (5.5.3b)$$

$$-\left(\frac{g}{h}\right)^{1/2} z_1^n + u_1^n = -\left(\frac{g}{h}\right)^{1/2} z_2^n + u_2^n \qquad (5.5.3c)$$

$$\left(\frac{g}{h}\right)^{1/2} z_N^n + u_N^n = \left(\frac{g}{h}\right)^{1/2} z_{N-1}^n + u_{N-1}^n. \qquad (5.5.3d)$$

$f(n)$ is driving function such as (5.2.5c). These boundary conditions are equivalent to (5.4.4).

(5.5.3a) and (5.5.3b) approximate the absorbing left and driving-absorbing right boundaries respectively. They are called *physical* conditions. The other two conditions are required so that the system of equations which determines the numerical solution at each time step, has full rank. They are called *artificial* conditions.

(5.5.3c) and (5.5.3d) are obtained through zeroth order spatial extrapolation of the leftward characteristic variable at the left boundary, and the rightward characteristic variable at the right boundary. They are by no means the only pair of artificial conditions. Many other possibilities exist. In this section, we demonstrate that a dispersion analysis can be used to choose the pair that is both stable and most accurate.

The accuracy analysis is similar to that for the RS scheme. Defining

$$\mathbf{X}^n = (u_1^n, z_1^n, \ldots, u_N^n, z_N^n), \qquad (5.5.4)$$

(5.5.1) and (5.5.3) can be expressed in matrix form as

$$A\mathbf{X}^{n+1} = B\mathbf{X}^n + \mathbf{X}_D f(n+1), \tag{5.5.5}$$

where the matrices A and B define the finite element operations, and $\mathbf{X}_D$ is the vector which locates the driving conditions. Since a system of equations must be solved at each time step, the GFEM is implicit. The matrix A is nonsingular, otherwise some of the row equations would be redundant and the system of equations would not have full rank.Consequently, A^{-1} exists and the steady state calculation will proceed as with the RS scheme.

Assume the forcing condition (5.2.5c). Provided $e^{i\omega\Delta t}$ is not an eigenvalue of $A^{-1}B$, the numerical solution after $n+1$ time steps is

$$\mathbf{X}^{n+1} = (A^{-1}B)^n\mathbf{X}^0 + Re\{ae^{i[(n+1)\omega\Delta t-\phi]}[I-(e^{-i\omega\Delta t}A^{-1}B)^{n+1}]\mathbf{Y}\} \tag{5.5.6a}$$

where

$$\mathbf{Y} = [I-e^{-i\omega\Delta t}A^{-1}B]^{-1}A^{-1}\mathbf{X}_D. \tag{5.5.6b}$$

When all eigenvalues of $A^{-1}B$ are strictly inside the unit circle, (5.5.6a) becomes

$$\mathbf{X}^{n+1} = Re\{ae^{i[(n+1)\omega\Delta t-\phi]}\mathbf{Y}\}. \tag{5.5.7}$$

As with the RS scheme, $\mathbf{Y}$ contains the spatial profile of the steady state solution.

The precise form of $\mathbf{Y}$ is found by assuming that (5.5.1) and (5.5.3) have separable solutions of the form

$$\begin{pmatrix} z_j^n \\ u_j^n \end{pmatrix} = \begin{pmatrix} \varsigma_0 \\ \mu_0 \end{pmatrix}\lambda^n\kappa^j. \tag{5.5.8}$$

This is essentially the same substitution that was used to form the resolvent solutions in the GKS stability analysis of Section 5.4. (5.5.1) has nontrivial solutions of the form (5.5.8) when

$$\begin{aligned}(\lambda-1)^2(\kappa^2+4\kappa+1)^2 + \tfrac{1}{2}\tau\Delta t(\lambda^2-1)(\kappa^2+4\kappa+1)^2 \\ -\tfrac{9}{4}gh\left(\frac{\Delta t}{\Delta x}\right)^2(\lambda+1)^2(\kappa^2-1)^2 = 0.\end{aligned} \tag{5.5.9}$$

This dispersion relationship (or resolvent equation) has four roots for each value of

$$\lambda = e^{i\omega\Delta t}. \tag{5.5.10}$$

The numerical solution is therefore

$$\begin{pmatrix} z_j^n \\ u_j^n \end{pmatrix} = \lambda^n \sum_{\ell=1}^{4} \begin{pmatrix} \varsigma_\ell \\ \mu_\ell \end{pmatrix} \kappa_\ell^j \tag{5.5.11}$$

where precise values of the coefficients ς_ℓ, μ_ℓ are determined by the boundary conditions (5.5.3).

The dispersion curve for (5.5.9) illustrates the relationship between λ and κ. When $\tau = 0$, this curve is the composite of Fig. 5.9 and Fig. 5.10. Positive values of $\omega\Delta t$ less than the cutoff frequency are associated with four wavenumbers. Two of these wavenumbers are positive and correspond to rightward waves, while the other two are negative and correspond to leftward waves. As Platzman [Pl81] points out, this means that forced periodic motion will generate a short wavelength noise component in addition to the longer wavelength appropriate to the forcing frequency. The energy in these four waves is determined by the boundary conditions. Since energy may be transferred from one wavelength to another at each reflection, the calculation of reflection coefficients is much more complicated here than with the RS scheme. However an indication of boundary condition accuracy can be obtained by calculating the steady state values of the coefficients ς_ℓ and μ_ℓ in (5.5.11).

The calculation of these coefficients is similar to that for the RS scheme. Assume N, f_1, f_2 are constant, and check that all the eigenvalues of $A^{-1}B$ are inside the unit circle. This ensures that the steady state numerical solution is given by (5.5.7). For any driving frequency $\omega\Delta t$, the spatial profile of the steady state solution is then found through the following steps:

i) using (5.5.9) solve for κ_1, κ_2, κ_3, κ_4;

ii) re-arrange (5.5.6b) to

$$\left[A - e^{-i\omega\Delta t}B\right]\mathbf{Y} = \mathbf{X}_D \tag{5.5.12}$$

and solve for $\mathbf{Y}$;

iii) do a least squares fit on the z (and u) components of $\mathbf{Y}$ to find the complex coefficients ς_ℓ and μ_ℓ in (5.5.11).

The analysis technique is now illustrated with five pairs of artificial boundary conditions. These conditions are used in combination with the physical conditions (5.3.3a), and (5.3.3b) with $f(n)$ specified as in (5.3.6c). Initial conditions are assumed to be zero, and the parameters (f_1, f_2, N) are fixed at (0.,1.,10). Analysis results are presented in figures similar to those of Section 5.3.

The first pair of artificial conditions is (5.5.3c) and (5.5.3d). Fig. 5.11 displays the analysis results.

The two lowest diagrams show the relationships between the eigenvalues and eigenvectors of matrix $A^{-1}B$. The diagram on the lower left is a dispersion relationship. The dotted line is the analytic relationship while the solid line is the numerical dispersion relationship for a ring domain. Asterisks plot $\omega\Delta t$ versus $k\Delta x$ (i.e., the arguments of λ versus those for κ). Only positive values of $\omega\Delta t$ and $k\Delta x$ are shown. Notice that all the asterisks lie along the solid line. This implies that the phase and group velocities of all transient and random signals are the same as they would be for some ring domain.

The diagram on the lower right plots $|\lambda|$ and $|\kappa|$ versus $k\Delta x$. The $|\lambda|$ values are shown as circles while the $|\kappa|$ values are asterisks. All the eigenvalue amplitudes are strictly less than unity so a steady state solution does exist. Lagrangian amplitudes are designated by crosses. Denote the $k\Delta x$ value when $\omega\Delta t = \omega_c$ as the folding wavenumber k_f [Pl81]. Then the Lagrangian amplitudes are close to unity when $k\Delta x < k_f$, and decrease monotonically when $k\Delta x > k_f$. Consequently, short transient waves decrease in amplitude as they propagate, while longer transient waves propagate with little amplitude change.

The middle row of diagrams shows model response to various driving frequencies. Again the left diagram is a dispersion curve with the same solid and dotted lines as the diagram below it. Asterisks now denote the numerical values obtained when twenty equally spaced values of $\omega\Delta t$ are assumed in (5.5.10) and substituted into (5.5.9). Again these asterisks lie along the numerical dispersion curve for a ring domain.

Notice that when $\omega\Delta t < \omega_c$, two wavenumbers are generated by the driving frequency. One is larger than k_f and one is smaller. Both associated κ values have amplitude unity.

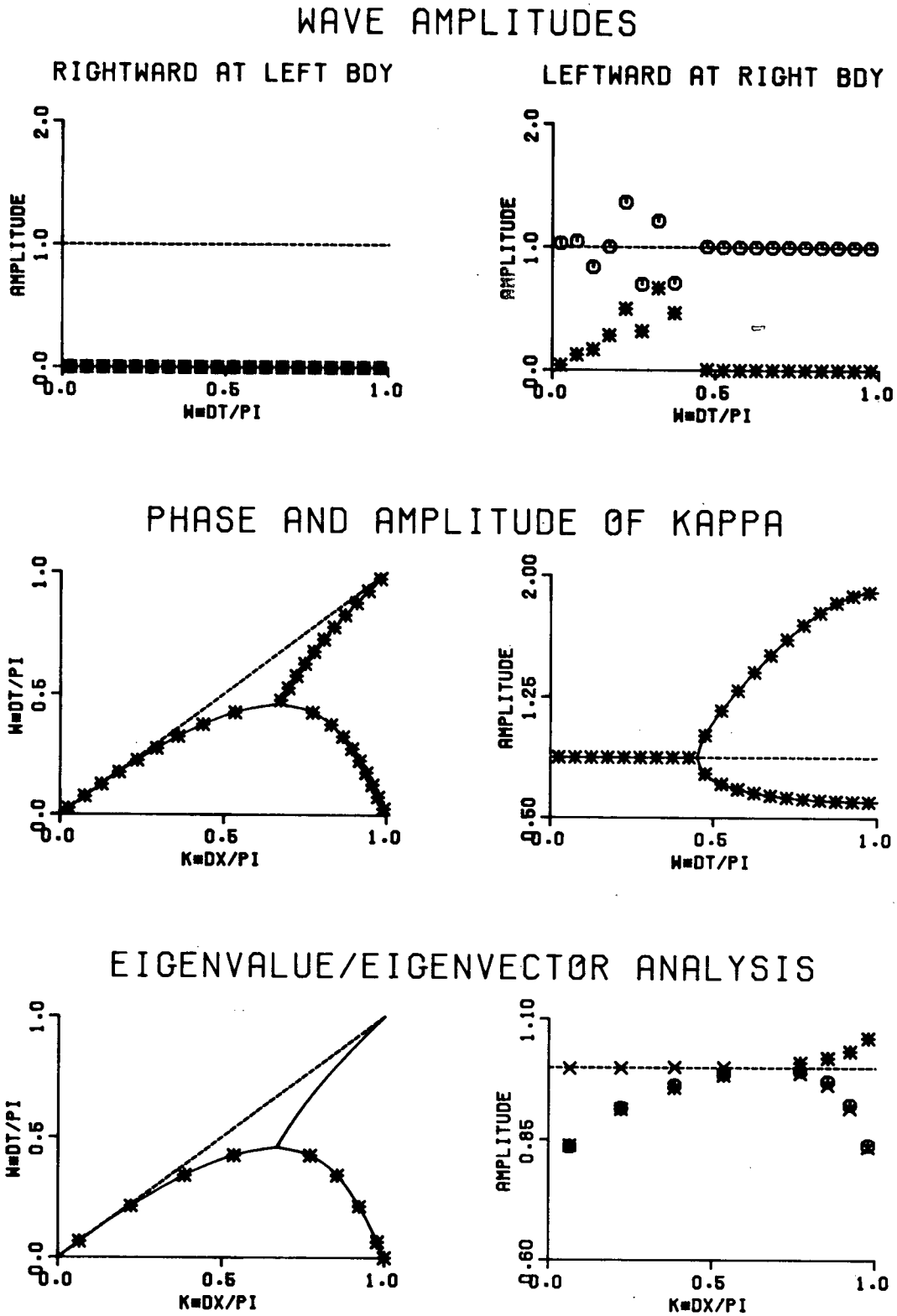

Fig. 5.11. Dispersion analysis for the GFEM with $f_1 = 0$, $f_2 = 1.$, $N = 10$; physical boundary conditions (5.5.3a), (5.5.3b); and artificial conditions (5.5.3c), (5.5.3d).

(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. The least squares analysis for the coefficients ς_ℓ

and μ_ℓ indicates that the κ value with the larger magnitude dominates. Consequently, an evanescent signal emanating from the right boundary will decay as it propagates leftward. This is predicted by Vichnevetsky [Vi80], 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 top diagrams show the amplitudes of the two rightward waves at the left boundary, and the two leftward waves at the right boundary. When $\omega\Delta t < \omega_c$, circles denote the longer wave and asterisks the shorter wave. When $\omega\Delta t > \omega_c$, both waves have the same wavelength with circles denote the wave with the larger $|\kappa|$. Conditions (5.5.3a) and (5.5.3c) are seen to produce rightward wave amplitudes that are very close to zero. Since the left boundary is absorbing, there should not be any reflected rightward waves. Therefore, the left boundary conditions are very accurate.

The right boundary condition determines how energy at the forced boundary is distributed between the two leftward waves. Since the driving frequency was assumed to have amplitude 1.0, an ideal boundary condition should assign all this energy to the longer wavelength. Clearly (5.5.3b) and (5.5.3d) are not ideal. For small $\omega\Delta t$, the amplitude of the short wave is close to zero and the amplitude of the long wave is close to 1. This means that the short wave has only a small amount of energy. However, as $\omega\Delta t$ increases, so do the amplitudes of the short wave. For example, when $\omega\Delta t = .225\pi$, amplitudes of the short and long waves are .50 and 1.37 respectively. Notice that both amplitude patterns seem to oscillate as $\omega\Delta t$ increases. When $\omega\Delta t > \omega_c$, both waves have the same length. All the energy is now 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 [Vi80].

The second set of artificial conditions is also obtained through zeroth order spatial extrapolation. However, in this case the extrapolation is applied to z rather than the characteristic variables. The conditions are

$$z_1^n = z_2^n \qquad (5.5.13a)$$

$$z_N^n = z_{N-1}^n. \qquad (5.5.13b)$$

Fig. 5.12 shows the analysis results.

Although these results seem quite similar to those of Fig. 5.11, there is one major difference. The artificial conditions have made the GFEM unstable. Close examination of the lower right diagram reveals that the eigenvalue amplitudes associated with $k\Delta x/\pi =$ .5339 and .7739 are larger than unity. The eigenvalues are $\lambda = 1.00576e^{\pm i1.33441}$. Although these values will cause only slow growth, a numerical model run has confirmed that it does occur.

The lower right diagram also reveals that $\lambda = 1$ is an eigenvalue. It may also contribute toward instability since its associated eigenvector has a spatial profile that is constant in z and a $2\Delta x$ wave in u. The group velocity of this $2\Delta x$ wave is positive when we consider the leftward dispersion curve (e.g., Fig. 5.10). So Trefethen's theory implies that there should be an instability arising from the left boundary. This is difficult to confirm with a numerical model since the eigenvalue with modulus 1.00576 dominates the unstable growth.

Instability invalidates our analysis of model response to a driving frequency. In particular, we cannot assume that the numerical solution (5.5.6) converges to (5.5.7). Wave amplitudes at the two boundaries now consist of two components, only one of which arises from the steady state solution given by (5.5.7). The top diagrams only show wave amplitudes of this convergent component. As might be expected, they are less accurate than those of Fig. 5.11. Amplitudes of the leftward waves at the right boundary indicate that more energy is assigned to the shorter wave. Furthemore, rightward waves are seen to be generated by reflections at the left boundary.

The third set of artificial conditions arises from zeroth order space-time extrapolation of the characteristic variables. The conditions are

$$-\left(\frac{g}{h}\right)^{1/2} z_1^{n+1} + u_1^{n+1} = -\left(\frac{g}{h}\right)^{1/2} z_2^n + u_2^n \qquad (5.5.14a)$$

$$\left(\frac{g}{h}\right)^{1/2} z_N^{n+1} + u_N^{n+1} = \left(\frac{g}{h}\right)^{1/2} z_{N-1}^n + u_{N-1}^n. \qquad (5.5.14b)$$

Fig. 5.13 shows the analysis results.

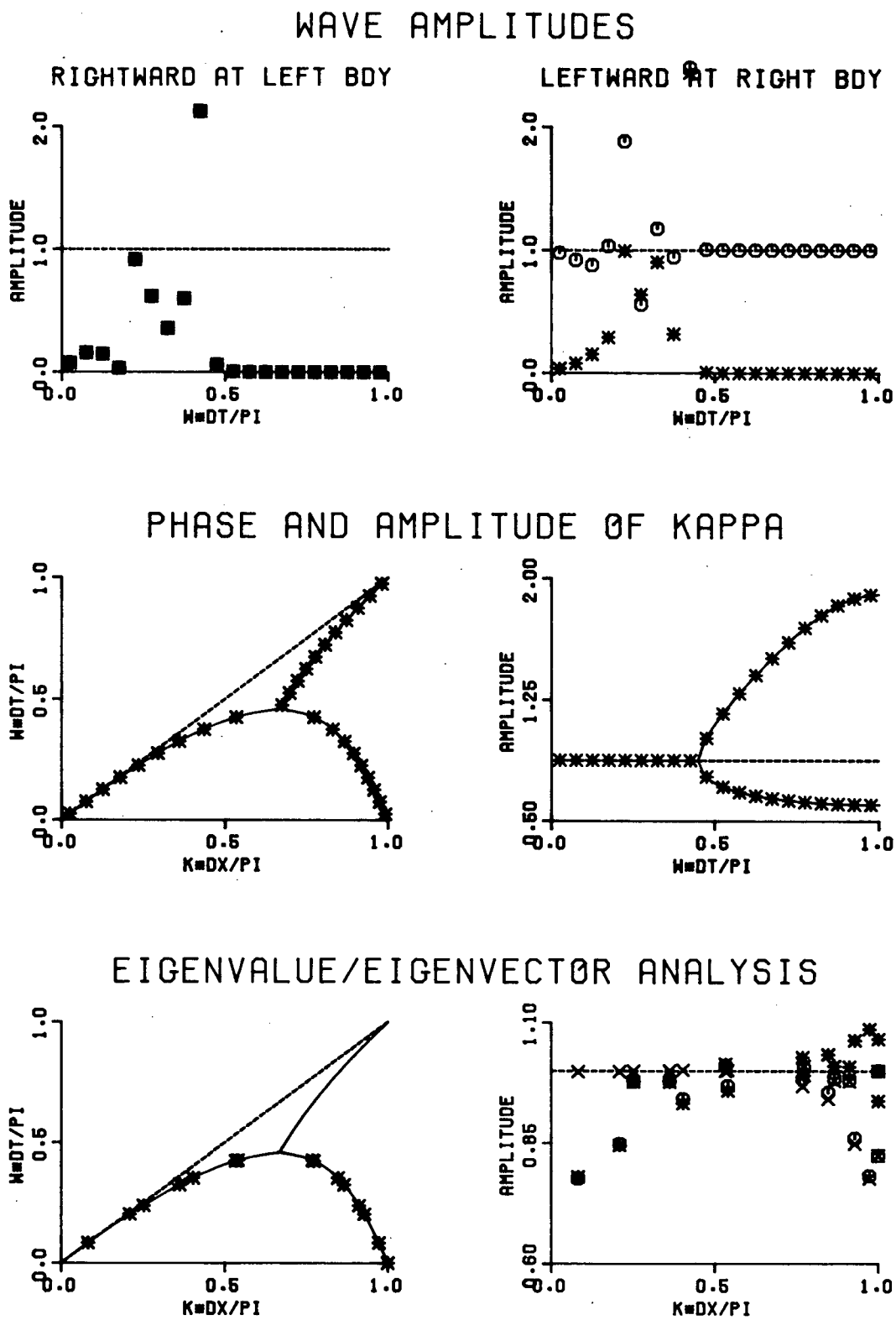

Fig. 5.12 Dispersion analysis for the GFEM with $f_1 = 0$, $f_2 = 1.$, $N = 10$; physical boundary conditions (5.5.3a), (5.5.3b); and artificial conditions (5.5.13).

All eigenvalues are inside the unit circle so this pair of boundary conditions is stable. The lower left, upper left, and two middle diagrams are quite similar to those of Fig. 5.11.

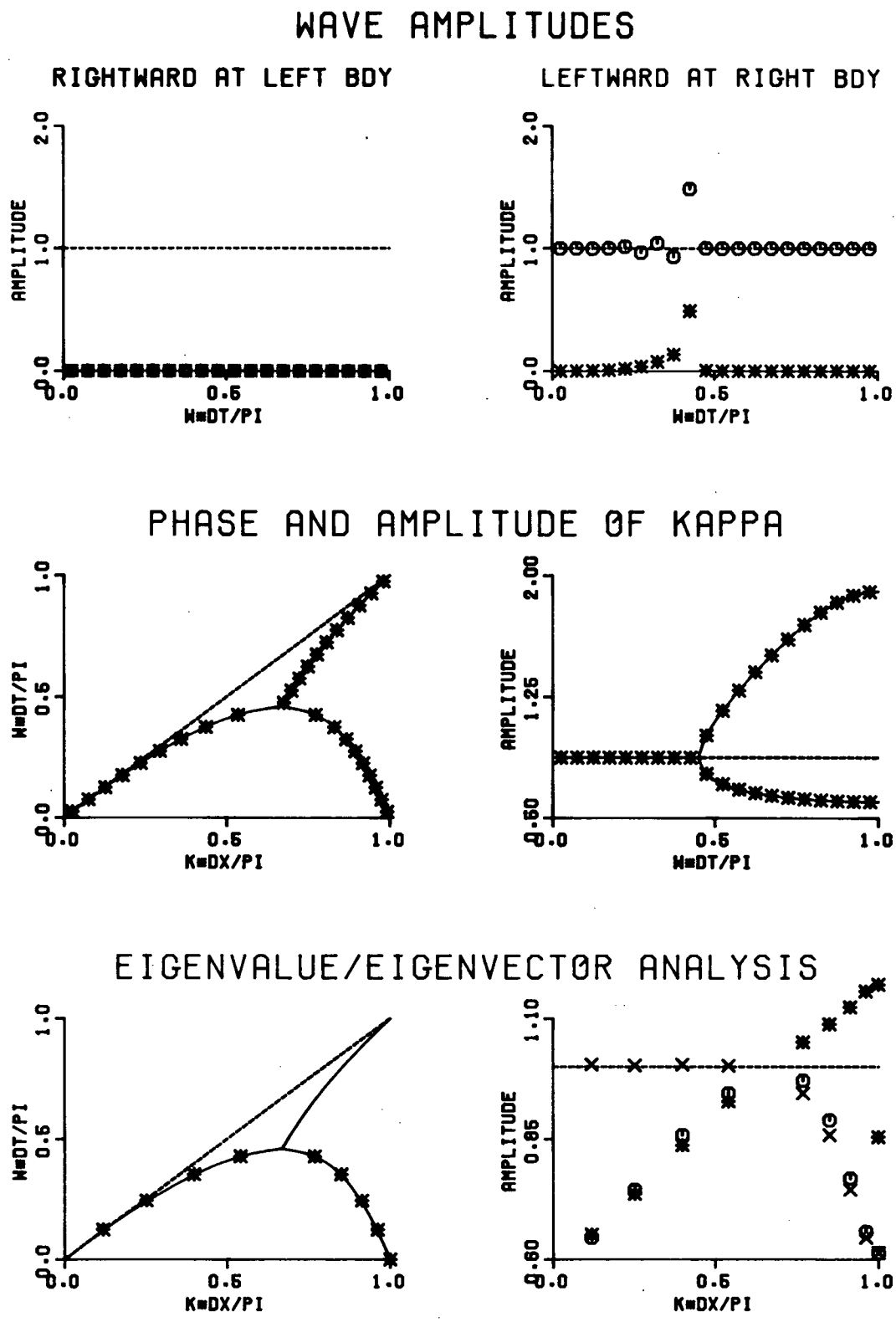

Fig. 5.13 Dispersion analysis for the GFEM with $f_1 = 0$, $f_2 = 1.$, $N = 10$; physical boundary conditions (5.5.3a), (5.5.3b); and artificial conditions (5.5.14).

All comments arising from those earlier diagrams can be repeated here. The upper right diagram, however, is significantly different. Long wave amplitudes are closer to 1.0, and

short wave amplitudes are closer to zero. Since less energy is assigned to short waves at the right boundary, we conclude that artificial condition (5.5.14b) is more accurate than (5.5.3d).

The previous three sets of artificial boundary conditions are simple mathematical extrapolations. Intuitively, we would expect that conditions which retain some of the problem physics should be more accurate. The fourth set of artificial conditions tests this intuition by applying the Box scheme to the continuity equation. The *artificial* conditions are now

$$(z_1^{n+1} - z_1^n) + (z_2^{n+1} - z_2^n) + \frac{h\Delta t}{\Delta x}\left[(u_2^n - u_1^n) + (u_2^{n+1} - u_1^{n+1})\right] = 0 \quad (5.5.15a)$$

$$(z_N^{n+1} - z_N^n) + (z_{N-1}^{n+1} - z_{N-1}^n) + \frac{h\Delta t}{\Delta x}\left[(u_N^n - u_{N-1}^n) + (u_N^{n+1} - u_{N-1}^{n+1})\right] = 0. \quad (5.5.15b)$$

Fig. 5.14 shows the analysis results. They are very similar to those of Fig. 5.13. The method is stable, and has an accurate right boundary condition. In fact, for some values of $\omega\Delta t < \omega_c$, condition (5.5.15b) is slightly more accurate than (5.5.14b). Waves at the left boundary are now seen to be greater than zero for some $\omega\Delta t$. This implies that condition (5.5.15a) causes some reflection at the left boundary and is less accurate than (5.5.3c) and (5.5.14a).

The final set of artificial boundary conditions arise from first order spatial extrapolation of the characteristic variables. The conditions are

$$-\left(\frac{g}{h}\right)^{1/2}(z_1^n - 2z_2^n + z_3^n) + u_1^n - 2u_2^n + u_3^n = 0 \quad (5.5.16a)$$

$$\left(\frac{g}{h}\right)^{1/2}(z_N^n - 2z_{N-1}^n + z_{N-2}^n) + u_N^n - 2u_{N-1}^n + u_{N-2}^n = 0. \quad (5.5.16b)$$

The analysis results, as shown in Fig. 5.15, are seen to be quite similar to those in Fig. 5.11. In fact, it is instructive to observe if higher order spatial extrapolation has increased the boundary condition accuracy. The diagram in the top right shows that it has. The longer leftward wave has an amplitude closer to 1.0 and the shorter leftward wave has amplitude closer to zero. However, (5.5.16b) is not as accurate as (5.5.14b) or (5.5.15b). The method is again stable.

The preceding analysis results were partially confirmed with numerical tests similar to those for the RS scheme. Each of the five pairs of artificial boundary conditions were used

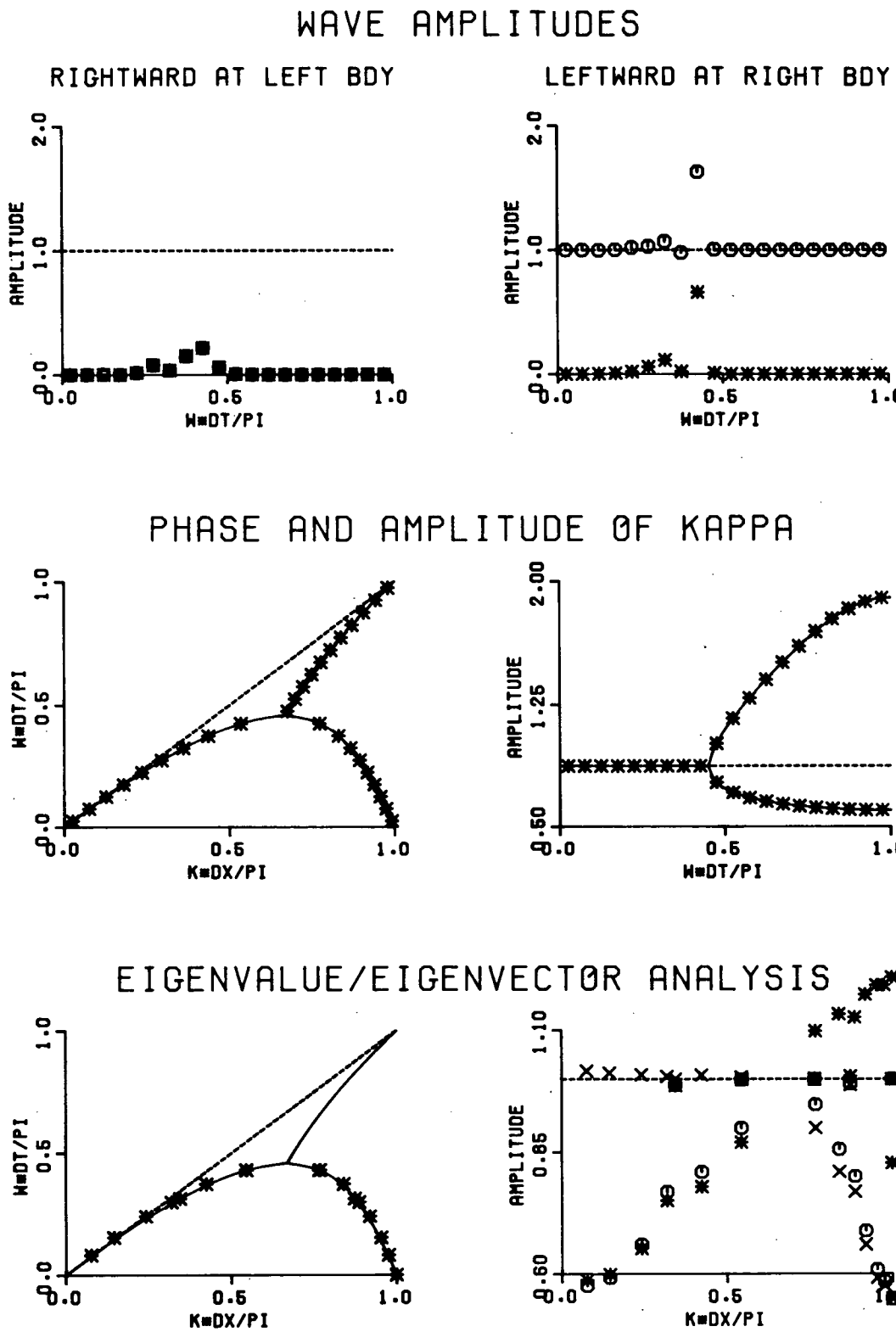

Fig. 5.14 Dispersion analysis for the GFEM with $f_1 = 0$, $f_2 = 1.$, $N = 10$; physical boundary conditions (5.5.3a), (5.5.3b); and artificial conditions (5.5.15).

in conjunction with (5.5.1), (5.5.3a), and (5.5.3b). The driving frequency $\omega\Delta t = .70685835$ was assumed. Parameter values for the model runs were identical to those in the analysis.

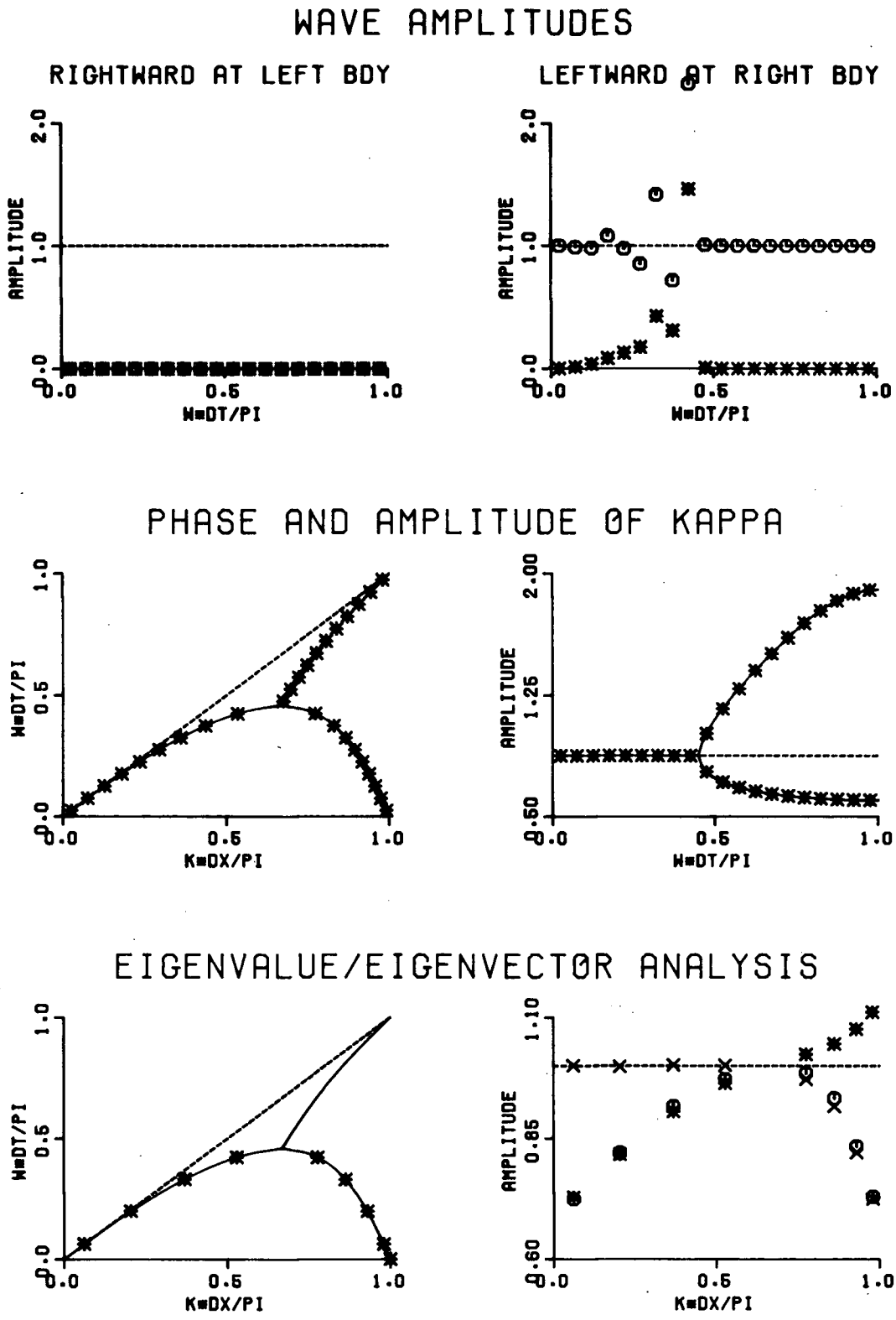

Fig. 5.15 Dispersion analysis for the GFEM with $f_1 = 0$, $f_2 = 1.$, $N = 10$; physical boundary conditions (5.5.3a), (5.5.3b); and artificial conditions (5.5.16).

Each test was run for 270 time steps. Least squares analyses over the successive time step ranges [135,167], [168,201], [202,235], [236,270] were used to calculate coefficients of

the leftward and rightward waves. As predicted by the analysis, the artificial conditions (5.5.13) caused a slow unstable growth. This growth is evident through residual errors which increased with successive fits. In all other cases, residuals decreased with each successive fit, and the fitted coefficients seemed to be converging. The coefficients ς_1, ς_2, ς_3, ς_4 obtained from the fourth fit were identical, to at least 2 digits, with the analysis results. Had the model calculations been done in double precision, as they were with the RS tests, analysis and model results would have been closer.

The relative accuracy of the preceding five pairs of artificial boundary conditions can be summarized as follows. For the absorbing left boundary, (5.5.3c), (5.5.14a) and (5.5.16a) were all very accurate when combined with (5.5.3a). (5.5.15a) became considerably less accurate as $\omega\Delta t$ increased and (5.5.13a) contributed toward model instability. At the absorbing-driving right boundary, conditions (5.5.14b) and (5.5.15b) were the most accurate when combined with (5.5.3b). There was little difference between the two. Listed in terms of decreasing accuracy, the other three conditions are (5.5.16b), (5.5.3d) and (5.5.13b). Again, the last of these contributed toward model instability.

Since no attempt has been made to generalize these accuracy results, they may be problem and parameter specific. As with the RS scheme, analyses with $f_1 > 0$, and with a closed left boundary should also be considered. However, they will not be done here.

5.6 Summary

This chapter has demonstrated that dispersion analyses can be extended to include boundary conditions. It has been shown that the accuracy of boundary conditions can be examined through physical concepts such as wave amplitude profiles and reflection coefficients. By applying Trefethen's [Tr83] theory, it has also been shown that a dispersion analysis can be used to indicate the GKS instability of an IBV problem.

Here is a summary of the highlights of this chapter.

In Section 5.1, the importance of boundary conditions was discussed and previous work was reviewed. In particular, the recent work of Beam, Warming, and Yee [Be82] was summarized because of its close relationship to the analysis of two-step methods in

Chapter 2.

In Section 5.2, the one dimensional channel problem was defined mathematically and the boundary conditions were shown to be well-posed. It was also shown that with $\tau = 0$, our radiation conditions are identical to the absorbing boundary conditions proposed by Engquist and Majda [En77].

In Section 5.3, the boundary condition analysis was developed for the Richardson-Sielecki FDM. Time-space staggering of the z and u variables means that several implementations of the radiating conditions are possible. Four implementations were analysed. Listed in terms of decreasing accuracy, they are:

i) first order space-time extrapolation,

ii) linear spatial extrapolation with phase velocity,

iii) zeroth order space-time extrapolation,

iv) linear time extrapolation followed by linear spatial extrapolation.

The last implementation causes instability. Implementations for nonzero τ were also discussed, and the analysis results were partially confirmed with numerical tests.

In Section 5.4, the GKS stability of a Galerkin FEM with selected boundary conditions was analysed. Trefethen's interpretation of GKS theory was also discussed and illustrated with a pair of unstable boundary conditions.

In Section 5.5, the dispersion analysis was applied to a Galerkin FEM. Five pairs of artificial boundary conditions were examined in combination with physical conditions that approximate an absorbing left boundary, and a driving-absorbing right boundary. Listed in terms of decreasing accuracy, the best conditions for the right boundary are:

i) zeroth order space-time extrapolation of the outgoing characteristic variables,

ii) the Box scheme applied to the continuity equation,

iii) first order spatial extrapolation of the outgoing characteristic variables,

iv) zeroth order spatial extrapolation of the outgoing characteristic variables,

v) zeroth order spatial extrapolation of z.

Actually, there was little difference between implementations i) and ii). Implementation

v) caused instability. At the left boundary, implementations i), iii), and iv) were all very accurate whereas implementation ii) became considerably less accurate as $\omega\Delta t$ increased. Implementation v) again caused instability.

BIBLIOGRAPHY

[Ab65] M. ABRAMOWITZ AND I.A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[Ad74] R.A. ADEY, *Numerical Prediction of Transient Water Quality and Tidal Motion in Estuaries and Coastal Waters*, Ph.D. Thesis, University of Southampton.

[Be79] R. BEAM AND R.F. WARMING, in *Proceedings, AIAA 4th Computational Fluids Dynamics Conference, Williamsburg, Virginia, July 1979*, Paper No. 79-1466.

[Be82] R.M. BEAM, R.F. WARMING, AND H.C. YEE, *Stability Analysis of Boundary Conditions and Implicit Difference Approximations for Hyperbolic Equations*, J. Comput. Phys. 48 (1982), 200-222.

[Br76] C.A. BREBBIA AND P.W. PARTRIDGE, *Finite element simulation of water simulation in the North Sea*, Appl. Math. Modelling, 1 (1976), 101-107.

[Br53] L. BRILLOUIN, *Wave Propagation in Periodic Structures*, Dover, 1953.

[Br60] L. BRILLOUIN, *Wave Propagation and Group Velocity*, Academic Press, New York, 1960.

[Ch75] R.C.Y. CHIN, *Dispersion and Gibbs Phenomenon Associated with Difference Approximations to Initial Boundary-Value Problems for Hyperbolic Equations*, J. Comput. Phys. 18 (1975), 233-247.

[Ch78] R.C.Y. CHIN AND G.W. HEDSTROM, *A Dispersion Analysis for Difference Schemes: Tables of Generalized Airy Functions*, Math. Comp. 32 (1978), 1163-1170.

[Ch79] R.C.Y. CHIN, G.W. HEDSTROM, AND K.E. KARLSSON, *A Simplified Galerkin Method for Hyperbolic Equations*, Math. Comp. 33 (1979), 647-658.

[Co74] J.J. CONNOR AND J.D. WANG, *Finite element modelling of hydrodynamic circulation*, Numerical Methods in Fluid Dynamics, ed. C.A. Brebbia and J.J. Connor, Pentech Press, London, 1974.

[Cr76] P.B. CREAN, *Numerical Model Studies of Tides Between Vancouver Island and the Mainland Coast*, J. Fish. Res. Board Can. 33 (1976), 2340-2344.

[Cu76] M.J.P. CULLEN, *The Application of Finite Element Methods to the Primitive Equations of Fluid Motion*, Finite Elements in Water Resources, ed. W.G. Gray et al., Pentech Press, London, 1976.

[Cu82] M.J.P. CULLEN, *The Use Of Quadratic Finite Element Methods and irregular Grids in the Solution of Hyperbolic Problems*, J. Comput. Phys. 45 (1982), 221-245.

[Da63] G. DAHLQUIST, *A Special Stability Problem for Linear Multistep Methods*, BIT 3 (1963), 27-43.

[En77] B. ENGQUIST AND A. MAJDA, *Absorbing Boundary Conditions for the Numerical Simulation of Waves*, Math. Comp. 31 (1977), 629-651.

[Fl76] R.A. FLATHER, *A tidal model of the north west European continental shelf*, Mem. Soc. R. Sci. Liege, Ser. 6, 10 (1976), 141-164.

[Fo83] M.G.G. FOREMAN, *An Analysis of Two-Step Time Discretizations in the Solution of the Linearized Shallow Water Equations*, J. Comput. Phys. 51 (1983), 454-483.

[Fo83b] M.G.G. FOREMAN, *An Analysis of the "Wave Equation" Model for Finite Element Tidal Computations*, J. Comput. Phys. 52 (1983), 290-312.

[Fo84] M.G.G. FOREMAN, *A Two Dimensional Dispersion Analysis of Selected Methods for Solving the Linearized Shallow Water Equations*, J. Comput. Phys., to appear.

[Ge71] C.W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice Hall, Englewood Cliffs, 1971.

[Gr77] W.G. GRAY, *An efficient finite element scheme for two dimensional surface water computation*, Finite Elements in Water Resources, ed. W.G. Gray et al., Pentech Press, London, 1974.

[Gr77b] W.G. GRAY AND D.R. LYNCH, *Time-Stepping Schemes for Finite Element Tidal Model Computations*, Adv. Water Resources 1 (1977), 83-95.

[Gr78] W.G. GRAY AND M. TH. VAN GENUCHTEN, *Economical Alternatives to Gaussian Quadrature over Isoparameteric Quadrilaterals*, Int. J. Numer. Meth. Engrg. 12 (1978), 1478-1484.

[Gr79] W.G. GRAY AND D.R. LYNCH, *On the Control of Noise in Finite Element Tidal Computations: A Semi-Implicit Approach*, Computers and Fluids 7 (1979), 47-67.

[Gr76] D.A. GREENBERG, *Mathematical description of the Bay of Fundy-Gulf of Maine numerical model*, Tech. note 16, Marine Env. Data Service, Environment Canada, Ottawa, 1976.

[Gr72] G. GROTKOP, *Die Berechnung von Flachwasserwellen nach der Methode der finiten Elemente*, Ph. D. Thesis, Jahresbericht 1971 des Sonderforschungsbereiches 79 der Techn. Univ. Hannover 2, 1972.

[Gr73] G. GROTKOP, *Finite element analysis of long period water waves*, Comp. Meth. in Appl. Mech. and Engng. 2 (1973), 147-157.

[Gu72] B. GUSTAFSSON, H. KREISS, AND A. SUNDSTRÖM, *Stability Theory of Difference Approximations for Mixed Initial Boundary Value Problems. II*, Math. Comp. 26 (1972), 649-686.

[Gu75] B. GUSTAFSSON, *The Convergence Rate for Difference Approximations to Mixed Initial Boundary Value Problems*, Math. Comp. 29 (1975), 396-406.

[Gu79] B. GUSTAFSSON AND H. KREISS, *Boundary Conditions for Time Dependent Problems with an Artificial Boundary*, J. Comput. Phys. 30 (1979), 333-351.

[Gu80] B. GUSTAFSSON AND J. OLIGER, *Stable Approximations for a Class of Time Discretizations of* $u_t = AD_0u$, Report No. 87, Uppsala University, Dept. Computer Sciences, 1980.

[Gu82] B. GUSTAFSSON, *The Choice of Numerical Boundary Conditions for Hyperbolic Systems*, J. Comput. Phys. 48 (1982), 270-283.

[Ha80] G.J. HALTINER AND R.T. WILLIAMS, *Numerical Prediction and Dynamic Meteorology, Second Edition* , Wiley and Sons, New York, 1980.

[Ha62] W. HANSEN, *Hydrodynamical Methods Applied to Oceanographic Problems*, Pro. Symp. Math. Hydrodynamical Methods of Phys. Oceanography, Institut fur Meereskunde der Universitat Hamburg, (1961), 25-34.

[Ha66] W. HANSEN, *The Reproduction of the Sea by means of Hydrodynamical Numerical Methods*, Mitt. Inst. Meereskunde Universitat Hamburg (1966), No. 5.

[He69] N.S. HEAPS, *A Two-Dimensional Numerical Sea Model*, Phil. Trans. R. Soc. London A 265 (1969), 93-137.

[He75] G. HEDSTROM, *Models of Difference Schemes for* $u_t + u_x = 0$ *by Partial Differential Equations*, Math. Comp. 29 (1975), 969-977.

[He76] R.F. HENRY AND N.S. HEAPS, *Storm Surges in the Southern Beaufort Sea*, J. Fish. Res. Board Can. 33 (1976), 2362-2376.

[He78] R.F. HENRY, *Computation of Shallow Water Waves by the Method of Finite Elements* (a translation of [Gr72]), IOS Note 8, Institute of Ocean Sciences, Patricia Bay, Sidney B.C., 1978.

[He81] R.F. HENRY, *Richardson-Sielecki Schemes for the Shallow-Water Equations, with Applications to Kelvin Waves*, J. Comput. Phys. 41 (1981), 389-406.

[Hi82] D. E. HINSMAN, R. T. WILLIAMS, AND E. WOODWARD, *Recent Advances in the Galerkin Finite Element Method as Applied to the Meteorological Equations on Variable Resolution Grids*, Finite Element Flow Analysis (ed. Tadahiko Kawai), University of Tokyo Press, Tokyo, 1982.

[Ho74] P. HOOD AND C. TAYLOR, *Navier Stokes Equations Using Mixed Interpolation*, Finite Element Methods in Fluid Problems, ed. J.T. Oden et al., UAH Press, Huntsville, 1974.

[Ja80] B.M. JAMART AND D.F. WINTER, *Finite element solution of the shallow water wave equations in Fourier space with application to Knight Inlet, British Columbia*, Proceedings of the Third International Conference on Finite Elements in Flow Problems, Vol. 2, University of Calgary, 1980, 103-112.

[Ka78] M. KAWAHARA AND K. HASEGAWA, *Periodic Galerkin Finite Element Method of Tidal Flow*, Int. J. Num. Meth. Engng. 2 (1978), 115-127.

[Ka78b] M. KAWAHARA, N. TAKEUCHI, AND J. YOSHIDA, *Two step explicit finite element method for tsunami wave propagation analysis*, Int. J. Num. Meth. Engng. 12 (1978) 331-351.

[Ka80] M. KAWAHARA, S. NAKAZAWA, S. OHMORI, AND T. TAGAKI, *Two step explicit finite element method for storm surge propagation analysis*, Int. J. Num. Meth. Engng. 15 (1980), 1129-1148.

[Ki75] I.P. KING, W.R. NORTON, AND K.R. ICEMAN, *A finite element solution for two-dimensional stratified flow problems*, Finite Elements in Fluids, ed. R.H. Gallagher et al., Wiley, London, 1975.

[Kr73] H. KREISS AND J. OLIGER, *Methods for the Approximate Solution of Time Dependent Problems*, WMO-ICSU Joint Organizing Committee, GARP Publication Series No. 10, 1973.

[La32] H. LAMB, *Hydrodynamics*, Dover, New York, 1932(sixth ed.).

[La73] J.D. LAMBERT, *Computational Methods in Ordinary Differential Equations*, Wiley, London, 1973.

[La71] L. LAPIDUS AND J.H. SEINFELD, *Numerical Solution of Ordinary Differential Equations*, Academic Press, New York, 1971.

[La82] L. LAPIDUS AND G.F. PINDER, *Numerical Solution of Partial Differential Equations in Science and Engineering*, Wiley, New York, 1982.

[Le78] P.H. LeBLOND AND L.A. MYSAK, *Waves in the Ocean*, Elsevier, Amsterdam, 1978.

[Le67] J.J. LEENDERTSE, *Aspects of a Computational Model for Long-Period Water-Wave Propagation*, Rand Memorandum RM-5294-PR, 1967.

[Le81] C. LE PROVOST, G. ROUGIER, AND A. PONCET, *Numerical Modeling of the Harmonic Constituents of the Tides, with Application to the English Channel*, J. Phys. Oceanogr. 11 (1981), 1123-1138.

[Li78] J. LIGHTHILL, *Waves in Fluids*, Cambridge University Press, Cambridge, 1978.

[Ly78] D.R. LYNCH, *Finite Element Solution of the Shallow Water Equations*, Ph. D. Thesis, Dept. of Civil Engineering, Princeton University, 1978.

[Ly78b] D.R. LYNCH AND W.G. GRAY, *Analytic Solutions for Computer Flow Model Testing*, J. Hydraulics Division ASCE 104(H 10) (1978), 1409-1428.

[Ly79] D.R. LYNCH AND W.G. GRAY, *A Wave Equation Model for Finite Element Tidal Computations*, Computers and Fluids 7 (1979), 207-228.

[Ly80] D.R. LYNCH AND W.G. GRAY, *On the Analysis of Accuracy for Two-Equation Transient Problems*, Int. J. Num. Meth. Engng. 15 (1980), 55-62.

[Me76] F. MESINGER AND A. ARAKAWA, *Numerical Methods Used in Atmospheric Models, Vol. 1*, WMO-ICSU Joint Organizing Committee, GARP Publication Series No. 17, 1976.

[Mu82] R. MULLEN AND T. BELYTSCHKO, *Dispersion Analysis of Finite Element Semidiscretizations of the Two-Dimensional Wave Equation*, Int. J. Numer. Methods Engrg. 18 (1982), 11-29.

[Mu77] T.S. MURTY, *Seismic Sea Waves - Tsunamis*, Department of Fisheries and the Environment, Ottawa, 1977.

[Na79] I.M. NAVON, *Numerical methods for the solution of the shallow-water equations in meteorology*, CSIR Special Report SWISK 10, National Research Institute for Mathematical Sciences, Pretoria, 1979.

[No73] W.R. NORTON, I.P. KING, AND J.T. ORLOB, *A Finite Element Model for Lower Granite Resevoir*, Water Resources Engineers Inc. 1973, Walnut Creek Ca., Prepared for Walla Walla District, U.S. Army Corps of Engineers.

[Pa76] P.W. PARTRIDGE AND C.A. BREBBIA, *Quadratic finite elements in shallow water problems*, J. Hydraulics Division, ASCE 102 (1976), 1299-1313.

[Pe77] C.E. PEARSON AND D.F. WINTER, *On the calculation of tidal currents in homogeneous estuaries*, J. Phys. Oceanogr. 7 (1977), 520-531.

[Pi77] G.F. PINDER AND W.G. GRAY, *Finite Element Simulation in Surface and Subsurface Hydrology*, Academic Press, New York, 1977.

[Pl63] G.W. PLATZMAN, *The Dynamical Prediction of Wind Tides on Lake Erie*, Meteor. Monogr. 4, No. 26 (1963).

[Pl81] G.W. PLATZMAN, *Some Response Characteristics of Finite Element Tidal Models*, J. Comput. Phys. 40 (1981), 36-63.

[Po78] S. POND AND G.L. PICKARD, *Introductory Dynamic Oceanography*, Pergamon Press, Oxford, 1978.

[Pr79] N. PRAAGMAN, *Numerical Solution of the Shallow Water Equations by a Finite Element Method*, Ph. D. Thesis, Delft University of Technology, 1979.

[Pu76] N.J. PULLMAN, *Matrix Theory and Its Applications, Selected Topics*, Marcel Dekker Inc., New York, 1976.

[Ri67] R.D. RICHTMYER AND K.W. MORTON, *Difference Methods for Initial-Value Problems*, Wiley-Interscience, New York, 1967.

[Sc80] A.L. SCHOENSTADT, *A Transfer Function Analysis of Numerical Schemes Used to Simulate Geostrophic Adjustment*, Mon. Weather Rev. 108 (1980), 1248-1259.

[Se65] S. M. SELBY AND B GIRLING, *Standard Mathematical Tables*, The Chemical Rubber Company, Cleveland, 1965.

[Si68] A. SIELECKI, *An Energy-Conserving Difference Scheme for the Storm Surge Equations*, Mon. Weather Rev. 96 (1968), 150-156.

[Sk75] G. SKÖLLERMO, *How the boundary conditions affect the stability and accuracy of some implicit methods for hyperbolic equations*, Report No. 62, Uppsala University, Dept. Computer Sciences, 1975.

[Sk79] G. SKÖLLERMO, *Error Analysis of Finite Difference Schemes Applied to Hyperbolic Initial Boundary Value Problems*, Math. Comp. 33 (1979), 11-35.

[St57] J.J. STOKER, *Water Waves*, Interscience Publishers, New York, 1957.

[St73] G. STRANG AND G.J. FIX, *An Analysis of the Finite Element Method*, Prentice Hall, Englewood Cliffs N.J., 1973.

[Su79] A. SUNDSTRÖM AND T. ELVIUS, *Computational Problems Related to Limited-Area Modelling*, Numerical Methods Used in Atmospheric Models II, GARP Publication Series No. 17, 379-416, 1979.

[Ta75] C. TAYLOR AND J. DAVIS, *Tidal and Long Wave Propagation—A Finite Element Approach*, Computers and Fluids 3 (1975), 125-148.

[Th77] W.C. THACKER, *Irregular Grid Finite-Difference Techniques: Simulations of Oscillations in Shallow Circular Basins*, J. Phys. Oceanog. 7 (1977), 284-292.

[Th78] W.C. THACKER, *Comparison of Finite-Element and Finite Difference Schemes. Part I: One-Dimensional Gravity Wave Motion*, J. Phys. Oceanog. 8 (1978), 676-679.

[Th78b] W.C. THACKER, *Comparison of Finite-Element and Finite Difference Schemes. Part II: Two-Dimensional Gravity Wave Motion*, J. Phys. Oceanog. 8 (1978), 680-689.

[Tr82] L.N. TREFETHEN, *Group Velocity in Finite Difference Schemes*, SIAM Rev. 24 (1982), 113-136.

[Tr82b] L.N. TREFETHEN, *Wave Propagation and Stability for Finite Difference Schemes*, Ph. D. Thesis, Dept. of Computer Sci., Stanford University, 1982.

[Tr83] L.N. TREFETHEN, *Group Velocity Interpretation of the Stability Theory of Gustafsson, Kreiss, and Sundström*, J. Comput. Phys. 49 (1983), 199-217.

[Vi75] R. VICHNEVETSKY AND B. PEIFFER, *Error waves in finite element and finite difference methods for hyperbolic equations*, Advances in Computer Methods for Partial Differential Equations (R. Vichnevetsky, ed.), Assoc. Int. Calcul. Analogique, Ghent, Belgium, 1975.

[Vi80] R. VICHNEVETSKY, *Propagation Properties of Semi-Discretizations of Hyperbolic Equations*, Math. Comput. Simulation 22 (1980), 98-102.

[Vi82] R. VICHNEVETSKY AND J. BOWLES, *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, SIAM, Philadelphia, 1982.

[Wa79] R.A. WALTERS AND R.T. CHENG, *A two-dimensional hydrodynamic model of a tidal estuary*, Adv. Water Resources 2 (1979), 177-184.

[Wa80] R.A. WALTERS AND R.T. CHENG, *Accuracy of an Estuarine Hydrodynamic Model Using Smooth Elements*, Water Resources Research 16 (1980), 187-195.

[Wa83] R.A. WALTERS AND G.F. CAREY, *Analysis of Spurious Oscillation Modes for the Shallow Water and Navier-Stokes Equations*, Computers and Fluids 11 (1983), 51-68.

[Wa83b] R.A. WALTERS, *Numerically Induced Oscillations in Finite Element Approximations to the Shallow Water Equations*, Int. J. Num. Meth. Fluids 3 (1983), 591-604.

[Wa75] J.D. WANG AND J.J. CONNOR, *Mathematical Modelling of Near Coastal Circulation*, MIT Parsons Laboratory Report No. 200, 1975.

[Wa77] J.D. WANG, *Comments on "Irregular Grid Finite-Difference Techniques: Simulations of Oscillations in Shallow Circular Basins"*, J. Phys. Oceanog. 7 (1977), 932-933.

[Wa74] R. WARMING AND B. HYETT, *The modified equation approach to the stability and accuracy analysis of finite-difference models*, J. Comput. Phys. 14 (1974), 159 179.

[We76] T.J. WEARE, *Finite Element or Finite Difference Methods for the Two-Dimensional Shallow Water Equations*, Computer Methods Appl. Mech. Eng. 7 (1976), 351-357.

[Wh74] G.B. WHITHAM, *Linear and Nonlinear Waves*, Wiley-Interscience, New York, 1974.

[Wi81] R.T. WILLIAMS, *On the Formulation of Finite-Element Prediction Models*, Mon. Weather Rev. 109 (1981), 463-466.

[Wi81b] R.T. WILLIAMS AND O.C. ZIENKIEWICZ, *Improved Finite Element Forms for the Shallow-Water Wave Equations*, Int. J. Numer. Methods Fluids 1 (1981), 81-97.

[Zi77] O.C. ZIENKIEWICZ, *The Finite Element Method (3rd edition)*, McGraw-Hill, London, 1977.