JOURNAL OF COMPUTATIONAL PHYSICS 122, 129–142 (1995)

Discretization of a Model for the Formation of Longshore Sand Ridges

JUAN MARIO RESTREPO* AND JERRY L. BONA†

*Mathematics and Computer Science Division, Argonne National Laboratory, Argonne Illinois 60439; and †Mathematics Department, The Pennsylvania State University, University Park, Pennsylvania 16802

Received December 20, 1993; revised November 29, 1994

This paper presents and evaluates the numerical solution of a coupled system of equations that arises in a model for the formation and evolution of three-dimensional longshore sand ridges. The model is based on the interaction between surficial or internal weakly nonlinear shallow-water waves, having weak spanwise spatial dependence, and the deformable bottom topography. The presentation of the details concerning the discretization of the model is primarily motivated by: (1) the model involves equations for which little is known regarding its solutions; (2) we believe that the methodology used in simplifying the solution to the coupled sand ridge model may be of interest to other researchers in the geophysical community; and (3) the predictor-corrector scheme presented here, which combines finite difference techniques and fixed-point methods, is simple, fast, and general enough to be used in the discretization of other partial differential equations with local nonlinearities whose solutions are smooth and bounded.

1. INTRODUCTION

The dynamics of sand ridges are not well understood. Sand ridges are underwater barlike features of the continental shelf, composed of loose granular sediment. Hundreds of meters long and up to a few meters high, sand ridges are usually found in groups, arranged in more or less parallel rows separated from each other by hundreds of meters. They may be loosely classified as either tidal ridges or longshore sand ridges. Tidal ridges are oriented more or less parallel to the prevailing direction of the local ocean currents, whereas longshore sand ridges are oriented normal to the direction in which the overlying water waves propagate. In [1, 2], the authors proposed a model for the formation and evolution of three-dimensional longshore sand ridges on the continental shelf, based on the nonlinear interaction of long water waves and a movable bottom topography, a mechanism first proposed in [3].

Briefly, sand ridges may be formed by the highly organized second-order Stokes flow generated by the passage of the long gravity waves. We believe that the chief morphological characteristics of sand ridges, namely, their cross-sectional asymmetry, their spacing and height, as well as their orientation, cannot be fully accounted for by either the action of violent storms and/or linear gravity waves. We have proposed that the dispersive and weakly nonlinear effects of energetic long gravity waves traveling over a featured bottom will lead to the above-mentioned features. We also conjecture that the bottom topography, which influences the propagation of the overlying water waves, evolves in time frames that are significantly larger than those of the water waves. The interactive nature of the waves and the sediment-laden boundary layer immediately above the movable bottom topography leads to a coupled nonlinear evolutionary model.

In scaled variables a model for nonlinear, dispersive shallow-water waves of amplitude $z = \eta(x, y, t)$ and tranverse velocity $\mathbf{u}(x, y, t)$, traveling over a bottom topography $z = -h(x, y, T) = O(1)$ is the Boussinesq system [4, 1]

$$\eta_t + \nabla \cdot [(h + \alpha\eta)\mathbf{u}] - \frac{1}{3}\beta^2 \nabla \cdot [\nabla(h^2\eta_t)] = 0 \qquad (1)$$

$$\mathbf{u}_t + \alpha(\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla\eta = 0, \qquad (2)$$

where coordinate x increases in the shoreward direction, y is the spanwise direction, time is indicated by t, and α and β are real parameters described below. The evolution of the bottom topography is assumed to occur in time scales typified by $T \gg t$, and to be approximately described by the mass transport equation

$$\frac{\partial h(x, y, T)}{\partial T} = \frac{K}{\rho_0}\left(\frac{\partial \mu}{\partial x} + \frac{\partial \nu}{\partial y}\right), \qquad (3)$$

with K/ρ_0 a constant of proportionality; μ and ν are the shoreward and spanwise mass fluxes, respectively, defined as

$$\mu \equiv \int_0^{\delta_{bl}} \rho(x, z')\mathcal{U}(x, z'))\, dz'$$

$$\nu \equiv \int_0^{\delta_{bl}} \rho(x, z')\mathcal{V}(x, z'))\, dz',$$

[1] Preprint MCS-P408-1293, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1993.

0021-9991/95 $12.00

where the integrations over the bottom-following vertical coordinate z' span the thickness of the boundary layer δ_{bl}; ρ is the suspended sediment concentration, $\mathcal{U}$ and $\mathcal{V}$ are the x and y components of the Stokes drift generated by the surface water waves. For simplicity, the lossy mechanisms of bottom drag and kinetic dissipation in the surface waves and the bottom topography respectively have been omitted.

The solution and analysis of the above system are quite challenging. In order to make the analysis of the model more tractable, we opted to make the following simplification: assume that the surface waves are composed solely of an incident wave field. Since most of the energy of these weakly nonlinear waves is in the neighborhood of a few distinct peaks in the spectrum, a WKB approximation for the wave packet solution of the Boussinesq system is

$$u(x, X, y, t) = \sum_{j=1}^{2} [a_j(X, y) + O(\alpha)]e^{i(k_j x - \omega_j t)} + \text{c.c.}, \tag{4}$$

where c.c. stands for complex conjugate of the expression immediately preceding its appearance. The a's are the complex amplitudes of wave packets centered at wavenumber k_j and with support $\Delta k_j < k_j$. The real parameter $\alpha \ll 1$ characterizes the degree of nonlinearity of the waves. The relation between the frequency ω_j and the wavenumber is

$$\omega_j^2 - \frac{k_j^2}{1 + \beta^2(k_j^2/3)} = 0, \tag{5}$$

where the real constant $\beta \ll 1$ is a dimensionless parameter related to the degree of dispersiveness in the water waves. Similar wave packet expressions are assumed for the spanwise velocity and the amplitude of the water wave. Let $k_2 = 2k_1 - \delta$, where $\delta \leq 0$ is the detuning parameter. A compatibility condition results when the Boussinesq system is expanded in terms of a series in α using the wave packet representation. This compatibility condition leads us to the equations that control the spatial variation of the wave packets in the domain of interest. Further, we assume that the waves have only weak y-dependence, hence a parabolic approximation is adopted [1, 5].

The following system of equations constitute a crude but very useful version of the model for the formation and evolution of three-dimensional sand ridges on the continental shelf,

$$\begin{aligned} a_{1X} - iK_1 a_{1yy} + iK_3 f(x,y,T)a_1 + iK_5 e^{-i\Delta X} a_1^* a_2 &= 0 \\ a_{2X} - iK_2 a_{2yy} + iK_4 f(x,y,T)a_2 + iK_6 e^{+i\Delta X} a_1^2 &= 0 \\ a_1(X = 0, y) &= \mathcal{A}_1(y, T) \\ a_2(X = 0, y) &= \mathcal{A}_2(y, T), \end{aligned} \tag{6}$$

and

$$\begin{aligned} h_T &= \frac{K}{\rho_0}[\mu_X(h, a_1, a_2) + \nu_y(h, a_1, a_2)] \\ h(X, y, 0) &= \mathcal{H}(X, y), \end{aligned} \tag{7}$$

with appropriate boundary conditions on $y = 0$ and $y = N$. The scaled detuning parameter appears in $\Delta = \delta/\alpha$. Also, note that $X = \alpha x$. The real constant coefficients K are given in the Appendix. The bottom is $h(X, y, T) = 1 + \varepsilon f(X, y, T)$, where $\varepsilon \ll 1$ typifies the size of the slopes of the bottom topography. Equation (6) describes the spatial structure of the complex amplitudes of the two most energetic wave packets of weakly nonlinear dispersive ocean waves traveling in the shoreward direction x over a deformable bottom topography. The bottom evolution is given by the mass transport relation, Eq. (7).

As mentioned above, the model assumes that the evolution of the bottom topography has characteristic time scales that are much longer than the time scales in which surface waves adjust to changes in the bottom topography: an incident wave field senses a bottom which is essentially fixed in time, from the moment it enters the purview of the model to the time it eventually leaves it. The bottom deforms slowly after the passage of many waves. Owing to the widely discrepant time scales between the (fast) evolution of the water waves and the (slow) bottom topography, this coupled system may be solved iteratively: Given an initial bottom configuration $\mathcal{H}(X, y)$, we seek a solution to the water waves using Eq. (6). The bottom is then updated using the mass transport equation, and the surface equations are solved using the new bottom configuration. The whole procedure is repeated until some prescribed final time T_f, say.

The input to the model is composed of an initial bottom configuration and the wave-packet amplitudes at the line $X = 0$. The required dynamic parameters are the fundamental frequency; an estimate of the size of the parameters $\alpha \ll 1$ and $\beta \ll 1$, and the dimensions of the rectangular patch, $0 \leq X \leq M$, $0 \leq y \leq N$, of ocean on which the solution is to be computed.

The similarity of Eq. (6) to the nonlinear Schrödinger equation (NLSE) leads us to guess that the numerical technique presented here applies to the NLSE in a straightforward manner. In fact, any equation or system with solutions of sufficient regularity with local nonlinearity may be solved by the method described below.

Several issues have motivated the particular choice of the scheme to be presented: (1) an efficient, simple, and sufficiently accurate method is desired to implement the above nonlinear system numerically; (2) the accuracy requirements are not very sophisticated since the main objective at present is the exploration of phenomenological questions; (3) a uniform grid is preferred over a variable one, so that both the surface and mass transport equations may be easily computed on the same grid;

and (4) the computational domain is fairly large for the sort of problem presented in this study, hence we would like to use a numerical method that is computationally efficient. The first level of simplification in the solution of the coupled system is the application of the decoupling algorithm described above. The next level is to adopt an efficient method for the solution of the surface and mass transport equations. The mass transport equation will be solved using standard methods. The coupled nonlinear surface system will be solved by combining finite difference techniques and fixed point methods. We refer to the numerical scheme adopted in this study as the fixed-point method (FPM). Among its best features are low storage requirements, high speed and simplicity.

The use of the FPM acronym does not imply that the method is completely novel. The use of fixed point techniques in the solution of hyperbolic systems is not as common as it is in the solution of other types of equations; an example of a successful application of fixed point methods to the solution of hyperbolic equations is Tappert's spectral split-step method for the solution of the Korteweg–de Vries equation and the NLSE [6]. We wish to show, using the sand ridge evolution model as an example, how finite difference and fixed point methods may be combined to solve nonlinear systems of equations with bounded solutions.

The following difference operators pertain to the discretization,

$$\begin{aligned} \Delta_q u &= u(q_{j-1}) - u(q_j) && \text{forward difference} \\ \nabla_q u &= u(q_j) - u(q_j - 1) && \text{backward difference} \\ \delta_q u &= u(q_{j-1/2}) - u(q_{j-1/2}) && \text{central difference} \\ A_q u &= u(q_{j-1}) + u(q_j) && \text{forward average} \end{aligned} \tag{8}$$

in the independent variable q, say. The physical space is given by $\mathbf{R}^2 \times T_f \equiv [0 \le X \le M,\ 0 \le y \le N] \times \{T_f \ge 0\}$. Define $\mathbf{R}^2_\Delta \times T_\Delta \equiv (X_r, y_s) \times T_n = (r\,\Delta x, s\,\Delta y) \times n\,\Delta T \in \mathbf{R}^2 \times T_f$. Furthermore, there are integers m and n, such that $M = m\,\Delta x$, $N = n\,\Delta y$.

2. DISCRETIZATION OF THE MASS TRANSPORT EQUATION

The mass transport equation is solved numerically by the well-known two-step Lax–Wendroff scheme [7], which is second-order accurate in time and space. The constant factor K/ρ_0 is set to 1 for simplicity in what follows. Equation (7) is approximated by the following computational module,

$$\begin{aligned} h^{n+1/2}_{r+1/2,s+1/2} &= \frac{1}{4}(A_x + A_y)h^n_{r,s} + \frac{\Delta T}{2\,\Delta x}\Delta_x \mu^n_{r,s} + \frac{\Delta T}{2\,\Delta y}\Delta_y \nu^n_{r,s} \\ h^{n+1}_{r,s} &= h^n_{r,s} + \frac{\Delta T}{\Delta x}\delta_x \delta_T \mu^n_{r,s} + \frac{\Delta T}{\Delta y}\delta_y \delta_T \nu^n_{r,s} \end{aligned} \tag{9}$$

on $\mathbf{R}^2_\Delta \times T_\Delta$. It may be shown [8] that the formal linearized stability condition is that the magnitude of the growth factor

$$|\xi| = \left|\frac{\mathbf{v}\,\Delta T}{2\,\Delta x}\right| < 1$$

where $\mathbf{v}$ is the local phase velocity of the conservation law. Hence, we need to identify $\mathbf{v}$ in the mass transport equation: Since

$$\mu_X = \frac{\partial \mu}{\partial h}\frac{\partial h}{\partial X} + \frac{\partial \mu}{\partial a_i}\frac{\partial a_i}{\partial X} + \frac{\partial \mu}{\partial a_i^*}\frac{\partial a_i^*}{\partial X},$$

with $i = 1, 2$. The first term is $\hat{x}\cdot\mathbf{v}h_X$, and the remaining terms can be thought of as forcing terms. Using Eq. (70), the expression for $\hat{x}\cdot\mathbf{v}$ is given by

$$\mu_h = \sum_{j=1}^{2} \frac{2\beta^2 h k_j^3 C_j |a_j|^2}{\sigma_j}\left(\frac{2}{\omega_j}\mathcal{J}_{ij} + \frac{\beta}{\sigma_j}\mathcal{J}_{2j}\right) + \text{c.c.}$$

Using the same argument, we can find $\hat{y}\cdot\mathbf{v}$. For the spanwise component

$$\nu_h = \sum_{j=1}^{2} \frac{2C_j\beta^2 k_j^2 h}{\sigma_j \omega_j} a_j^* a_{j\gamma}\mathcal{J}_j + \text{c.c.}$$

The size of ν_h and μ_h can be estimated to be $\beta^2 h|a_j|^2$, assuming that the model's other parameters are of $O(1)$. The higher of either wave packet amplitude is taken here. Hence, for stability the grid size is determined by the constraint

$$\begin{aligned} h\frac{\Delta T}{\Delta x} &\le \beta^{-2}\|a_j|^{-2}, \quad \text{and} \\ h\frac{\Delta T}{\Delta y} &\le \beta^{-2}|a_j|^{-2}. \end{aligned} \tag{10}$$

Dissipation effects would manifest themselves in the solution of the model primarily by attenuating the amplitude of the water waves and of the drift velocity. Since the model is nonlinear, attenuation would then lead to drastic changes in the morphology of the bars, namely, smaller bars with longer interbar spacing. Dissipation is known to occur in the two-step Lax–Wendroff scheme, except when $|\xi| = 1$. The effect, however, can be quite small—fourth order in Δx and Δy—if the grid size is restricted to being much smaller than the wavelengths.

3. SOLUTION OF THE SURFACE EQUATIONS

Since exact solutions for (6) are unknown at present, we will make use of the two-dimensional version of the model, which was studied by Boczar-Karakiewicz *et al.* in [3], to check the

numerical solution, albeit in a limited way, of the fully three-dimensional model. The two-dimensional case is

$$\begin{aligned} a_{1X} &= -iK_3 f(X)a_1 - iK_5 e^{-i\Delta X} a_1^* a_2 \\ a_{2X} &= -iK_4 f(X)a_2 - iK_6 e^{+i\Delta X} a_1^2 \\ a_1(X=0) &= \mathcal{A}_1 \\ a_2(X=0) &= \mathcal{A}_2, \end{aligned} \tag{11}$$

where $\mathcal{A}_1$ and $\mathcal{A}_2$ are constants. In [3], it was found that the numerical solution of Eq. (11) could be successfully carried out using an explicit fourth-order Runge–Kutta scheme. We adopt the same computational scheme and rely on their confidence in their method to enable us to use its solutions as the approximately correct solutions to the three-dimensional case under appropriate circumstances.

For the surface equations in the three-dimensional case, Eq. (6) is rewritten as

$$\begin{aligned} a_{1X} - iK_1 a_{1yy} + iK_3 f(X,y)a_1 &= -iK_5 e^{-i\Delta X} a_1^* a_2 \\ a_{2X} - iK_2 a_{2yy} + iK_4 f(X,y)a_2 &= -iK_6 e^{+i\Delta X} a_1^2 \\ a_1(X=0,y) &= \mathcal{A}_1(y,T) \\ a_2(X=0,y) &= \mathcal{A}_2(y,T) \\ a_{1y}(X,y=0) &= 0 \\ a_{2y}(X,y=0) &= 0 \\ a_{1y}(X,y=N) &= 0 \\ a_{2y}(X,y=N) &= 0 \end{aligned} \tag{12}$$

to separate the linear and nonlinear parts. The first two boundary conditions are inherent in the physics of the problem. The remaining boundary conditions are artificial. These Neumann boundary conditions, combined with a computational procedure that will be explained presently, ensures that the overall structure of the solutions remains negligibly affected by the choice of lateral boundary conditions. We call this technique the "zero-flux procedure."

To justify the need for such procedure, we spell out what sort of problem we are faced with: Since we need to compute a solution over a finite but large domain, care must be exercised in imposing boundary conditions on the lateral sides so as to avoid the introduction of structure in the solution that is strictly mathematical rather than physical in nature. The use of Neumann boundary conditions make the problem well-posed, but these hard barriers will reflect waves back into the domain, a situation that does not mimic an effectively laterally unbounded ocean. Another possible way to compute a solution of the problem over an effectively unbounded domain on a finite grid is to impose periodic boundary conditions. However, periodicity imposes unwanted symmetries on the structure of the computed solutions. To avoid this situation, we opt to use the lateral boundary conditions that make the problem well-posed and, in addition, place restrictions on the initial bottom configuration and the boundary condition at $X = 0$ so that we can compute an oceanic event on a swath of what amounts to be an effectively unbounded domain. We have found that when used with care, the zero flux procedure is simpler to use, and as effective as other synthetic boundary conditions in minimizing unwanted structure in the solutions.

A posteriori we know that the solution to the model is two-dimensional if neither the bottom, nor the boundary condition at $X = 0$, have y dependence. In such a case the Neumann boundary conditions have no effect on the solution over any part of the domain (i.e., it does not lead to y-dependent solutions). To carry out the zero flux condition we calculate the system over a computational domain that is divided into three regions. The large central region, flanked by two sufficiently wide lateral strips, is one in which y variation in the initial bottom or in the boundary condition at $X = 0$ is possible. In the lateral strips no y dependence in the above-mentioned quantities is permitted. The solution in these lateral strips is discarded. The initial bottom and the boundary condition at $X = 0$ are connected smoothly in all three regions so that a minimal amount of structure is introduced in the solutions. The flux at the boundaries can be calculated for monitoring purposes, however, it is not hard to guess what size to use for the lateral strips so that nearly zero-flux conditions are met.

Before considering the discretization of the three-dimensional system, we introduce some notation that will make the presentation more concise. Define the following vectors, with the superscript T meaning transpose,

$$\begin{aligned} \mathbf{k} &= i[K_1, K_2]^T \in \mathcal{C}^2 \\ \mathbf{k}_f &= if(X,y)[K_3, K_4]^T \in \mathcal{C}^2 \\ \phi &= [a_1(X,y), a_2(X,y)]^T \in \mathcal{C}^2 \end{aligned} \tag{13}$$

with $(X, y) \in \mathbf{R}^2_+$. With this notation Eq. (12), is

$$[\partial_X - \mathbf{k}\partial_{yy} + \mathbf{k}_f]\phi = b(X, y, \phi), \tag{14}$$

with the linear part on the left-hand side and the nonlinear terms on the right of the equals sign, plus boundary conditions,

$$\begin{aligned} \phi_y &= 0 \quad \text{on } y = 0, y = N, \\ \phi &= \phi_0 \quad \text{on } X = 0. \end{aligned} \tag{15}$$

The term $b(X, \phi)$ represents the nonlinear terms. Succinctly, the above equation may be written as

$$\mathcal{L}\phi = b, \tag{16}$$

where $\mathcal{L}$ is the linear operator. Let L be a suitable discretization of the linear operator. Suppose the value of the vector ϕ at level r for all s is known. The value of the vector at level $r+1$ may be found making use of fixed point methods.

Equation (16) is discretized

$$L\phi_{r,s}^{l+1} = B(r\Delta x, s\Delta y, \phi_{r,s}^{l}), \tag{17}$$

where B is a discretization of the right hand side of Eq. (16), and l is the index of the iteration. To start the iteration, we use the value of the field variables at the rth level in X (i.e., $\phi^0 = \phi_r$).

If so desired, the calculation may be performed on a computer in two steps: let $\overline{\phi}$ be an intermediate result. Then the following computational scheme may be used:

$$\begin{aligned} L\overline{\phi}_{r,s} &= B(\Delta x, \Delta y, \phi^{l}) \\ L\phi_{r,s}^{l+1} &= B(r\Delta x, s\Delta y, \overline{\phi}). \end{aligned} \tag{18}$$

The condition for convergence of Eq. (17) is found by appealing to the fixed-point theorem [9, Chap. 5]. For the purpose of determining the convergence criterion, define $\mathfrak{X}$, a region in $\mathcal{C}^4$, a four-vector complex space. Let $\mathbf{\Phi}$ and $\mathbf{h} \in \mathfrak{X}$ be two vectors in that space. For some $\mathbf{A} \in \mathfrak{X}$ the derivative of $\mathbf{A}$ with respect to $\mathbf{\Phi}$ is

$$\mathbf{A}_{\Phi} \equiv \mathbf{J}(\mathbf{\Phi}) = \frac{\partial A_i}{\partial \Phi_j}. \tag{19}$$

If the second derivative is continuous for all $\mathbf{\Phi} \in \mathfrak{X}$, then it satisfies

$$\|\mathbf{A}_{\Phi\Phi}(\mathbf{\Phi}, \mathbf{h}, \mathbf{h})\| \leq R\|\mathbf{h}\|^2 \tag{20}$$

for all $\mathbf{\Phi}$, where R is a constant. Furthermore, let $\|\mathbf{A}\|_p$, with $p = 1, 2, \infty$, represent the l_1, l_2, and l_∞ induced norms,

$$\begin{aligned} \|\mathbf{A}\|_1 &= \max_{1\leq j\leq n}\left\{\sum_{i=1}^{n} |A_{ij}|\right\} \quad \|\mathbf{A}\|_2 = \left\{\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}A_{ij}^{*}\right\}^{1/2} \\ \|\mathbf{A}\|_\infty &= \max_{1\leq i\leq n}\left\{\sum_{j=1}^{n} |A_{ij}|\right\}. \end{aligned} \tag{21}$$

Equations (14) and (15), are used to construct a supersystem

$$[\partial_X - \mathbf{K}\partial_{yy} + \mathbf{K}_f]\mathbf{\Phi} = \mathbf{b}(X, y, \mathbf{\Phi}), \tag{22}$$

with boundary conditions

$$\begin{aligned} \mathbf{\Phi}_y &= 0 \quad \text{on } y = 0, y = N \\ \mathbf{\Phi} &= \mathbf{\Phi}_0 \quad \text{on } X = 0 \end{aligned} \tag{23}$$

and

$$\begin{aligned} \mathbf{K} &= [\mathbf{k}, \mathbf{k}^*]^{\mathrm{T}} \in \mathfrak{X} \\ \mathbf{K}_f(X, y) &= [\mathbf{k}_f, \mathbf{k}_f^*]^{\mathrm{T}} \in \mathfrak{X} \\ \mathbf{\Phi} &= [a_1(X, y), a_2(X, y), a_1^*(X, y), a_2^*(X, y)]^{\mathrm{T}} \in \mathfrak{X}. \end{aligned} \tag{24}$$

Let $\mathbf{L}$ be the resulting discrete operator of the supersystem on $\mathbf{R}_3$, composed of L and its complex conjugate, and $\mathbf{B}$ be the discrete counterpart of the nonlinear terms in the supersystem. Choosing a consistent nonsingular discretization for L (hence $\mathbf{L}$ will be nonsingular as well), and multiplying both sides of (22) by $\mathbf{L}^{-1}$, we have

$$\mathbf{\Phi} = \mathbf{A}(X, y, \mathbf{\Phi}), \tag{25}$$

where $\mathbf{A} = \mathbf{L}^{-1}\mathbf{B}$. We wish to solve Eq. (25) by the procedure suggested in Eq. (17). That is,

$$\mathbf{\Phi}^{l+1} = \mathbf{A}(X, y, \mathbf{\Phi}^{l}). \tag{26}$$

Define the iteration discrepancy as

$$\|\delta\mathbf{\Phi}^{l+1}\|_p \equiv \|\mathbf{\Phi}^{l+1} - \mathbf{\Phi}^{l}\|_p. \tag{27}$$

Appealing to the fixed-point theorem, it can be surmised that

$$\begin{aligned} \|\delta\mathbf{\Phi}^{l+1}\|_p &= \|\mathbf{A}(\mathbf{\Phi}^{l}) - \mathbf{A}(\mathbf{\Phi}^{l-1})\|_p \\ &\leq \|\mathbf{J}(\mathbf{\Phi}^{l-1})\delta^{l}\|_p \\ &\leq \|\mathbf{J}(\mathbf{\Phi}^{l-1})\|_p\|\mathbf{J}(\mathbf{\Phi}^{l-2})\|_p\|\delta\mathbf{\Phi}^{l-2}\|_p \leq \cdots \\ &\leq \prod_{l=0}^{k-1} \|\mathbf{J}(\mathbf{\Phi}^{l})\|_p\|\delta\mathbf{\Phi}^{0}\|_p, \end{aligned} \tag{28}$$

the inequality holds provided

$$0 < \|\mathbf{J}(\mathbf{\Phi}^{l})\|_p < 1. \tag{29}$$

Equation (29) yields the convergence criterion for the iteration process. It must be emphasized that inversion of the operator $\mathbf{L}$ is not required in the actual computation. What we are trying to do is to find the conditions required for convergence of the scheme.

An estimate of the rate of convergence can be found by considering the continuous problem, projected onto the grid.

Let $q > 0$ be given such that the set of vectors $\mathcal{S} = \{\Phi : \|\Phi - \mathbf{s}\|_p < q\}$ contains a fixed point $\mathbf{s}$ of $\mathbf{a}(\mathbf{s})$, that is,

$$\mathbf{s} = \lim_{l\to\infty} \Phi^l = \lim_{l\to\infty} \mathbf{a}(\Phi^l) = \mathbf{a}(\mathbf{s}), \tag{30}$$

where $\mathbf{a}(\mathbf{s}) = \mathcal{L}^{-1}\mathbf{b}(\mathbf{s})$. Further, let $\mathcal{S} \subseteq \mathcal{Z}$, $\mathbf{J}(\mathbf{s})$ continuous on $\mathcal{S}$ and $\|\mathbf{J}(\mathbf{s})\|_p < 1$. Then there exists an $\varepsilon > 0$ such that the fixed point iteration is convergent whenever $\|\Phi^0 - \mathbf{s}\|_p < \varepsilon$. Define $\|\mathbf{e}^{l+1}\|_p$, the measure of difference between the $(l+1)$th iterate and the root. Hence

$$\begin{aligned}\|\mathbf{e}^{l+1}\|_p = \|\Phi^{l+1} - \mathbf{s}\|_p &\approx \|\mathbf{J}(\mathbf{s})\mathbf{e}^l + \mathbf{A}''(\zeta; \mathbf{e}^l, \mathbf{e}^l)\|_p \\ &\le \|\mathbf{J}(\mathbf{s})\mathbf{e}^l\|_p + R\|\mathbf{e}^l\|_p^2.\end{aligned} \tag{31}$$

Quadratic convergence is possible if $\mathbf{J}(\mathbf{s}) = 0$,

$$\lim_{l\to\infty} \frac{\|\mathbf{e}^{l+1}\|_p}{\|\mathbf{e}^l\|_p^2} \le R. \tag{32}$$

For the problem in question, however, the best rate of convergence will be linear since $\mathbf{J}(\mathbf{s}) \neq 0$:

$$\lim_{l\to\infty} \frac{\|\mathbf{e}^{l+1}\|_p}{\|\mathbf{e}^l\|_p} \le \|\mathbf{J}(\mathbf{s})\|_p. \tag{33}$$

An estimate of the distance between the discrete and the continuous fixed point on the grid can be established as follows. Let Ξ be the fixed point of the discrete problem, and assume $\Xi = \mathbf{s} - \omega$. Also note that $\mathbf{B}(\mathbf{s}) = \mathbf{b}(\mathbf{s})$, and remember that the nonlinear term is quadratic. Then

$$\mathbf{L}\Xi - \mathcal{L}\mathbf{s} = \mathbf{B}(\Xi) - \mathbf{b}(\mathbf{s}) \tag{34}$$

hence the truncation error $\tau\mathbf{s}$ is related to the distance ω by

$$\tau\mathbf{s} = (\mathbf{L} - \mathbf{b}'(\mathbf{s}))\omega + \tfrac{1}{2}\mathbf{b}''(\mathbf{s})\omega\cdot\omega. \tag{35}$$

Assume that $\|\omega\| < 1$ is small, so that the quadratic term may be neglected. Then

$$\|(\mathbf{I} + \mathbf{L}^{-1}\mathbf{b}'(\mathbf{s}))\mathbf{L}^{-1}\tau\mathbf{s}\|_p \ge \|\omega\|_p, \tag{36}$$

thus the distance between the fixed points shrinks as the discretization is refined, as long as the discretization is consistent.

We also show that the cumulative error $\mathcal{S}^l \le |\mathcal{S}|$ in the fixed point iteration procedure remains bounded, so long as the condition in Eq. (29) is satisfied, and that the error is controlled by the grid size. Equation (26) may be recast as

$$\Phi^{l+1} = \mathbf{a}(\Phi^l) + \mathcal{S}^l. \tag{37}$$

Using the same argument in [10, pp. 92–93], one can show that

$$|\Phi^{l+1} - \mathbf{s}| \le \frac{\mathcal{S}}{1-J} + J^{l+1}\left(\mathbf{s} - \Phi^0 - \frac{\mathcal{S}}{1-J}\right), \tag{38}$$

where $1 > J \ge |\mathbf{J}(\Phi^{l+1})|$, for all l. In terms of the iterates for the continuous system,

$$\mathcal{S}^l = (\mathbf{L}^{-1} - \mathcal{L}^{-1})[\mathbf{b} - \mathbf{b}'(\mathbf{s})\omega^l + \tfrac{1}{2}\mathbf{b}''(\mathbf{s})\omega^l\cdot\omega^l], \tag{39}$$

so that

$$\mathcal{S}^l = (\mathbf{I} - \mathcal{L}^{-1}\mathbf{L})(\mathbf{s} - \omega^l). \tag{40}$$

It is assumed that the discretization of the linear operator is consistent and, as that as shown in Eq. (36), the distance between the discrete and the continuous fixed points gets smaller as the grid is refined. In conclusion the bounded iteration error depends on the size of J and on the grid size.

We now describe the particular discretization used in the model. There is flexibility in the choice of discretization for the linear operator L. The most economical discretizations are those that lead to a tridiagonal or quintadiagonal matrix. One possibility is to adopt the scheme [7, p. 138]

$$\begin{aligned}L\phi_{r,s} = &\left(\frac{3}{2\,\Delta x}\Delta_x - \frac{1}{2\,\Delta x}\nabla_x\right)\phi_{r,s} \\ &- \left(\frac{\mathbf{k}}{\Delta y^2}\delta_s^2 - \mathbf{k}_{\mathbf{f}_{r-1}^t}\right)\phi_{r+1,s}\end{aligned} \tag{41}$$

which leads to two $n \times n$ tri-diagonal matrix. Thus L has eigenvalues

$$\begin{aligned}\lambda_s = &-(3 + 2\rho + 2\,\Delta x\mathbf{k}_{\mathbf{f}}) \\ &+ 2\rho\cos\left[\frac{s\pi}{n+1}\right] \quad s = 1, \ldots, n,\end{aligned} \tag{42}$$

where $\rho = 2\,(\Delta x/\Delta y^2)\,\mathbf{k}$, and the eigenfunctions

$$\left\{\sin\frac{s\pi}{n+1}, \sin\frac{2s\pi}{n+1}, \cdots, \sin\frac{s\pi}{n+1}\right\}^{\mathrm{T}} \quad s = 1, \ldots, n. \tag{43}$$

Furthermore, the operator L is diagonally dominant, since

$$\sum_{j\neq i}^{n+1} |L_{ij}| \le |L_{ii}| \quad i = 1, \ldots, 2n,$$

the L_{ij}'s being the entries of the matrix L, and nonsingular since

$$\begin{aligned}&|L_{ii}| > |L_{ii+1}| > 0 && i = 1, \ldots, 2n-1 \\ &|L_{ii}| \ge |L_{ii+1}| + |L_{ii-1}| \quad L_{ii+1}L_{ii-1} \neq 0 && i = 2, \ldots, 2n-1 \\ &|L_{ii}| > |L_{ii-1}| && i = 2, \ldots, 2n.\end{aligned}$$

If $\phi = \xi^r e^{i\theta}$, where $\theta = \alpha\,\Delta y s$, upon substituting these quantities in L the magnification factor is

$$\xi = \frac{1}{2\rho(1-\cos\theta) + 2\,\Delta x \mathbf{k}_t + 3 \{2 + \sqrt{1 - 2\rho(1-\cos\theta) - 2\,\Delta x \mathbf{k}_t}\}}, \tag{44}$$

from which it is clear that $|\xi| \leq 1$. Thus the discretization of the linear operator is unconditionally stable.

An estimate of the accuracy of the discretization of the linear operator, as well as a check on its consistency with the continuous operator on the grid, is given by

$$\tau\phi \equiv (L - \mathcal{L})\phi = -\frac{\Delta x^2}{3}\phi_{xxx} + \mathbf{k}\frac{\Delta y^2}{12}\phi_{yyyy} + \cdots. \tag{45}$$

Equation (45) impies that the scheme is $O(\Delta x^2 + \Delta y^2)$ accurate.

Consistency of the discretization is readily established by comparing the continuous problem with its discretization in the limit as the grid size gets smaller. It can be shown that the discretization approaches the continuous operator on the grid uniformly.

Since the scheme is inapplicable at $r = 0$, a standard backward Euler scheme,

$$\frac{1}{\Delta x}\Delta_x\phi_{r,s} - \frac{1}{\Delta y^2}\mathbf{k}\delta^2\phi_{r+1,s} + A_x(\mathbf{k}_t\phi)_{r+1,s}, \tag{46}$$

is used to discretize L for the first step in X, which can be shown to be unconditionally stable as well.

Having made a choice on the particular form of the operator L, the condition that $\|\mathbf{J}(\Phi)\|_p < 1$ for the surface system must be determined explicitly, so that convergence is established for the sand ridge problem. To estimate the size of $\mathbf{J}(\Phi)$ we use the supersystem, Eq. (25), to find that

$$\begin{aligned} \delta\Phi^{l+1} &\approx \mathbf{J}(\Phi^l)\delta\Phi^l \\ \delta\Phi^l &\approx \mathbf{J}(\Phi^{l-1})\delta\Phi^{l-1} \\ &\text{etc.} \end{aligned} \tag{47}$$

with

$$\mathbf{J} = \mathbf{L}^{-1}\mathbf{B}'(\Phi), \tag{48}$$

where

$$\mathbf{B}'(\Phi^l) = \begin{pmatrix} 0 & -iK_5 e^{-i\Delta X_{r+1}} a_1^{*l} & -iK_5 e^{-i\Delta X_{r+1}} a_2^l & 0 \\ -i2K_6 e^{-i\Delta X_{r+1}} a_1^l & 0 & 0 & 0 \\ iK_5 e^{-i\Delta X_{r-1}} a_2^{l*} & 0 & 0 & iK_5 e^{+i\Delta X_{r+1}} a_1^l \\ 0 & 0 & i2K_6 e^{+i\Delta X_{r+1}} a_1^{l*} & 0 \end{pmatrix}, \tag{49}$$

for the l^{th} iterate. In Eq. (49), it is understood that the a's are defined only on the grid.

From Eq. (48), $\|\mathbf{J}\|_p < 1$ if the size of $\mathbf{L}$ is greater than the size of $\mathbf{B}'$. In the l^2 norm, the convergence condition is

$$\|\mathbf{J}\|_2 = \|\mathbf{L}^{-1}\mathbf{B}'\|_2 \leq \|\mathbf{L}^{-1}\|_2\|\mathbf{B}'\|_2 \leq 1. \tag{50}$$

Since $\mathbf{L}\mathbf{L}^\dagger = \mathbf{L}^\dagger\mathbf{L}$, where $\mathbf{L}^\dagger$ is the Hermitian matrix of $\mathbf{L}$, then the spectral radius gives a measure of the two norm. Using this information, we have that

$$\|\mathbf{L}^{-1}\|_2 \leq \min_{s=1,n} |\lambda_s|, \tag{51}$$

or, using Eq. (42),

$$+ \sqrt{(K_5^2 + 4K_6^2)|a_1|^2 + K_5^2|a_2|^2}/(3 + 2\,\Delta x \mathbf{k}_t) < 1, \tag{52}$$

where the l_∞ in y is used to estimate the size of the vectors, that is, $a_i = \max_{1\leq s\leq n} a_i^s$, $i = 1, 2$. Hence, Eq. (52) gives strict constraints on ρ, Δx, and a_i, to be satisfied in order to guarantee convergence in the solution. Phase error may be possible if $\Delta x > 2\pi/\delta$. However, since the size of $\delta \ll 1$ phase error constraints are not difficult to meet.

4. PERFORMANCE EVALUATION OF THE NUMERICAL SCHEMES

4.1. *Evaluation of the Mass Transport Equation Scheme*

We ran a few test cases in order to confirm qualitatively the stability, consistency and accuracy of the Lax–Wendroff scheme, checking for agreement with the well-established theoretical results. Of more concern to us was the issue of damping and of phase drift. We wish to determine the upper bounds for the spatial and temporal discretization of the mass transport equation scheme that will ensure good qualitative results without requiring excessive computational resources. To quantify the scheme's dissipation and drift, we integrated a model problem for which an exact solution is known over time scales comparable to those used in the sand bar problem.

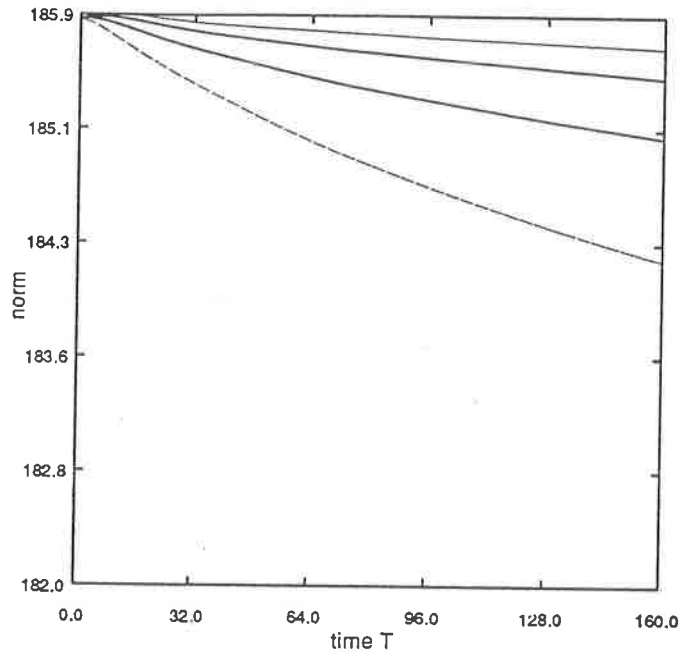

FIG. 1. Dissipation of the norm $c(T, \theta)$ with $k = 0.1$ for the Lax–Wendroff scheme. From top to bottom, $\theta = 0.05$, 0.1, 0.2, and 0.4, respectively.

The model problem was

$$h_T + khh_x = 0, \quad x \in \mathcal{R}^1, T > 0, \tag{53}$$

with initial condition

$$h(x, 0) = \begin{cases} 1 & x < 0 \\ 1 + \varepsilon x & 0 \le x \le l \\ 1 + \varepsilon l & x > l \end{cases} \tag{54}$$

in which $0 < k < 1$, and $\varepsilon < 1$. The exact solution of Eqs. (53) and (54) is

$$h(x, T) = \begin{cases} 1 & x < kT \\ 1 + \varepsilon \dfrac{x - kT}{1 + \varepsilon kT} & kT \le x \le l + k(1 + \varepsilon l)T \\ 1 + \varepsilon l & \text{otherwise.} \end{cases} \tag{55}$$

Different values of k were tried—it scales the time step—but we report our results for $k = 0.1$, which is in the range of values for constants in the mass transport equation. For such a case, convergence is possible if $h\ \Delta T/\Delta x < 10$ in the time interval 0 to T. Since Eq. (53) conserves a quantity proportional to h^s, where s is an integer, we compared the computed value h^s_Δ with the theoretical value h^s as a function of $\theta \equiv k\ \Delta T/\Delta x$ and as a function of time T, to get an idea of the scheme's dissipation. Specifically, we monitored the constant of motion

$$c(T, \theta) = \sum_{r=0}^{M/\Delta x} h^2_\Delta(T, r\,\Delta x) r\,\Delta x + \frac{2}{3} kT[h^3_\Delta(T, M) - h^3_\Delta(T, 0)], \tag{56}$$

where M is a very large value in x_r. For an estimation of the phase drift, we computed

$$e^2(\theta, T) = \sum_r |h_\Delta(T, x_r) - h(T, x_r)|^2 \Big/ \sum_r |h(T, x_r)|^2. \tag{57}$$

Figures 1 and 2 show parametric plots of c and e^2, respectively. The main source of error is the resultant oscillations in the calculated solutions in the neighborhood of the discontinuity in the derivative in the solution. These oscillations lead to the noisy character of the phase calculation evidenced in Fig. 2, which was seen to disappear with smoother initial conditions. The plots suggest that $\theta = 0.2$ is a reasonable upper limit for the discretization of the sand bar problem.

4.2. *Performance of the Runge–Kutta Scheme*

The accuracy and dissipation of the explicit fourth-order Runge–Kutta was investigated using a flat bottom and $\mathcal{A}_j$ con-

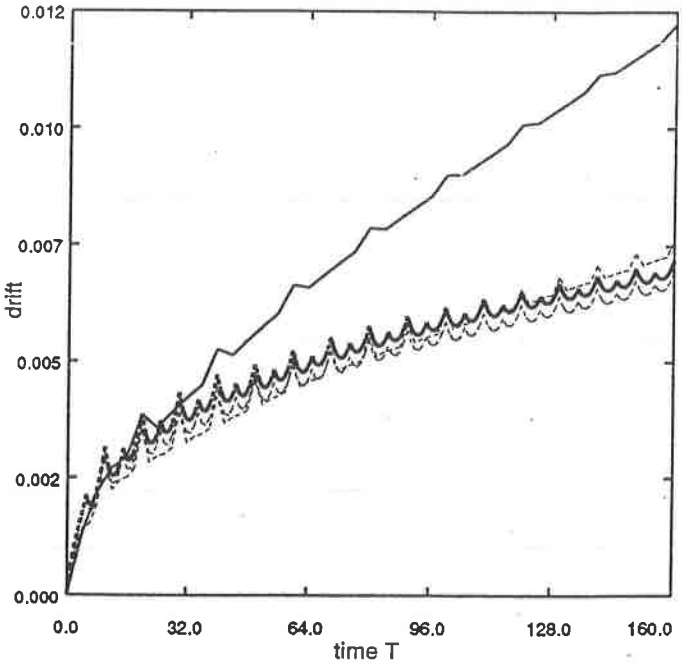

FIG. 2. Phase drift $e^2(T, \theta)$ for the Lax–Wendroff scheme with $k = 0.1$. From top to bottom, $\theta = 0.4$, 0.2, 0.1, and 0.05, respectively.

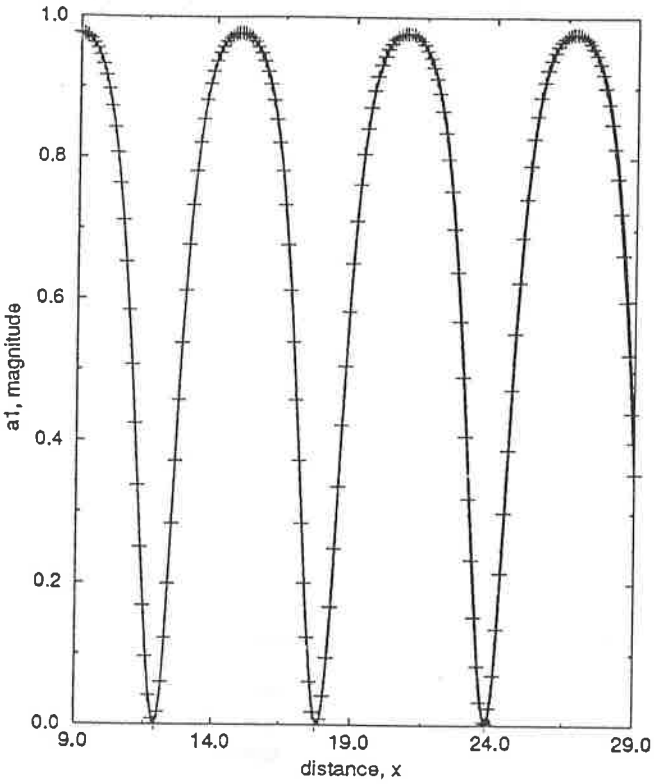

FIG. 3. Comparison of a portion of the exact (solid) versus Runge-Kutta (crosses) solution for $\Delta x = 1$. Magnitude of a_1.

stants. The domain was 60 units in extent, or roughly 15 interaction lengths. The energy, which is proportional to the sum of the mode amplitudes squared, was conserved by all trials: the increased error with grid size Δx was only 0.2% for $\Delta x = 2.0$. The two-dimensional solution will be used later on in making an assessment of the FPM solution. We wish to determine an appropriate range for Δx that will enable us to perform such comparison.

An exact solution to Eq. (11) is known when $f(X) = 0$. To estimate the error in the scheme we compared the outcome of the numerical solution with the exact solution of this special case (cf. [11, 12]). When $\mathcal{A}_1 = 0.5$ and $\mathcal{A}_1 = 0.0$, in terms of Jacobi elliptic functions "sn," the exact solution is

$$a_1^2(\tilde{x}) = v_b^2 \operatorname{sn}^2[v_b^{-1/2}\tilde{x}; v_b]$$

$$a_2^2(\tilde{x}) = 1 - a_1^2(\tilde{x}),$$

where v_b is a constant that depends on the detuning parameter δ, and $\tilde{x}$ is a normalized distance coordinate. The Jacobi Elliptic functions need to be calculated. Hence, by exact solution we mean a series solution.

The following measures were used to estimate the accuracy of the discretization:

$$l_2 = \frac{[\sum_r |\chi(r\,\Delta x) - \chi'(x_r)|^2]^{1/2}}{[\sum_r |\chi'(r\,\Delta x)|^2]^{1/2}}, \tag{58}$$

$$\text{norm} = [\sum_r |\chi(r\,\Delta x)|^2]^{1/2}. \tag{59}$$

Here χ is the calculated value of a_i, and χ' the exact value at the grid location. The exact solution χ', was computed using the algorithm given in [13, p. 189]. The error manifests itself in the solution as spatial phase drift and as a degradation in the ability to capture sharp features. Figure 3 shows the superimposition of the computed and the series solution, for $\Delta x = 1$. The error as a function of grid size is shown in Fig. 4 on a log-log plot. The upper plot in the graph shows the l_2 error norm described above, showing convergence, albeit not fourth-order in Δx. This slower convergence is due to the nature of the series solution. The lower curve is the plot of the norm of the Runge–Kutta solution as a function of the grid size. For these trials the $\mathcal{A}_1 = 0.5$, $\mathcal{A}_2 = 0$, in Eq. (11), a flat bottom and parameters $\alpha = 0.3$, $\beta = 0.1$, $\omega_1 = 0.5$, were used.

4.3. *Fixed-Point Method Performance and Evaluation*

Since an exact solution to the three-dimensional surface wave system is as yet unknown, we sought to discern the accuracy of the fixed-point method (FPM) using local analysis. Let Δ be the size in X or y of each grid element. A comparison of the computed solution at a particular point, using Δ, with a solution with grid size $\Delta/2$ yields

$$|\chi_\Delta - \chi_{\Delta/2}| \equiv k_1 = CO'[(\Delta/2)^p]. \tag{60}$$

Halving the grid size again

$$|\chi_{\Delta/2} - \chi_{\Delta/4}| \equiv k_2 = CO[(\Delta/4)^p]. \tag{61}$$

Thus, using Eq. (60) and Eq. (61) one can solve for p to get an estimate of the order of accuracy of the scheme:

$$p = \frac{\log k_1 - \log k_2}{\log 2}. \tag{62}$$

Using the same parameters and boundary conditions as those used in connection with the Runge–Kutta scheme evaluation trials, and a domain with length of 128 and span of 32, we found that FPM yields an average value of $p = 1.8$, with a standard deviation of 0.5 for the X-discretization and about $p = 2$ for the y-discretization with some degradation in the immediate neighborhood of the lateral boundaries. Values of both field quantities were used to estimate p, and they were taken from various points in the domain.

Convergence of the FPM scheme was checked by examining the solution as the grid was refined. The domain was a square of 60 units per side. Since comparisons of the computed solu-

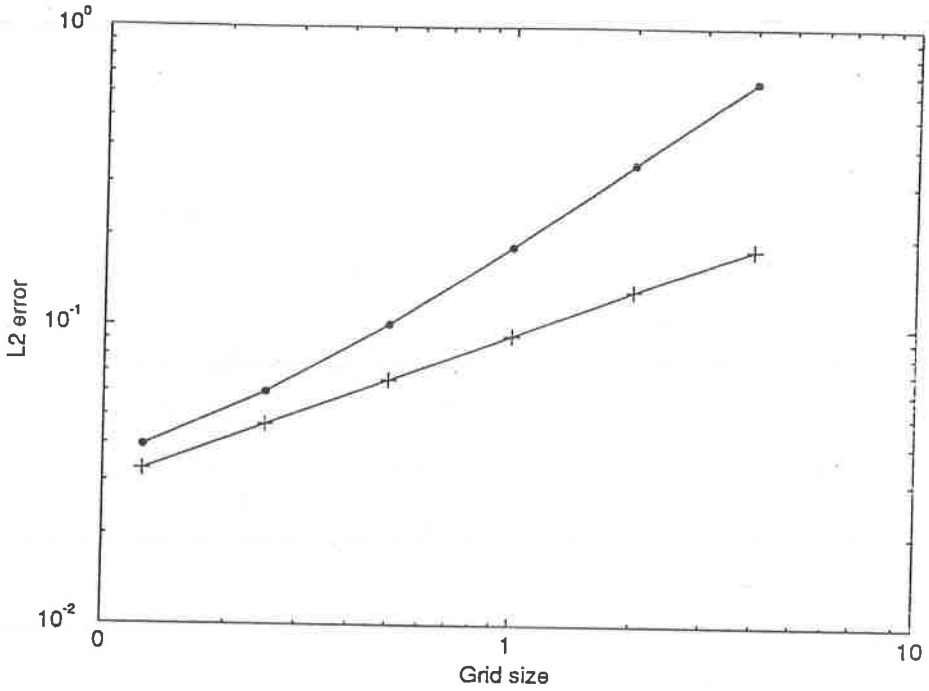

FIG. 4. l_2 error as a function of grid size Δx for the Runge–Kutta method (circles). Comparison with the series solution. l_2 norm of the Runge–Kutta solution as a function of grid size (crosses).

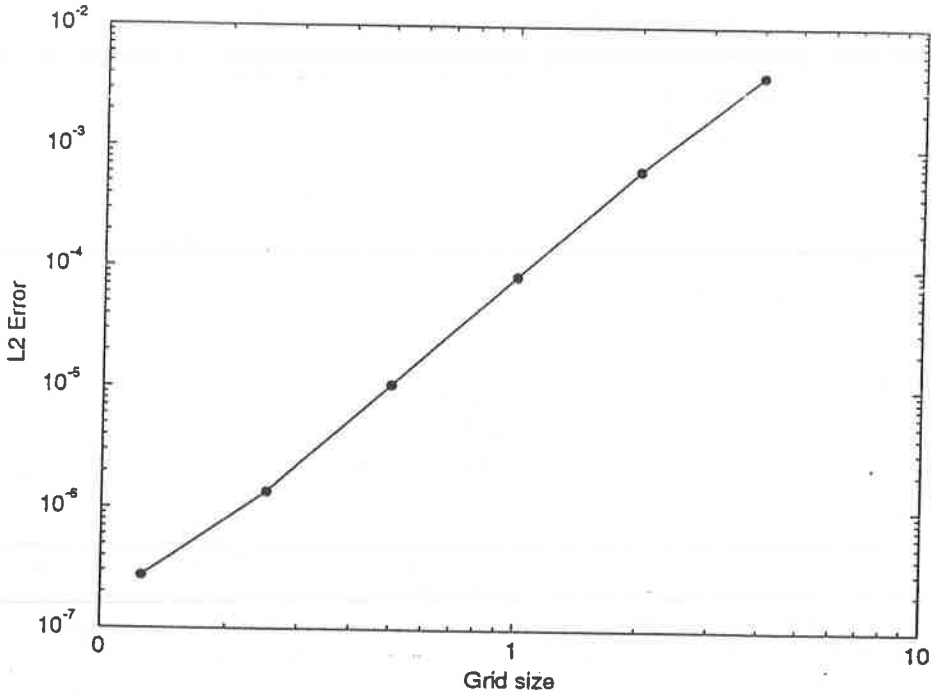

FIG. 5. l_2 error as a function of grid size, with $\Delta x = \Delta y$. Comparison between FPM and Runge–Kutta solution.

tions with an exact expression for the three-dimensional case were not possible, a comparison of the cross-sectional values of an effectively two-dimensional solution computed using FPM along the whole length in X and midway in the spanwise direction y, with a solution computed using the Runge–Kutta method with a very fine grid spacing was made to ascertain the qualitative correctness of the FPM scheme. A measure of the error is given by the norm

$$l_2(\Delta x, \Delta y) = \frac{[\Sigma_r |\chi(r\,\Delta x, \text{mid}) - \chi'(x_r)|^2]^{1/2}}{[\Sigma_r |\chi'(r\,\Delta x)|^2]^{1/2}}, \qquad (63)$$

where χ represents the solution obtained using FPM and χ' the solution computed with the Runge–Kutta scheme.

The result for the case $\Delta x = \Delta y$ is shown in Fig. 5. Note that there is no y dependence in the solution for this particular trial. The same outcome is obtained when $\Delta y = 0.25$ and Δx is varied. On the other hand, when $\Delta x = 0.25$ is fixed and Δy is varied, very little change in the norm is observed, as it should. In this last case, the norm had an approximate value of 1.35×10^{-6} for all grid sizes in the y direction that were used, which lead us to conclude that the fixed point procedure converged uniformly along y.

The rate at which the iteration procedure converges in FPM as a function of the grid size was also investigated. With $\alpha = 0.3$, $\beta = 0.08$, $\omega_1 = 0.5$, and boundary conditions $\mathcal{A}_1 = 0.5$ and $\mathcal{A}_2 = 0.1$, and a flat bed, the iteration discrepancy

$$\log_{10}\left[\max\left\{\sum_{s=0}^{n} |\phi^{l+1}(X, s\,\Delta y) - \phi^{l}(X, s\,\Delta y)|\right\}\right] \qquad (64)$$

was monitored at a particular value of X in a fairly large domain. It was found that the number of iterations required to meet a certain iteration tolerance decreased as the grid was refined. Figure 6 shows how the iteration discrepancy, as defined in Eq. (64), drops after each iteration l for a number of different grid sizes. It is evident from the graph that a finite and small number of iterations are required to reach adequate error tolerances using reasonably sized grids.

The iteration convergence of the solution at the first step in X was examined as well. Recall that for the first step a backwards Euler scheme was used to discretize the linear operator instead of the second-order scheme. The finding is that the number of iterations was roughly double the number required elsewhere in the domain, where the second-order scheme is used.

An example of a solution computed using the FPM/Lax–Wendroff scheme is shown in Fig. 7. This example shows how an initially flat bottom topography $f(X, y) = 0$ at $T = 0$ will eventually develop a refracting pattern when acted upon by water waves boundary conditions are $\mathcal{A}_1 = 0.5 + 0.001y$ and $\mathcal{A}_2 = 0.02 + 0.001y$, corresponding to an incoming gravity wave that has slightly higher amplitude at one end than at the other. The parameters for this run were $\alpha = 0.1$, $\beta = 0.08$, $\omega_1 = 1.2$, and $\varepsilon = 0.2$.

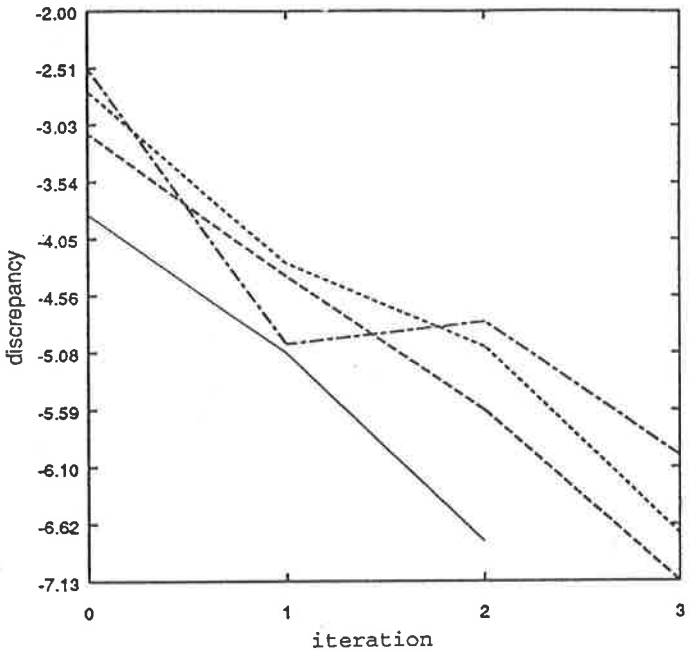

FIG. 6. Iteration discrepancy as a function of grid spacing. The number of iterations drops as $\Delta = 4$, 2, 1, and 0.5 respectively.

4.4. *Storage and Speed of the FPM*

An estimate for the operation count for the FPM is as follows. Eq. (18) leads to the problem of solving a tri-diagonal system L for the unknown ϕ at each r, where L is a $2n \times 2n$ matrix. This system is solved m times to cover all values of X in the domain. The efficient way to solve the tri-diagonal system is

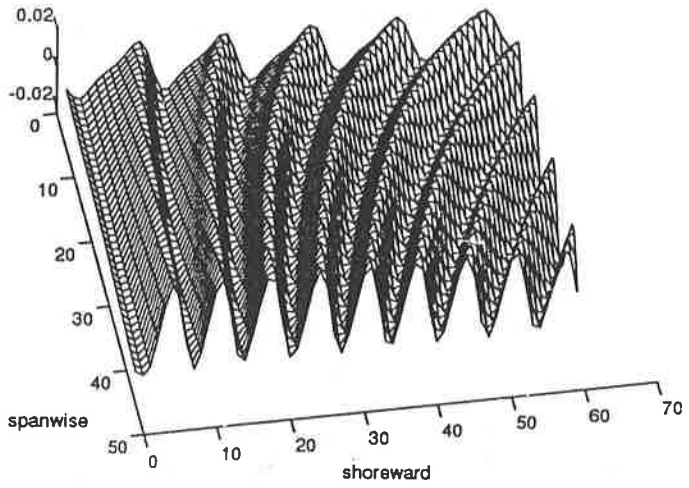

FIG. 7. Fate of an initially flat bottom after $T = 100\,\Delta T$ under the action of a refractive water wave field.

to decompose the problem in two steps: let $L = WU$, where W is a lower triangular matrix and U and upper triangular matrix. Then

$$Wg = \tilde{b} \tag{65}$$

is solved for g, followed by

$$U\phi = g, \tag{66}$$

to finally obtain ϕ at each r. The total operation count for the solution of Equations (65) and (66) is $2(5n - 4)$ multiplies and $2(3n - 3)$ adds. All told, $O(16n)$ operations. In turn, this process is performed l times to compute the $(l + 1)$th iterate, where l is usually a small number that depends on the iteration discrepancy tolerance; finally, m times to cover all values of X. The total is then $m \times l \times O(n)$.

The storage requirements of the FPM may be estimated as follows: the old and the new vector at each X, and another vector for the iteration process, need to be stored. Hence $6n$ values are stored. In addition, all the entries of a tridiagonal matrix of size $2n \times 2n$, or roughly $6n$ values need to be stored. The total is thus $12n$ values. In fact, we could be even more economical and use multipliers in the entries of L, so that only one half of the tri-diagonal matrix entries need to be stored. Note that the economy of resources hinges upon the simplicity of the matrix that the discretization generated. Obviously higher-order schemes could lead to more storage requirements. The point being made here is that FPM does not increase the storage cost over the amount required already for the discretized linear part of the model. This advantage is somewhat offset by having to compute more, but as we have shown the iteration leads to an acceptably low number of extra calculations.

A somewhat unavoidable problem with the FPM is that the discretization has significant dissipation. The attenuation we know is inherent in the discretization of the linear operator. However, the dissipation can be made tolerable at the expense of greater computational resources, that is, by refining the grid. To illustrate the degree of dissipation in the surface system FPM implementation, we used the same parameters and domain that was used in connection with the iteration issue, and we fixed the iteration discrepancy tolerance at 10^{-6}. Two types of trials were carried out, both using a flat bottom. In the three-dimensional trial we assumed the boundary conditions were $\mathcal{A}_1 = 0.5 + 0.01y$ and $\mathcal{A}_2 = 0.1 + 0.01y$ and monitored the conserved Hamiltonian along the length in the X direction, midway in the spanwise direction. With $\tilde{a}_2 = a_2 e^{-i\Delta X}$, the Hamiltonian density [11] is

$$\tilde{H} = \Re[(a_1^*)^2 \tilde{a}_2] + K_1|a_{1y}|^2 + K_2|\tilde{a}_{2y}|^2 + \delta|\tilde{a}_2|^2/2, \tag{67}$$

for a flat bottom. In the two-dimensional trial, we set $\mathcal{A}_1 = 0.5$ and $\mathcal{A}_2 = 0.1$ and monitored the reduced Hamiltonian density which results when y dependence in Eq. (67) is dropped. The outcome of both trials was qualitatively similar: the computed conserved quantity oscillated with a period equal to the interaction length, i.e., the length in which energetic exchanges recur in the modes. The difference between the peak value and the minimum value increased as the grid size was made larger. In addition, dissipation (i.e., the drop of the peak value as a function of position X) increased as the grid size was made larger, and as a result, the local interaction length grew since the amplitude of the modes were attenuated. The dissipation and oscillation of the conserved quantities can be made negligible by refining the grid. We also found that the effect is much more pronounced when $\mathcal{A}_2 = 0$ exactly, which yields solutions with very sharp minimas in the field variables. Table I shows the difference between successive maxima and minima for the second trial as a function of grid size, with $\Delta x = \Delta y$. We also report the outcome of fixing $\Delta x = 0.25$ and varying Δy, in Table II, and the opposite settings are illustrated in Table III. The two-dimensional trials for $\Delta x = 0.25$ and $\Delta y = 4$ showed

TABLE I

Energy Fluctuation vs Grid Size (Equilateral Grid Case)

Grid size Δ	Fluctuation
4.00	0.1002
2.00	0.0627
1.00	0.0168
0.50	0.0050
0.25	0.0014

TABLE II

Energy Fluctuation vs Δy ($\Delta x = 0.25$ Fixed)

Grid size Δy	Fluctuation
4.00	0.0018
2.00	0.0013
1.00	0.0013
0.50	0.0012
0.25	0.0014

TABLE III

Energy Fluctuation vs Δx ($\Delta y = 0.25$ Fixed)

Grid size Δx	Fluctuation
4.00	0.1415
2.00	0.0628
1.00	0.0198
0.50	0.0049
0.25	0.0014

TABLE IV

Wall-Clock Times in Seconds vs Grid Size (Number of Grid Points per Domain) for the Computation of the Surface System over the Whole Domain Using the Fixed-Point Method

Machine	$\Delta = 1$, (50 × 50)	$\Delta = 0.5$, (100 × 100)	$\Delta = 0.25$, (200 × 200)
Sun Sparc SLC	7.43	25.42	78.8
Sun Sparc 2	2.29	7.81	23.13
Ardent Titan 2X P1	3.9	13.9	44.81

significant discrepancies when compared with the Runge–Kutta calculation, and the energy for this case oscillated in a somewhat regular pattern.

To conclude this section, we report the wall-clock times for three runs of the surface wave equations, as discretized using FPM. The code was written in Fortran 77, using recursion in the iteration procedure. For the size of these runs, the use of recursion was probably marginally slower than having opted for repeated subroutine calls. No machine optimization or floating-point accelerators were used. The time trials were carried out with an initial bottom configuration $f = 0.01X$. All other parameters and physical quantities were the same as those used previously. The domain was a square with 50 units on its side. The wall clock times appear in Table IV, corresponding to the total time required to find the field variables everywhere in the domain.

5. SUMMARY

The model for the formation and evolution of three-dimensional sand ridges on the continental shelf described in [1, 2] has been shown to be adequately discretized using finite difference techniques and fixed point methods. The mass transport equation is implemented by using a standard Lax–Wendroff scheme, while the surface system was discretized using a second-order scheme for the linear part and iterative correction for the nonlinear terms.

The schemes' performance was evaluated in detail. It was found that both schemes are second-order accurate in time and space. As a result of the small number of iterations required in the fixed point procedure, the FPM scheme is also found to be efficient in both storage and speed. The schemes were found to converge as the mesh size was diminished. In order to reduce the phase error and dissipation in the computed solutions, the mesh size must be small and comparable in each dimension. The Lax–Wendroff scheme was found to have significant phase drift, especially when the mesh size is increased. The FPM was shown to have significant diffusion for large grid spacings. Since the model is a nonlinear hyperbolic system this damping will introduce phase errors in the waves, especially if the domain is quite large. An outcome of the evaluation is that confidence in the qualitative outcome of the coupled model can be assured with spatial grids as large as 1 and time grids roughly twice as large, assuming that the extent of the computational grid is in the order of 1000 × 1000 and the number of time integrations is in the mid hundreds.

The size of the solutions to the wave system must be monitored to insure proper convergence of the iteration procedure. Included in this study is a prescription to monitor the stability of the solutions. This condition was monitored a posteriori in all trial runs. The condition poses a restriction on the size of the computed solutions, but it has been found to be large enough to encompass most physically relevant situations.

In order to not introduce unwanted structure in the solution of the surface system due to the boundary conditions, a "zero flux condition" was introduced to handle the boundary conditions on the lateral sides of the domain. The condition amounts to placing Neumann boundary conditions on the lateral sides of the domain, sufficiently far away from the region of interest. The central region is connected smoothly to the lateral swaths of computational space. In the lateral swaths the three-dimensionality of the solutions is gradually collapsed into two dimensions. For domains that are very long in the X direction in which considerable refraction in the waves is possible, the computational domain must be supplemented with fairly significant auxiliary computational swaths in order to avoid the effects from the lateral hard barriers on the solution, making the calculation more expensive. Nevertheless, it was preferred over other alternatives that would complicate the problem or pose severe symmetry conditions on the solutions.

APPENDIX

The following are real constants associated with Eq. (6):

$$\begin{aligned} K_1 &= F_1 \\ K_2 &= F_2 \\ K_3 &= D_1 E_1 \\ K_4 &= D_2 E_2 \\ K_5 &= D_1 S_1 \\ K_6 &= D_2 S_2 \end{aligned} \tag{68}$$