Finite-element simulation of the shallow-water equations model on a limited-area domain

I. M. Navon

National Research Institute for Mathematical Sciences, CSIR, PO Box 395, Pretoria, 0001 South Africa

(Received July 1978; revised February 1979)

A Galerkin finite-element model of the shallow-water equations on a limited domain is presented.

The evolutionary equations of continuity and momentum are coupled at each time step using an extrapolated Crank–Nicolson method to quasilinearize the nonlinear advective terms. The coupling allows time steps to be used larger than those possible with an uncoupled model.

A linear one-dimensional stability analysis of the finite-element model is presented. Three mass matrix schemes, the consistent mass (CM), the lumped mass (LM) and a generalized mixed mass (GMM) scheme were used for numerical tests and for comparing the accuracy of the finite-element model both against a refined mesh solution and a highly accurate nonlinear ADI finite difference model when the same test problem was solved. The accuracy of the GMM mass matrix scheme was found to be greater than that of both the LM and CM schemes.

Integral invariants of the shallow-water equations conserved almost perfectly for long-term runs. Extensive comparisons with results of other investigators using two different initial conditions for the shallow-water equations showed the results with the GMM mass scheme to have fourth-order accuracy in both amplitude and phase.

A compact storage scheme is provided in which advantage has been taken of the sparsity of the global matrices.

Introduction

The shallow-water equations are used when tidal effects and surface runoff are modelled; they can also be used in numerical weather prediction to study large-scale waves in the atmosphere and ocean. In this latter domain they are often called the barotrophic primitive equations and are frequently used to test new numerical schemes.

Galerkin finite-element techniques have been applied to the shallow-water equations by many writers (see references 1–13).

Here we are concerned with the solution of the evolutionary shallow-water equations for a limited-area domain on a β-plane.

A Galerkin finite-element method (FEM) is employed for the space discretization using three-noded triangular finite elements, while a time-extrapolated Crank–Nicolson numerical time integration scheme is employed to quasilinearize the nonlinear advective terms.

We here describe three finite element models differing in the treatment of the mass matrix. Special consideration has been given to the accuracy of the various models, and their accuracy is compared with that of a highly accurate nonlinear ADI finite-difference method, as well as by integrating the same models with double resolution in both space dimensions.

0307-904X/79/050337-12/$02.00

To obtain an estimate of the behaviour of the numerical scheme a linear stability analysis is performed on a similar linearized model.

Results of short-term and long-term numerical test calculations on a rectangular domain using a regular grid are compared and discussed.

Finally conclusions are drawn, based on numerical experience with this model, and suggestions made regarding areas for further research.

Shallow-water equations

The primitive equations describing divergent barotropic motion in an incompressible inviscid fluid with a free surface are often called the shallow-water equations.

Using a Cartesian coordinate system with the x-axis running from West to East and the y-axis from South to North, the equations for the model can be written as follows:

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}+\frac{\partial \phi}{\partial x}-fv=0 \tag{1a}$$

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+\frac{\partial \phi}{\partial y}+fu=0 \tag{1b}$$

$$\frac{\partial \phi}{\partial t}+\frac{\partial(\phi u)}{\partial x}+\frac{\partial(\phi v)}{\partial y}=0 \tag{1c}$$

$$0\leqslant x\leqslant L \qquad 0\leqslant y\leqslant D \qquad t\geqslant 0$$

here L and D are the dimensions of a rectangular domain of area $A = LD$.

Here u and v are the velocity components in the x- and y-directions respectively; $\phi = gh$ is the geopotential; h is the depth of the fluid; g is the acceleration of gravity; and f is the Coriolis parameter, required when we consider a fluid in a rotating frame of reference.

The Coriolis term f is given by:

$$f=\hat{f}+\beta(y-D/2) \qquad \beta=\frac{\partial f}{\partial y} \tag{2}$$

with $\hat{f}$ and β constants.

Boundary and initial conditions

The solution of equations (1a) to (1c) requires a knowledge of the corresponding boundary and initial conditions.

Periodic boundary conditions are assumed in the x-direction, while in the y-direction the boundary condition is:

$$v(x,0,t)=v(x,D,t)=0 \tag{3}$$

With these boundary conditions and with the initial condition:

$$w(x,y,0)=\varphi(x,y) \tag{4}$$

where w is the vector function:

$$w=(u,v,\phi)^T \tag{5}$$

and $\varphi(x,y)$ is an initial condition to be specified later, the total energy:

$$E=\tfrac{1}{2}\int_0^L\int_0^D (u^2+v^2+\phi)\frac{\phi}{g}\,\mathrm{d}x\,\mathrm{d}y \tag{6}$$

is independent of time.

Also the average value of the height, which is proportional to the total mass:

$$\bar{h}=\frac{1}{A}\int_0^L\int_0^D h\,\mathrm{d}x\,\mathrm{d}y \tag{7}$$

is independent of time.

Test problem

The test problem used here for determining the initial conditions is the initial height field condition No. 1 of Grammeltvedt,[14] viz.:

$$h(x,y)=H_0+H_1\tanh\left(\frac{9(D/2-y)}{2D}\right)+H_2\operatorname{sech}^2\left(\frac{9(D/2-y)}{D}\right)\sin\left(\frac{2\pi x}{L}\right) \tag{8a}$$

The initial velocity fields were derived from the initial height field using the geostrophic relationship:

$$u=-\left(\frac{g}{f}\right)\frac{\partial h}{\partial y} \qquad v=\left(\frac{g}{f}\right)\frac{\partial h}{\partial x} \tag{9}$$

The constants used were:

$$\begin{array}{ll} L = 4400\,\text{km} & g = 10\,\text{m s}^{-2} \\ D = 6000\,\text{km} & H_0 = 2000\,\text{m} \\ \bar{f} = 10^{-4}\,\text{s}^{-1} & H_1 = 220\,\text{m} \\ \beta = 1.5\times10^{-11}\,\text{s}^{-1}\text{m}^{-1} & H_2 = 133\,\text{m} \end{array} \tag{10}$$

The space increments used were:

$$\Delta x=\Delta y=400\,\text{km} \tag{11}$$

while the time increments varied between $\Delta t = 900$ s and $\Delta t = 2700$ s.

Another initial height field condition, i.e. initial condition II of Grammeltvedt[14] viz.:

$$h(x,y)=H_0+H_1\tanh\left(\frac{9(D/2-y)}{2D}\right)+H_2\operatorname{sech}^2\left(\frac{9(D/2-y)}{D}\right)\times\left[0.7\sin\frac{2\pi x}{L}+0.6\sin\left(\frac{6\pi x}{L}\right)\right] \tag{8b}$$

was also experimented with.

Initial condition (I) initially has energy in wave number one in the x-direction, whereas initial condition (II) initially contains energy in wave numbers one and three in the x-direction.

Initial condition (II) was employed by Gerrity *et al.*[43] with a fourth-order accurate space differencing scheme and by Cullen[6] with a finite element scheme and thus provides a basis for comparison.

Formulation of finite-element model

We approximate the shallow-water equations model (equations 1a–1c) by the Galerkin FEM. The rectangular domain is subdivided into triangular elements forming a regular grid. Linear piecewise polynomial interpolation functions were employed to save computing time and also for the sake of simplicity. Over a given triangular element, each variable was represented as a linear sum of the interpolation functions, i.e.:

$$u_{\mathrm{el}} = \sum_{j=1}^{3} u_j(t)\, v_j \qquad (12)$$

where $u_j(t)$ represents the scalar nodal value of the variable u at the node j of the triangular element and v_j is the basis function (interpolation function) which can be defined by the coordinates of the nodes.

The advection terms in the continuity equation (1c) are usually integrated by parts (using Green's theorem) to shift from derivatives of the variable to derivatives of the basis function.

This permits the use of basis functions with lower-order interelement continuity and often offers a convenient way of introducing the natural boundary conditions that must be satisfied on some portion of the boundary.

This integration gives:

$$\left\langle \frac{\partial \phi}{\partial t}, V_i \right\rangle + \int (\phi u V_i)_x \Big|_0^L \mathrm{d}y - \left\langle (\phi u), \frac{\partial V_i}{\partial x} \right\rangle + \int (\phi v V_i)_y \Big|_0^D \mathrm{d}x - \left\langle (\phi v), \frac{\partial V_i}{\partial y} \right\rangle = 0 \qquad (13)$$

where the notation:

$$\langle f(x,y), V_i \rangle = \sum^{M}_{\text{element}} \iint f(x,y)\, V_i \,\mathrm{d}x\,\mathrm{d}y = \iint_{\text{global}} f(x,y)\, V_i \,\mathrm{d}x\,\mathrm{d}y \qquad (14)$$

defines the inner product when a function is multiplied by the trial function. Taking into account the cyclic boundary conditions in the x-direction and the boundary condition on v, the component of velocity in the y-direction, the second and fourth terms of equation (13) vanish.

The final expression for the continuity equation is:

$$\left\langle \frac{\partial \phi}{\partial t}, V_i \right\rangle - \left\langle (\phi u), \frac{\partial V_i}{\partial x} \right\rangle - \left\langle (\phi v), \frac{\partial V_i}{\partial y} \right\rangle = 0 \qquad (15)$$

Following the Galerkin FEM, the momentum equations (1a) and (1b) become:

$$\left\langle \frac{\partial u}{\partial t}, V_i \right\rangle + \left\langle u\frac{\partial u}{\partial x}, V_i \right\rangle + \left\langle v\frac{\partial u}{\partial y}, V_i \right\rangle - \langle fv, V_i \rangle + \left\langle \frac{\partial \phi}{\partial x}, V_i \right\rangle = 0 \qquad (16)$$

$$\left\langle \frac{\partial v}{\partial t}, V_i \right\rangle + \left\langle u\frac{\partial v}{\partial x}, V_i \right\rangle + \left\langle v\frac{\partial v}{\partial y}, V_i \right\rangle + \langle fu, V_i \rangle + \left\langle \frac{\partial \phi}{\partial y}, V_i \right\rangle = 0 \qquad (17)$$

We assume that over an element the same basis functions V apply for the u, v, unknowns, i.e.:

$$u = \sum_{j=1}^{3} u_j(t)\, V_j$$

$$v = \sum_{j=1}^{3} v_j(t)\, V_j \qquad (18)$$

$$\phi = \sum_{j=1}^{3} \phi_j(t)\, V_j$$

where $u_j(t)$, $v_j(t)$, $\phi_j(t)$ are the time-dependent nodal values of the variables u, v, ϕ respectively.

Upon substituting these expressions into equations (15) to (17) one obtains:

$$\left\langle \frac{\partial \phi_j}{\partial t} V_j, V_i \right\rangle - \left\langle \phi_j u_k V_j V_k, \frac{\partial V_i}{\partial x} \right\rangle - \left\langle \phi_j v_k V_j V_k, \frac{\partial V_i}{\partial y} \right\rangle = 0 \qquad (19)$$

$$\left\langle \frac{\partial u_j}{\partial t} V_j, V_i \right\rangle + \left\langle u_k V_k u_j \frac{\partial V_j}{\partial x}, V_i \right\rangle + \left\langle v_k V_k u_j \frac{\partial V_j}{\partial y}, V_i \right\rangle - \langle f v_k V_k, V_i \rangle + \left\langle \phi_k \frac{\partial V_k}{\partial x}, V_i \right\rangle = 0 \qquad (20)$$

$$\left\langle \frac{\partial v_j}{\partial t} V_j, V_i \right\rangle + \left\langle u_k V_k v_j \frac{\partial V_j}{\partial x}, V_i \right\rangle + \left\langle v_k V_k v_j \frac{\partial V_j}{\partial y}, V_i \right\rangle + \langle f u_k V_k V_i \rangle + \left\langle \phi_k \frac{\partial V_k}{\partial y}, V_i \right\rangle = 0 \qquad (21)$$

Time integration

The time-extrapolated Crank–Nicolson method was used to integrate in time the system of ordinary differential equations resulting from the application of the Galerkin FEM to the shallow-water equations model.

In this method, previously used by Douglas and Dupont,[15] Wang *et al.*,[2] Neuman[16] and Hinsman,[12] an average is taken at time levels N and $N+1$ of the expressions involving space derivatives, while the nonlinear advective terms are quasilinearized by estimating them at time level $N+½$ using the following second-order approximation in time:

$$\begin{aligned} u^{N+1/2} &= u^* = \tfrac{3}{2}u^N - \tfrac{1}{2}u^{N-1} + 0(\Delta t^2) \\ v^{N+1/2} &= v^* = \tfrac{3}{2}v^N - \tfrac{1}{2}v^{N-1} + 0(\Delta t^2) \\ \phi^{N+1/2} &= \phi^* = \tfrac{3}{2}\phi^N - \tfrac{1}{2}\phi^{N-1} + 0(\Delta t^2) \end{aligned} \qquad (22)$$

The shallow-water equations system was coupled at every time step, i.e. the solution of each equation after one iteration at a given time step was used to solve the other two equations for the same iteration for the same time step.

It was found experimentally that coupling the equations makes it possible to extend the allowable time step, in contrast to an uncoupled system.

The advantage of coupling the three equations at any one time step would be that the equations would be more accurate and consistent and larger time steps would be possible.[12]

Upon introducing time discretization into the continuity equation (19), which is the first to be solved at a given time step, one obtains:

$$\begin{aligned}&\langle(\phi_j^{n+1}-\phi_j^n)V_j, V_i\rangle\\&-\frac{\Delta t}{2}\left[\left\langle\phi_j^{n+1}u_k^*V_jV_k, \frac{\partial V_i}{\partial x}\right\rangle+\left\langle\phi_j^{n+1}v_k^*V_jV_k, \frac{\partial V_i}{\partial y}\right\rangle\right]\\&-\frac{\Delta t}{2}\left[\left\langle\phi_j^{n}u_k^*V_jV_k, \frac{\partial V_i}{\partial x}\right\rangle+\left\langle\phi_j^{n}v_k^*V_jV_k, \frac{\partial V_i}{\partial y}\right\rangle\right]=0\end{aligned}\tag{23}$$

By defining the matrices:

$$M=\iint V_jV_i\,\mathrm{d}A$$

and:

$$K_1=\iint V_jV_ku_k^*\frac{\partial V_i}{\partial x}\,\mathrm{d}A+\iint_A V_jV_kv_k^*\frac{\partial V_i}{\partial y}\,\mathrm{d}A\tag{24}$$

the continuity equation can be written as:

$$M(\phi_j^{n+1}-\phi_j^n)-\frac{\Delta t}{2}K_1(\phi_j^{n+1}+\phi_j^n)=0\tag{25}$$

Introducing time discretization in the same way into the momentum equations (20) and (21), one obtains:

$$\begin{aligned}&\langle(u_j^{n+1}-u_j^n)V_j, V_i\rangle+\frac{\Delta t}{2}\left[\left\langle u_j^{n+1}u_k^*V_k\frac{\partial V_j}{\partial x}, V_i\right\rangle\right.\\&\left.+\left\langle u_j^{n+1}v_k^*V_k\frac{\partial V_j}{\partial y}, V_i\right\rangle+\left\langle\phi_k^{n+1}\frac{\partial V_k}{\partial x}, V_i\right\rangle\right]\\&+\frac{\Delta t}{2}\left[\left\langle u_j^nu_k^*V_k\frac{\partial V_j}{\partial x}, V_i\right\rangle+\left\langle u_j^nv_k^*V_k\frac{\partial V_j}{\partial y}, V_i\right\rangle\right.\\&\left.+\left\langle\phi_k^n\frac{\partial V_k}{\partial x}, V_i\right\rangle\right]-\Delta t\langle fv_k^*V_k, V_i\rangle=0\end{aligned}\tag{26}$$

and:

$$\begin{aligned}&\langle(v_j^{n+1}-v_j^n)V_j, V_i\rangle+\frac{\Delta t}{2}\left[\left\langle v_j^{n+1}u_k^{n+1}V_k\frac{\partial V_j}{\partial x}, V_i\right\rangle\right.\\&\left.+\left\langle v_j^{n+1}v_k^*V_k\frac{\partial V_j}{\partial x}, V_i\right\rangle+\left\langle\phi_k^{n+1}\frac{\partial V_k}{\partial y}, V_i\right\rangle\right]\\&+\frac{\Delta t}{2}\left[\left\langle v_j^nu_k^{n+1}V_k\frac{\partial V_j}{\partial x}, V_i\right\rangle+\left\langle v_j^nv_k^*V_k\frac{\partial V_j}{\partial y}, V_i\right\rangle\right.\\&\left.+\left\langle\phi_k^n\frac{\partial V_k}{\partial y}, V_i\right\rangle\right]+\Delta t\langle fu_k^{n+1}V_k, V_i\rangle=0\end{aligned}\tag{27}$$

Using the matrix definitions:

$$\begin{aligned}K_2&=\iint_A u_kV_kV_i\frac{\partial V_j}{\partial x}\,\mathrm{d}A+\iint_A v_kV_kV_i\frac{\partial V_j}{\partial y}\,\mathrm{d}A\\K_{21}&=\iint_A\phi_k\frac{\partial V_k}{\partial x}V_i\,\mathrm{d}A\\P_2&=-\iint_A fv_k^*V_kV_i\,\mathrm{d}A\end{aligned}\tag{28}$$

the u-momentum equation becomes:

$$\begin{aligned}&M(u_j^{n+1}-u_j^n)+\frac{\Delta t}{2}K_2(u_j^{n+1}+u_j^n)\\&\qquad+\frac{\Delta t}{2}(K_{21}^{n+1}+K_{21}^n)+\Delta tP_2=0\end{aligned}\tag{29}$$

where:

$$K_{21}^{n+1}=\iint_A\phi_k^{n+1}\frac{\partial V_k}{\partial x}V_i\,\mathrm{d}A\quad\text{etc.}\tag{30}$$

while by defining:

$$\begin{aligned}K_3&=\iint_A u_k^{n+1}V_k\frac{\partial V_j}{\partial x}V_i\,\mathrm{d}A+\iint_A v_kV_k\frac{\partial V_j}{\partial y}\,\mathrm{d}A\\K_{31}&=\iint_A\phi_k\frac{\partial V_k}{\partial y}V_i\,\mathrm{d}A\\P_3&=\iint_A fu_k^{n+1}V_kV_i\,\mathrm{d}A\end{aligned}\tag{31}$$

The v-momentum equation becomes:

$$\begin{aligned}&M(v_j^{n+1}-v_j^n)+\frac{\Delta t}{2}K_3(v_j^{n+1}+v_j^n)\\&\qquad+\frac{\Delta t}{2}(K_{31}^{n+1}+K_{31}^n)+\Delta tP_3=0\end{aligned}\tag{32}$$

In order to implement boundary conditions in the Galerkin FEM we have here adopted an approach suggested by Payne and Irons[18] and mentioned by Huebner.[17] This approach consists of modifying the diagonal terms of the global matrix associated with the nodal variables by multiplying them by a large number, say 10^{16} (chosen with a view to the significant number of digits possible with the given computer and the size of the field variables), while the corresponding term in the right-hand vector R in the system of linear equations:

$$KX=R\tag{33}$$

where K is the global matrix, is replaced by the specified boundary nodal variable multiplied by the same large factor times the corresponding diagonal term. This procedure is repeated until all prescribed boundary nodal variables have been treated.

After these modifications have been made, it is possible to proceed with the solution of the set of equations, using the modified matrix K and the modified vector R.

For instance, if in the matrix K we wish to implement the boundary condition:

$$X_r=\beta_r$$

Then the modification is:

$$\begin{bmatrix}k_{11}&k_{12}&&k_{1N}\\&\ddots&&\\\vdots&&&\\k_{r1}&k_{r2}&k_{rr}10^{16}&k_{rN}\\&&\ddots&\\k_{N1}&k_{N2}&&k_{NN}\end{bmatrix}\begin{bmatrix}x_1\\x_2\\\vdots\\x_r\\\vdots\\x_N\end{bmatrix}=\begin{bmatrix}R_1\\R_2\\\vdots\\\beta_rk_{rr}10^{16}\\\vdots\\R_N\end{bmatrix}$$

If the rth equation is then considered it can be observed that the desired boundary condition has been imposed as:

$$k_{r1}x_1 + k_{r2}x_1 + \ldots + k_{rr}10^{16}x_r + \ldots + k_{rN}x_N = \beta_r k_{rr}10^{16}$$

i.e.

$$x_r = \beta_r$$

Since:

$$k_{ri} \leqslant k_{rr}10^{16} \qquad i = 1, 2, \ldots, N \qquad i \neq r$$

The global ($N \times N$) coefficient matrix generated by the assembly process is very sparse, as the maximum number of triangles incident on any one point is six. Therefore each row in the global $N \times N$ matrix has at most seven entries and it is an advantage to store the global matrix in a compact manner to save fast-core storage. An efficient scheme was devised to compact the ($N \times N$) matrix into an ($N \times 7$) matrix (see also Hinsman[12] and Navon and Müller[19]).

The method adopted in this paper for solving the system of linear equations was the iterative one of Gauss–Seidel which has the virtue of simplicity and requires only diagonal dominance of the coefficient matrix.

Finite element method with the CM, LM and GMM mass matrix schemes

In the previous section we saw that the application of the Galerkin FEM to the shallow-water equations model reduced the problem to solving a set of matrix equations whose term involving derivatives of time is the mass matrix M (equation (24)).

Using linear basis functions over triangular elements and introducing the well known area coordinates[20] one can obtain exact integrations using the following formula for area integrals[21]:

$$\iint_A L_1^a L_2^b L_3^c \, \mathrm{d}x \, \mathrm{d}y = \frac{a!b!c!}{(a+b+c+2)!} \qquad (34)$$

The mass-element (3 x 3) matrix is then:

$$M_{el} = \frac{A}{12}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \qquad (35)$$

where A is the area of the element triangle.

The assembled mass matrix is called the consistent mass matrix M. A lumped-mass element matrix M_L is one in which the mass of the elements is equally distributed at the three corner nodes. By lumping the element mass matrix before assembling the elements, a diagonal global mass matrix M_L is obtained.

The convenience of employing a lumped-mass system is that M_L is a diagonal matrix and its inverse is immediate. Ishara[22] proposed a generalized mixed-mass (GMM) scheme for a second-order hyperbolic wave equation. He defines the GMM mass matrix as:

$$M_G = \alpha M + (1 - \alpha) M_L \qquad (36)$$

where α is a parameter such that:

$$0 \leqslant \alpha \leqslant 1$$

The GMM scheme includes the CM scheme ($\alpha = 1$) and the LM scheme ($\alpha = 0$).

All three mass schemes were used in the numerical experiments in which the accuracy of each scheme was compared with that of a highly accurate nonlinear ADI finite difference model due to Gustafsson[23] when the same test problem was solved, as well as by integrating the same models with double resolution in both space dimensions.

Lumped mass matrices – convergence and accuracy considerations

The use of lumped or diagonal mass matrices has been first adopted for its considerable computational convenience. Key and Beisinger,[44] Hinton *et al.*[45] among others have experimented with different lumping schemes, which often give improved results over those attainable with consistent mass matrix formulations.[42] Fried[37] and Fried and Malkus[39] have shown such schemes for several finite elements and demonstrated not only that convergence order is maintained, but that the accuracy is often improved.

An important theorem concerning the order of numerical integration which does not affect the convergence rate when using lumped mass matrices states that if p is the order of polynomials used in the shape functions and m the order of differentiation present in the variational functional, then any integration exact to the order of $2(p - m)$ will not affect the rate of convergence.

If thus an integration scheme which uses only nodal points for sampling is devised and which possesses the correct order of integration, then the lumping process will not affect the convergence rate.

Fujii[31] as well as Oden and Fost[36] show that for nonlinear hyperbolic equations the use of the lumped mass formulation results, for regular space grids, in an increase of $\sqrt{3}$ in the time-step allowed by stability criteria of Courant–Friedrichs–Levy (CFL), while the same rates of convergence for the consistent mass formulation are also obtained for the lumped mass formulation. Tong[33] has observed the added stability with lumping for hyperbolic problems.

Mock[34] observed that in hyperbolic problems it is the direction rather than the magnitude of the lumping perturbation which is important and lumping is intimately related to the stability of the methods we construct. He also showed that lumping the mass matrix is achieved by the addition of a differential operator which for smooth splines is dissipative and strongly enhances the stability properties of the discretization scheme. This is achieved by broadening the domain of dependence of the discrete solution, which, in view of the Courant–Friedrichs–Levy criteria, is also in the direction of increasing the stability.

The same approach was used by Holtz.[40] Schreyer[35] proposed a new approach to obtain consistent mass matrices through the combined use of orthogonal base functions and a mixed variational formulation.

Linear stability analysis

In order to gain some insight into the behaviour of our finite-element numerical solution of the shallow-water equations model we shall examine the FEM discretized equations of a one-dimensional simple system with gravity waves, i.e.:

$$\begin{aligned} \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial \phi}{\partial x} &= 0 \\ \frac{\partial \phi}{\partial t} + u\frac{\partial \phi}{\partial x} + \phi\frac{\partial u}{\partial x} &= 0 \end{aligned} \qquad (37)$$

A linearized form of equations (37) obtained by applying a perturbation technique, is:

$$\frac{\partial u}{\partial t} + U\frac{\partial u}{\partial x} + \frac{\partial \phi}{\partial x} = 0$$

$$\frac{\partial \phi}{\partial t} + U\frac{\partial u}{\partial x} + \bar{\phi}\frac{\partial u}{\partial x} = 0 \tag{38}$$

where U is the constant basic flow speed and $\bar{\phi}$ the mean geopotential.

Using a regular one-dimensional finite-element grid in space, with linear basis functions, the expansions for ϕ and u in terms of the basis functions can be made such that:

$$u \simeq \sum_{j=1}^{M} u_j(t)V_j(x)$$

$$\phi \simeq \sum_{j=1}^{M} \phi_j(t)V_j(x) \tag{39}$$

where $u_j(t)$ and $\phi_j(t)$ are the approximations to u and ϕ respectively at node j and time t, and $V_j(x)$ is the basis function associated with node j.

The nodes are assumed to have been numbered consecutively, with node $j+1$ adjacent to node j in the positive x-direction. Application of Galerkin's method to the system (38) by weighting with respect to the ith basis function, yields:

$$\sum_{j=1}^{M}\int_x \left[\frac{\mathrm{d}\phi_j}{\mathrm{d}t}V_jV_i + U\phi_j\frac{\mathrm{d}V_j}{\mathrm{d}x}V_i + \bar{\phi}u_j\frac{\mathrm{d}V_j}{\mathrm{d}x}V_i\right]\mathrm{d}x = 0 \tag{40}$$

$$\sum_{j=1}^{M}\int_x \left[\frac{\mathrm{d}u_j}{\mathrm{d}t}V_jV_i + Vu_j\frac{\mathrm{d}V_j}{\mathrm{d}x}V_i + \phi_j\frac{\mathrm{d}V_j}{\mathrm{d}x}V_i\right]\mathrm{d}x = 0 \tag{41}$$

Denoting the length of each element by Δx, the various integrals are non-zero only for $j = i-1, i$ or $i+1$, and integrating we obtain:

$$\frac{1}{6}\left(\frac{\mathrm{d}\phi_{i-1}}{\mathrm{d}t} + 4\frac{\mathrm{d}\phi_i}{\mathrm{d}t} + \frac{\mathrm{d}\phi_{i+1}}{\mathrm{d}t}\right) + U\frac{\phi_{i+1}-\phi_{i-1}}{2\Delta x} + \bar{\phi}\frac{u_{i+1}-u_{i-1}}{2\Delta x} = 0 \tag{42}$$

$$\frac{1}{6}\left(\frac{\mathrm{d}u_{i-1}}{\mathrm{d}t} + 4\frac{\mathrm{d}u_i}{\mathrm{d}t} + \frac{\mathrm{d}u_{i+1}}{\mathrm{d}t}\right) + U\frac{u_{i+1}-u_{i-1}}{2\Delta x} + \frac{\phi_{i+1}-\phi_{i-1}}{2\Delta x} = 0 \tag{43}$$

Introducing the extrapolated Crank–Nicolson time-differencing scheme while the time derivatives are finite-differenced over the time step Δt, we have (taking into account coupling between equations (42) and (43)):

$$\frac{1}{6}\left[\frac{\phi_{i-1}^{n+1}-\phi_{i-1}^{n}}{\Delta t} + 4\frac{\phi_i^{n+1}-\phi_i^{n}}{\Delta t} + \frac{\phi_{i+1}^{n+1}-\phi_{i+1}^{n}}{\Delta t}\right] + \frac{U}{2}\left[\frac{\phi_{i+1}^{n+1}-\phi_{i-1}^{n+1}}{2\Delta x} + \frac{\phi_{i+1}^{n}-\phi_{i-1}^{n}}{2\Delta x}\right] + \frac{\bar{\phi}}{2}\left[\frac{(3u_{i+1}^{n}-u_{i+1}^{n-1})-(3u_{i-1}^{n}-u_{i-1}^{n-1})}{2\Delta x}\right] = 0 \tag{44}$$

$$\frac{1}{6}\left[\frac{u_{i-1}^{n+1}-u_{i-1}^{n}}{\Delta t} + 4\frac{u_i^{n+1}-u_i^{n}}{\Delta t} + \frac{u_{i+1}^{n+1}-u_{i+1}^{n}}{\Delta t}\right] + \frac{U}{2}\left[\frac{u_{i+1}^{n+1}-u_{i-1}^{n+1}}{2\Delta x} + \frac{u_{i+1}^{n}-u_{i-1}^{n}}{2\Delta x}\right] + \frac{1}{2}\left[\frac{\phi_{i+1}^{n+1}-\phi_{i-1}^{n+1}}{2\Delta x} + \frac{\phi_{i+1}^{n}-\phi_{i-1}^{n}}{2\Delta x}\right] = 0 \tag{45}$$

For u and ϕ we take Fourier components:

$$u = u^0 \exp(\mathrm{i}\omega t + \mathrm{i}kx) \tag{46}$$

$$\phi = \phi^0 \exp(\mathrm{i}\omega t + \mathrm{i}kx) \tag{47}$$

where u^0 and ϕ^0 are the amplitudes; $k = 2\pi/L$ is the wave number; L is the wave length; $\omega = kc$ is the frequency and c is the phase speed.

Inserting (46) and (47) into equations (44) and (45) (and using the notation $\exp \mathrm{i}\omega t = \lambda$), one obtains:

$$\frac{1}{6\Delta t}(\lambda-1)[\exp(-\mathrm{i}k\Delta x)+4+\exp(\mathrm{i}k\Delta x)]\phi^0 + \frac{U}{4\Delta x}(\lambda+1)[\exp(\mathrm{i}k\Delta x)-\exp(-\mathrm{i}k\Delta x)]\phi^0 + \frac{\bar{\phi}}{4\Delta x}(3-\lambda^{-1})[\exp(\mathrm{i}k\Delta x)-\exp(-\mathrm{i}k\Delta x)]u^0 = 0 \tag{48}$$

$$\frac{1}{6\Delta t}(\lambda-1)[\exp(-\mathrm{i}k\Delta x)+4+\exp(\mathrm{i}k\Delta x)]u^0 + \frac{U}{4\Delta x}(\lambda+1)[\exp(\mathrm{i}k\Delta x)-\exp(-\mathrm{i}k\Delta x)]u^0 + \frac{1}{4\Delta x}(\lambda+1)[\exp(\mathrm{i}k\Delta x)-\exp(-\mathrm{i}k\Delta x)]\phi^0 = 0 \tag{49}$$

Using basic trigonometric identities, equations (48) and (49) can be simplified to take the form:

$$\frac{1}{6\Delta t}(\lambda-1)[2\cos(k\Delta x)+4]\phi^0 + \frac{U}{4\Delta x}(\lambda+1)[2\mathrm{i}\sin k\Delta x]\phi^0 + \frac{\bar{\phi}}{4\Delta x}[(3-\lambda^{-1})\,2\mathrm{i}\sin k\Delta x]u^0 = 0 \tag{50}$$

$$\frac{1}{6\Delta t}(\lambda-1)[2\cos(k\Delta x)+4]u^0 + \frac{U}{4\Delta x}(\lambda+1)[2\mathrm{i}\sin k\Delta x]u^0 + \frac{1}{4\Delta x}(\lambda+1)[2\mathrm{i}\sin k\Delta x]\phi^0 = 0 \tag{51}$$

There are two equations in the two coefficients ϕ^0 and u^0, which can be eliminated to obtain an expression for λ which is the eigenvalue of the amplification matrix. The well known Von Neumann necessary condition for stability states that for all wave numbers the eigenvalues λ_i of the amplification matrix must satisfy:

$$|\lambda_i| \leqslant 1 + \mathrm{O}(\Delta t)$$

By equating the determinant of the two equations to zero, i.e.:

$$\begin{vmatrix} \frac{1}{3\Delta t}(\lambda-1)(\cos(k\Delta x)+2) + \frac{U}{2\Delta x}(\lambda+1)\,\mathrm{i}\sin k\Delta x & \frac{\bar{\phi}}{2\Delta x}(3-\lambda^{-1})\,\mathrm{i}\sin k\Delta x \\ \frac{1}{2\Delta x}(\lambda+1)\,\mathrm{i}\sin k\Delta x & \frac{1}{3\Delta t}(\lambda-1)(\cos(k\Delta x)+2) + \frac{U}{2\Delta x}(\lambda+1)\,\mathrm{i}\sin k\Delta x \end{vmatrix} = 0 \qquad (52)$$

a complex cubic equation for λ – the eigenvalues of the amplification matrix – is obtained as follows:

$$\begin{aligned} &\lambda^3\left[\frac{4\Delta x^2}{9\Delta t^2}(\cos(k\Delta x)+2)^2 - U^2\sin^2 k\Delta x + \mathrm{i}U\frac{4\Delta x}{3\Delta t}(\cos k\Delta x\,\sin k\Delta x + 2\sin k\Delta x)\right] \\ &+\lambda^2\left[\frac{-8\Delta x^2}{9\Delta t^2}(\cos(k\Delta x)+2)^2 - 2U^2\sin^2 k\Delta x + 3\bar{\phi}\sin^2 k\Delta x\right] \\ &+\lambda\left[\frac{4\Delta x^2}{9\Delta t^2}(\cos(k\Delta x)+2)^2 - \mathrm{i}U\frac{4\Delta x}{3\Delta t} \times(\cos k\Delta x\,\sin k\Delta x + 2\sin k\Delta x) - U^2\sin^2 k\Delta x + 2\bar{\phi}\sin^2 k\Delta x\right] - \bar{\phi}\sin^2 k\Delta x = 0 \qquad (53) \end{aligned}$$

This equation was solved numerically for the roots λ using a computer subroutine (CO 2A DF OF NAG library, Vol. 1) and various values of the wavelength L ranging from 100 km to 5000 km. The time step employed was Δt = 1800 s, while the constants U and $\bar{\phi}$ were:

$$U = 30\,\mathrm{m\,s^{-2}} \qquad \bar{\phi} = 2\times 10^4\,\mathrm{m^2\,s^{-2}} \qquad (54)$$

The space interval was Δx = 400 km. The results are presented in *Table 1*. Here the solutions associated with λ_1 and λ_2 are physical modes, while the solution associated with λ_3 is a computational mode.

The results suggest a numerical behaviour similar to that when the Adams–Bashforth time-differencing scheme is used, i.e.:

$$U^{n+1} = U^{(n)} + \Delta t(\tfrac{3}{2}f^{(n)} - \tfrac{1}{2}f^{(n-1)}) \qquad (55)$$

when the equation:

$$\frac{\mathrm{d}U}{\mathrm{d}t} = f(U,t) \qquad (56)$$

is solved. If we take $f = \mathrm{i}\omega U$, equation (56) describes the oscillation equation (see Mesinger and Arakawa[24]) and its eigenvalues are:

$$\begin{aligned} |\lambda_1| &= 1 + \tfrac{1}{4}p^4 + \ldots \\ |\lambda_2| &= \tfrac{1}{2}p + \ldots \end{aligned} \qquad (57)$$

$$p = \omega\Delta t = kc\Delta t = \frac{2\pi}{L}c\Delta t \qquad (58)$$

Except for wavelengths less than L = 600 km, the scheme is stable.

A similar analysis was conducted for the lumped-mass scheme. The equations for ϕ^0 and u^0 corresponding to equations (48) and (49) are:

$$(\lambda-1)\phi^0 + \frac{U}{4\Delta x}(\lambda+1)[\exp(\mathrm{i}k\Delta x) - \exp(-\mathrm{i}k\Delta x)]\phi^0 + \frac{\bar{\phi}}{4\Delta x}(3-\lambda^{-1})[\exp(\mathrm{i}k\Delta x) - \exp(-\mathrm{i}k\Delta x)]u^0 = 0 \qquad (59)$$

$$(\lambda-1)u^0 + \frac{U}{4\Delta x}(\lambda+1)[\exp(\mathrm{i}k\Delta x - \exp(-\mathrm{i}k\Delta x)]u^0 + \frac{1}{4\Delta x}(\lambda+1)[\exp(\mathrm{i}k\Delta x) - \exp(-\mathrm{i}k\Delta x)]\phi^0 = 0 \qquad (60)$$

The resulting complex cubic equation for $\lambda = \exp(\mathrm{i}\omega t)$ is:

$$\begin{aligned} &\lambda^3\left[\frac{4\Delta x^2}{\Delta t^2} - U^2\sin^2 k\Delta x + \frac{4\Delta x}{\Delta t}\mathrm{i}U\sin k\Delta x\right] \\ &+\lambda^2\left[\frac{-8\Delta x^2}{\Delta t^2} - 2U^2\sin^2 k\Delta x + 3\bar{\phi}\sin^2 k\Delta x\right] \\ &+\lambda\left[\frac{4\Delta x^2}{\Delta t^2} - U^2\sin^2 k\Delta x - \frac{4\Delta x}{\Delta t}\mathrm{i}U\sin k\Delta x + 2\bar{\phi}\sin^2 k\Delta x\right] - \bar{\phi}\sin^2 k\Delta x = 0 \qquad (61) \end{aligned}$$

The results obtained for various wavelengths L when the same constants were used as for the consistent mass scheme linear analysis, and the same computer subroutine was used to solve numerically for the roots of λ, are summarized in *Table 2*.

The results of a similar analysis of the generalized mixed mass scheme (equation (36)) with the same constants and α = 0.5 are summarized in *Table 3*.

Accuracy tests

The three mass matrix schemes CM, LM and GMM were used for comparing the accuracy of the Galerkin FEM with that of a highly accurate nonlinear ADI finite difference method due to Gustafsson.[23]

Table 1 Variation of eigenvalues of amplification matrix as function of wavelength (L) for consistent mass method

L (km)	λ_1	λ_2	λ_3
100	1.156	1.066	0.1784
200	1.085	1.030	0.1373
300	1.019	1.006	0.06872
400	1.006	1.002	0.03953
500	1.002	1.001	0.02046
600	1.001	1.000	0.01772
700	1.001	1.000	0.01303
800	1.000	1.000	0.009983
900	1.000	1.000	0.007890
1000	1.000	1.000	0.006392
1200	1.000	1.000	0.004440
1500	1.000	1.000	0.002842
2000	1.000	1.000	0.001599
3000	1.000	1.000	0.7106×10^{-3}
4000	1.000	1.000	0.2997×10^{-3}
5000	1.000	1.000	0.2558×10^{-3}

Table 2 Variation of eigenvalues of amplification matrix as function of wavelength (*L*) for lumped mass method

L (km)	λ_1	λ_2	λ_3
100	1.0046	1.0013	0.034722
200	1.0318	1.0098	0.087540
300	1.0118	1.0034	0.054936
400	1.0045	1.0013	0.034722
500	1.0020	1.0005	0.023413
600	1.0010	1.0003	0.016715
700	1.0005	1.0001	0.012484
800	1.0003	1.0000	0.96601×10^{-2}
900	1.0002	1.0000	0.76878×10^{-2}
1000	1.0001	1.0000	0.62590×10^{-2}
1200	1.0000	1.0000	0.43755×10^{-2}
1500	1.0000	1.0000	0.28154×10^{-2}
2000	1.0000	1.0000	0.15903×10^{-2}
3000	1.0000	1.0000	0.708926×10^{-3}
4000	1.0000	1.0000	0.399186×10^{-3}
5000	1.0000	1.0000	0.27732×10^{-3}

Table 3 Variation of eigenvalues of amplification matrix as function of wavelength (*L*) for GMM method

L (km)	λ_1	λ_2	λ_3
100	1.1737	1.0827	0.18880
200	1.4015	1.2654	0.25432
300	1.1858	1.0846	0.19199
400	1.0799	1.0281	0.13341
500	1.0362	1.0112	0.09305
600	1.0180	1.0053	0.06718
700	1.0098	1.0028	0.05032
800	1.0057	1.0016	0.03895
900	1.0036	1.0010	0.03098
1000	1.0022	1.0006	0.02520
1200	1.0011	1.0003	0.01759
1500	1.0004	1.0001	0.01130
2000	1.0001	1.0000	0.6375×10^{-2}
3000	1.0000	1.0000	0.2838×10^{-2}
4000	1.0000	1.0000	0.1597×10^{-2}
5000	1.0000	1.0000	0.1109×10^{-2}

In order to obtain the difference between the true solution and the approximate solution, it was assumed that the true solution of the shallow-water equations model was represented by w_{QN3}, where w is the vector equation given by equation (5) and QN3 is a quasi-Newton method of solution for the nonlinear ADI finite-difference method.[19,23]

Representing the Galerkin FEM solution by w_G the error is given by:

$$\epsilon_G = w_G - w_{QN3} \qquad (62)$$

and the relative error by:

$$\|\epsilon_G\|/\|w_{QN3}\| \qquad (63)$$

where the norm $\|\ \|$ is defined as follows.

Define a Hilbert space H by considering all vector functions satisfying:

$$w_{jk} = w_{j,N_x+k} \qquad v_{j,0} = v_{j,N_y} = 0$$

The inner product of two vectors α, β and the norm is defined by:

$$(\alpha,\beta) = \Delta x\,\Delta y \sum_{j=1}^{N_x} \left\{ \sum_{k=1}^{N_y-1} \alpha_{jk}^T \beta_{jk} + \tfrac{1}{2}(\alpha_{j0}^T\beta_{j0} + \alpha_{jN_y}^T \beta_{jN_y}) \right\}$$
$$\|\alpha\|^2 = (\alpha,\alpha) \qquad (65)$$

where:

$$N_x \Delta x = L \qquad N_y \Delta y = D \qquad (66)$$

and L and D are given by equation (10).

The test problem of equation (8) was now solved, using the coupled Galerkin FEM with the three different mass-matrix schemes and with $\Delta x = \Delta y = 400$ km and a time-step $\Delta t = 30$ min. The comparative results summarized in *Table 4* were then obtained by employing the QN3 nonlinear ADI Gustafsson method with identical data, and integrating for 2 days.

It is evident from the results that the run (LM) – i.e. that in which the masses were lumped – is less accurate than the CM scheme. The accuracy of the generalized mixed mass (GMM) scheme with $\alpha = 0.5$ is, however, greater than that of both the LM and CM schemes. For the sake of comparison, the accuracy is also shown of the result obtained by using the nonlinear ADI finite-difference method with one nonlinear iteration per time-step (QNEX1) and with the LU decomposition of the Jacobian matrix J updated every 12 time-steps.[23,25]

Another set of accuracy tests was conducted by integrating the same finite element models with double resolution in both space dimensions ($\Delta x = \Delta y = 200$ km) and a time-step $t = 15$ min, and assuming the refined grid FEM solution to be the true solution. Representing the coarse mesh Galerkin FEM solution by W_G and the Galerkin FEM refined mesh solution by W_{FG}, the error is given by:

$$\epsilon_{2G} = W_G - W_{FG} \qquad (67)$$

with the norm defined by equation (65).

The comparative results summarized in *Table 5* were then obtained after a 2 days' numerical integration.

Accuracy merits of the GMM scheme – tentative explanation

Although error analyses for the GMM method applied to hyperbolic partial differential equations exist[22,29,30] they are all of the form

$$\|\text{error}\| \leqslant C(h^2 + \Delta t^2) \qquad (68)$$

where C is a constant independent of h and Δt, not known *a priori*, and do not directly suggest an explanation of the fact that the GMM scheme is most accurate.

Table 4 Error between approximate and true solution for different finite-element methods

Method	Relative error $\|\epsilon_G\|/\|w_{QN3}\|$, $t = 2$ days, $\Delta t = 1800$ s, resolution $\Delta x = \Delta y = 400$ km
CM	4.5×10^{-4}
LM	5.2×10^{-4}
GMM ($\alpha = 0.5$)	1.1×10^{-4}
QNEX1 (M = 12)	4.1×10^{-4}

Table 5 Error between approximate and true solution for different finite-element methods

Method	Relative error $\|\epsilon_G\|/\|w_{QN3}\|$, $t = 2$ days, $\Delta t = 900$ s, resolution of fine mesh $\Delta x = \Delta y = 200$ km
CM	3.7×10^{-4}
LM	4.6×10^{-4}
GMM ($\alpha = 0.5$)	0.8×10^{-4}

A tentative explanation is to be found in a survey paper by Morton.[41] Remarking about the significance of the role of the mass matrix in assessment of accuracy, Morton points out that for regular linear elements, the coefficients of the mass matrix $(\frac{1}{6}, \frac{2}{3}, \frac{1}{6})$ correspond to an operator $(1 + \delta_x^2/6)$ acting on U_j where:

$$\delta_x^2 U_j = U_{j+1} - 2U_j + U_{j-1} \tag{69}$$

The operator $(1 + \delta_x^2/6)$ is often inverted by iteration, and gives a Numerov-type scheme which is fourth-order accurate in space. The approximation:

$$\left(1 + \frac{\delta_x^2}{6}\right)^{-1} = \left(1 - \frac{\delta_x^2}{6}\right) + 0(h^4) \tag{70}$$

which is characteristic of fourth-order compact difference schemes is equivalent to a 'half-lumped' mass matrix (Morton[41]). This connects the GMM mass scheme with the fourth-order compact implicit schemes[46,47] and explains its higher accuracy.

It is worthwhile to note at this point that Ishihara[22,29,30] finds that the CM mass scheme gives the upper bound and the LM mass scheme gives the lower bound for the exact values of the solution. The numerical results obtained by Ishihara with $\alpha = 0.5$ for the GMM scheme give approximations located between the CM and LM scheme results. Donea *et al.*[38] solving an advection diffusion problem proposed a two-stage explicit technique which resembles the GMM scheme. In their approach a lumped mass matrix is used to derive a first approximation of the time-derivatives:

$$(\dot{T}_i)_1 = \frac{\bar{F}_i}{\bar{M}_{ii}} \tag{71}$$

where $\bar{M}_{ii}$ is the lumped mass matrix and $\{\bar{F}\}$ are global load nodes accounting for convection, diffusion and boundary contributions.

Then a second approximation is sought by using the consistent mass matrix M_{ij}:

$$(\dot{T}_i)_2 = \frac{\bar{F}_i - \sum\limits_{j \neq i} M_{ij}(\dot{T}_j)_1}{M_{ii}} \tag{72}$$

The final values of the time derivatives are computed as a weighted average of the above approximations:

$$\{\dot{T}\} = \gamma\{\dot{T}_1\} + (1 - \gamma)\{\dot{T}_2\} \tag{73}$$

Donea *et al.*[38] found by a one-dimensional analysis of the numerical phase speeds and using numerical experimentation, that the optimum value of γ is 0.5.

Results

Many tests were run with the three different FEM mass schemes and various time steps. A guideline for the success of the model was the conservation of the two integral invariants of the shallow-water equations model, viz. the total energy and the average height.

We expected an approximate linear stability criterion of the form:

$$c\frac{\Delta t}{\Delta x} \leqslant 0.707$$

due to the Courant–Friedrichs–Levy (CFL) criterion, where c is the phase speed of the fastest gravity waves and Δx the minimum Δx in the finite-element grid:

$$c = \sqrt{gh} = \sqrt{2} \cdot 10^2 \text{ ms}^{-1}$$

and:

$$\Delta x = 400 \text{ km}$$

the maximum allowable time step is 30 min.

The coupled Galerkin FEM using the CM mass scheme and a time step of 40 min became unstable after 48 h but when a time step of 35 min was used, yielded stable integrations for up to 5 days.

Figure 1 shows the initial height field contours drawn at 50 m intervals for initial condition (I).

Figures 2–4 show the height field after 2, 6 and 10 days of simulation respectively, using the CM mass scheme with a time-step of $\Delta t = 2100$ sec.

The coupled Galerkin FEM using the LM scheme with a 50-min time step became unstable after 31 h but yielded stable long-term integrations when the time step was reduced to 45 min.

Figures 5–7 show the height field after 1.5, 3 and 5 days of simulation respectively, using the LM mass matrix scheme with initial condition (I) and a time step of $t = 45$ min. The coupled Galerkin FEM with the GMM mass scheme gave long-term stable integrations only when a time step of 30 min was employed.

The height field after 2 and 5 days of simulation (respectively) using the GMM mass matrix scheme is shown in *Figures 8* and *9* respectively. A time step of $\Delta t = 30$ min was used.

A test was also conducted by running an uncoupled version of the Galerkin FEM. The uncoupled model remained stable with a 15-min time step, but became unstable after 24 h when a 20-min time step was used.

In all cases the onset of instability was marked by a sudden increase in the total energy, and the solutions 'blew up' regardless of the iteration technique.

All the figures in this paper display isoline contour plots of the height field, with a contour interval of 50 m.

Another set of numerical experiments was conducted, using this time the initial height field condition (II) of Grammeltvedt (equation (8b)) and only for the GMM mass scheme.

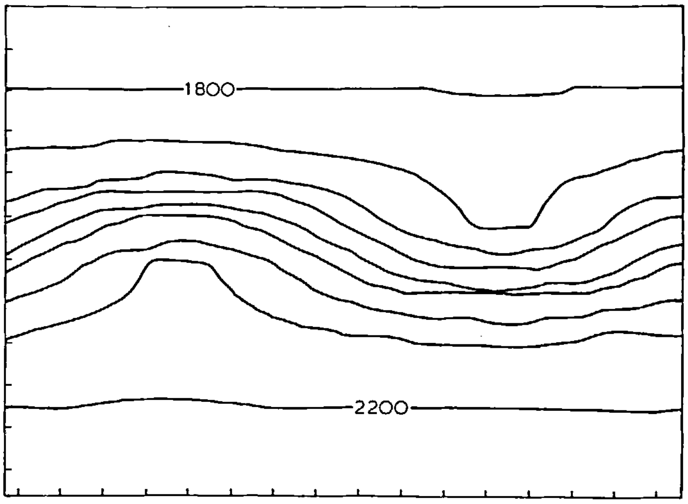

Figure 1 Initial height field contours (every 50 m). $\Delta x = \Delta y = 400$ km; $H_{\text{mean}} = 2000$ m; $E_{\text{tot}} = 6.2504 \times 10^{20}$, CM scheme initial condition (IC) 1

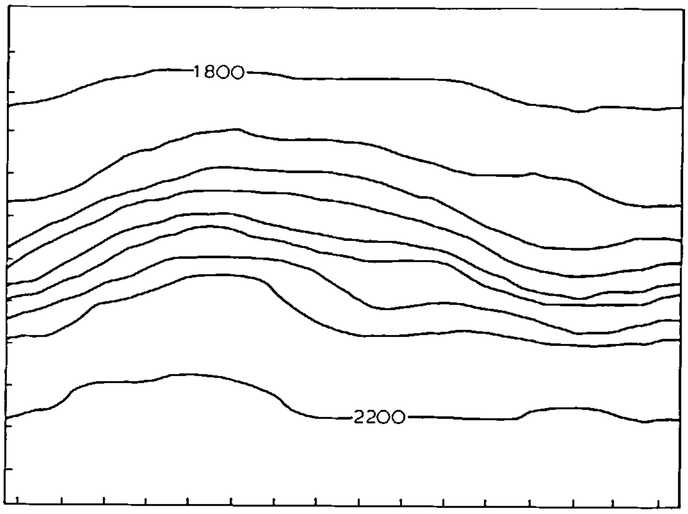

Figure 2 Height field contours after 2 days. $\Delta x = \Delta y = 400$ km; $\Delta t = 1800$ sec; $H_{\text{mean}} = 1998.38$ m; $E_{\text{tot}} = 6.2413 \times 10^{20}$; CM scheme IC 1

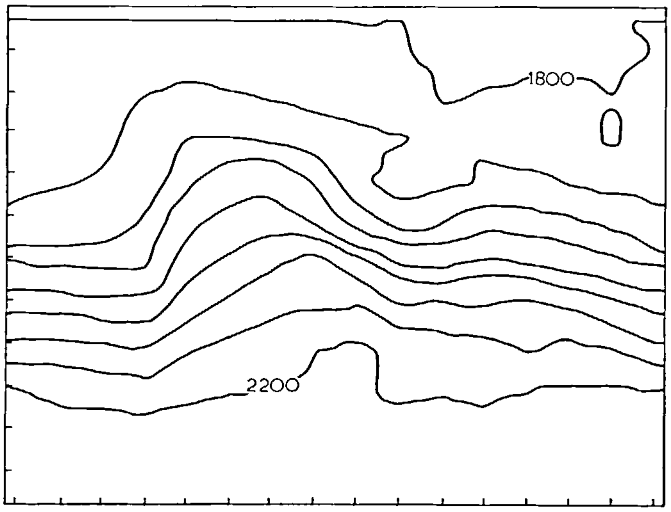

Figure 3 Height field contours after 6 days. $\Delta x = \Delta y = 400$ km; $\Delta t = 1800$ sec; $H_{\text{mean}} = 1998.32$ m; $E_{\text{tot}} = 6.2463 \times 10^{20}$; CM scheme IC 1

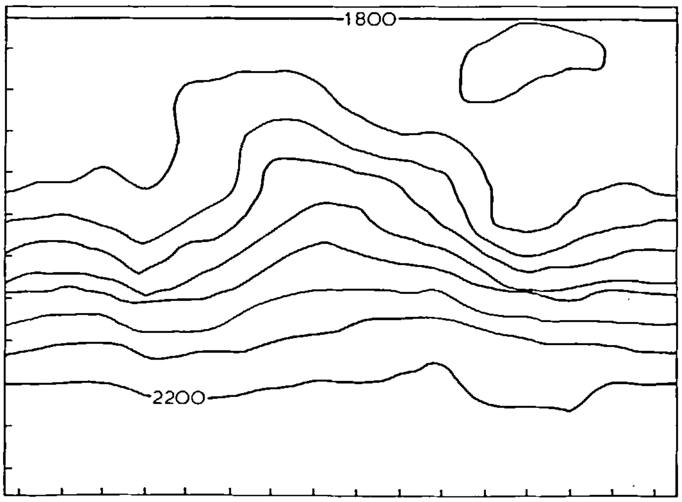

Figure 4 Height field contours after 10 days. $\Delta x = \Delta y = 400$ km; $\Delta t = 1800$ sec; $H_{\text{mean}} = 1999.05$ m; $E_{\text{tot}} = 6.2529 \times 10^{20}$; CM scheme; IC 1

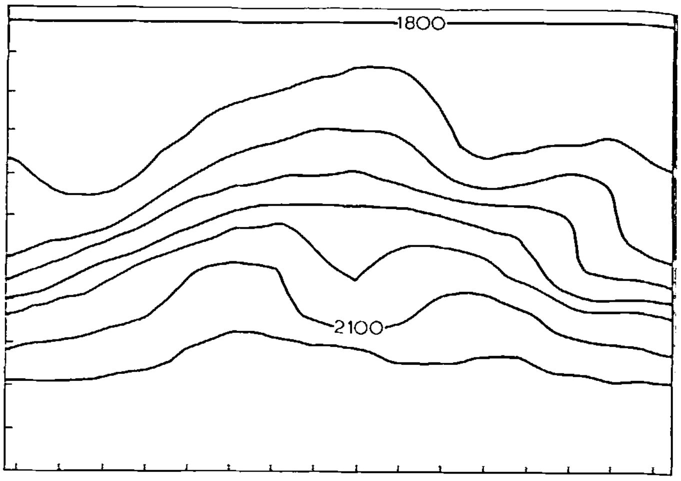

Figure 5 Height field contours after 36 h. $\Delta x = \Delta y = 400$ km; $\Delta t = 2700$ sec; $H_{\text{mean}} = 1998.78$ m; $E_{\text{tot}} = 6.2404 \times 10^{20}$; LM scheme; IC 1

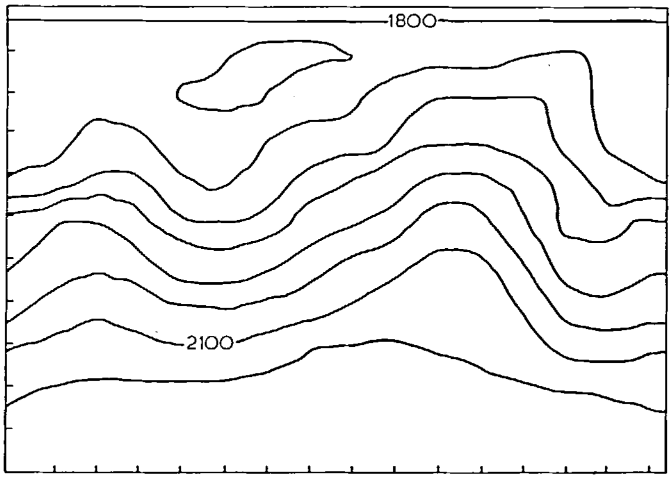

Figure 6 Height field contours after 3 days. $\Delta x = \Delta y = 400$ km; $\Delta t = 2700$ sec; $H_{\text{mean}} = 1998.35$ m; $E_{\text{tot}} = 6.2416 \times 10^{20}$; LM scheme; IC 1

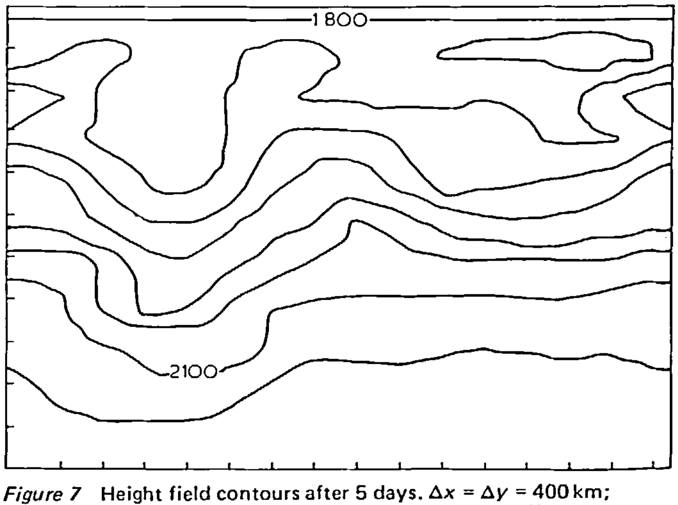

Figure 7 Height field contours after 5 days. $\Delta x = \Delta y = 400$ km; $\Delta t = 2700$ sec; $H_{\text{mean}} = 2001.43$ m; $E_{\text{tot}} = 6.2535 \times 10^{20}$; LM scheme; IC 1

Figure 10 shows the initial height field contours drawn at 50 m intervals for initial condition (II).

We then compared our results with the results of Gerrity *et al.*[32] after 2 days, and also with the results obtained by Cullen.[6] *Table 6* gives the extreme amplitude values of the height field in each trough and ridge at the midpoint of the channel after 2 days for different methods, including a fourth-order compact method due to Navon *et al.*[45]

Table 7 gives the corresponding positions as a fraction of the distance along the channel of the corresponding extreme values of troughs and ridges for the different methods. *Figure 11* shows the height field after 2 days of integration using the GMM mass matrix scheme in conjunction with initial condition (II) with a time step of Δt = 1800 sec. The results obtained show that the FEM integrations using the GMM mass matrix scheme match the Gerrity results with a spatial resolution Δx = 100 km as far as the amplitudes and the detailed positions of the troughs and ridges are concerned.

A good correspondence with the Cullen[6] two-stage Galerkin FEM and the compact fourth-order ADI method is observed.

Conclusions

A method for solving the nonlinear shallow-water equations using finite elements has been applied to a limited-area domain.

For the particular data used for comparison, it was experimentally found that the most accurate method was the coupled Galerkin FEM employing a generalized mixed mass (GMM) for the time (mass) matrix.

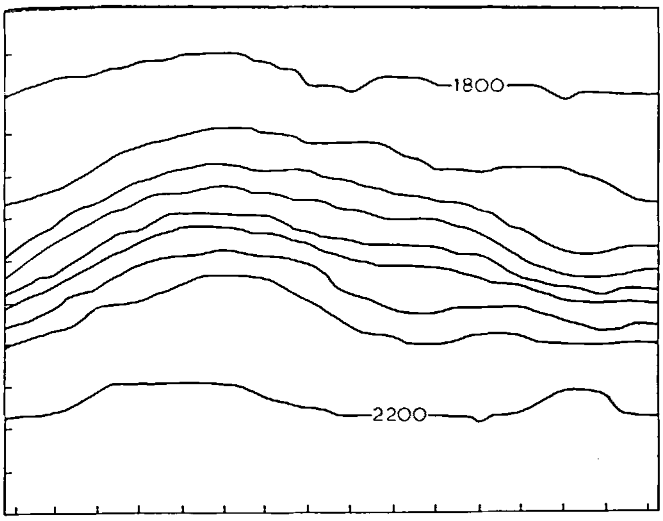

Figure 8 Height field contours after 2 days. $\Delta x = \Delta y$ = 400 km; Δt = 1800 sec; H_{mean} = 1998.15 m; $E_{\text{tot}} = 6.2426 \times 10^{20}$; GMM scheme; IC 1

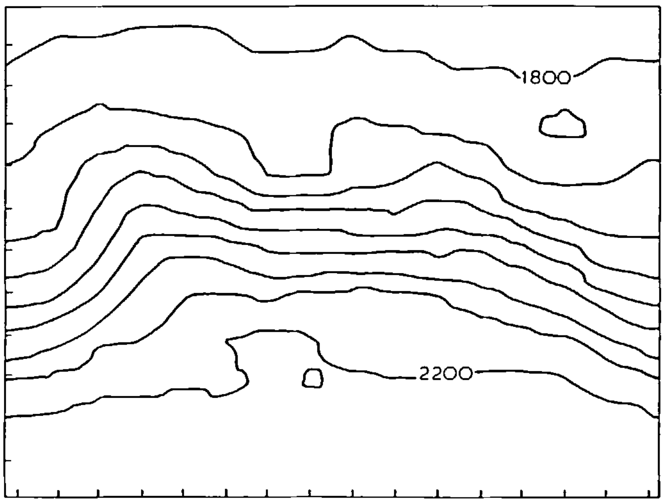

Figure 9 Height field contours after 5 days. $\Delta x = \Delta y$ = 400 km; Δt = 1800 sec; H_{mean} = 1997.43 m; $E_{\text{tot}} = 6.2375 \times 10^{20}$; GMM scheme; IC 1

Table 6 Amplitudes (after 2 days) in decametres

Method	Amplitude of troughs and ridges in middle of channel					
FEM with						
GMM mass matrix scheme ($\Delta x = \Delta y$ = 400 km)	210	204	207	192	197	187
Finite element ($\Delta x = \Delta y$ = 400 km) using the two-stage Galerkin method (Cullin[6])	210	204	205	193	197	186
Finite difference (Δx = 100 km) (Gerrity *et al.*[43])	208	204	206	192	197	189
Compact fourth-order ADI method ($\Delta x = \Delta y$ = 200 km) (Navon and Riphagen[46])	208	204	207	193	198	189

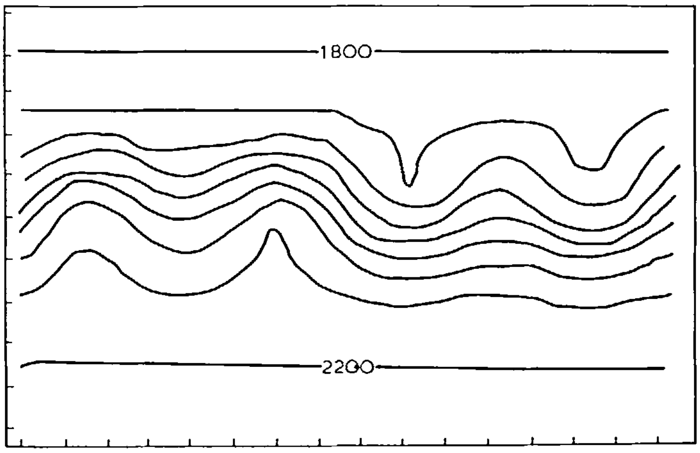

Figure 10 Initial height field contours (every 50 m). $\Delta x = \Delta y$ =. 400 km; H_{mean} = 2000 m; $E_{\text{tot}} = 6.2613 \times 10^{20}$. IC 2

Table 7 Phases after 2 days

Method	Phases					
FEM with						
GMM mass matrix scheme ($\Delta x = \Delta y$ = 400 km)	0.221	0.410	0.500	0.689	0.812	0.986
Finite element ($\Delta x = \Delta y$ = 400 km) using the two-stage Galerkin method (Cullen[6])	0.225	0.419	0.475	0.668	0.775	0.968
Finite difference ($\Delta x = \Delta y$ = 100 km) (Gerrity *et al.*[43])	0.235	0.399	0.499	0.730	0.857	1.000
Compact fourth-order ADI method ($\Delta x = \Delta y$ = 200 km) (Navon and Riphagen[46])	0.225	0.373	0.497	0.716	0.854	1.000

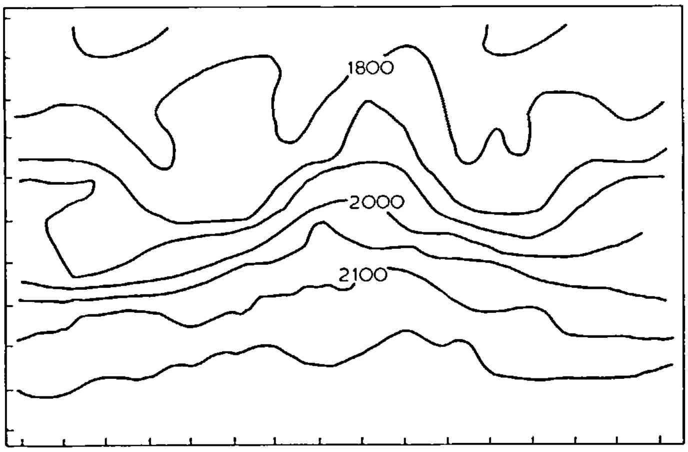

Figure 11 Height field contours after 2 days. $\Delta x = \Delta y = 400$ km, $\Delta t = 1800$ sec. $H_{mean} = 1999.98$ m; $E_{tot} = 6.2564 \times 10^{20}$; GMM scheme; IC 2

When the same Galerkin FEM was used with the LM scheme, the time step could be increased by damping the short gravity waves, and the procedure proved to be highly economic in computer time. No other numerical smoothing or damping was included in the model.

When accuracy was tested by comparison with a highly accurate nonlinear ADI scheme, the viability of this simple model was demonstrated, a good degree of accuracy being achieved although simple linear basis functions were used on three noded triangles. The computer time was further reduced by use of a compact storage scheme for sparse matrices.[19]

The accuracy could be improved if the method suggested by Cullen[8] were employed, in finite-element approximation of the products.

A final comment by the author is that the coefficient α in the GMM scheme should be further optimized and its connection with rational Padé approximants further investigated.[27]

Acknowledgement

The author thanks the reviewer for his helpful comments.

References

1 Wang, H. H. *et al. Mon. Weather Rev.*, 1972, 100-(10), 738
2 Baker, A. J. *Int. J. Num. Meth. Eng.*, 1973, 6 (1), 89
3 Baker, A. J. 'A finite-element solution algorithm for the Navier–Stokes equations', NASA report CR2391, 1974
4 Cullen, M. J. P. *J. Inst. Math. Applics.*, 1973, 11, 15
5 Cullen, M. J. P. *J. Inst. Math. Applics.*, 1974, 13, 233
6 Cullen, M. J. P. *PhD thesis*, University of Reading, UK, 1975
7 Cullen, M. J. P. *Quart. J. Roy. Met. Soc.*, 1976, 102, 77
8 Cullen, M. J. P. In 'Finite elements in water resources' (Ed. Gray, Pinder and Brebbia), Pentech Press, London and Plymouth, 1977
9 Brebbia, C. A. and Partridge, P. W. *Appl. Math. Modelling*, 1976, 1, 101
10 Connor, J. J. and Brebbia, C. A. 'Finite-element techniques for fluid flow', Newnes–Butterworths, London and Boston, 1976
11 Smith, S. L. and Brebbia, C. A. *J. Comp. Phys.*, 1975, 17, 235
12 Hinsman, D. E. *MSc thesis*, Naval Postgraduate School, Monterey, California, 1975
13 Staniforth, A. N. and Mitchell, H. L. *Mon. Weather Rev.*, 1977, 105 (2), 154
14 Grammeltvedt, A. *Mon. Weather Rev.*, 1969, 97 (5), 384
15 Douglas, J. and Dupont, T. *SIAM J. Numer. Anal.*, 1970, 7 (4), 575
16 Neuman, S. P. In 'Finite elements in fluids', Vol. 1 (Ed. Gallagher *et al.*), John Wiley & Sons, Chichester, 1975
17 Huebner, K. H. 'The finite-element method for engineers', John Wiley & Sons, Chichester, 1975
18 Payne, N. A. and Irons, B. M. Private communication to O. Zienkiewicz, 1963
19 Navon, I. M. and Müller, U. *Adv. Eng. Software* 1979, 1, 77
20 Desai, C. S. and Abel, J. F. 'Introduction to the finite-element method', Van Nostrand Reinhold Co., New York, 1972
21 Eisenberg, M. A. and Malvern, L. E. *Int. J. Num. Meth. Eng.*, 1973, 7, 574
22 Ishihara, K. *Mem. Numer. Math.*, 1977, 4, 1
23 Gustafsson, B. *J. Comp. Phys.*, 1971, 7, 239
24 Mesinger, F. and Arakawa, A. 'Numerical methods used in atmospheric models', Vol 1, GARP Publications Series, No. 17, 1976
25 Navon, I. M. Submitted to *Comput. Geosci.*
26 Mercer, J. W. and Faust, R. C. In 'Finite elements in water resources' (Ed. Gray, Pinder and Brebbia), Pentech Press, London and Plymouth, 1977
27 Nassif, N. R. *Calcolo*, 1975, 12 (1), 51
28 Tong, P. *et al. Comput. Struct.*, 1971, 1, 623
29 Ishihara, K. *Publ. Res. Inst. Math. Sci.*, Kyoto Univ., 1977 (13)
30 Ishihara, K. Dept of Mathematics, Faculty of Science, Ehime University, Japan (Submitted to *Numer. Math.*)
31 Fujii, H. 'Finite element schemes: stability and convergence', *Second US–Japan Seminar*, 1972
32 Gerrity, J. P. *et al. Mon. Weather Rev.*, 1972, 100 (8), 637
33 Tong, P. 'On the numerical problems of the finite-element methods', 1971, Waterloo Conference
34 Mock, M. S. *Numer. Math.* 1976, 26, 367
35 Schreyer, H. L. *Int. J. Numer. Meth. Eng.*, 1978, 12, 1171
36 Oden, J. T. and Fost, R. B. *Int. J. Numer. Meth. Eng.*, 1973, 6, 357
37 Fried, J. *J. Sound Vib.*, 1972, 22 (4), 407
38 Donea, J. *et al.* In 'Numerical methods in laminar and turbulent flow' (C. Tayor *et al.*, Eds), Pentech Press, London and Plymouth, 1978
39 Fried, I. and Malkus, D. S. *Int. J. Solid Struct.*, 1975, 11, 461
40 Holtz, K. P. *ZAMM*, 1978, 58, T277
41 Morton, K. W. In 'The state of the art in numerical analysis' (D. Jacobs, Ed), Academic Press, London and New York, 1977
42 Zienkiewicz, O. C. 'The finite-element method', McGraw-Hill Book Company (UK) Ltd, 1978
43 Key, S. W. and Beisinger, Z. E. *Proc. Third Conf. Matrix Meth. Struct. Eng.*, Air Force Inst. of Technology, Wright Patterson AFB, Ohio, 1971
44 Hinton, E., Rock, A. and Zienkiewicz, O. C. *Int. J. Earthquake Eng. Struct. Dynam.*, 1976, 4, 245
45 Navon, I. M. and Riphagen, H. A. Submitted to the *Mon. Weather Rev.*
46 Ciment, M. and Leventhal, S. H. *Math. Comp.*, 1975, 29 (132), 985
47 Swartz, B. K. and Wendroff, B. *SIAM J. Numer. Anal.*, 1974, 11 (5), 979