JOURNAL OF COMPUTATIONAL PHYSICS 107, 51–65 (1993)

Energy-Bounded Flow Approximation on a Cartesian-Product Grid over Rough Terrain

DON K. PURNELL AND MICHAEL J. REVELL

Atmospheric Division, National Institute of Water & Atmospheric Research Ltd., Wellington, New Zealand

Received August 6, 1991

We construct a method for modelling of three-dimensional, time dependent, compressible fluid flow in a gravitational field on a rotating cartesian-product grid with a spatially rough metric that bounds solutions by the total initial physical energy. Specifically: (1) the total physical energy is an l_2 norm on the model state and (2) this total energy cannot increase provided the timestep does not exceed CFL limits. In particular, the first property means that our measure of the energy is always positive unless the mass, momentum, and internal energy are all everywhere zero. These conditions guarantee that no error can grow unchecked. This is thought to be a desirable property, although only in the case of linear systems is it sufficient for convergence of a consistent approximation to the true solution. The great merit of this choice of norm is that the method is applicable to a wide variety of real physical problems because, even in complex circumstances, the total physical energy is conserved and each component of this energy is in limited supply. We first note that conservation of energy is equivalent to antisymmetry of a particular tendency operator. Energy-bounded approximations of fluid flow are then constructed either from antisymmetric finite difference operators, or from antisymmetric Galerkin operators. The method may be particularly useful when reliability in difficult conditions is needed. For example, when the viscosity must be small in order to simulate flow separation or turbulence, a model of viscous dissipation may be chosen purely from physical considerations, uncompromised by any requirements of numerical stability. We demonstrate this for an "internal hydraulic jump" flow over a bell-shaped mountain, simulating an internal wave as it steepens and breaks to form a turbulent jump.

1. INTRODUCTION

Finite difference methods which instantaneously conserve energy are well known (e.g., [2, 3, 4]), but these are neither sufficient to guarantee net conservation of total energy because of additional error due to time-differencing, nor do they limit the supply of all forms of energy. See [15] for an analysis of instability associated with this phenomenon in "leapfrog" integration of meteorological models. Schemes which approximate the "conservation-law form" of the equations [13], where energy is one of the state variables, are frequently designed to conserve energy, but then the total energy cannot be a complete norm on the system state, so that conservation of energy in such schemes does not provide a bound on the system state.

We derive an explicit method in which the numerical approximation errors can only cause a decrease in the total energy, provided that the timestep does not exceed CFL limits. An implicit version of this conserves energy exactly. In this scheme the initial total energy provides a bound on the system state, just as it does in a real fluid. This bound is related to the possibility of choosing state variables for which the total energy is the square of a euclidean distance on the system state. Each of these state variables represents a form of energy in limited supply. For example, the supply of heat, or of kinetic energy, is limited since neither absolute temperature, nor kinetic energy, may be negative.

In practice the total energy is limited only to the accuracy to which computations are performed, and in fact numerical truncation error is a source of uncorrellated noise, and hence a source of energy. For the hydraulic jump experiment we used 32-bit floating-point arithmetic (24-bit mantissa). This experiment was done with and without Rayleigh friction (9.1). Similar experiments were done on a three-dimensional $20 \times 20 \times 9$ grid. In no case did mean energy increase faster than one part in 10^6 per timestep when CFL limits were met.

Our design is intended for modelling flow over rough terrain, for which detail of the initial conditions is not well known, but the main objective is to simulate the effects of the terrain. The flows of interest evolve slowly relative to the timestep, so that a first-order accurate, dissipative time-difference scheme is suitable. We choose a compressible ideal gas model because this is simpler than, for example, the hydrostatic approximation.

We choose to stagger the state variables so that wind is represented at cell faces, while energy and mass are represented by cell-integral values. If the state variables are unstaggered the shortest wavelength acoustic and gravity waves have zero frequency in the model, because wind and mass become decoupled at the shortest scale. In our discrete approximation, we define the kinetic energy in a way which ensures that it is positive unless the velocity is everywhere

0021-9991/93 $5.00

zero. One disadvantage of this particular finite difference scheme is that it cannot exactly conserve momentum on an irregular grid. Another disadvantage is that it does not apply to a moving, flexible grid, although a splitting approximation might be employed to separate grid motion from fluid motion since the scheme is well suited to splittings in which the sub-steps are each bounded by the total physical energy.

2. ANTISYMMETRY AND CONSERVATION LAWS

2.1. *Conservation of Energy*

Define the following real scalar fields: $\mathscr{V} \to \Re$ on a three-dimensional domain $\mathscr{V}$ at time t; a gravitational + centrifugal potential Φ, a fluid density ρ, and a fluid pressure p. Define the following real vector fields: $\mathscr{V} \to \Re^3$; a rigid-rotation vector Ω, a fluid momentum density u, and a fluid velocity v; $v(s) = u(s)/\rho(s)$.

We start with the following description of an ideal gas with gas constant R, specific heat C_v, and thermodynamic parameter $\kappa = R/C_v$:

$$\begin{aligned} \partial/\partial t(u) &= -(u\cdot\nabla)v - v(\nabla\cdot u) - 2\Omega\times u - \nabla p - \rho\nabla\Phi \\ \partial/\partial t(p) &= -v\cdot\nabla p - (1+\kappa)\,p\nabla\cdot v \\ \partial/\partial t(\rho) &= -\nabla\cdot u. \end{aligned} \tag{2.1}$$

We may add an arbitrary constant to the potential Φ in (2.1), and we now choose one such that $0 < \Phi < \Phi_{\max}$, where $\Phi_{\max}$ is as small as possible, so as to avoid computational problems due to finite numerical precision. If E is the total physical energy of the system, then we make $E^{1/2}$ an L_2 norm (the euclidean distance between states of the system) by defining new state variables,

$$\begin{aligned} w &= \rho^{-1/2}u \\ c &= (2p/\kappa)^{1/2} \\ h &= (2\rho\Phi)^{1/2} \\ g &= -\nabla(2\Phi)^{1/2}, \end{aligned} \tag{2.2}$$

so that the total kinetic plus internal plus gravitational energy in the domain $\mathscr{V}$ is

$$E = \int_{\mathscr{V}} \tfrac{1}{2}(w\cdot w + c^2 + h^2). \tag{2.3}$$

Substituting (2.2) into (2.1) and noting that $gh = -\rho^{1/2}\nabla\Phi$,

$$\begin{aligned} \partial/\partial t(w) = &-\tfrac{1}{2}(\nabla\cdot v + v\cdot\nabla)w - 2\Omega\times w \\ &-\tfrac{1}{2}\rho^{-1/2}\nabla\kappa c^2 + gh \end{aligned} \tag{2.4}$$

$$\partial/\partial t(c) = -\tfrac{1}{2}(\nabla\cdot v + v\cdot\nabla)c - \tfrac{1}{2}c\kappa\nabla\cdot\rho^{-1/2}w \tag{2.5}$$

$$\partial/\partial t(h) = -\tfrac{1}{2}(\nabla\cdot v + v\cdot\nabla)h - g\cdot w \tag{2.6}$$

which we can write in the form

$$\partial/\partial t\begin{bmatrix} w \\ c \\ h \end{bmatrix} = A\begin{bmatrix} w \\ c \\ h \end{bmatrix} = A\xi, \tag{2.7}$$

where

$$A = \begin{bmatrix} -\tfrac{1}{2}(\nabla\cdot v + v\cdot\nabla) - 2\Omega\times & -\tfrac{1}{2}\rho^{-1/2}\nabla\kappa c & g \\ -\tfrac{1}{2}c\kappa\nabla\cdot\rho^{-1/2} & -\tfrac{1}{2}(\nabla\cdot v + v\cdot\nabla) & 0 \\ -g\cdot & 0 & -\tfrac{1}{2}(\nabla\cdot v + v\cdot\nabla) \end{bmatrix}. \tag{2.8}$$

On a domain $\mathscr{V}$ which is periodic or has suitable boundary conditions, A is an antisymmetric operator; i.e., $A = -A^*$, where A^* is the adjoint of A (since by inspection $A_{ij} = -A^*_{ji}$). This implies that the energy (2.3) is conserved because, using the notation $a^*b \equiv a\cdot b$ for the euclidean inner product of any two vectors a and b associated with a point in $\mathscr{V}$,

$$\begin{aligned} \partial/\partial t(2E) &= \partial/\partial t\int_{\mathscr{V}} \xi^*\xi \\ &= \int_{\mathscr{V}} (\xi^*\partial/\partial t(\xi) + (\partial/\partial t(\xi))^*\,\xi) \\ &= \int_{\mathscr{V}} \xi^*(A + A^*)\xi = 0. \end{aligned} \tag{2.9}$$

Conversely, the reason why A is antisymmetric is that energy conservation requires it. We can conclude this from the same argument used to demonstrate that energy conservation implies antisymmetry of the numerical approximation of A, by replacing the finite dimensional inner product in (5.13) with $\langle a, b\rangle = \int_{\mathscr{V}} a^*b$. The form of A is not unique, however. For example, another antisymmetric form is

$$A = \begin{bmatrix} -\tfrac{1}{2}(\nabla\cdot v + v\cdot\nabla) - 2\Omega\times & -\rho^{-1/2}c\,\nabla\kappa & g \\ -\kappa\nabla\cdot c\rho^{-1/2} & -(1-\kappa)\,\tfrac{1}{2}(\nabla\cdot v + v\cdot\nabla) & 0 \\ -g\cdot & 0 & -\tfrac{1}{2}(\nabla\cdot v + v\cdot\nabla) \end{bmatrix}. \tag{2.10}$$

2.2. *Conservation of Momentum*

In the case of constant potential Φ and zero rotation Ω, total momentum is conserved if the pressure force conserves momentum and if the diagonal terms in (2.8) which correspond to the transport of w and h are identical and

antisymmetric. To see this, note that if Φ is constant then the rate of change of momentum is

$$\partial/\partial t(u) = (h\partial/\partial t(w) + w\partial/\partial t(h))/(2\Phi^{1/2}).$$

Define a projection operator Π such that for any fields over $\mathscr{V}$ of vectors v and scalars a and b

$$\Pi\begin{bmatrix} v \\ a \\ b \end{bmatrix} = \begin{bmatrix} v \\ b \end{bmatrix}.$$

Define similarly a permutation operator Ξ such that

$$\Xi\begin{bmatrix} v \\ b \end{bmatrix} = \begin{bmatrix} b \\ v \end{bmatrix}.$$

Then the rate of change of total momentum may be written

$$\partial/\partial t \int_{\mathscr{V}} u = \frac{1}{2\Phi^{1/2}} \int_{\mathscr{V}} \left[\begin{bmatrix} w \\ h \end{bmatrix}^* \Xi\Pi A\Pi^* \begin{bmatrix} w \\ h \end{bmatrix} - \nabla p \right] = 0$$

since A is antisymmetric with identical diagonal elements, which implies $\Xi\Pi A\Pi^*$ is antisymmetric, so that the integral is zero as in (2.9). If the transport terms for w and h were not identical, then A and $\Xi\Pi A\Pi^*$ could not both be antisymmetric. So from the argument (5.13) we conclude that conservation of momentum will require identical approximations of the transport terms for w and h.

2.3. *General Coordinates*

We approximate (2.4)–(2.6) in non-orthogonal coordinates where $\delta r = \mu(\delta s)$ is the coordinate displacement vector corresponding to a physical displacement δs, with the matrices $[\mu]_{ij} = \partial r_i/\partial s_j$ and $[\eta]_{ij} = \partial s_j/\partial r_i$, $i, j \in \{1, 2, 3\}$. Define vectors μ_i, with components $[\mu_i]_j = [\mu]_{ij}$, normal to coordinate i-surfaces $\mu_i \cdot \delta s = 0$. Define vectors η_i, with components $[\eta_i]_j = [\eta]_{ij}$. Then we define contravariant wind components ω_i, and covariant wind components α_i,

$$\begin{aligned} \omega_i &= \rho^{-1/2}\mu_i \cdot u \\ \alpha_i &= \rho^{-1/2}\eta_i \cdot u, \end{aligned} \tag{2.11}$$

so that the kinetic energy density is $\omega \cdot \alpha/2$.

3. FINITE DIFFERENCE APPROXIMATION ON A STAGGERED GRID

We plan to construct our antisymmetric tendency operator by composition of a few primitive operators that can be tailored to suit an application. On staggered grids we need only the same number of operators required for the unstaggered case if, for any operator on one grid, we use its transpose, or adjoint, to compute the corresponding operation on the other grid, in each dimension. This device also helps to enforce the rules of energy conservation. On a domain $\mathscr{V}$ which is periodic or has suitable boundary conditions, the gradient ∇ is an antisymmetric operator; i.e., $\nabla = -\nabla^*$ since

$$\langle a, (\nabla + \nabla^*)b \rangle = \langle a, \nabla b \rangle + \langle \nabla a, b \rangle = \int_{\mathscr{V}} \nabla(ab) = 0.$$

We can achieve a similar property for staggered finite difference approximations of gradient on the real interval $[0, M]$, for integral $M > 0$, as follows. Cell faces are at points from the set

$$\mathscr{F}(M+1) = \{j; j = 0, 1, 2, ..., M\},$$

and cell centres are at points from the set

$$\mathscr{C}(M) = \{k; k = 0.5, 1.5, 2.5, ..., M - 0.5\}.$$

Let $\mathscr{W}$ be a vector space. We will be mostly interested in the cases $\mathscr{W} = \Re$ (real scalars) and $\mathscr{W} = \Re^3$ (vectors with three real components). For a smooth real vector field $f: [0, M] \to \mathscr{W}$, with gradient $f': [0, M] \to \mathscr{W}$, let $f_{\mathscr{F}}: \mathscr{F}(M+1) \to \mathscr{W}$ be the restriction of f to cell faces; $(f_{\mathscr{F}})(j) = f(j)$, $j \in \mathscr{F}(M+1)$. Let $(\mathscr{F}(M+1), \mathscr{W})$ be the linear space of all such functions $f_{\mathscr{F}}$, defined by pointwise addition of two functions and multiplication of a function by a scalar. Let $f_{\mathscr{C}}: \mathscr{C}(M) \to \mathscr{W}$ be the restriction of f to cell centres; $f_{\mathscr{C}} \in (\mathscr{C}(M), \mathscr{W})$. The transpose or adjoint of a linear operator T on a linear space X to a linear space Y will be denoted T^*. This is the unique operator $T^*: Y \to X$ such that, for any $x \in X$ and $y \in Y$, $\langle x, T^*y \rangle = \langle Tx, y \rangle$, where $\langle\ ,\ \rangle$ is a euclidean inner product on either X or Y. We define an operator $\triangle$ to approximate the gradient of f using (combinations of) differences between cell faces. Its transpose $\triangle^*$ operates the other way, using reversed differences between cell centres to estimate a reversed gradient:

$$\triangle: (\mathscr{F}(M+1), \mathscr{W}) \to (\mathscr{C}(M), \mathscr{W}), (\triangle f_{\mathscr{F}})(k) \approx f'(k)$$

$$\triangle^*: (\mathscr{C}(M), \mathscr{W}) \to (\mathscr{F}(M+1), \mathscr{W}), (\triangle^* f_{\mathscr{C}})(j) \approx -f'(j).$$

In a similar way we can define an operator $\square$ for interpolation from cell centres to cell faces so that its transpose is an interpolation the other way, from cell faces to cell centres:

$$\square: (\mathscr{C}(M), \mathscr{W}) \to (\mathscr{F}(M+1), \mathscr{W}), (\square f_{\mathscr{C}})(j) \approx f(j)$$

$$\square^*: (\mathscr{F}(M+1), \mathscr{W}) \to (\mathscr{C}(M), \mathscr{W}), (\square^* f_{\mathscr{F}})(k) \approx f(k).$$

The simplest such approximations are, for $k \in \mathscr{C}(M)$,

$$\begin{aligned}(\triangle f_{\mathscr{F}})(k) &= f_{\mathscr{F}}(k+\tfrac{1}{2}) - f_{\mathscr{F}}(k-\tfrac{1}{2}) \\ (\square^* f_{\mathscr{F}})(k) &= \tfrac{1}{2}(f_{\mathscr{F}}(k-\tfrac{1}{2}) + f_{\mathscr{F}}(k+\tfrac{1}{2}))\end{aligned} \tag{3.1}$$

so that, for $j \in \mathscr{F}(M+1) - \{0, M\}$,

$$\begin{aligned}(\triangle^* f_{\mathscr{C}})(j) &= f_{\mathscr{C}}(j-\tfrac{1}{2}) - f_{\mathscr{C}}(j+\tfrac{1}{2}) \\ (\triangle^* f_{\mathscr{C}})(0) &= f(0) - f_{\mathscr{C}}(\tfrac{1}{2}) \\ (\triangle^* f_{\mathscr{C}})(M) &= f_{\mathscr{C}}(M-\tfrac{1}{2}) - f(M) \\ (\square f_{\mathscr{C}})(j) &= \tfrac{1}{2}(f_{\mathscr{C}}(j-\tfrac{1}{2}) + f_{\mathscr{C}}(j+\tfrac{1}{2})) \\ (\square f_{\mathscr{C}})(0) &= \tfrac{1}{2}(f(0) + f_{\mathscr{C}}(\tfrac{1}{2})) \\ (\square f_{\mathscr{C}})(M) &= \tfrac{1}{2}(f_{\mathscr{C}}(M-\tfrac{1}{2}) + f(M)).\end{aligned} \tag{3.2}$$

3.1. *Notation*

We find that the above notation translates very easily into computer code. It allows a model to be programmed at a high level of abstraction, avoiding a maze of details, by using sequences of procedure calls that match the sequences of operators that we will use to describe the model.

To simplify notation, we associate operators with all finite-dimensional fields as follows. We use the same symbol to represent both a scalar field $f \in (\mathscr{C}(M), \Re)$ on a grid of M points, and its associated operator $f: (\mathscr{C}(M), \Re) \to (\mathscr{C}(M), \Re)$ defined by pointwise multiplication with any other scalar field $g \in (\mathscr{C}(M), \Re)$; $(fg)(k) = f(k)\, g(k)$. Since our fields are always real, the adjoint of this operator is $f^* = f$. Similarly, for a field of three-vectors $v \in (\mathscr{C}(M), \Re^3)$, its associated operator is $v: (\mathscr{C}(M), \Re) \to (\mathscr{C}(M), \Re^3)$; $(vg)(k)_i = g(k)\, v(k)_i$, and the transpose is $v^*: (\mathscr{C}(M), \Re^3) \to (\mathscr{C}(M), \Re)$ such that, for any other field of three-vectors $w \in (\mathscr{C}(M), \Re^3)$, $v^* w = v \cdot w$ is the conventional vector dot product, but $v^* w$ is a more natural notation for constructing adjoints of compositions of operators.

3.2. *Finite Differences in Three Dimensions*

Define an unstaggered grid G_0 and three staggered grids G_1, G_2, G_3:

$$\begin{aligned}G_0 &= \mathscr{C}(L) \times \mathscr{C}(M) \times \mathscr{C}(N) \\ G_1 &= \mathscr{F}(L+1) \times \mathscr{C}(M) \times \mathscr{C}(N) \\ G_2 &= \mathscr{C}(L) \times \mathscr{F}(M+1) \times \mathscr{C}(N) \\ G_3 &= \mathscr{C}(L) \times \mathscr{C}(M) \times \mathscr{F}(N+1).\end{aligned}$$

We use the notation $\triangle_i$ for an operator which gives differences in the i-direction between cell i-faces, $\triangle_i: (G_i, \mathscr{W}) \to (G_0, \mathscr{W})$, $i \in \{1, 2, 3\}$;

$$\begin{aligned}(\triangle_1 q)(l, m, n) &= (\triangle q(\cdot, m, n))(l) \\ (\triangle_2 q)(l, m, n) &= (\triangle q(l, \cdot, n))(m) \\ (\triangle_3 q)(l, m, n) &= (\triangle q(l, m, \cdot))(n).\end{aligned} \tag{3.3}$$

A similar notation $\square_i: (G_0, \mathscr{W}) \to (G_i, \mathscr{W})$ is used for an operator which interpolates in the i-direction to cell i-faces, $i \in \{1, 2, 3\}$, and its transpose $\square_i^*: (G_i, \mathscr{W}) \to (G_0, \mathscr{W})$ interpolates back to cell centres.

4. TENSOR-PRODUCT FINITE ELEMENT APPROXIMATION

Our scheme is naturally suited to the Galerkin approximation method (e.g., [6, 10, 12]) because, for our state variables (2.2), the Galerkin projection minimises error in the energy norm (2.3), and furthermore, it will never increase this total energy because the projection is an orthogonal one onto a linear subspace (with the same origin). This also means that we will naturally be able to construct an antisymmetric Galerkin approximation of A (2.8). Here we do this for a tensor-product spline basis, so that the required operators may be constructed by composition of one-dimensional operators on a product of one-dimensional splines [6]. One way to ensure dynamical coupling at the shortest scale is to stagger the knots of splines representing the wind w, relative to the knots of splines representing mass-like variables h, c. Another way of coupling wind and mass is to use an odd-order spline for wind, and an even-order spline for the mass. This second scheme has only one set of knots $\mathscr{F}(M+1)$, all at cell faces, which will simplify the computation of products of mass-like and wind splines because a three-dimensional integral of such a product can then be evaluated by a single polynomial per cell, instead of the sum of eight similar polynomials for the subdivision of each cell implied by staggered knots. Given a one-dimensional spline basis e_m, $m = 1, ..., M$, for even-order polynomial splines $q: [0, M] \to \Re$ with knots $\mathscr{F}(M+1)$ and coefficients $\mathbf{q}$, $q = \pi_e(\mathbf{q})$, π_e defined by

$$q(x) = \sum_{m=1}^{M} \mathbf{q}_m e_m(x) \tag{4.1}$$

and similarly for a one-dimensional spline basis b_m, $m = 0, ..., M$, for odd-order polynomial splines with the same knots $\mathscr{F}(M+1)$. Then

$$(\pi_e^* q)_m = \int_{x=0}^{M} e_m(x)\, q(x).$$

And we can define the components of the energy norm (2.3) using π:

$$\int_{x=0}^{M} q^2(x) = \sum_{m=1}^{M} \mathbf{q}_m (\pi_e^* \pi_e \mathbf{q})_m. \tag{4.2}$$

If $\{e_m\}$ is a B-spline basis [6], then the coefficients $\mathbf{q}_m$ (4.1) are a type of "local average" of q, so it follows from (4.2) that $(\pi_e^* \pi_e \mathbf{q})_m$ must also be (another) type of "local average." The Galerkin approximation of a function f: $[0, M] \to \mathfrak{R}$ is the projected function $\pi_e (\pi_e^* \pi_e)^{-1} \pi_e^* f$.

4.1. *Representation in Three Dimensions*

We represent a mass-like field h: $[0, L] \times [0, M] \times [0, N] \to \mathfrak{R}$ by a three-dimensional tensor-product of even-order splines with coefficients $\mathbf{h}$, $h = \Theta(\mathbf{h})$, Θ defined by

$$\begin{aligned} h(r_1, r_2, r_3) &= \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} e_l(r_1)\, e_m(r_2) \\ &\quad \times e_n(r_3)\, \mathbf{h}_{lmn} \\ (\Theta^* h)_{lmn} &= \int_{r_1=0}^{L} \int_{r_2=0}^{M} \int_{r_3=0}^{N} e_l(r_1)\, e_m(r_2)\, e_n(r_3) \\ &\quad \times h(r_1, r_2, r_3) \end{aligned} \tag{4.3}$$

and similarly for $c = \Theta(\mathbf{c})$. We represent the contravariant wind components ω_i (2.11) by a three-dimensional tensor-product of odd-order and even-order splines with coefficients $\boldsymbol{\omega}_i$, $\omega_i = \Psi_i(\boldsymbol{\omega}_i)$, Ψ_i defined by

$$\begin{aligned} \omega_1(r_1, r_2, r_3) &= \sum_{l=0}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} b_l(r_1) \\ &\quad \times e_m(r_2)\, e_n(r_3)\, \boldsymbol{\omega}_{1,lmn} \\ (\Psi_1^* \omega_1)_{lmn} &= \int_{r_1=0}^{L} \int_{r_2=0}^{M} \int_{r_3=0}^{N} b_l(r_1)\, e_m(r_2) \\ &\quad \times e_n(r_3)\, \omega_1(r_1, r_2, r_3) \\ \omega_2(r_1, r_2, r_3) &= \sum_{l=1}^{L} \sum_{m=0}^{M} \sum_{n=1}^{N} e_l(r_1) \\ &\quad \times b_m(r_2)\, e_n(r_3)\, \boldsymbol{\omega}_{2,lmn} \\ (\Psi_2^* \omega_2)_{lmn} &= \int_{r_1=0}^{L} \int_{r_2=0}^{M} \int_{r_3=0}^{N} e_l(r_1)\, b_m(r_2) \\ &\quad \times e_n(r_3)\, \omega_2(r_1, r_2, r_3) \\ \omega_3(r_1, r_2, r_3) &= \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=0}^{N} e_l(r_1) \\ &\quad \times e_m(r_2)\, b_n(r_3)\, \boldsymbol{\omega}_{3,lmn} \\ (\Psi_3^* \omega_3)_{lmn} &= \int_{r_1=0}^{L} \int_{r_2=0}^{M} \int_{r_3=0}^{N} e_l(r_1)\, e_m(r_2) \\ &\quad \times b_n(r_3)\, \omega_3(r_1, r_2, r_3) \end{aligned} \tag{4.4}$$

and similarly for the covariant wind components $\alpha_i = \Psi_i(\boldsymbol{\alpha}_i)$.

5. FINITE DIFFERENCE APPROXIMATION ON A FIXED GRID

5.1. *Mapping onto Physical Coordinates*

Approximations required for the general coordinates (2.11) on the physical domain $\mathscr{V} \subset \mathfrak{R}^3$ are developed as follows. For any differentiable mapping $\mathscr{X}$: $[0, L] \times [0, M] \times [0, N] \to \mathscr{V}$; $\mathscr{X}(r) = s$, we define a tangent mapping $\eta = \mathscr{D}_r \mathscr{X}$, with inverse $\mu^* = \eta^{-1}$ as in (2.11). We use the following notation for restrictions of these mappings to staggered grids: Define a vector field on a grid of cell i-faces, μ_i: $G_i \to \mathfrak{R}^3$, with components $[\mu_i]_j$, $j \in \{1, 2, 3\}$,

$$[\mu_i(r)]_j = \partial r_i / \partial s_j |_{r \in G_i}. \tag{5.1}$$

Define similarly a vector field η_i: $G_i \to \mathfrak{R}^3$,

$$[\eta_i(r)]_j = \partial s_j / \partial r_i |_{r \in G_i}. \tag{5.2}$$

The component of displacement in the i-direction at cell i-faces is δr_i: $G_i \to \mathfrak{R}$ corresponding to a physical displacement vector field at grid points δs: $G_i \to \mathfrak{R}^3$ is

$$\begin{aligned} \delta r_i &= \mu_i^* \, \delta s \\ \delta s &= \eta_i \, \delta r_i. \end{aligned}$$

Define the cell volume scalar field at grid points v: $G_0 \to \mathfrak{R}$,

$$v = |\eta|, \tag{5.3}$$

and the "staggered volume" at i-faces $v_i = \square_i v$.

5.2. *Kinetic Energy*

Given the momentum vector field u_i: $G_i \to \mathfrak{R}^3$ on cell i-faces, define new state variables

$$\alpha_i = v_i^{1/2} \rho_i^{-1/2} \eta_i^* u_i. \tag{5.4}$$

In order to define the kinetic energy of the system, we also define an approximation of the corresponding contravariant component at an i-face; $\omega_i \approx v_i^{1/2} \rho_i^{-1/2} \mu_i^* u_i$. A variety of adequate approximations of ω are possible. A computationally favourable one is

$$\begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \square_1 \lambda_{12} \square_2^* & \square_1 \lambda_{13} \square_3^* \\ \square_2 \lambda_{21} \square_1^* & \lambda_{22} & \square_2 \lambda_{23} \square_3^* \\ \square_3 \lambda_{31} \square_1^* & \square_3 \lambda_{32} \square_2^* & \lambda_{33} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \Lambda \alpha, \tag{5.5}$$

where

$$\lambda_{ij} = \begin{cases} \bar{\mu}_i^* \bar{\mu}_j & \text{if } i \neq j \\ \square_i(\bar{\mu}_i^* \bar{\mu}_i) & \text{if } i = j \end{cases}$$

and

$$\bar{\mu}_i = \square_i^*(\mu_i).$$

A consistent definition of the total kinetic energy of the system is then

$$K = \tfrac{1}{2} \sum_{i=1}^{3} \langle \alpha_i, \omega_i \rangle = \tfrac{1}{2} \langle \alpha, \Lambda \alpha \rangle, \tag{5.6}$$

where, for any vector fields $a: G_i \to \mathscr{W}$ and $b: G_i \to \mathscr{W}$, the inner product $\langle a, b \rangle \equiv \sum_{r \in G_i} a(r)^* b(r)$. By inspection of (5.5) the matrix Λ is symmetric: $\Lambda^* = \Lambda$. We also want this measure of the kinetic energy to be positive definite, so that the energy (2.3) is an l_2 norm. We verify that (5.6) is positive for any nonzero α as follows. First, note that for (3.1) and (3.2),

$$\tfrac{1}{4} \triangle_i^* \rho \triangle_i = \square_i(\rho) - \square_i \rho \square_i^*,$$

where ρ and $\square_i(\rho)$ are both scalar fields, so that

$$\begin{aligned} \langle \alpha, \Lambda \alpha \rangle &= \left\langle \left(\sum_{i=1}^{3} \mu_i \square_i^* \alpha_i \right), \left(\sum_{j=1}^{3} \mu_j \square_j^* \alpha_j \right) \right\rangle \\ &\quad + \tfrac{1}{4} \sum_{i=1}^{3} \langle (\mu_i \triangle_i \alpha_i), (\mu_i \triangle_i \alpha_i) \rangle \\ &> 0. \end{aligned} \tag{5.7}$$

Here we assume cell volumes are positive so that μ_i is nonsingular. The expression (5.7) separates long and short wavelength contributions to the kinetic energy, since $\square$ is a low pass filter, whereas $\triangle$ is a high pass filter. Together, these contributions ensure that all waves are represented and (5.6) is positive definite.

5.3. *A Tendency Operator That Does No Work*

Define new state variables c and h,

$$c = (2vp/\kappa)^{1/2} \tag{5.8}$$

$$h = (2v\rho\Phi)^{1/2}, \tag{5.9}$$

where Φ is the gravitational potential. Define J, $\check{c}$, $\check{h}$

$$\begin{bmatrix} \omega \\ c \\ h \end{bmatrix} = \begin{bmatrix} \Lambda & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \check{c} \\ \check{h} \end{bmatrix} = J \begin{bmatrix} \alpha \\ \check{c} \\ \check{h} \end{bmatrix}. \tag{5.10}$$

Since Λ (5.5) is symmetric and positive definite, we can define a state vector ξ,

$$\xi = J^{1/2} \begin{bmatrix} \alpha \\ c \\ h \end{bmatrix}. \tag{5.11}$$

The sum of kinetic, internal, and gravitational energy of the system is

$$\begin{aligned} &\tfrac{1}{2} \left(\sum_{i=1}^{3} \langle \alpha_i, \omega_i \rangle + \langle (p/\kappa), v \rangle + \langle \Phi, v\rho \rangle \right) \\ &\quad = \tfrac{1}{2} (\langle \alpha, \Lambda \alpha \rangle + \langle c, c \rangle + \langle h, h \rangle) = \tfrac{1}{2} \langle \xi, \xi \rangle. \end{aligned}$$

We show that if Λ is not changing, then no net work is done by the system

$$\partial/\partial t \begin{bmatrix} \alpha \\ c \\ h \end{bmatrix} = A \begin{bmatrix} \omega \\ c \\ h \end{bmatrix}$$

if and only if A is antisymmetric. First, antisymmetry implies zero work rate because if $A = -A^*$ then $(J^{1/2} A J^{1/2}) = -(J^{1/2} A J^{1/2})^*$ so that

$$\begin{aligned} \partial/\partial t \langle \xi, \xi \rangle &= \langle \xi, \partial/\partial t(\xi) \rangle + \langle \partial/\partial t(\xi), \xi \rangle \\ &= \langle \xi, J^{1/2} A J^{1/2} \xi \rangle + \langle J^{1/2} A J^{1/2} \xi, \xi \rangle \\ &= 0. \end{aligned} \tag{5.12}$$

Conversely, if the work rate is zero for all values of the real state vector ξ, then for any real state vector ζ in the domain of A there is a ξ for which $\zeta = J^{1/2} \xi$ and

$$\begin{aligned} 0 &= \langle \xi, J^{1/2} A J^{1/2} \xi \rangle + \langle J^{1/2} A J^{1/2} \xi, \xi \rangle \\ &= \langle \zeta, (A + A^*) \zeta \rangle. \end{aligned}$$

And therefore for all real fields a and b in the domain of A we have the identity

$$\begin{aligned} 0 &= \langle (a+b), (A + A^*)(a+b) \rangle \\ &\quad - \langle a, (A + A^*) a \rangle - \langle b, (A + A^*) b \rangle \\ &= 2 \langle b, (A + A^*) a \rangle \\ &\Rightarrow A = -A^*. \end{aligned} \tag{5.13}$$

We choose an antisymmetric operator $A(\bar{\omega}, \bar{c}, \bar{h})$ which approximates (2.4)–(2.6) when the parameters $\bar{\omega}$, $\bar{c}$, $\bar{h}$

approximate ω, c, h at time t, so that the errors $|\omega^t - \bar{\omega}|$, $|c^t - \bar{c}|$, and $|h^t - \bar{h}|$ are of the order of one timestep,

$$\partial/\partial t \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \check{c} \\ \check{h} \end{bmatrix} = A(\bar{\omega}, \bar{c}, \bar{h}) \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ c \\ h \end{bmatrix}$$

$$= \begin{bmatrix} \mathscr{R}_{11} & \mathscr{R}_{12} & \mathscr{R}_{13} & \mathscr{P}_1 & \mathscr{H}_1 \\ \mathscr{R}_{21} & \mathscr{R}_{22} & \mathscr{R}_{23} & \mathscr{P}_2 & \mathscr{H}_2 \\ \mathscr{R}_{31} & \mathscr{R}_{32} & \mathscr{R}_{33} & \mathscr{P}_3 & \mathscr{H}_3 \\ -\mathscr{P}_1^* & -\mathscr{P}_2^* & -\mathscr{P}_3^* & \mathscr{T} & 0 \\ -\mathscr{H}_1^* & -\mathscr{H}_2^* & -\mathscr{H}_3^* & 0 & \mathscr{T} \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ c \\ h \end{bmatrix}, \tag{5.14}$$

where an estimate of the volume displaced at an i-face is

$$\varepsilon_i = v_i^{1/2} \rho_i^{-1/2} \bar{\omega}_i$$

and transport of a field f is described by the antisymmetric operator $\mathscr{T}$:

$$\mathscr{T}(f) = -\tfrac{1}{2} \sum_{i=1}^{3} (v^{-1/2} \triangle_i \varepsilon_i \square_i v^{-1/2} f - v^{-1/2} \square_i^* \varepsilon_i \triangle_i^* v^{-1/2} f). \tag{5.15}$$

Define $W = v^{1/2} \rho^{-1/2} u$, where u is the momentum three-vector. We use the following estimate of W:

$$W = \sum_{j=1}^{3} v^{1/2} \square_j^* \eta_j v_j^{-1/2} \omega_j. \tag{5.16}$$

The tendency of W due to transport and rigid-rotation of the coordinates is

$$\partial/\partial t(W) = \mathscr{T} W - 2\Omega \times W.$$

Then the effect of transport and rigid-rotation of the coordinates on the tendency of α is described by the antisymmetric operator $\mathscr{R}$,

$$\mathscr{R}_{ij} = v_i^{-1/2} \eta_i^* \square_i (v^{1/2} \mathscr{T} v^{1/2} - 2v\Omega \times) \square_j^* \eta_j v_j^{-1/2}. \tag{5.17}$$

$\mathscr{P}$ is defined so that the rates of change of covariant components of velocity due to pressure differences between cells approximate (2.4),

$$\mathscr{P}_i(c) = \tfrac{1}{2} v_i^{1/2} \rho_i^{-1/2} \triangle_i^* v^{-1} \kappa \bar{c} c, \tag{5.18}$$

and $\mathscr{H}$ is defined

$$\mathscr{H}_i(h) = v_i^{1/2} g_i \square_i v^{-1/2} h. \tag{5.19}$$

Given that the net work rate is zero and, hence, that $A(\bar{\omega}, \bar{c}, \bar{h})$ must be antisymmetric, we can now derive the numerical approximation of energy-tendency from (5.18) and (5.19) simply by reversing compositions and taking adjoints of the operators as

$$\begin{aligned} \partial/\partial t(c) &= -\sum_{i=1}^{3} \mathscr{P}_i^*(\omega_i) + \mathscr{T} c \\ &= -\tfrac{1}{2} \kappa v^{-1} \bar{c} \sum_{i=1}^{3} \triangle_i \rho_i^{-1/2} v_i^{1/2} \omega_i \\ &\quad - \tfrac{1}{2} \sum_{i=1}^{3} (v^{-1/2} \triangle_i \varepsilon_i \square_i v^{-1/2} c \\ &\quad - v^{-1/2} \square_i^* \varepsilon_i \triangle_i^* v^{-1/2} c) \end{aligned} \tag{5.20}$$

which is a discrete approximation of (2.5). Similarly, an estimate of change of gravitational energy due to mass flux is

$$\begin{aligned} \partial/\partial t(h) &= -\sum_{i=1}^{3} v^{-1/2} \square_i^* g_i v_i^{1/2} \omega_i \\ &\quad - \tfrac{1}{2} \sum_{i=1}^{3} (v^{-1/2} \triangle_i \varepsilon_i \square_i v^{-1/2} h \\ &\quad - v^{-1/2} \square_i^* \varepsilon_i \triangle_i^* v^{-1/2} h). \end{aligned} \tag{5.21}$$

6. GALERKIN APPROXIMATION ON A FIXED GRID

An antisymmetric operator A (5.14) can be constructed from the Galerkin approximation operators Θ, Ψ_i (4.3), (4.4) as follows. The variables h, c, ω_i, ρ, v, μ_i, η_i, g are now piecewise polynomials defined on each cell, and finite differences are replaced by differential operators, but the scheme is similar to the finite difference approximation. The transport of a field f is described by the antisymmetric operator T:

$$T(f) = -\tfrac{1}{2} \sum_{i=1}^{3} (v^{-1/2} \partial/\partial r_i (\rho^{-1/2} \bar{\omega}_i f) + \bar{\omega}_i \rho^{-1/2} \partial/\partial r_i (v^{-1/2} f)).$$

Then the operators $\mathscr{T}$, $\mathscr{R}$, $\mathscr{P}$, $\mathscr{H}$ in (5.14) are re-defined:

$$\begin{aligned} \mathscr{T} &= \Theta^* T \Theta \\ \mathscr{R}_{ij} &= \Psi_i^* \eta_i^* (T - 2\Omega \times) \eta_j \Psi_j \\ \mathscr{P}_i &= -\tfrac{1}{2} \Psi_i^* v^{1/2} \rho^{-1/2} \partial/\partial r_i v^{-1} \kappa \bar{c} \Theta \\ \mathscr{H}_i &= \Psi_i^* g \Theta. \end{aligned}$$

The Galerkin projection can then be completed by the operator J (5.10) if J is re-defined so that

$$\boldsymbol{\omega} = (\Psi^* \eta^* \eta \Psi)^{-1}\, \boldsymbol{\alpha}$$
$$\mathbf{c} = (\Theta^* \Theta)^{-1}\, \check{\mathbf{c}}$$
$$\mathbf{h} = (\Theta^* \Theta)^{-1}\, \check{\mathbf{h}}.$$

7. TIME INTEGRATION

7.1. *Euler-Backward Explicit Integration*

This scheme damps high-frequency waves. If ξ is a real vector representing the state of a system which evolves as

$$\partial/\partial t(\xi) = -B\xi, \tag{7.1}$$

where B is an antisymmetric operator, i.e., $B^* = -B$, then the "Euler-backward" time-integration scheme for this is a predictor–corrector method as follows:

$$\begin{aligned} \hat{\xi}^{t+1} &= \xi^t - B\xi^t \\ \xi^{t+1} &= \xi^t - B\hat{\xi}^{t+1}. \end{aligned} \tag{7.2}$$

Then the consequent change in the euclidean norm $\|\xi\|$ follows from

$$\begin{aligned} &\langle \xi, \xi \rangle^{t+1} - \langle \xi, \xi \rangle^t \\ &\quad = \langle (\xi^{t+1} + \xi^t), (\xi^{t+1} - \xi^t) \rangle \\ &\quad = \langle (2\xi^t - B\xi^t - B^* B\xi^t), (-B\xi^t - B^* B\xi^t) \rangle \\ &\quad = -\langle \xi^t, (B^* B)\, \xi^t \rangle + \langle \xi^t, (B^* B)(B^* B)\, \xi^t \rangle \end{aligned}$$

since antisymmetry implies $\langle \xi, B\xi \rangle = 0$ and $B^*(B^* B) + (B^* B)B = 0$. Hence we have

$$\|\xi\|^{t+1} \leqslant \|\xi\|^t, \tag{7.3}$$

provided that

$$\|B\| \leqslant 1, \tag{7.4}$$

and we are free to choose a small enough timestep (define one unit of time) so as to guarantee that (7.4) is true. From (5.14), (5.11), and (7.1) we have

$$B = -J^{1/2} A(\bar{\omega}, \bar{c}, \bar{h})\, J^{1/2} \tag{7.5}$$

and B is antisymmetric as required in the Euler-backward scheme above.

7.1.1. THE EXPLICIT ALGORITHM. In general, convergence of the Euler-backward scheme requires a forward estimate of $\bar{\omega}$, $\bar{c}$, $\bar{h}$ in (7.5):

$$\begin{bmatrix} \bar{\omega} \\ \bar{c} \\ \bar{h} \end{bmatrix} = \begin{bmatrix} \omega \\ c \\ h \end{bmatrix}^t + JA(\omega, c, h)^t \begin{bmatrix} \omega \\ c \\ h \end{bmatrix}^t. \tag{7.6}$$

The forward step is then

$$\begin{bmatrix} \hat{\omega} \\ \hat{c} \\ \hat{h} \end{bmatrix}^{t+1} = \begin{bmatrix} \omega \\ c \\ h \end{bmatrix}^t + JA(\bar{\omega}, \bar{c}, \bar{h}) \begin{bmatrix} \omega \\ c \\ h \end{bmatrix}^t. \tag{7.7}$$

The corrector step is

$$\begin{bmatrix} \omega \\ c \\ h \end{bmatrix}^{t+1} = \begin{bmatrix} \omega \\ c \\ h \end{bmatrix}^t + JA(\bar{\omega}, \bar{c}, \bar{h}) \begin{bmatrix} \hat{\omega} \\ \hat{c} \\ \hat{h} \end{bmatrix}^{t+1}. \tag{7.8}$$

7.2. *Implicit Time Integration*

Instead of the Euler-backward scheme (7.2), energy is conserved exactly by the implicit scheme

$$\begin{aligned} \xi^{t+1} &= \xi^t - B(\xi^{t+1} + \xi^t)/2 \\ &= (1 + \tfrac{1}{2}B)^{-1} (1 - \tfrac{1}{2}B)\, \xi^t \end{aligned} \tag{7.9}$$

since then

$$\begin{aligned} \langle \xi, \xi \rangle^{t+1} - \langle \xi, \xi \rangle^t &= \langle (\xi^{t+1} + \xi^t), (\xi^{t+1} - \xi^t) \rangle \\ &= -2\langle (\xi^{t+1} + \xi^t)/2, B(\xi^{t+1} + \xi^t)/2 \rangle \\ &= 0. \end{aligned}$$

This ensures exact conservation of energy but requires the solution of a large set of linear equations for $(1 + \frac{1}{2}B)^{-1}$.

7.3. *Time Integration Diagrams*

The above schemes are constructed from an antisymmetric, and hence normal operator B. Therefore the behaviour of these schemes can be shown by a two-dimensional diagram in the eigenspace of B corresponding to eigenvalue ib, since the phase space of the model is spanned by such eigenspaces. Figure 1 shows the construction of one timestep of a state vector ξ for a Euler-backward scheme (a), an Implicit scheme (b), and a leapfrog scheme (c). We emphasise that the eigenspaces will change from one timestep to the next, but what does not change is the energy norm. So although we cannot show several timesteps in two-dimensions, the behaviour of the length of the state vector over a series of timesteps can be inferred from the diagrams. In (a) it is evident that the vector ξ^{t+1} will be

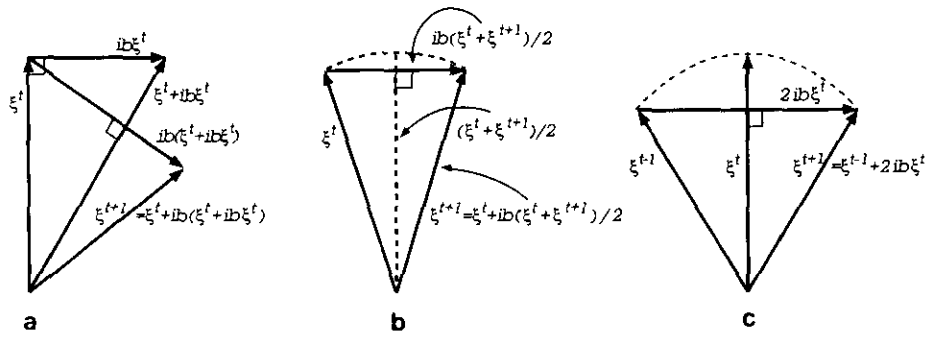

FIG. 1. A graphical construction of one timestep of a state vector ξ for (a) a Euler-backward scheme, (b) an Implicit scheme, and (c) a leapfrog scheme in the eigenspace of B corresponding to eigenvalue ib.

shorter than ξ^t, provided b is not too large. Since b is proportional to the timestep (one unit of time), we infer that the Euler-backward scheme is dissipative. In (c) we are assuming a symmetry corresponding to the absence of the leapfrog computational mode. From the diagram we expect that lack of this symmetry would cause instability in the system energy.

8. INCLUSION OF MORE COMPLEX PHYSICS

The general theoretical picture is simple: All physical processes conserve total energy, and all state variables are to be formulated so as to participate in a single budget for a total physical energy that is an l_2 norm on the system as a whole. In a wet model, for example, each phase of water could be accounted according to the amount of latent heat it stores. This scheme will both enforce a balance in the net budget of energy transformations and also deny credit on any particular form of energy.

The following sections are intended to show that the energy-bounded method could be used in a wide variety of applications, by providing simple examples of ways that complex physical processes such as diffusion of momentum by Reynolds stresses, and of phase changes of water in a wet model, could be incorporated into an energy-bounded model.

8.1. *Viscous Stresses and Heating*

We derive an energy bounded approximation of the Reynolds stresses due to turbulent momentum exchange on spatial scales too small to be resolved by the model grid. This approximation assumes that stress has been determined somehow, for example, from the rate of strain (e.g., [8]), and that any kinetic energy dissipated by Reynolds stresses is converted directly into heat. Reynolds stresses are modelled by a term R added to (2.7),

$$\partial/\partial t(\xi) = A\xi + R\xi, \tag{8.1}$$

where

$$R = \begin{bmatrix} 0 & \frac{1}{2}\rho^{-1/2}\sum_{i=1}^{3}\frac{\partial}{\partial s_i}\tau_i c^{-1} & 0 \\ \frac{1}{2}c^{-1}\sum_{i=1}^{3}\tau_i^*\frac{\partial}{\partial s_i}\rho^{-1/2} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

(s_1, s_2, s_3) are physical rectangular cartesian coordinates and τ_i is the stress vector across planes of constant s_i corresponding to stress tensor τ.

The force on the jth face is calculated from the stress on the directed area $\mu_j v_j$ of the face. To construct an antisymmetric approximation at time t we approximate this force as a linear function $\bar{F}(\square_j c)$ of the internal energy variable c, where, as in (5.14), $\bar{c}$ is a zeroth-order approximation of c at time t,

$$\bar{F}_j(\square_j c) = \bar{\tau}_j \mu_j v_j \frac{\square_j c}{\square_j \bar{c}}. \tag{8.2}$$

The viscous component of the tendency of the covariant wind components α_i at cell i-faces can now be written

$$(\partial/\partial t)_R(\alpha_i) = \tfrac{1}{2}\rho_i^{-1/2} v_i^{-1/2} \eta_i^* \square_i \sum_{j=1}^{3} \triangle_j \bar{F}_j \square_j c. \tag{8.3}$$

Reversing compositions and taking adjoints of each operator gives the corresponding approximation of the heating due to viscous dissipation,

$$(\partial/\partial t)_R(c) = -\tfrac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3} \square_j^* \bar{F}_j^* \triangle_j^* \square_i^* \eta_i v_i^{-1/2} \rho_i^{-1/2} \omega_i. \tag{8.4}$$

8.2. *Phase Changes of Water*

Phase changes of water could be incorporated into an energy-bounded model as follows. Transfer of heat between

air, ice, water, and water vapour could be represented by an antisymmetric operator on the energies stored by the air and by each phase of water. Other accounting schemes may also be suitable, depending on the application. Given the masses per unit volume of ice ρ_I, water ρ_W, and water vapour ρ_V, and corresponding energy densities corresponding to latent heats L_f for the freezing of liquid water, L_c for condensation of water vapour, and $L_s = L_c + L_f$ for the formation of ice from water vapour, and with the addition of a constant $L_0 > 0$ to ensure that the energy measures are positive definite:

$$\tfrac{1}{2}q_I^2 = \rho_I L_0$$
$$\tfrac{1}{2}q_W^2 = \rho_W(L_f + L_0)$$
$$\tfrac{1}{2}q_V^2 = \rho_V(L_c + L_f + L_0).$$

In circumstances near thermal equilibrium, where differences in the temperature of various components of the mixture are negligible, we can generalize the internal energy variable c to represent the total sensible heat of the mixture; then the conversion rates of q are represented by the antisymmetric matrix Q,

$$\partial/\partial t \begin{bmatrix} q_V \\ q_W \\ q_I \\ c \end{bmatrix} = \begin{bmatrix} \mathscr{T}(q_V) \\ \mathscr{T}(q_W) \\ \mathscr{T}(q_I) \\ \mathscr{T}(c) - \sum_{i=1}^{3} \mathscr{P}_i^*(\omega_i) \end{bmatrix} + Q \begin{bmatrix} q_V \\ q_W \\ q_I \\ c \end{bmatrix}, \qquad (8.5)$$

where

$$Q = \begin{bmatrix} 0 & -r_c & -r_s & -(r_c L_c + r_s L_s) \\ r_c & 0 & -r_f & -r_f L_f \\ r_s & r_f & 0 & 0 \\ (r_c L_c + r_s L_s) & r_f L_f & 0 & 0 \end{bmatrix}.$$

The conversion rate factors r_c, r_f, r_s could be estimated using standard "physics package" software as follows. The partial tendencies of q due only to phase changes could be written

$$Q \begin{bmatrix} q_V \\ q_W \\ q_I \\ c \end{bmatrix} = \begin{bmatrix} -(q_W + cL_c) & -(q_I + cL_s) & 0 \\ q_V & 0 & -(q_I + cL_f) \\ 0 & q_V & q_I \\ q_V L_c & q_V L_s & q_I L_f \end{bmatrix} \times \begin{bmatrix} r_c \\ r_s \\ r_f \end{bmatrix}. \qquad (8.6)$$

A standard "physics package" would predict all the partial tendencies on the left side of (8.6), and consequently overdetermine the conversion rate factors. One could ignore one of the predictions of the "physics package" (preferably the least reliable), solve the resulting partition of (8.6) to determine the conversion rate factors, and then use (8.5) to estimate the full tendencies. This procedure would provide an interface to force antisymmetry despite any incompatibilities of the package.

Another vital consideration, to guarantee the energy bound, is the generalised CFL requirement (7.4). Although a small enough timestep would suffice, a guarantee for long timesteps would be desirable. One way towards providing this could be via a condition of the kind $\|Q\| \leqslant e < 1$.

9. COMPUTATIONAL EXPERIMENTS

In this section we validate the numerical method for simulations of mountain waves. Using the simplest finite-difference approximations (3.1), (3.2) we have computed the flow over a bell-shaped mountain. We compare our results for a small hill with the analytical linear solutions of Queney in [1] and numerical experiments in [7]. For a large hill, with conditions in which a lee-wave steepens and breaks to produce a turbulent "hydraulic-jump" wake, we compare our results with similar numerical experiments by [14]. Phenomena of this kind are observed in the real atmosphere as pictured in [5] and described in [14].

Our simulations are in two-dimensions because of the computational cost of a 3D simulation, and this will be one of the limitations on the realism of our turbulence. Another limitation is that the only mechanism provided for dissipation at the shortest scales is via the damping of high frequencies inherent in the Euler-backward scheme. This damping is too weak to be physically realistic, because the timestep of 0.3 s is much shorter than the timescale of turbulence on the scales resolved by the model. However, for the present purpose we wish to avoid the complication of specifying a parameterisation of turbulence, and we wish to demonstrate that the choice of parameterisations is not prejudiced by any consideration of numerical stability.

For these simulations an explicit time scheme has been used, with a time step Δt of 0.3 s required to remain within the CFL limit. To simplify comparisons the Earth rotation rate has been set to zero. The horizontal grid spacing Δx is 2 km in the 160 km wide central region and a horizontal stretching (expansion factor 1.1) of the grid outside this region has enabled the boundaries to be pushed so far away that they do not affect the simulation. The vertical grid consists of 36 layers of equal mass giving an effective resolution (Δz) of 200 m in the lowest few kilometers. All boundaries of the model domain are rigid, fixed walls with zero normal flux. To prevent this causing spurious reflection at the top, a region of Rayleigh friction has been incorporated. The

form of this friction closely follows that of [14] and consists of adding a tendency term of the form

$$D\phi = \left\{1 + \cos\left[\pi\left(\frac{H-z}{H-z_d}\right)\right]\right\}\frac{(\bar{\phi}-\phi)\,\Delta t}{500\text{ s}} \qquad (9.1)$$

to all variables ϕ; z_d is the height above which the damping is imposed—here it is 8.5 km. $\bar{\phi}$ is the initial, upstream value of ϕ and H is the height of the model domain—here it is 16 km.

These experiments simulate flow over a mountain of the form

$$z_s = \frac{ha^2}{x^2 + a^2},$$

where z_s is the surface elevation, x is the distance from the centre of the mountain, a is the halfwidth—here it is 12 km—and h is the maximum mountain height. The model atmosphere is isothermal with surface temperature (T_s) 273 K and surface pressure (p_s) 1013 hPa. Density and pressure both decrease exponentially with height as

$$p = p_s e^{-gz/RT_s}, \qquad \rho = \rho_s e^{-gz/RT_s},$$

where here g is the gravitational acceleration, and they are initially in approximate hydrostatic balance. This implies a constant static stability parameter

$$N = \frac{g}{\sqrt{C_p T_s}} = 0.0187 s^{-1}.$$

9.1. *Perturbation of Flow by a Small Hill*

Here we compare the model with analytical solutions for flow over a small hill. For this experiment the mountain height $h = 100$ m and there is initially a uniform W–E flow of $U = 15.04\ \text{ms}^{-1}$ throughout the model domain, making the inverse Froude number

$$F_r^{-1} = \frac{Nh}{U} = 0.125,$$

putting this simulation into the regime of linear solutions. The analytical and model solutions after 4 h of integration, corresponding to a nondimensional time of

$$\frac{Ut}{a} = 18,$$

are compared in Fig. 2 and show good agreement. Vertical momentum flux calculations are displayed in Fig. 3 at hourly intervals as the simulation progresses. Except in

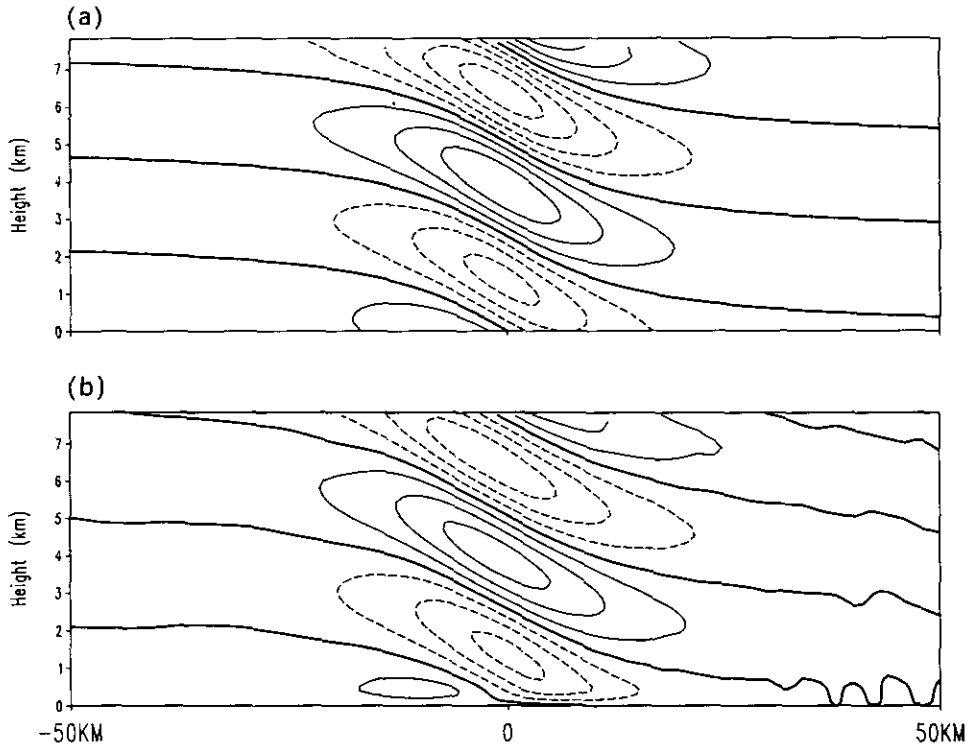

FIG. 2. Vertical cross section of the vertical velocity from $x = -50$ km to $x = +50$ km for the 100 m bell-shaped mountain experiment. Contours (negative dashed, zero highlighted) are every 0.04 ms^{-1} for (a) the analytical solution, (b) the model solution after 4 h.

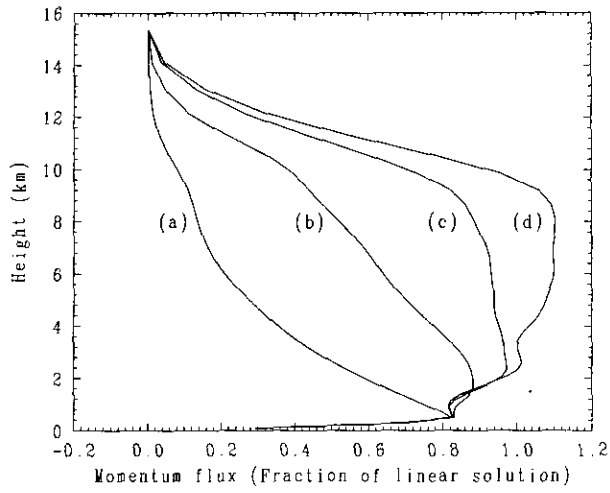

FIG. 3. Momentum flux profiles for $h = 100$ m at (a) 1 h, (b) 2 h, (c) 3 h, and (d) 4 h, normalised by their hydrostatic, linear analytical values.

the lowest kilometers, where weak nonlinearity is apparent, the vertical momentum flux is within 5 to 10% of the hydrostatic, linear analytical value consistent with the results of [7] for similar parameters. Above 8.5 km it is clear that Rayleigh friction is absorbing all the upward propagating gravity waves.

These simulations for a small hill were done using 64-bit floating point arithmetic. In this experiment we are interested in relatively small perturbations of a large mean flow, and we found these small differences to be noisy if computed with 32-bit arithmetic.

9.2. *Hydraulic-Jump over a Bell-Shaped Mountain*

We now consider a large amplitude case where non-linear effects are important. For this experiment the mountain height $h = 1050$ m and there is initially a uniform W–E flow of $U = 7.52\ \mathrm{ms}^{-1}$ throughout the model domain. Saito

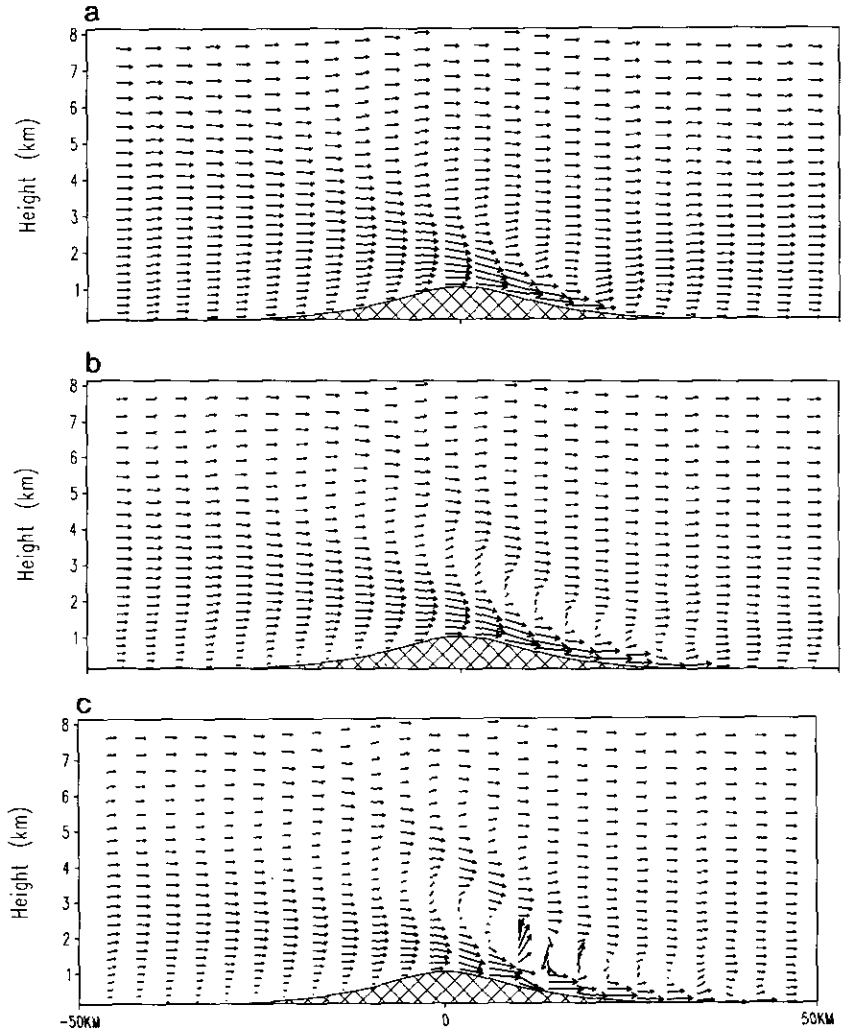

FIG. 4. Vertical cross section of the wind from $x = -50$ km to $x = +50$ km for the 1-km bell-shaped mountain experiment. (u, w) wind vectors are plotted every second point in the horizontal corresponding to a spacing of 4 km. The scaling of length of the wind vectors differs between plots; the unmodified vectors at the top of each plot correspond to speeds of $7.52\ \mathrm{ms}^{-1}$: (a) after 30 min, (b) after 1 h, (c) after 1.5 h.

and Ikawa [14] have used $N = 0.01\ s^{-1}$, $U = 4\ ms^{-1}$ and $h = 1050$ m, thus giving similar values to the initial key parameters, the vertical wavelength

$$L = \frac{2\pi U}{N} = 2.51 \text{ km}$$

and the inverse Froude number

$$F_r^{-1} = \frac{Nh}{U} = 2.611,$$

putting this simulation well into the nonlinear regime, and just into the region of parameter space where flow reversal should develop in the lee of the mountain peak. The main differences between our simulation setup and that of [14] are that we have not made the Boussinesq or anelastic approximations, but we have allowed our experiment to start abruptly with the full mountain from step one, with any resulting shocks rapidly propagating out of the area, and we do not have any explicit dissipation mechanism.

Figure 4a shows the mountain wave after 30 min beginning to get established with some acceleration of the flow in the lee of the mountain peak. A decelerated layer has also begun to appear above this. The wind vectors are plotted at every point in the vertical and every second point in the horizontal, showing the terrain-following coordinate system. In order to compare results with those in [14] and to highlight the regions of accelerated and reversed flow, Fig. 5 shows contours of the U field for the same experiment. By 60 min Figs. 4b and 5b show several accelerated and

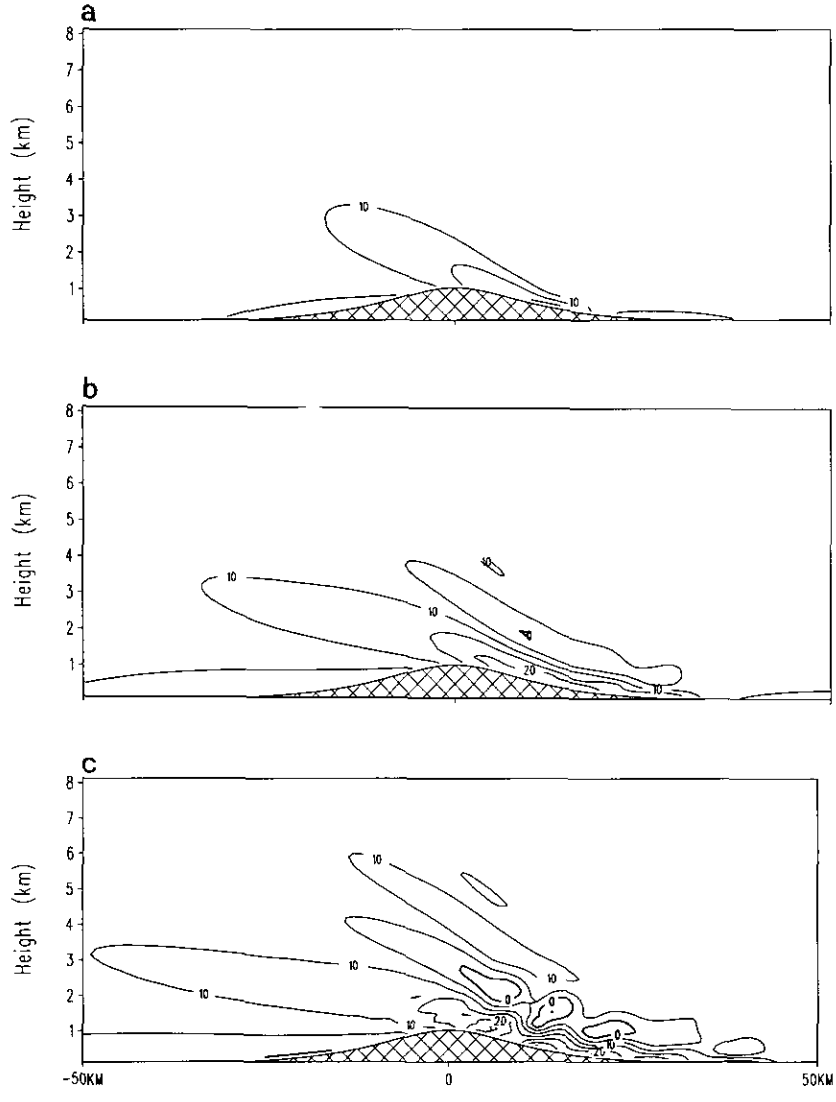

FIG. 5. Vertical cross section of the U component of the wind from $x = -50$ km to $x = +50$ km for the 1-km bell-shaped mountain experiment. Wind contours are every 5 ms^{-1} with the zero contour highlighted: (a) after 30 min, (b) after 1 h, (c) after 1.5 h.

decelerated layers with a region of reversed flow (indicating wave breaking) at a height of 2 km in the lee of the peak. The maximum low level wind is now 3.5 times the mean initial wind, which is the same ratio as in [14]. As we continue the simulation, unstable shear layers develop, and the model tries to adjust these on its smallest available scale. After 1.5 h from Fig. 4c it is clear that a hydraulic jump has formed in the lee of the mountain peak, and Fig. 5c indicates that a low level area of reversed flow is now developing upstream of the mountain, in accordance with [14].

9.3. *Modelling of Turbulence*

We do not intend to suggest that diffusion should be neglected in a simulation of turbulent flow. Rather, we have omitted it from the experiments just to demonstrate that the method is stable for the least possible amount of diffusion, even in severe circumstances, as asserted by the theory.

Figures 4c and 5c show production of turbulence on the scale of the hydraulic jump, about 2 km in the vertical and 8 km in the horizontal and, as the integration progresses further, on the scale of the grid cells as well. The rate of production of turbulent kinetic energy is clearly greater on the scale of the jump than on the smaller scale of the grid cells.

It should therefore be possible to choose a turbulence parameterisation which dissipates turbulence on small scales, but which still allows overturning on the scale of the jump. However, Saito and Ikawa [14] use a subgrid-scale turbulence parameterisation scheme of the form given by [11, 9]. This appears to prevent overturning of the streamlines and allows a steady state to develop, but looking at pictures like that of [5] it is not clear to us that this is the correct physical solution. Hydraulic jumps usually have a turbulent wake, where excess kinetic energy is dissipated and overturning on the scale of the jump height is likely. Some combination of overturning and mixing on smaller scales is probably more accurate. Small scale mixing could be modelled within the energy bounded method, using the scheme described in Section 8.1, for example.

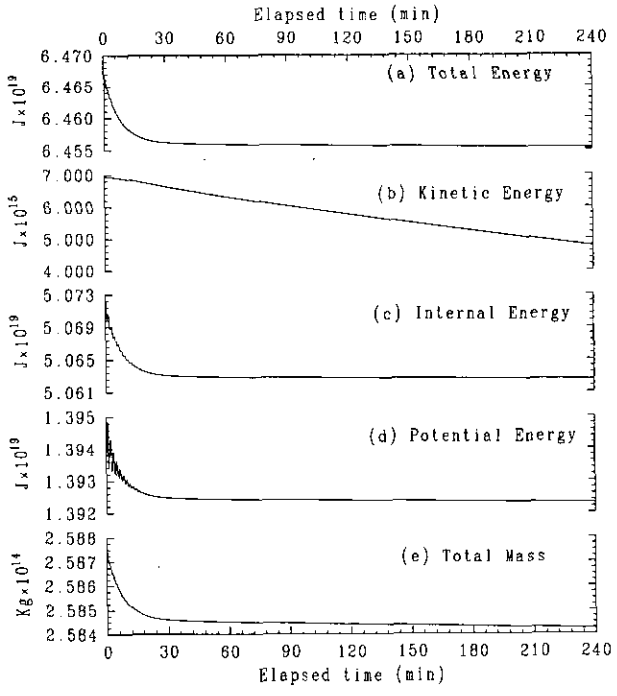

FIG. 6. Time series of (a) total energy, (b) kinetic energy, (c) internal energy, (d) potential energy, and (e) total mass for the 1-km bell-shaped mountain experiment using mks units.

9.4. *Energy Budgets*

Figure 6 shows time series of the total energy, its kinetic, internal, and potential components, and the total mass in the model domain. The oscillations during the first half hour are mainly vertically propagating acoustic waves due to the initial conditions. The Euler-backward method damps these, resulting in balanced gravitational and pressure forces. The steady loss of kinetic energy is due to an intense acoustic wave propagating inwards from the lateral boundaries (not shown). These lateral boundaries are rigid, fixed, impermeable walls, so that behind the acoustic wave the initial uniform velocity of 7.52 ms^{-1} is reduced to zero, with a conversion of kinetic to internal energy. A calculation of this effect, modified by the Rayleigh friction applied above 8.5 km and accurate to 5%, explains a reduction of kinetic energy to 90% after 1 h and 83% after 2 h, compared to the observed reductions of 91% after 1 h and 83% after 2 h shown in Fig. 6b. The Rayleigh friction is introduced in order to prevent reflection of gravity waves from the top boundary, but, on the other hand, this device complicates our total energy statistics. The scheme is evidently stable for a long period of turbulent flow, without any diffusion terms.

10. DISCUSSION AND SUMMARY

The general theoretical picture is simple: All physical processes conserve total energy, and in the "energy bounded" scheme this total energy is an l_2 norm that provides a bound on the system state. The great merit of this choice of norm is that the method is applicable to a wide variety of real physical problems because, even in complex circumstances, the total physical energy is conserved and each component of this energy is in limited supply. It may be practical to further restrict these supplies of available energy by tightening the definitions of the state variables. We have briefly outlined an extension of this theory to incorporate approximations of diffusion of momentum and phase changes of water, and we do not anticipate any real difficulty in extending the l_2 norm to include further forms of

energy. Rather, we expect that this approach will be helpful, by enforcing balance in the net budget of energy transformations, while also refusing credit on any particular form of energy. All state variables are to be formulated so as to participate in a single budget for a total energy that is an l_2 norm on the system as a whole.

The method was demonstrated for an "internal hydraulic jump" flow over a bell-shaped mountain, simulating an internal wave as it steepened into a turbulent jump. We suggest that simulation of turbulent or separating flows are applications for which the method may be particularly useful. Here the use of the word "turbulent" implies that one is not interested in much of the detail of the flow, such as the instantaneous flow within a turbulent wake, yet one may want to estimate some of the statistics of the turbulence, such as the intensity, duration, and frequency of gusts at the surface in the hydraulic jump example. While some form of "parameterisation" of mixing will be needed to allow for processes which are not resolved by the model, it may be a good strategy to resolve as much as possible, and the method allows this by avoiding the need for spatial filtering to maintain stability. We do not suggest that diffusion should be neglected in a simulation of turbulent flow. Rather, we have omitted it from the experiments just to demonstrate that the method is stable for the least possible amount of diffusion, even in severe tests, as asserted by the theory.

The "energy bounded" method guarantees that no error can grow unchecked, provided the timestep does not exceed CFL limits. This was achieved by a choice of state variables for which the total energy is an l_2 norm on the model state. Conservation of energy is then equivalent to antisymmetry of a tendency operator, both for continuous variables and for discrete approximations. This antisymmetry provides a simple rule for construction of discrete approximations that limit the total energy. The Galerkin method naturally suits this design because the Galerkin projection in these state variables cannot increase the system energy. Transposed difference operators provide a natural notation for finite-difference approximation on staggered grids and for construction of antisymmetric tendency operators.

ACKNOWLEDGMENTS

Much of this work was funded by the former New Zealand Meteorological Service. We thank Mark Sinclair and Jim Renwick of the New Zealand Meteorological Service for their help in producing the diagrams.

REFERENCES

1. M. A. Alaka, W.M.O. Tech. Note 34, 1960 (unpublished).
2. A. Arakawa, *J. Comput. Phys.* **1**, 119 (1966).
3. A. Arakawa and V. R. Lamb, *Mon. Weather Rev.* **109**, 18 (1981).
4. A. Arakawa and Y. G. Hsu, *Mon Weather Rev.* **118**, 1960 (1990).
5. A. H. Auer, Cover photograph of hydraulic jump, in *Bull. Amer. Meteor. Soc.* **54**, No. 7 (1973).
6. C. de Boor, *A Practical Guide to Splines* (Springer-Verlag, New York, 1978).
7. M. J. Weissbluth and W. R. Cotton, *Mon. Weather Rev.* **117**, 2518 (1989).
8. W. R. Cotton and G. J. Tripoli, *J. Atmos. Sci.* **35**, 1503 (1978).
9. J. W. Deardorff, *Boundary-Layer Meteor.* **18**, 495 (1980).
10. A. Greenbaum and J. M. Ferguson, *J. Comput. Phys.* **64**, 97 (1986).
11. J. B. Klemp and R. B. Wilhelmson, *J. Atmos. Sci.* **35**, 1070 (1978).
12. K. W. Morton and P. K. Sweby, *J. Comput. Phys.* **73**, 203 (1987).
13. R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems* (Interscience Wiley, New York, 1967), p. 300.
14. K. Saito and M. Ikawa, *J. Met. Soc. Jpn* **69**, 31 (1991).
15. J. Zhongzhen and Z. Quigcun, in *Short- and Medium-Range Numerical Weather Prediction, 1986*, edited by T. Matsuno (*J. Met. Soc. Japan*, Tokyo, 1987).