Combined Penalty Multiplier Optimization Methods to Enforce Integral Invariants Conservation

I. M. NAVON AND R. DE VILLIERS

National Research Institute for Mathematical Sciences, CSIR, P.O. Box 395, Pretoria 0001 South Africa

(Manuscript received 6 July 1982, in final form 15 March 1983)

ABSTRACT

An augmented Lagrangian multiplier–penalty method is applied for the first time to solving the problem of enforcing simultaneous conservation of the nonlinear integral invariants of the shallow water equations on a limited-area domain. The method approximates the nonlinearly constrained minimization problem by solving a series of unconstrained minimization problems.

The computational efficiency and accuracy of the method is tested using two finite-difference solvers of the nonlinear shallow water equations on a β-plane. The method is also compared with a pure quadratic penalty approach. The updating of the Lagrangian multipliers and the penalty parameters is done using procedures suggested by Bertsekas. The method yielded satisfactory results in the conservation of the integral constraints while the additional CPU time required did not exceed 15% of the total CPU time spent on the numerical solution of the shallow water equations. The methods proved to be simple in their implementation and they have a broad scope of applicability to other problems involving nonlinear constraints; for instance, the variational nonlinear normal mode initialization.

1. Introduction

It has become evident through the work of Arakawa and Lamb (1981), Fjörtöft (1953), Arakawa (1966), Lilly (1965), Sadourny (1980, 1975), that the maintenance, in the discrete representation, of the integral constraints satisfied by invariants associated with partial differential equations can help inhibit or prevent nonlinear instability.

There also seems to be a general consensus in the finite-difference literature that discretization schemes which assure conservation of quadratic properties are more stable than other possible discretization combinations. On the other hand, criticism of such formulations centers on two issues:

1) Conservative discretization schemes do not necessarily imply more accurate results.

2) How many, and which integral invariants should be conserved in the discretized form [see Lee *et al.* (1980)].

Kalnay-Rivas (1979) using the GLAS fourth-order global atmospheric model showed that schemes that do not formally conserve enstrophy in practice do so with high accuracy if waves shorter than four times the grid size are periodically removed before they attain significant amplitudes. However this method encounters problems in the case of a limited-area domain (see Navon, 1981). Many modellers have taken this approach similar to the one proposed by Phillips (1959). Considerable effort has been dedicated to the design of spatial finite-difference schemes for the shallow-water equations that retain the integral constraints of the continuous system. The methods currently available for achieving this can be divided into two main categories:

On the one hand we have what are called *a priori* schemes which utilize finite difference combinations which can be shown to conserve the integral invariants of interest for the time continuous case. These *à priori* numerical schemes result, however, in rather complicated finite-difference expressions which are difficult to generalize to fluid dynamics problems of interest. On the other hand the recently developed *à posteriori* methods enforce the required discretized quantities explicitly by modifying the forecast field values after every so many time-steps following some prescribed criterion.

In this second category we have the Sasaki (1976, 1977) and Sasaki *et al.* (1979, 1980) variational approach and the Bayliss and Isaacson (1975) method which makes a given finite-difference scheme conservative with respect to any given quantity. The Bayliss–Isaacson technique was tested by Isaacson (1977) as well as by Kalnay-Rivas *et al.* (1977). Navon (1981) tested both a modified Sasaki variational approach and modified Bayliss–Isaacson technique, both designed to enforce conservation of potential enstrophy and total mass in two ADI finite-difference approximations of the nonlinear shallow-water equations, Navon (1978), Navon and Riphagen (1979). In the present research a new general *à posteriori*

© 1983 American Meteorological Society

approach is proposed. This approach is derived from viewing the enforcing of conservation of integral invariants as a nonlinearly constrained optimization problem, with nonlinear equality constraints. These constraints require that at each time step three discretized integral invariants of the shallow water equations, i.e. the mass, energy and enstrophy invariants should be approximately equal to their initial values.

Both an augmented Lagrangian multiplier penalty method and a pure quadratic penalty method are employed. The multiplier methods were first developed by Powell (1969) and Hestenes (1969) and further developed by Bertsekas (1973, 1975a, 1975b, 1975c, 1976a, 1976b, 1980). These methods approximate the nonlinearly constrained minimization problem by solving a series of unconstrained minimization problems. The augmented Lagrangian function approach was first studied by Arrow and Solow (1958) and then by Fiacco and McCormick (1968).

2. The augmented Lagrangian multiplier method—motivation and theory

a. The penalty function

When solving a general nonlinear programming problem in which the constraints cannot be easily eliminated, it is necessary to balance the aims of reducing the objective function and staying inside or close to the feasible region when we consider the problem:

$$\text{minimize } f(\mathbf{x})$$

$$\text{subject to equality constraints } \mathbf{e}(\mathbf{x}) = 0. \quad (1)$$

This inevitably leads to the idea of a penalty function which is a combination of the objective function f and the constraints $\mathbf{e}(\mathbf{x})$, which enables the minimization of the objective function f whilst controlling constraint violations by penalizing them (Fletcher, 1981).

For equality constraints the earliest penalty function is due to Courant (1943) and takes the form

$$L(\mathbf{x}, r) = f(\mathbf{x}) + \frac{1}{2r}|e(\mathbf{x})|^2 = f(\mathbf{x}) + \frac{\sigma}{2}|e(\mathbf{x})|^2. \quad (2)$$

This method tends to become numerically unstable in the final stages of the computation as it involves the product of a large number σ_k by a short vector $e(\mathbf{x}_k)$, a procedure that is subject to considerable round-off errors.

It is also hampered by slow convergence and numerical instabilities associated with ill-conditioning, induced by very large values of the penalty parameter. This is expressed mathematically by the condition number of the Hessian matrix $\nabla^2 L$ approaching infinity.

b. Primal–dual method (Lagrange multipliers method)

Here again we consider $p\mathbf{b}(1)$, associated with the augmented Lagrangian

$$L(\mathbf{x}, \mathbf{u}) = f(x) + \mathbf{u}^T e(\mathbf{x}) \quad (3)$$

such that $(x_0, \mathbf{u}_0)$ is the solution of the equations

$$L_{\mathbf{x}}(\mathbf{x}, \mathbf{u}) = f_{\mathbf{x}} + \mathbf{u}^T e_{\mathbf{x}}(\mathbf{u}), \quad (4)$$

$$L_{\mathbf{u}}(\mathbf{x}, \mathbf{u}) = e(\mathbf{x}) = 0. \quad (5)$$

One can use a method of sequential minimization of the Lagrangian function, where the vector u_k is updated by

$$u_{k+1}^i = u_k^i + \alpha_k e_i(\mathbf{x}_k). \quad (6)$$

This iteration may be viewed as a steepest descent iteration aimed at finding an optimal solution of an associated dual problem. This is the reason that this algorithm is also called a primal–dual method. The disadvantages of this method come first from the fact that $p\mathbf{b}(1)$ must have a locally convex structure in order for iteration (6) to be meaningful.

Second it is necessary to minimize the Lagrangian function (3) a large number of times since the ascent iteration (6) converges only moderately fast. In the last few years the methods of multipliers were proposed, in which the penalty idea is merged with the primal–dual philosophy.

To illustrate these points we will consider the following simple example (Gill *et al.*, 1982): Minimize the function

$$f(\mathbf{x}) = x^3, \quad x \in \mathcal{R}^1 \quad (7)$$

$$\text{subject to the constraint } x + 1 = 0. \quad (8)$$

The unique solution is $x^* = -1$ and $u^* = 3$ where u^* is the Lagrange multiplier.

Thus an augmented Lagrangian (with Lagrange multiplier only) is

$$L(x, u^*) = x^3 - 3(x + 1). \quad (9)$$

However x^* *is not* a local minimum of $L(x, u^*)$. However if we add to (9) the penalty term $\rho|e(x)|^2$ then for all $\rho > \bar{\rho}$ ($\bar{\rho} = 6$), x^* is a local minimum of the new augmented Lagrangian (combined penalty and multiplier)

$$L(x, u^*, \rho) = x^3 - 3(x + 1) + \frac{\rho}{2}(x + 1)^2. \quad (10)$$

c. The theoretical set-up

We consider the problem

$$\text{minimize } F(\mathbf{x}), \quad \mathbf{x} \in \mathcal{R}^n, \quad (1^*)$$

subject to $\mathbf{e}(\mathbf{x}) = 0$ or $e_i(\mathbf{x}) = 0$, $i = 1, \ldots, m$. The gradient vector of the constraints function $e_i(\mathbf{x})$ is denoted by the vector $a_i(\mathbf{x})$ and its Hessian will be denoted $\hat{\mathbf{G}}_i(\mathbf{x})$.

Given a set of constraint functions $[e_i(\mathbf{x}),\ i = 1, \ldots, m]$ the $m \times n$ matrix $\hat{\mathbf{A}}(\mathbf{x})$, whose i-th row is $a_i(\mathbf{x})$, is called the Jacobian matrix of the constraints.

If $\mathbf{x}^*$ is the optimal point for the nonlinear equality constraint problem, let $\mathbf{Z}(\mathbf{x}^*)$ denote a matrix whose columns form a basis for the set of vectors orthogonal to the rows of $\hat{\mathbf{A}}^* = \hat{\mathbf{A}}(\mathbf{x}^*)$.

Define then the Lagrangian function as

$$L(\mathbf{x}, \mathbf{u}) = F(\mathbf{x}) + \mathbf{u}^T\mathbf{e}(\mathbf{x}). \tag{11}$$

The Hessian with respect to $\mathbf{x}$ of the Lagrangian function is

$$\mathbf{W}(\mathbf{x}, \mathbf{u}) = \mathbf{G}(\mathbf{x}) + \sum_{i=1}^{m} u_i\hat{\mathbf{G}}_i(\mathbf{x}), \tag{12}$$

where

$$\mathbf{G}(\mathbf{x}) = \nabla^2_{,}F(\mathbf{x}) \tag{13}$$

and is the Hessian matrix of $F(\mathbf{x})$.

The following conditions are necessary conditions for x^* to be optimal (i.e. a minimum) for the nonlinear equality constrained problem:

1) $$\mathbf{e}(\mathbf{x}^*) = 0, \tag{14}$$

2) $$\mathbf{Z}(\mathbf{x}^*)^T g(\mathbf{x}^*) = 0 \quad \text{or} \quad g(\mathbf{x}^*) = \hat{\mathbf{A}}(\mathbf{x}^*)^T\mathbf{u}^*, \tag{15}$$

where

$$g(\mathbf{x}) = \nabla F(\mathbf{x}) \tag{16}$$

is the gradient vector of $F(\mathbf{x})$,

3) $\mathbf{Z}(\mathbf{x}^*)^T\mathbf{W}(\mathbf{x}^*, \mathbf{u}^*)\mathbf{Z}(\mathbf{x}^*)$ is positive semi-definite while the sufficient conditions imply again 1) and 2) while they require that 3) be positive definite.

$\mathbf{Z}(\mathbf{x}^*)^T\mathbf{W}\mathbf{Z}(\mathbf{x}^*)$ is called the *projected Hessian of the Lagrangian function* and $\mathbf{x}^*$ is a minimum of the Lagrangian function within the subspace of vectors orthogonal to the active constant gradients.

This suggests the construction of the combined penalty multiplier augmented Lagrangian by augmenting the Lagrangian function L with a quadratic penalty term, i.e.

$$\tilde{L}(\mathbf{x}, \mathbf{u}, r) = F(\mathbf{x}) + \mathbf{u}^T\mathbf{e}(\mathbf{x}) + \frac{1}{2r}\mathbf{e}(\mathbf{x})^T e(\mathbf{x}), \tag{17}$$

which retains the stationarity properties of x^* but alters the Hessian in the subspace of vectors defined by $\hat{\mathbf{A}}(\mathbf{x}^*)$ (Gill *et al.*, 1981b).

As the quadratic penalty term and its gradient vanish at x^* if $\mathbf{u} = \mathbf{u}^*$, then x^* is a stationary point of (17). The Hessian matrix of the penalty term in (17) is

$$\sum_i e_i(\mathbf{x})\hat{\mathbf{G}}_i(\mathbf{x}) + \hat{\mathbf{A}}(\mathbf{x})^T\hat{\mathbf{A}}(\mathbf{x}). \tag{18}$$

But at x^* where $e(\mathbf{x}^*) = 0$ we get that the Hessian of the quadratic penalty term reduces to $\hat{\mathbf{A}}(\mathbf{x}^*)^T\hat{\mathbf{A}}(\mathbf{x}^*)$ which is a positive semi-definite matrix with strictly positive eigenvalues which correspond to eigenvectors in the range of $\hat{\mathbf{A}}(\mathbf{x}^*)^T$.

Hence the adding of the positive quadratic penalty term increases only the eigenvalues of $\mathbf{W}$ corresponding to eigenvectors in the range-space of $\mathbf{A}(\mathbf{x}^*)^T$ and leaves the other eigenvalues unchanged. As Gill *et al.* point out, it can thus be shown under mild conditions that there exists a finite threshold penalty $\bar{\rho} = 1/r$ such that for all $\rho > \bar{\rho}$, $\mathbf{x}^*$ is an unconstrained minimum of $\tilde{L}(\mathbf{x}, \mathbf{u}^*, r)$ for all $\rho > \bar{\rho}$, and as $\mathbf{Z}(\mathbf{x})$ is a matrix orthogonal to the rows of $\hat{\mathbf{A}}(\mathbf{x})$ it holds that:

$$\mathbf{Z}(\mathbf{x}^*)^T\nabla^2_{xx}\tilde{L}(\mathbf{x}^*, \mathbf{u}^*, r)\mathbf{Z}(\mathbf{x}^*) = \mathbf{Z}(\mathbf{x}^*)^T\mathbf{W}(\mathbf{x}, \mathbf{u}^*)\mathbf{Z}(\mathbf{x}^*) \tag{19}$$

which means that the addition of the penalty term is not affecting the projected Hessian of the Lagrangian function at $\mathbf{x}^*$. Thus the effect of the penalty term is that of a convexification effect.

3. The augmented Lagrangian multiplier method—application

We start by defining a functional

$$f = \sum_{j=1}^{N_x}\sum_{k=1}^{N_y}[\tilde{\alpha}(u - \tilde{u})^2 + \tilde{\alpha}(v - \tilde{v})^2 + \tilde{\beta}(h - \tilde{h})^2]_{jk}, \tag{20}$$

where

$$N_x\Delta x = L, \quad N_y\Delta y = D, \tag{21}$$

and where $\Delta x = \Delta y = h$ is the grid size, n designates the time-level $t_n = n\Delta t$, where Δt is the time step and L and D are the respective dimensions of the rectangular domain.

$(\tilde{u}, \tilde{v}, \tilde{h})^n_{jk}$ are the predicted variables at the nth time-step using a finite-difference algorithm for solving the shallow-water equations—while $(u, v, h)^n_{jk}$ are the values adjusted by the nonlinear constrained optimization method so as to enforce the three conservation-laws.

Here $\tilde{\alpha}$ and $\tilde{\beta}$ are weights determined following Sasaki (1976)'s principle that the relative weights are so chosen as to make the fractional adjustment of variables proportional to the fractional magnitude of the truncation errors in the predicted variables. Here we took

$$\tilde{\alpha} = 1, \quad \tilde{\beta} = g/H, \tag{22}$$

H being the mean depth of the shallow fluid, and we adopt the same three basic principles as Sasaki (1976). We then define the following augmented Lagrangian function L by

$$L(\mathbf{x}, \mathbf{u}, r) = f(\mathbf{x}) + \mathbf{u}^T\mathbf{e}(\mathbf{x}) + \frac{1}{2r}|\mathbf{e}(\mathbf{x})|^2. \tag{23}$$

While considering the problem

$$\text{minimize } f(\mathbf{x}) \tag{24}$$

subject to the equality constraints $\mathbf{e}(\mathbf{x}) = 0$,

where

$$\mathbf{x} = (\tilde{u}_{11} \cdots \tilde{u}_{N_xN_y}, \quad \tilde{v}_{11} \cdots \tilde{v}_{N_xN_y}, \quad \tilde{h}_{11} \cdots \tilde{h}_{N_xN_y})^n, \tag{25}$$

$\mathbf{e}(\mathbf{x})$ is a vector of three nonlinear quantities given by

$$\mathbf{e}(\mathbf{x}) = \begin{cases} E^n - E^0 \\ Z^n - Z^0 \\ H^n - H^0, \end{cases} \tag{26}$$

where

$$E^n = \frac{1}{2} \sum_{j=1}^{N_x} \sum_{k=1}^{N_y} [\tilde{h}(\tilde{u}^2 + \tilde{v}^2) + g\tilde{h}^2]^n_{jk} \Delta x \Delta y, \tag{27a}$$

$$Z^n = \frac{1}{2} \sum_{j=1}^{N_x} \sum_{k=1}^{N_y} \left[\frac{\dfrac{\partial \tilde{v}}{\partial x} - \dfrac{\partial \tilde{u}}{\partial y} + f}{\tilde{h}} \right]^2_{n\,jk} \Delta x \Delta y, \tag{27b}$$

$$H^n = \sum_{j=1}^{N_x} \sum_{k=1}^{N_y} \tilde{h}_{jk} \Delta x \Delta y. \tag{27c}$$

Here E^n, Z^n and H^n are the values of the discrete integral invariants of total energy, potential enstrophy and mass at time $t_n = n\Delta t$, while E^0, Z^0 and H^0 are the values of the same integral invariants at the initial time $t = 0$. In general if we have m equality constraints then the vector $\mathbf{e}(\mathbf{x})$ is given by

$$\mathbf{e}(\mathbf{x}) = (e_1(\mathbf{x}) \cdots e_m(\mathbf{x})). \tag{28}$$

The vector $\mathbf{u}$ is the m component multiplier vector

$$\mathbf{u} = (u_1, u_2, \ldots, u_m), \tag{29}$$

while r is the penalty parameter (which can be different for each constraint).

The basic idea of the multiplier method is to solve the constrained minimization problem by performing a sequence of unconstrained minimizations of the following problem.

$$\underset{x \in R^n}{\text{minimize}}\ L_{r_k}(\mathbf{x}, \mathbf{u}_k)$$

$$= f(\mathbf{x}) + \sum_{i=1}^{n} u^i_k e_i(\mathbf{x}) + \frac{1}{2r_k} |\mathbf{e}(\mathbf{x})|^2. \tag{30}$$

This is based on the following proposition proved by Bertsekas (1975c, 1980):

a. Proposition [Bertsekas (1975)]

For $k = 0, 1, \ldots,$ let x_k be a global minimum of the problem

$$\text{minimize } L_{r_k}(\mathbf{x}, \mathbf{u}_k)$$

$$\text{subject to } x \in \mathcal{R}^n, \tag{31}$$

where $\{u_k\}$ is bounded and $0 < r_{k+1} < r_k$ for all k, and $r_k \rightarrow 0$. Then every limit point of the sequence $\{x_k\}$ is a global minimum of f subject to the equality constraints $\mathbf{e}(\mathbf{x}) = 0$.

The method, described in detail in the next sections, consists of a sequence of unconstrained minimizations of the Lagrangians $L_{r_k}(\mathbf{x}, u_k)$. Given a multiplier vector $\mathbf{u}_k$ and a penalty parameter r_k we minimize $L_{r_k}(\mathbf{x}, u_k)$ over R^n and obtain a vector $\mathbf{x}_k$. The variable $\mathbf{u}_k$, the vector of Lagrange multipliers and the penalty parameters are held fixed during the minimization and then updated prior to the next unconstrained minimization. The algorithm is typically terminated at a point $\mathbf{x}_k$ where

$$|\nabla_x L_{r_k}(\mathbf{x}_k, \mathbf{u}_k)| \leqslant \epsilon_k$$

or

$$|e_i(\mathbf{x}_k)| < \epsilon'_k, \ i = 1, \ldots, m, \tag{32}$$

where ϵ_k and ϵ'_k are some small scalars.

As proved by Bertsekas (1975c) the multiplier method has the advantage over the penalty method in that it does not require r_k to decrease to zero (i.e., to use very large values of r^{-1}) in order to induce convergence in the method of multipliers.

Thus, the difficulties associated with ill-conditioning are avoided. A second advantage of the method of multipliers is that its rate of convergence is considerably better than that of the penalty method (Bertsekas (1975c)).

b. Updating the multiplier vector **u** *in the method of multipliers*

Given a multiplier vector $\mathbf{u}_k$ and penalty parameters r^i_k $(i = 1, \ldots, m)$ one minimizes $L_{r_k}(\mathbf{x}, \mathbf{u}_k)$ over R^n obtaining a vector $\mathbf{x}_k$. Then, following Bertsekas (1982, 1975c) we used two methods to update the multiplier vector. The first is

$$\mathbf{u}_{k+1} = \mathbf{u}_k + r_k^{-1}\mathbf{e}(\mathbf{x}_k). \tag{33}$$

This method is derived from the fact that at the solution $\mathbf{x}_k$ of the problem

$$\underset{x \in \mathcal{R}^n}{\text{minimize}}\ \tilde{L}(\mathbf{x}, \mathbf{u}_k, r) \tag{34}$$

it holds that

$$g(\mathbf{x}_k) - \hat{\mathbf{A}}(\mathbf{x}_k)^T \mathbf{u}_k + \frac{1}{r} \hat{\mathbf{A}}(\mathbf{x}_k)^T \mathbf{e}(\mathbf{x}_k) = 0. \tag{35}$$

Hence the new multiplier estimate (33). Since from (35) we observe that the vector on the rhs of (33) contains the coefficients in the expansion of $g(\mathbf{x}_k)$ as a linear combination of the rows of $\hat{\mathbf{A}}(\mathbf{x}_k)$.

A second-order multiplier iteration (see Bertsekas (1980) is given by

$$\mathbf{u}_{k+1} = \mathbf{u}_k - [\nabla^2 d_{r_k}(\mathbf{u}_k)]^{-1}\nabla d_{r_k}(\mathbf{u}_k), \qquad (36)$$

where

$$\nabla d_{r_k}(\mathbf{u}_k) = e[x(u_k, r_k)], \qquad (37)$$

$$\nabla^2 d_{r_k}(u_k) = \hat{A}[x(u_k, r_k)]^T \times \{\nabla^2_{xx} L_{r_k}[x(u_k, r_k)u_k]\}^{-1}\hat{A}[x(u_k, r_k)]. \qquad (38)$$

The second method is a step-size rule proposed by Bertsekas (1975) and used at every second unconstrained minimization iteration. Given $\mathbf{u}_{2k}$ and r_{2k} one obtains $\mathbf{x}_{2k}$ and $e(\mathbf{x}_{2k})$ by unconstrained minimization of the augmented Lagrangian and we set

$$u_{2k+1} = u_{2k} + r_{2k}^{-1}\mathbf{e}(\mathbf{x}_{2k}). \qquad (39)$$

Again $\mathbf{x}_{2k+1}$ and $e(\mathbf{x}_{2k+1})$ are obtained by the unconstrained minimization of the augmented Lagrangian.

However, we now set

$$\mathbf{u}_{2k+2} = \mathbf{u}_{2k+1} + \alpha_{2k+1}\mathbf{e}(\mathbf{x}_{2k+1}), \qquad (40)$$

where

$$\alpha_{2k+1} = r_{2k+1}^{-1}\frac{\mathbf{e}(\mathbf{x}_{2k+1})^T\mathbf{e}(\mathbf{x}_{2k})}{\mathbf{e}(\mathbf{x}_{2k+1})^T\mathbf{e}(\mathbf{x}_{2k}) - \|\mathbf{e}(\mathbf{x}_{2k+1})\|^2}. \qquad (41)$$

The initial multiplier vector $\mathbf{u}_0$ is chosen arbitrarily or on the basis of some linearized analysis (see Sasaki (1976), Sasaki *et al.* (1979), Sasaki and Reddy (1980)).

c. *Updating of the penalty parameters*

The main considerations for selecting an initial penalty parameter sequence and updating it have been clearly exposed by Bertsekas (1980) and Kort and Bertsekas (1976). The initial penalty parameters r_0^i (a separate parameter is chosen for each equality constraint) should be chosen so that they are not so small that ill-conditioning results in the first unconstrained minimization.

As for the updating process, a reasonable penalty parameter adjustment scheme is to decrease the penalty factors r_k^i by a factor $\beta < 1$, only if the constraint vector violation as measured by $|\mathbf{e}(\mathbf{x}(u_k, r_k^i)|$ is not decreased by a factor $\gamma < 1$ over the previous unconstrained minimization (Bertsekas, 1975c; Bertsekas, 1976b; Rockafellar, 1973). The factor $\beta < 1$ is chosen so that r_k^i do not decrease too fast (i.e., r_k^{i-1} increase too fast) resulting in ill-conditioning but do not decrease too slowly so that a poor convergence rate is induced.

The final penalty parameter adjustment scheme took the form:

$$r_{k+1} = \begin{cases} \beta_{r_k}, & \text{if} \quad |e[\mathbf{x}(\mathbf{u}_k, r_k)]| > \gamma|e[\mathbf{x}(\mathbf{u}_{k-1}, r_{k-1})]| \\ r_k, & \text{if} \quad |\mathbf{e}[\mathbf{x}(u_k, r_k)]| \le \gamma|e[\mathbf{x}(\mathbf{u}_{k-1}, r_{k-1})]|. \end{cases} \qquad (42)$$

By using separate penalty parameters for the different constraints we are already performing a certain scaling of the constraints [Bertsekas (1976b)]. In our runs we used $\beta = (0.4)^k$ and $\gamma = 0.25$.

d. *Inexact minimization of the augmented Lagrangian*

This method advocated by Bertsekas (1975c), Kort (1975) and Buys (1972) is a modification of the basic algorithm aimed at improving computational efficiency. Instead of requiring unconstrained minimization of the augmented Lagrangian to be carried out exactly, only moderate accuracy is demanded in the initial minimizations, i.e., one terminates the unconstrained minimization at a point $\mathbf{x}_k$ such that:

$$\|\nabla_k L_{r_k}(\mathbf{x}_k, u_k)\| \le \epsilon_k, \qquad (43)$$

and then the accuracy is increased at later iterations by using a preselected decreasing sequence $\{\epsilon_k\}$ tending to zero. The convergence analysis relating to multiplier methods with inexact minimizations is given in Bertsekas (1975c) and Kort (1975). Our computational experience indicates that in our case we could not achieve reasonable computational results without this method.

Bertsekas (1975c, 1976) showed that by using the stopping criterion

$$\|\nabla_x L_{r_k}(\mathbf{x}_k, u_k)\| \le \eta_k\|e(\mathbf{x}_k)\|, \qquad (44)$$

with $\{\eta_k\}$ a decreasing sequence tending to zero, a much better asymptotic rate of convergence is obtained than by using (43) [see Kort (1975)]. The asymptotic rate of convergence with (44) is identical to the one associated with exact unconstrained optimization [see Kort (1975)]. A typical choice of the sequence $\{\eta_k\}$ used as stop rule parameter was in our case $\eta_k = (0.8)^k$. Note that $\eta^k = 0$ corresponds to exact minimization, which, in our case, means setting the termination parameters of the unconstrained minimization routine to 10^{-6}.

e. *Scaling by transformation of variables and constraints*

1) Rough scaling

One of the crucial issues in the success of the solution of given nonlinear constrained optimization problem is the issue of scaling. Scaling by variable transformation converts the variables from units that reflect the physical nature of the problem to units that display desirable properties for the minimization process (Gill *et al.*, 1981a; Gill *et al.*, 1981b). Due to the different physical units used in our problem the integral invariants initially had enormously different magnitudes which varied over a range factor of 10^{20}. In order to improve the performance of the optimization procedure, i.e. to transform the constrained optimization problem into a better conditioned one, and thus one more amenable to solution, we scaled the variables x_i, and thus implicitly the constraints. The first basic rule of scaling is that the scaled variables should be of similar magnitude and of order unity in the region of interest. The same applies to the nonlinear equality constraints in order to avoid a situation where one constraint persistently dominates the other constraints.

We first determined typical scaling factors for the horizontal length L the velocity V and the time T by applying a dimensional analysis to the three integral constraints of mass, total energy and total enstrophy denoted by H, E and Z respectively. We obtained the following relationships for the system of the three equality constraints:

$$\left.\begin{aligned} H_S &= HL^{-1} \\ E_S &= L^{-2}(V^2L)^{-1}E = EV^{-2}L^{-3} \\ Z_S &= L^{-2}L(LV^{-1})^2Z = LV^{-2}Z \\ T &= LV^{-1} \end{aligned}\right\}, \tag{45}$$

where H_S, E_S and Z_S are the scaled values of the total mass, total energy and enstrophy, respectively. Solving the system (20) we chose the values

$$L = 5.5 \times 10^4, \quad V = 2 \times 10^3, \quad T = 26.5. \tag{46}$$

The variables in the vector **x** were then scaled as follows:

$$\left.\begin{aligned} &u_{ij}^s = u_{ij}V^{-1}, \quad v_{ij}^s = v_{ij}V^{-1} \\ &h_{ij}^s = h_{ij}L^{-1}, \quad i = 1, \ldots, N_x \\ &f_j^s = Tf_j, \quad g^s = gLV^{-2}, \quad j = 1, \ldots, N_y \end{aligned}\right\}. \tag{47}$$

In our case due to the fact that we do not allow large constraint violations, we knew a realistic range of values that a variable might assume during the minimization. After the scaling the initial values of the total mass enstrophy and total energy were

$$H = 0.020, \quad Z = 6.95, \quad E = 0.586, \tag{48}$$

respectively.

By using separate penalty parameters for each constant, which corresponds to a mere scaling of these constraints, (Bertsekas, 1976) we can refine the scaling.

2) The Haarhoff–Buys–Molendorff (1969) scaling procedure

This procedure is implemented after some rough scaling of the object function, constraints and variables has been carried out beforehand.

The scaling factors by which unscaled values are divided to obtain scaled values are determined as follows:

(i) The scaling factor for the object function is chosen as the absolute value of the initial function value

$$FSC = |F(x_1, \ldots, x_n)| \tag{49}$$

(ii) To scale the variables x_i we require as a first approximation typical coefficients in the Taylor series expansions of all x_i to be of order unity after scaling. Thus, if f^s, e^s and x^s are the scaled values of F, the constraints e_i and x_i the relation

$$\frac{\partial f^s}{\partial x^s} \leqslant 1 \tag{50}$$

implies that the scaling factor XSC_i for x_i will satisfy

$$XSC_i \leqslant FSC \Big/ \left|\frac{\partial F}{\partial x_i}\right| = (XSC_i)_1. \tag{51}$$

Similarly

$$XSC_i \leqslant FSC \Big/ \left[\frac{\partial^2 F}{\partial x_i^2}\right]^{1/2} = (XSC_i)_2. \tag{52}$$

If $\partial e_j/\partial x_i$ is not very small, the relation

$$\frac{\partial^2 e_j^s}{\partial x_i^{s2}} \leqslant \frac{\partial e_j^s}{\partial x_i^s} \tag{53}$$

may be used leading to

$$XSC_i \leqslant \left|\frac{\partial e_j}{\partial x_i}\right| \left|\frac{\partial^2 e_j}{\partial x_i^2}\right|^{-1} = (XSC_i)_{3j}, \quad j = 1, \ldots, m. \tag{54}$$

XSC_i is first taken to be

$$XSC_i = \min((XSC_i)_1, (XSC_i)_2, (XSC_i)_{3j, j=1,\ldots,m}) \tag{55}$$

provided

$$BOT \leqslant XSC_i \leqslant TOP, \tag{56}$$

where TOP and BOT are specified.

As the constraints e_i are zero at the optimum, $(XSC_i)_{3j}$ provides an estimate of the distance in which $\partial e_j/\partial x_i$ changes by an amount equal to its magnitude. $(XSC_i)_{3j}$ is omitted from (55) when the various derivatives of e_j^s are of same order of magnitude i.e. when

$$\left|\frac{\partial e_j^s}{\partial x_i^s}\right| \sim \left|\frac{\text{grad}e_j^s}{N^{1/2}}\right|, \quad N = 3N_xN_y, \tag{57}$$

so that $(XSC_i)_3$ may be omitted when

$$\left|\frac{\partial e_j^s}{\partial x_i^s}\right| < \left|\frac{\text{grad}e_j^s}{10N^{1/2}}\right|. \tag{58}$$

Also we will omit $(XSC_i)_{3j}$ when the gradient of e_j increases while its derivatives with respect to x_i decrease (in the unscaled case). This condition for omission is

$$\left|\frac{\partial e_j}{\partial x_i}\right| < \left(\frac{|\text{grad}e_j|}{10N^{1/2}}\right)\left(\frac{BOT}{TOP}\right). \tag{59}$$

(iii) The scaling factors ESG for the constraining functions e_j are found by requiring

$$|\mathrm{grad} e_j^s| = 1 \tag{60}$$

after scaling so that

$$ESG = \left[\left(\frac{\partial e_j}{\partial x_1^s}\right)^2 + \cdots \left(\frac{\partial e_j}{\partial x_N^s}\right)^2\right]^{1/2}$$

$$= \left[XSC_1^2\left(\frac{\partial e_j}{\partial x_1}\right)^2 + \cdots + XSC_N^2\left(\frac{\partial e_j}{\partial x_N}\right)^2\right]^{1/2}. \tag{61}$$

(d) The final estimate of the XSC_i is done as follows. For a given repair time we can write the augmented Lagrangian as

$$\tilde{L} = f^s(\mathbf{x}) - \sum_{j=1}^{m} u_j^s e_j^s + \frac{1}{2r_k}\sum_{j=1}^{m} (\tilde{e}_j^s)^2 \tag{62}$$

For a given unconstrained minimization iteration we can assume $u_i, \ldots, u_m$ and $1/2r_k = B$ to be constants.

As an approximation we assume e_j to be linear functions of k_i and then

$$\frac{\partial^2 \tilde{L}}{\partial x_i^{s2}} = \frac{\partial^2 f^s}{\partial x_i^{s2}} + 2B\sum_{j=1}^{m}\left(\frac{\partial e_j^s}{\partial x_i^s}\right)^2. \tag{63}$$

Then

$$XSC_i = \left[\frac{1}{FSC}\frac{\partial^2 f}{\partial x_i^2} + 2B\sum_{j=1}^{m}\left(\frac{1}{ESG}\frac{\partial e_j}{\partial x_i}\right)^2\right]^{-1/2}. \tag{64}$$

The first and second derivatives of f and g with respect to each x_i are required in the calculations.

4. The numerical algorithms

In all the algorithms to be described one starts by defining a set of relative error bounds for the discretized integral constraints enforcing conservation of mass, total energy and potential enstrophy, i.e.,

$$\left|\frac{E^{(n)} - E^{(0)}}{E^0}\right| \leqslant \delta_E, \quad \left|\frac{H^{(n)} - H^{(0)}}{H^0}\right| \leqslant \delta_H$$

$$\text{and} \quad \left|\frac{Z^{(n)} - Z^0}{Z^0}\right| \leqslant \delta_Z, \tag{65}$$

where the n superscript denotes the value of the discretized integral invariants at time $t_n = n\Delta t$, while the 0 superscript denotes the initial value of the same discretized invariant at time $t = 0$. Only when one of the relations (65) is violated is the nonlinear constrained optimization algorithm activated.

a. The quadratic penalty algorithm

The quadratic penalty method consists of sequential unconstrained minimization of the form

$$\min_{x\in R^n} L(\mathbf{x}, r) = f(\mathbf{x}) + \frac{1}{2r_k}|\mathbf{e}(\mathbf{x})|^2. \tag{66}$$

We use different penalty factors for each constraint, r^H, r^E and r^Z, and we start the penalty parameters sequence with initial penalty factors r_0^H, r_0^E and r_0^Z. There the algorithm proceeds as follows:

Step 1: Select penalty parameters $r_0^i > 0$, $i = 1, \ldots, m$ and a sequence $\{\eta_k\}$ with $\eta_0 \geqslant 0$, $\{\eta_k\} \to 0$.

Step 2: Solve the problem

$$\min_{x\in R^n} L(\mathbf{x}, r_k) \tag{67}$$

(stop when $|\nabla_x L_{r_k}\| \leqslant \{\eta_k\}\|e(\mathbf{x}_k)\|$), i.e. find $\mathbf{x}_k$ solving (67) by inexact unconstrained minimization.

Step 3:

$$\text{If } |e_i(\mathbf{x}_k)| < \epsilon_i, \quad i = 1, \ldots, n, \text{ stop.} \tag{68}$$

Otherwise go to 4.

Step 4: Update and select penalty parameters $r_{k+1}^i \in (0, r_k)$ following formula (42). Select $\eta_{k+1} \geqslant 0$ (following a formula of the form $\eta_k = (l)^k$, $0 < \rho < 1$) and return to step 1.

b. The multiplier algorithm

In this method a multiplier term is added to the Lagrangian in (67), i.e. the penalty function idea is merged with the multiplier method, and we minimize the following augmented Lagrangian

$$L_{r_k}(\mathbf{x}, u_k) = f(\mathbf{x}) + \mathbf{u}_k\mathbf{e}(\mathbf{x}) + \frac{1}{2r_k}|\mathbf{e}(\mathbf{x})|^2. \tag{69}$$

The algorithm proceeds as follows.

First select an initial vector of multipliers $\mathbf{u}_0$ based either on prior knowledge [see Sasaki (1976), Sasaki *et al.* (1979)] or start with a zero vector in the absence of such knowledge. Select penalty scalars $r_0^i > 0$ and a sequence $\{\eta_k\}$ with $\eta_0 \geqslant 0$.

Step 1: Given a multiplier vector u_k, penalty parameters r_k^i and η_k find a vector $\mathbf{x}_k$ satisfying

$$\|\nabla_k L_{r_k}(\mathbf{x}_k, \mathbf{u}_k)\| \leqslant \{\eta_k\}\|e(\mathbf{x}_k)\| \tag{70}$$

by solving an inexact unconstrained minimization problem.

Step 2:

$$\text{If } |e_i(\mathbf{x}_k)| < \epsilon_i, \quad i = 1, \ldots, m. \tag{71}$$

Stop. Otherwise go to 3.

Step 3: Update the multiplier vector using either

$$\mathbf{u}_{k+1} = \mathbf{u}_k + r_k^{-1}e(\mathbf{x}_k) \tag{72}$$

or the alternative updating in (39)–(41).

Update and select penalty parameters $r_{k+1}^i \in (0, r_k)$ following formula (42). Select $\eta_{k+1} \geqslant 0$ (following a formula of the form $\eta_k = (l)^k$, $0 < l < 1$, we chose $l = 0.8$) and return to Step 1.

For the unconstrained minimizations a simple minimization routine E04DBF of Numerical Algorithms Group, England (NAGLIB) was used which minimizes a general function $F(x)$ of N variables, i.e. $\mathbf{x}$ is an N component vector. The methods employed are the conjugate-gradient methods due to Fletcher and Reeves (1964), Powell (1977) and Shanno (1978).

We modified the routine so as to incorporate the following stopping criteria:

$$\mathrm{GNORM} = \|\nabla_x L_{r_k}(\mathbf{x}, \mathbf{u}_k)\| \leq \{\eta_k\}\|e_k(\mathbf{x})\|. \quad (73)$$

Formulae to calculate the value of the function (the augmented Lagrangian) and its first derivatives must be supplied by the user. This minimization method has the virtue of requiring relatively very few memory storage locations; only a few multiples of N are required where N in our case is

$$N = 3N_xN_y \sim 540. \quad (74)$$

By contrast, Newton-like minimization procedures require the Hessian matrix of second derivatives necessitating storage locations in multiples of $N(N-1)/2$. For our case this would have been prohibitive.

c. *A modified multiplier penalty method*

As we observed very good conservation of the total mass even in the absence of constrained optimization, we decided to adjust the heights forecast by the finite difference model at time steps where the relative error bound

$$\left|\frac{H^{(n)} - H^{(0)}}{H^0}\right| \leq \delta_H$$

is violated.

At such time-steps the heights forecast by the discretized model were adjusted by using the formula

$$h^n_{jk} = \tilde{\mathrm{h}}^n_{jk} - \Big(\sum_{j=1}^{N_x}\sum_{k=1}^{N_y} \tilde{h}^n_{jk} - H^0\Big)(N_xN_y)^{-1}. \quad (75)$$

We then applied the same multiplier-penalty algorithm as in (69)–(72), using only the total energy and potential enstrophy conservation constraints, two multipliers and two penalty factors. The same stopping criteria as in Section 4b were used. We also tested this approach with the quadratic penalty method. Our numerical experiments were conducted using these modified algorithms.

TABLE 1. Relative accuracy of the penalty and penalty-multiplier constrained optimization methods using the GUSTAF model.

Time = 10 days	Z_0 (potential enstrophy). Ratio between final and initial values)	E_0 (total energy) Ratio between final and initial values	(H mass) Ratio between final and initial values
Combined penalty multiplier method quadratic method	1.00003	1.0006	0.99994
Penalty alone	1.0001	1.001	1.00001
No constraint	1.001	1.12	1.002

TABLE 2. Relative accuracy of the penalty and penalty-multiplier constrained optimization methods using the GUSTAF method.

Time = 20 days	Z_0 (potential enstrophy). Ratio between final and initial values)	E_0 (total energy). Ratio between final and initial values	(H mass). Ratio between final and initial values
Penalty multiplier	1.0001	1.002	1.00001
Penalty alone	1.0005	1.009	0.9980
No constraint	overflow	overflow	overflow

d. *Numerical results of long-term integrations*

In order to assess and compare the performance of the augmented Lagrangian penalty-multiplier method and the quadratic penalty method in enforcing the conservation of the integral invariants of the shallow water equations, we used the test problem described by Grammeltvedt (1969) (initial height field condition Number 1), see also Gustafsson (1971) and Navon and Riphagen (1979), Navon (1981), i.e.

$$h(x, y) = H_0 + H_1 \tanh\left[\frac{9(D/2 - y)}{2D}\right] + H_2 \operatorname{sech}^2\left[\frac{9(D/2 - y)}{D}\right] \sin\frac{2\pi x}{L}. \quad (76)$$

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

$$u = -(gf^{-1})\frac{\partial h}{\partial y}, \quad v = (gf^{-1})\frac{\partial h}{\partial x}. \quad (77)$$

The constants used were

$$\begin{aligned} L &= 4400 \text{ km} & g &= 10 \text{ m s}^{-2} \\ D &= 6600 \text{ km} & H_0 &= 2000 \text{ m} \\ f &= 10^{-4} \text{ s}^{-1} & H_1 &= 220 \text{ m} \\ \beta &= 1.5 \times 10^{-11} \text{ s}^{-1} \text{ m}^{-1} & H_2 &= 133 \text{ m}. \end{aligned} \quad (78)$$

For long-term runs the space and time increments used were

$$\Delta x = \Delta y = 400 \text{ km}, \quad \Delta t = 1800 \text{ s}. \quad (79)$$

Long-term runs were conducted using three different shallow water equations solvers GUSTAF, (Navon 1978), Navon (1979), ADIF and SHALL4 (Navon and Riphagen, 1979). We concentrated our attention only on the pure penalty algorithm and the modified penalty-multiplier algorithm using external mass-adjustment corrections for GUSTAF only.

TABLE 3. Number of line searches required by unconstrained minimizer and number of cycles of the algorithm (number of unconstrained minimizations carried out) for multiplier and penalty methods.

	Multiplier method				Penalty method	
Run	Penalty constant decrease rate	Stop rule parameter	Minimum cycle	Line search	Minimum cycle	Line search
GUSTAF	$(0.4)^k$	$(0.8)^k$	2	24	3	22
GUSTAF	$(0.4)^k$	$(0.4)^k$	3	32	3	35
GUSTAF	$(0.25)^k$	$(0.8)^k$	4	37	4	28
GUSTAF	$(0.25)^k$	$(0.4)^k$	4	38	4	30

e. Discussion of the numerical results

Each of the unconstrained minimizations using the subroutine E04DBF used no more than 26 functions calls to converge and the full process of nonlinear constrained optimization for a given time-step converged after using four values of the descending sequence $\{\eta_k\}$ (cycles). On the average a constraint violation occurred once every 10 time steps and only then did we use the constrained optimization algorithms. We found that considerable care has to be exercised in the choice of the initial multiplier vector $\mathbf{u}_0$ as well as in the choice of the initial penalty sequence. For the choice of the initial multipliers prior knowledge obtained by linearized analysis as suggested by Sasaki (1976), Sasaki *et al.* (1979) was used along with a first guess of $(0, 0, 0) = u_0$. In the choice of the initial penalty parameters we used the following suggestions of Bertsekas (1980):

1) The initial parameters r_0^i should not be too small to the point of inducing ill-conditioning in the first unconstrained minimization.

2) One should not increase the parameters r_k^i too fast to the point of inducing ill-conditioning in the unconstrained minimizations too early.

Our initial values of r_0^i were

$$r_0^Z = 1.10^{-1}, \quad r_0^E = 2.5.10^{-2}, \tag{80}$$

i.e., $(r_0^Z)^{-1} = 10$, $(r_0^E)^{-1} = 40$. We found, as expected, that the augmented Lagrangian penalty–multiplier method performed better than the pure penalty method.

After 10 days (using the GUSTAF model) we obtained the results given in Table 1. Similar results were obtained using the ADIF and SHALL4 models.

Similar results obtained after 20 days are shown in Table 2. Note that a finite time blow-up occurred at day 14 when the conservation of integral invariants was not enforced.

Table 3 shows the number of cycles of the algorithm (i.e., the number of unconstrained minimizations carried out) for a typical time step and the total number of line searchers required by the unconstrained minimizer using as shallow-water equations solver the program GUSTAF.

f. Numerical stability

Both the multiplier method and the quadratic penalty method were employed in integrations up to 20 days using a 30 min time-step. The multiplier method using the parameters enumerated before, behaved

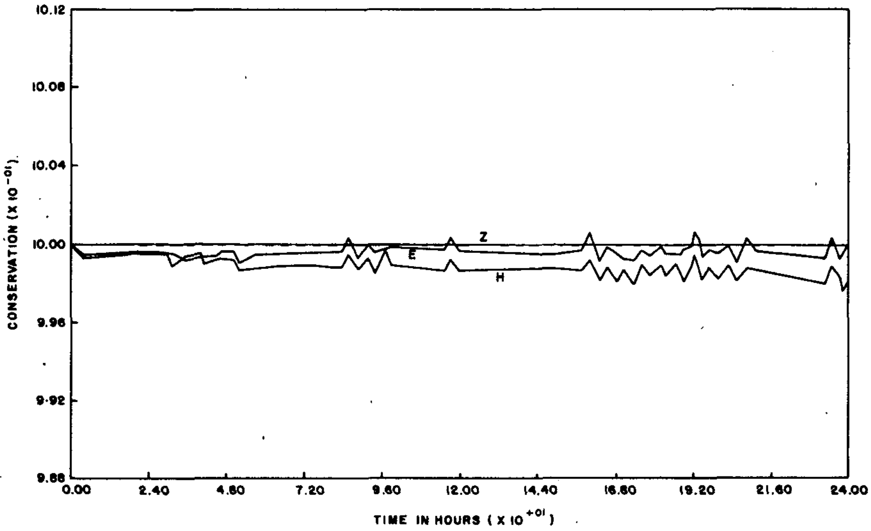

FIG. 1. Time variation of total mass, total energy and potential enstrophy as functions of their initial values with combined penalty multiplier constrained optimization using the GUSTAF model. The inflection points correspond to adjustment times.

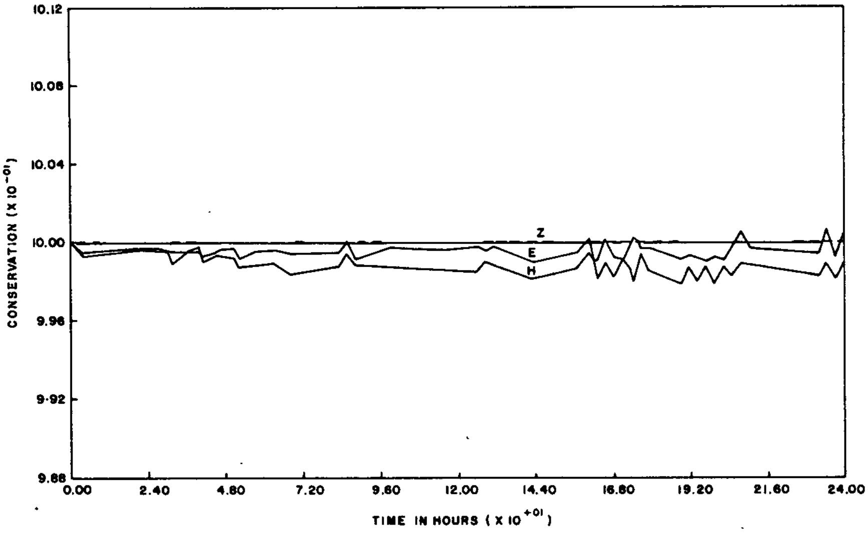

FIG. 2. Time variation of total mass, total energy and potential enstrophy as functions of their initial values using the quadratic penalty method. (GUSTAF model).

stably in conjunction with the three ADI models employed for solving the shallow water equations and there was no sign of impending numerical instability.

It is well known that for long-term integrations of the shallow water equations (more than 10 days) a finite-time "blow-up" is encountered if enstrophy is not exactly conserved (see Sadourny, 1975), Fairweather and Navon, 1980, Sasaki and Reddy, 1980). The advantage of the enforcement of the constraints to satisfy among others enstrophy conservation, is to extend the forecast period considerably beyond the critical "blow-up" time T_c which is roughly given as

$$T_c \sim Z_0^{-1/2}C(\Delta), \tag{81}$$

where Z_0 is the initial enstrophy and $C(\Delta)$ is a mesh dependent constant which increases with increasing mesh resolution (see Fairweather and Navon, 1980). The adjustments of the enforcement of integral constraints alters the spatial pattern very little and as pointed by Sasaki and Reddy (1980) visual inspection shows that improvement has occurred despite a slight increase in the rms norm error. In our case if enstrophy is not conserved the finite time "blow-up" occurs around the critical time $T_c \approx 14$ days.

The quadratic penalty method, however, was, in some cases, prone to numerical instability traceable to ill-conditioning due to a fast increase of the penalty parameters. In some cases the quadratic penalty

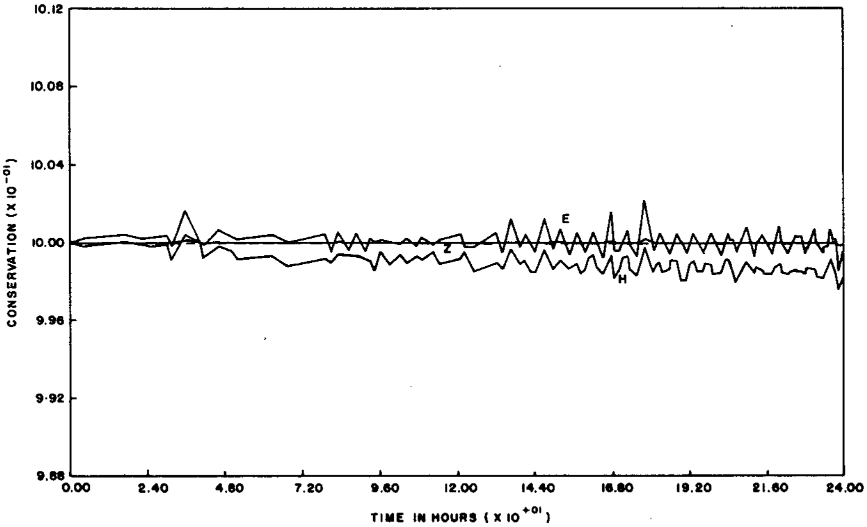

FIG. 3. As in Fig. 2 but as functions of their initial values with combined penalty multiplier. Mass conservation is used as a constraint. (GUSTAF model).

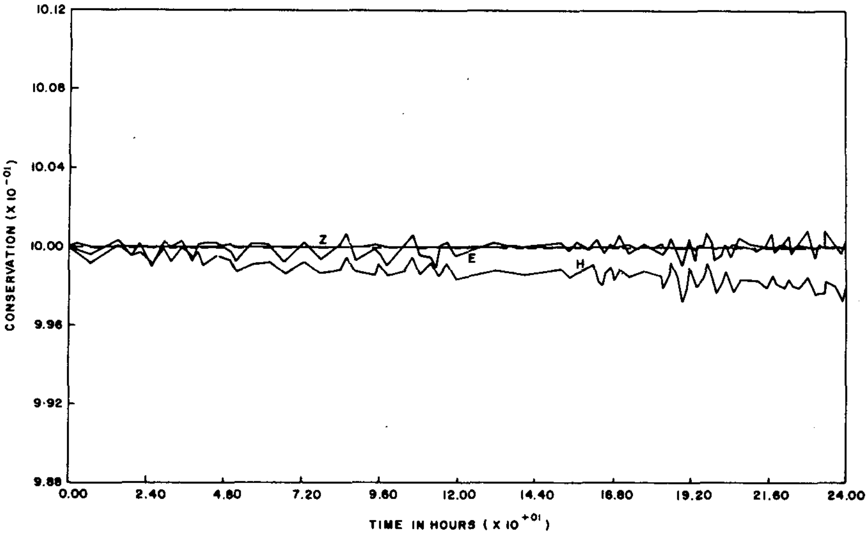

FIG. 4. As in Fig. 2 but as functions of their initial values with combined penalty multiplier constrained optimization. (ADIF model).

method required a smaller number of line searches and also a smaller number of corrections between different time-steps. In Figs. 1–10 we display the behaviour of the integral invariants versus time, using the penalty and the multiplier methods. An almost perfect conservation of potential enstrophy (Z) is obtained.

g. Accuracy tests

In order to provide a basis for comparison between the multiplier method of enforcing nonlinear constraints on the one hand and between the quadratic penalty on the other hand, we use the same method as in Navon (1981), i.e. we assume the exact solution w_{Ex} of the shallow-water equations solver, say GUSTAF, is the solution of GUSTAF computed with a fine-mesh discretization, viz. $\Delta x = \Delta y = 200$ km and $\Delta t = 900$ s (using the same method to enforce the discrete integral constraints).

In order to assess the influence of the nonlinear constrained optimization multiplier technique on the accuracy, we also computed the accuracy without any enforcement of constraints. In both cases we used the Gustafsson (1971) relative error norm (see Navon, 1981). The results are given in Table 4. The conservation enforcement of the integral constraints affects the relative error only slightly, but makes it possible

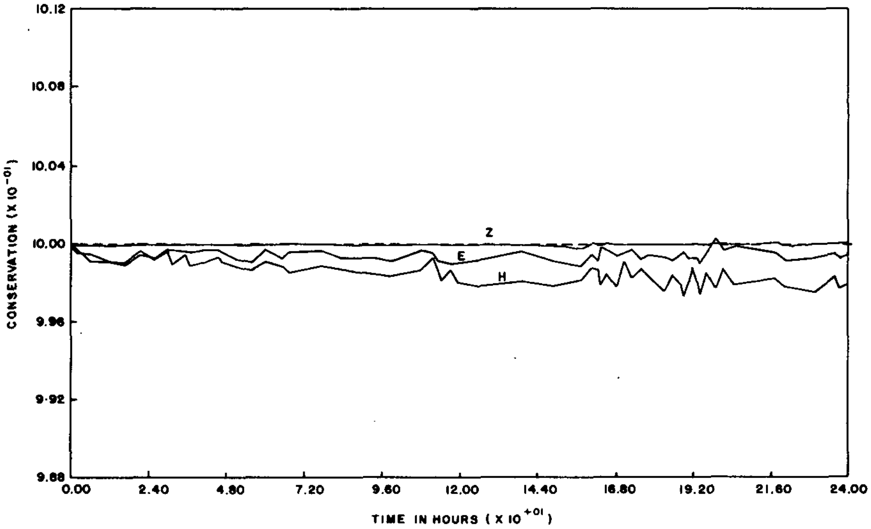

FIG. 5. As in Fig. 2 but as functions of their initial values with quadratic penalty method (ADIF model).

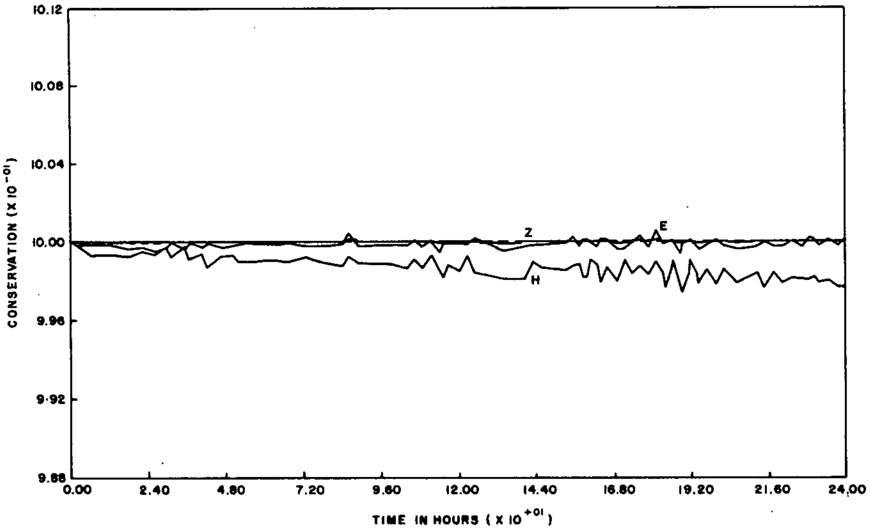

FIG. 6. As in Fig. 2 but as functions of their initial values with combined penalty multiplier method. Mass conservation is used as a constraint.

for the schemes to avoid computational instability (see Sasaki and Reddy, 1980, p. 160) and finite-time "blow-up" as we experienced at time T_c = 14 days when enstrophy conservation was not enforced.

h. Computational efficiency

The CPU time requested on the CDC 750 requested for a single time step with the GUSTAF shallow-water equations solver was 0.22 s per full time-step.

An augmented Lagrangian adjustment required 0.14 s to converge to the prescribed accuracy limits.

As the adjustment is done in the average only every 15 time steps, the additional time required by the implementation of the augmented Lagrangian method never exceeded 10% of the total integration CPU time.

i. Application of the method of Augmented Lagrangians to variational nonlinear normal mode initialization

In the variational formulation of the nonlinear normal mode initialization (Daley, 1978; Tribbia, 1982) we wish to minimize a functional of the form:

$$I = \int_A [(\mathbf{v}_0 - \mathbf{v}_c)^2 w_v + (\phi_0 - \phi_c)^2 w_\phi] dA, \qquad (82)$$

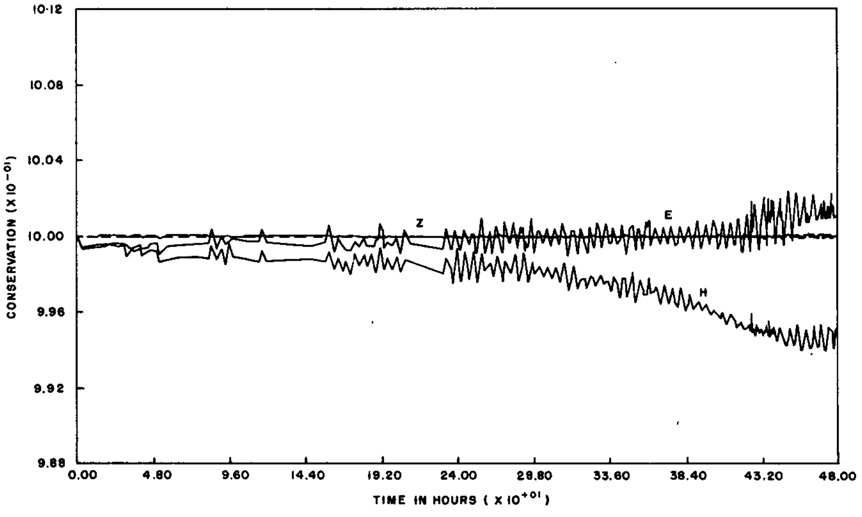

FIG. 7. Long-term (20 days) time variation of total mass, total energy and potential enstrophy as functions of their initial values with combined penalty multiplier method. (GUSTAF model).

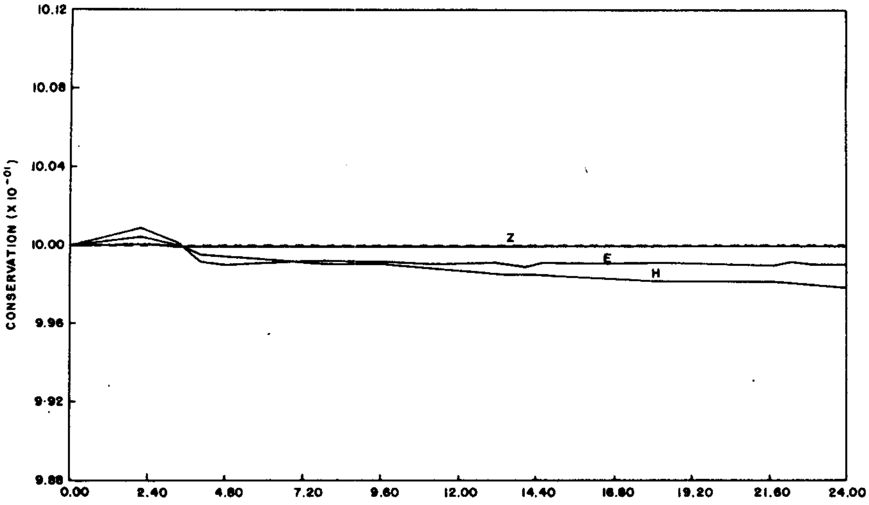

FIG. 8. Time variation of total mass, total energy and potential enstrophy as functions of their initial values with combined penalty multiplier method. Fine-mesh discretization, $\Delta x = \Delta y = 200$ km used. (GUSTAF method).

where w_v, w_ϕ are weight functions, $\mathbf{v}$ the velocity, ϕ the geopotential, $\mathbf{v}_0$, ϕ_0 are the observed values while $\mathbf{v}_c$ and ϕ_c indicate values after constrained initialization. The functional (82) is minimized subject to the constraint that the final state c lies on the slow manifold (i.e., that $\mathbf{v}_c$, ϕ_c be balanced).

Daley (1978) used Lagrange multipliers for the constrained initialization which led to a system of Euler-Lagrange equations which was solved iteratively.

It is proposed here to solve the constrained minimization problem using an augmented Lagrangian multiplier method with a combined penalty-multiplier approach, i.e. a functional of the form

$$J = I + \sum_{S} \sum_{j}^{G} a_j^s[\gamma_j^s(c) - \gamma_j^s(0)] + \frac{1}{2r} \sum_{S} \sum_{j} [\gamma_j^s(c) - \gamma_j^s(0)]^2, \qquad (83)$$

where a_j^s are the Lagrange multipliers and γ_j^s are complex free-mode expansion coefficients (see Daley, 1978) and the second term is a quadratic penalty term and apply the augmented Lagrangian algorithm.

5. Summary and conclusions

A general approach for solving nonlinearly constrained optimization problems using both a com-

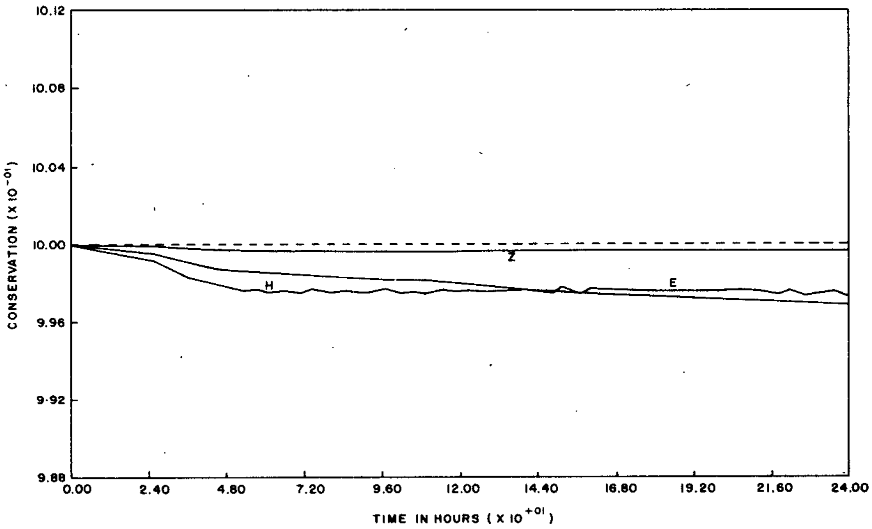

FIG. 9. As in Fig. 8 but as functions of their initial values with combined penalty multiplier method. Fine-mesh discretization $\Delta x = \Delta y = 200$ km used. (ADIF method).

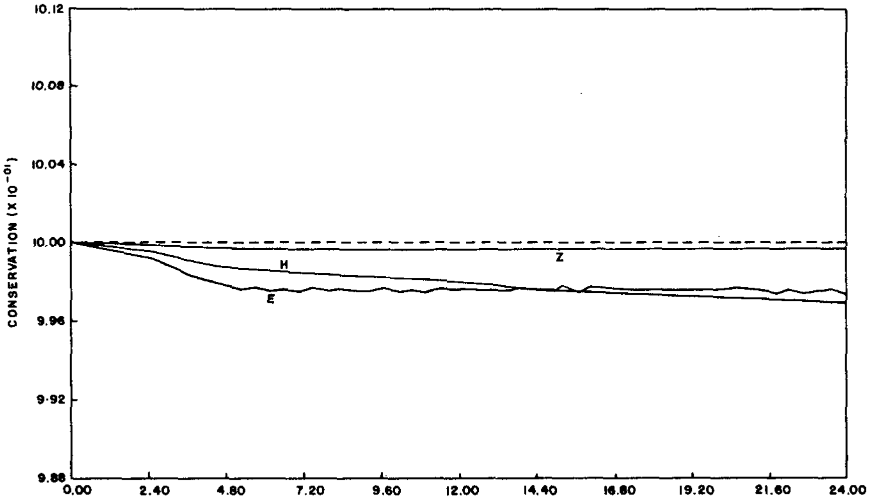

FIG. 10. As in Fig. 8 but as functions of their initial values with combined penalty multiplier method. Quadratic penalty method used. (ADIF method).

bined penalty multiplier method perfected by Bertsekas (1980) and a penalty method has been tested. This method generalizes the Sasaki variational approach, which in our context can be viewed as a pure multiplier method and is known to converge only moderately fast, and then only when the problem has a locally convex structure (Bertsekas 1976). It does not require any linearization in order to obtain and update the Lagrange multipliers and has a very good rate of convergence. Problems treated by Barker *et al.* (1977), Haltiner *et al.* (1975) and Haltiner and Barker (1976) as well as the problem of variational nonlinear normal mode initialization, (Daley, 1978; Phillips, 1981), seem to be prime candidates for testing this new general method. In the past the method has been successfully applied to optimal control problems with terminal state constraints, by Nakayama *et al.* (1975), amongst others, and is now used almost exclusively in optimal control problems involving state constraints (Bertsekas, 1980; Mond 1982).

Compared with the modified Sasaki and Bayliss–Isaacson methods (see Navon, 1981) the multiplier method is more efficient in that it is applied only at every 15 time steps on the average, and in that it has a very fast convergence rate. The implementation of this method requires less computer coding, and it has a broad scope of applicability. The method can be recommended for long time integrations of the shallow water equations on limited-area domains as well as for problems where nonlinear constraints have to be imposed. It should be mentioned that the Bayliss–Isaacson method is also quite efficient since it requires the solution of a small system of equations at each point. (See Isaacson *et al.*, 1979.)

TABLE 4. Relative errors of the multiplier and penalty methods for GUSTAF.

Shallow-water equations solver	Multiplier method $\Delta t = 1800$ s, $\epsilon_i = 10^{-3}$	Penalty method $\Delta t = 1800$, $\epsilon_i = 10^{-3}$	No constraint being enforced $\Delta t = 1800$ s
GUSTAF (QN3)			
After 1 day	1.518×10^{-3}	1.517×10^{-3}	1.510×10^{-3}
After 2 days	2.571×10^{-3}	2.579×10^{-3}	1.509×10^{-3}
After 10 days	2.294×10^{-3}	2.310×10^{-3}	2.102×10^{-3}
After 20 days	3.751×10^{-3}	4.218×10^{-3}	blow-up at 14 days

The method of augmented Lagrangian is however the most general framework not only for enforcing conservation of integral invariants but also for solving most of the partial differential equations appearing in the numerical weather prediction context by formulating them as the problem of minimization of a functional. For more insight the interested reader is referred to the book by Fortin and Glowinski (1981).

Acknowledgments. The first author would like to acknowledge numerous stimulating discussions with Professor B. Mond, of the Department of Pure Mathematics, La Trobe University, Melbourne, Australia and Visiting Scientist at the National Research Institute for Mathematical Sciences of the CSIR, Pretoria, while this work was under way.

The authors would like to thank an anonymous referee for his constructive criticism which led to a better clarification of the present paper.

REFERENCES

Arakawa, A., 1966: Computational design for long-term numerical integration of the equations of fluid-motion (I): Two-dimensional incompressible flow. *J. Comput. Phys.*, **1**, 119–143.

——, and V. R. Lamb, 1981: A potential enstrophy conserving scheme for the shallow-water equations. *Mon. Wea. Rev.*, **109**, 18–36.

Arrow, K. J., and R. M. Solow, 1958: Gradient methods for constrained maxima with weakened assumptions. *Studies in Linear and Nonlinear Programming,* K. Arrow, L. Hurwitz and H. Uzawa, Eds., Stanford University Press.

Bayliss, A., and E. Isaacson, 1975: How to make your algorithm conservative *No. Amer. Math. Soc.,* Aug., A594–A595.

Barker, E., G. Haltiner and Y. K. Sasaki, 1977: Three dimensional initialization using variational analysis. *Proc. Third Conference on Numerical Weather Prediction,* Omaha, Amer. Meteor. Soc., 169–181.

Bertsekas, D. P., 1973: Convergence rate of penalty and multiplier methods. *Proc. 1973 IEEE Conf. on Decision and Control,* San Diego, IEEE Publication No. 73, CHO 806-0, SMC, 260–264.

——, 1975a: On penalty and multiplier methods for constrained optimization in nonlinear programming. O. Mangasarian, R. Meyer and S. Robinson, Eds., Academic Press, 165–191.

——, 1975b: On the method of multipliers for convex programming. *IEEE Trans. Auto. Control.* **AC-20,** 385–388.

——, 1975c: Combined primal–dual and penalty methods for constrained minimization. *SIAM J. Control Optim.*, **13,** 521–544.

——, 1976a: On penalty and multiplier methods for constrained minimization. *SIAM J. Control Optim.*, **14,** 216–235.

——, 1976b: Multiplier methods: A survey. *Automatica,* **12,** 133–145.

——, 1980: Penalty and multiplier methods in nonlinear optimization, theory and algorithms. L. C. W. Dixon, E. Spedicato and G. P. Szegö, Eds., Birkhäuser, Boston, Chap. 11, 253–278.

——, 1982: *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press, 416 pp.

Buys, J. D., 1972: Dual algorithms for constrained optimization. Ph.D. thesis, Rijksuniversiteit de Leiden, The Netherlands. 107 pp. [Available from Bronder Offset, NV-Rotterdam, Holland].

Courant, R., 1943: Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.*, **49,** 1–23.

Daley, R., 1978: Variational nonlinear normal mode initialization. *Tellus,* **30,** 201–208.

Fairweather, G., and I. M. Navon, 1980: A linear ADI method for the shallow-water equations. *J. Comput. Phys.*, **37,** 1–18.

Fiacco, V., and G. P. McCormick, 1968: *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* Wiley and Sons, 210 pp.

Fletcher, R., 1981: *Practical Methods of Optimization,* Vol. 2, *Constrained optimization.* Wiley and Sons, 224 pp.

——, and C. M. Reeves, 1964: Function minimization by conjugate gradients. *Comput. J.*, **7,** 149.

Fortin, M., and R. Glowinski, 1981: Resolution Numerique de Problèmes aux limites par des methodes de Lagrangian Augmenté. Institut National de Recherche en Informatique et eu Automatique. Roquencourt, BP 105, 78150 Le Chesnay, France, 320 pp.

Fjørtoft, R., 1953: On the changes in the spectral distribution of kinetic energy for two dimensional non-divergent flow. *Tellus,* **5,** 225–230.

Gill, P. E., W. Murray, M. A. Saunders and M. H. Wright, 1981a: Aspects of mathematical modelling related to optimization. *Appl. Math. Modelling,* **5,** 71–83.

——, W. Murray and M. H. Wright, 1981b: *Practical Optimization.* Academic Press, 401 pp.

Grammeltvedt, A., 1969: A survey of finite-difference schemes for the primitive equations for a barotropic fluid. *Mon. Wea. Rev.*, **97,** 384–404.

Gustaffson, B., 1971: An alternating direction implicit method for solving the shallow-water equations. *J. Comput. Phys.*, **7,** 239–254.

Haarhoff, P. C., J. D. Buys and H. von Molendorff, 1969: Conmin-A. computer programme for the minimization of a nonlinear function subject to nonlinear constraints. Atomic Energy Board, Pelindaba, South Africa, PEL 190, 34 pp.

Haltiner, G. J., and E. H. Barker, 1976: Initial balancing with a variational method. *Ann. Meteor.*, **11,** 119–121.

——, Y. K. Sasaki and E. H. Barker, 1975: A variational procedure for obtaining global balanced winds. *Proc. J.O.C. Study Group Conference on Four-Dimensional Data Assimilation,* Paris, 198–223.

Hestenes, M. R., 1969: Multiplier and gradient methods. *J. Opt. Theory Appl.*, **4,** 303–320.

Isaacson, E., 1977: Integration schemes for long-term calculation advances in computer methods for partial differential equations II, A. Vichnevetsky, Ed., AICA, 251–255. IMACS (AICA) *Int. Symp. Computer Methods for Partial Differential Equations.* Lehigh University, 392 pp.

——, D. Marchesin and G. Zwas, 1979: Numerical methods for meteorology and climatology. *Fourth NASA Weather and Climate Program Science Review,* Earl R. Kreins, Ed., NASA Conf. Pub. 2076, Pap 31, 183–190.

Kalnay-Rivas, E., 1979: The effect of accuracy, conservation and filtering on numerical weather forecasting. *Proc. Fourth Conf. Numerical Weather Prediction,* Silver Spring, Amer. Meteor. Soc., 302–312.

——, A. Bayliss and J. Storch, 1977: The 4th-order GISS model of the global atmosphere. *Contrib. Atmos. Phys.*, **50,** 306–311.

Kort, B. W., 1975: *Rate of Convergence of the Method of Multipliers with Inexact Minimization in Nonlinear Programming* 2. O. Mangasarian, S. Robinson and R. Meyer, Eds., Academic Press, 193–214.

——, 1977: Combined prima-dual and penalty function algorithms for nonlinear programming. Ph.D. thesis, Stanford University, Palo Alto, CA, 216 pp.

——, and D. P. Bertsekas, 1976: Combined primal-dual and penalty methods for convex programming. *SIAM J. Control Optim.*, **14,** 268–294.

Lee, R. L., C. M. Gresho, S. T. Chan and R. L. Sani, 1980: A comparison of several conservative forms for finite-element formulations of the incompressible Navier–Stokes or Boussinesq equations. *Third Int. Conf. Finite-Elements in-flow Problems,* Banff, 216–227.

Lilly, D. K., 1965: On the computational stability of numerical solutions of time-dependent nonlinear geophysical fluid dynamics problems. *Mon. Wea. Rev.*, **93,** 11–26.

Mond, B., 1982: Techniques for constrained minimization. TWISK 252, Tech. Rep., National Research Institute for Mathematical Sciences, CSIR, P.O. Box 395, Pretoria 0001, South Africa, 11 pp.

Nakayama, H., H. Sayama and Y. Sawargi, 1975: Multiplier method and optimal control problems with terminal state constraints. *Int. J. Sys. Sci.*, **6,** 465–477.

Navon, I. M., 1978: GUSTAF: A nonlinear alternating direction implicit FORTRAN IV program for solving the shallow-water equations. TWISK 25, Tech. Rep., National Research Institute for Mathematical Sciences, CSIR, P.O. Box 395, Pretoria 0001, South Africa, 24 pp.

——, 1979: ADIF, a FORTRAN IV program for solving the shallow-water equations. *Comput. Geosci.*, **5,** 19–39.

——, and H. A. Riphagen, 1979: An implicit compact fourth-order algorithm for solving the shallow-water equations in conservation-law form. *Mon. Wea. Rev.*, **107,** 1107–1127.

——, 1981: Implementation of 'a posteriori' methods for enforcing conservation of potential enstrophy and mass in discretized shallow-water equations models. *Mon. Wea. Rev.*, **109,** 946–959.

Phillips, N. A., 1959: An example of nonlinear computational instability. *The Atmosphere and Sea in Motion, Rossby Memorial Volume,* Rockefeller Institute Press, 501–504.

——, 1981: Variational analysis and the slow manifold. *Mon. Wea. Rev.,* **109,** 2415–2426.

Powell, M. J. D., 1969: A method for nonlinear constraints in minimization problems. *Optimization,* R. Fletcher, Ed., Academic Press, Chap. 19, pp. 283–298.

——, 1977: Restart procedures for the conjugate-gradient method. *Math. Prog.,* **12,** 241–254.

——, Ed., 1982: *Nonlinear Optimization* 1981. Academic Press, 509 pp.

Rockafellar, R. T., 1973: The multiplier method of Hestenes and Powell applied to convex programming. *J. Optim. Theory Appl.,* **12,** 555–562.

Sadourny, R., 1975: The dynamics of finite-difference models of the shallow-water equations. *J. Atmos. Sci.,* **32,** 680–689.

——, 1980: Conservation-laws, quasi two-dimensional turbulence and numerical modelling of large-scale flows. *Seminar 1979,* Vol. 2: *Dynamic Meteorology and Numerical Weather Prediction,* European Centre for Medium Range Weather Forecasts, 167–195.

Sasaki, Y., 1976: Variational design of finite-difference schemes for initial value problems with an integral invariant. *J. Comput. Phys.,* **21,** 270–278.

——, 1977: Variational design of finite-difference schemes for initial value problems with a global divergent barotropic mode. *Contrib. Atmos. Phys.,* **50,** 284–289.

——, T. Barker and J. S. Goerss, 1979: *Dynamic Data Assimilation by the Noise Freezing Method.* Final Report No. F52551792, Naval Environmental Prediction Research Facility, Monterey, California 93940, 80 pp.

——, and J. N. Reddy, 1980: A comparison of stability and accuracy of some numerical models of two-dimensional circulation. *Int. J. Numer. Meth. Eng.,* **16,** 149–170.

Shanno, D. F., 1978: Conjugate-gradient methods with inexact searches. *Math. Oper. Res.,* **3,** 244–256.

Tribbia, J., 1982: On variational normal mode initialization. *Mon. Wea. Rev.,* **110,** 450–470.