Tellus (1992), *44A*, 273–296

Incomplete observations and control of gravity waves in variational data assimilation

By X. ZOU, *Supercomputer Computations Research Institute, Florida State University, Tallahassee, FL 32306, USA*, I. M. NAVON, *Department of Mathematics and Supercomputer Computations, Research Institute, Florida State University, Tallahassee, FL 32306, USA*, and F. X. LE DIMET, *Laboratoire de Modelisation et Calcul, Université Joseph Fourier, BP 53X, 38041 Grenoble Cedex, France*

(Manuscript received 12 June 1991; in final form 24 February 1992)

ABSTRACT

Methods aimed at the specification of suitably balanced initial fields, combined with a 4-D variational data assimilation, were examined and applied to a finite-difference limited-area shallow water equations model. Ageostrophic (noisy) initial conditions which lead to the generation of unrealistically large amplitude gravity-wave oscillations in the model, were used. A penalty function method is presented, aiming to achieve the elimination of unwanted gravity oscillations. The procedure consists in the addition of a quadratic penalty term to the cost function used in the variational data assimilation requiring that the time tendency of the geopotential height satisfy an inequality constraint. Numerical experiments show that using the penalized cost function results in the efficient elimination of all gravity oscillations. The results point to the fact that if used correctly, the penalty method can efficiently control gravity waves in the model. New results are also presented concerning the impact of incomplete observations in both the time and space dimensions on the uniqueness of the solution, on the convergence rate of the minimization process, on the condition number of the Hessian (which in turn affects the convergence rate), on the quality of the retrieved initial fields, and finally on the quality of the ensuing forecast.

1. Introduction

Recently, considerable research has focused on variational data assimilation as a means of controlling the initial conditions in a numerical weather prediction model by using observations distributed in a certain time window as well as the model itself. The problem may be viewed as that of finding the initial conditions that minimize a cost function or distance between model and observations over the time interval considered. The initial conditions determine the space-time trajectory of the system of partial differential equations which constitute the model and the integration of the adjoint equation model yields the value of the gradient of the cost function with respect to the initial conditions which are the control variables of this optimal control problem. This powerful method, which was described in the papers of Derber (1985), LeDimet and Talagrand (1986), Courtier and Talagrand (1987) and Talagrand and Courtier (1987), incorporates the time dimension in the assimilation process quite naturally. Further, by incorporating the physics of the problem in the definition of the cost function and the constraining dynamics of the model itself, this method turns out to be very versatile and essentially imposes no limitation on the characteristics of the data to be assimilated. However, due to its high computational cost, the operational application of variational data assimilation is still limited to research and development versions. Many issues, such as scaling, weighting, preconditioning, conditions for the uniqueness of the solution, efficient minimization, gravity wave control and incomplete observations need to be addressed before it becomes an operational tool.

In this paper we will address the two main issues, i.e., gravity wave control and the influence of incomplete observations in both the time and space dimensions. A number of theoretical issues and some numerical results related to the

previously mentioned issues using a limited-area shallow water equations model are presented.

Gravity oscillations with unrealistically large amplitudes may be generated in numerical models due to the insertion of erroneous data or due to errors in the numerical model. Methods aimed at the specification of suitably balanced initial fields, which will not give rise to spurious gravity oscillations, are known as initialization methods. The normal mode initialization (NMI) technique (Machenhauer, 1977), for instance, is characterized by a separation of the slow and fast mode components of the flow as a means of controlling the amplitude of the high-frequency gravity-inertia waves in a primitive equation model and gained a wide popularity during the last decade. An implicit NMI approach, stimulated by the desire to apply the NMI to limited area models, was also developed starting from the idea of the bounded derivative method (Kreiss, 1980; Browning and Kreiss, 1982; Ballish, 1981; Semazzi and Navon, 1986; Temperton, 1988, 1989). The implicit NMI approach may be implemented in models whose normal modes cannot readily be found. It also seems to be a more useful approach than the explicit NMI approach for initializing very high resolution spectral models.

Variational data assimilation, by its nature, can be used to assimilate data which was not previously initialized. Problems arise as to how to efficiently control high frequency gravity wave oscillations in variational data assimilation when assimilating noisy data.

Courtier and Talagrand (1990) indicated that gravity wave noise can be efficiently eliminated by adding a penalty term, defined by the normal modes of the model, to the cost function and combine this approach with a NMI procedure in the variational data assimilation. The penalty method follows the logic of tending to reduce the time tendency of the gravity wave component of the flow to a very small value. When the penalty term was included in the cost function, the minimization process converged to a solution which looked acceptable in terms of the amount of gravity waves it contained, but the convergence rate turned out to be extremely slow. Only by combining the penalty method with the adjoint of NMI have properly balanced conditions been obtained in an acceptable number of descent steps. Inclusion of the penalty term in the cost function while performing the minimization on uninitialized fields leads to better numerical conditioning and ensures that the minimizing solution will not contain an unacceptable amount of gravity waves. However this approach involves a sizable computational effort, namely, it requires the availability of both the normal modes of the model and that of the adjoint of the NMI procedure.

In this paper we propose both a short cut penalty function method and an augmented Lagrangian method aimed at controlling the total time tendency of the flow while achieving a reasonable convergence rate. The proposed method requires less memory and computational effort than the method of combined penalty and NMI proposed by Courtier and Talagrand (1990).

Another issue directly related to variational data assimilation is the case of incomplete observations in either the space or time dimensions. An important question is whether the solution is unique or not if there are less observations in space. An additional related question concerns the influence of the density of observations in time on the convergence rate of the minimization procedure. In Thacker (1988), the process of fitting a model to inadequate data is discussed. Results show that for the simple three wave model of Thacker and Long (1988), a reasonable fit can be obtained even if the number of observations is less than the number of degrees of freedom of the model. A theoretical derivation of a sufficient adequate condition for the solution to be unique in the case of incomplete observations was given in Appendix B of the paper. Several numerical experiments to clarify this idea using a shallow water equations model as a test problem were provided.

The plan of the paper is the following. In Section 2, a brief description of the limited-area shallow-water equations model, its adjoint model, the definition of the cost function, the calculation of the gradient of the cost function, and different test problems are given. In Section 3, the penalty method and augmented Lagrangian method, the penalized cost function and the corresponding changes in the gradient calculation using the adjoint model are described. Application of the penalty and augmented Lagrangian methods are presented in Sections 4 and 5, respectively. Theoretical conclusions and results assessing the impact of insufficient observational data are

presented in Section 6. Summary and conclusions are presented in Section 7.

2. Brief description of the model and the variational problem

2.1. *The shallow-water equations model and its adjoint system*

The shallow-water equations may be written in the form

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}-fv+\frac{\partial \phi}{\partial x}=0, \tag{2.1}$$

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+fu+\frac{\partial \phi}{\partial y}=0, \tag{2.2}$$

$$\frac{\partial \phi}{\partial t}+u\frac{\partial \phi}{\partial x}+v\frac{\partial \phi}{\partial y}+\phi\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right)=0, \tag{2.3}$$

with periodic boundary conditions in east-west direction and a solid smooth wall condition on the north and south boundaries, respectively. Here u, v, and ϕ are the two components of the velocity field and the geopotential field, respectively, both spatially discretized with a centered difference scheme in space and an explicit leapfrog time-differencing scheme. Numerical experiments were carried out on a rectangular domain of size $L=6000$ km, $D=4400$ km. The parameters of the space and time discretization are $\Delta x=300$ km, $\Delta y=220$ km and $\Delta t=600$ s.

To derive the adjoint system of this shallow-water equations model, the linear tangent system to eqs. (2.1)–(2.3) is first derived. This system can be written as:

$$\frac{\partial u'}{\partial t}+\frac{\partial (uu')}{\partial x}+v'\frac{\partial u}{\partial y}+v\frac{\partial u'}{\partial y}-fv'+\frac{\partial \phi'}{\partial x}=0, \tag{2.4}$$

$$\frac{\partial v'}{\partial t}+u'\frac{\partial v}{\partial x}+u\frac{\partial v'}{\partial x}+\frac{\partial vv'}{\partial y}+fu'+\frac{\partial \phi'}{\partial y}=0, \tag{2.5}$$

$$\frac{\partial \phi'}{\partial t}+\frac{\partial (\phi' u)}{\partial x}+\frac{\partial (\phi u')}{\partial x}+\frac{\partial (\phi v')}{\partial y}+\frac{\partial (\phi' v)}{\partial y}=0, \tag{2.6}$$

or in matrix form as

$$\frac{\partial \boldsymbol{X}}{\partial t}+A\boldsymbol{X}=0, \tag{2.7}$$

where the vector $\boldsymbol{X}$ and the operator matrix A are given by

$$\boldsymbol{X}=(u', v', \phi')^{\mathrm{T}}, \tag{2.8a}$$

$$A=\begin{pmatrix} \frac{\partial (u(\cdot))}{\partial x}+v\frac{\partial ((\cdot))}{\partial y} & (\cdot)\frac{\partial u}{\partial y}-f & \frac{\partial (\cdot)}{\partial x} \\ (\cdot)\frac{\partial v}{\partial x}+f & u\frac{\partial (\cdot)}{\partial x}+\frac{\partial (v(\cdot))}{\partial y} & \frac{\partial (\cdot)}{\partial y} \\ \frac{\partial (\phi(\cdot))}{\partial x} & \frac{\partial (\phi(\cdot))}{\partial y} & \frac{\partial (u(\cdot))}{\partial x}+\frac{\partial (v(\cdot))}{\partial y} \end{pmatrix}, \tag{2.8b}$$

respectively. The vector $\boldsymbol{X}$ represents a perturbation around the state u, v, Φ.

Let us assume that the adjoint problem to (2.7) is

$$-\frac{\partial \hat{\boldsymbol{X}}}{\partial t}+A^*\hat{\boldsymbol{X}}=0. \tag{2.9}$$

The adjoint operator A^* satisfies

$$\langle A\boldsymbol{X}, \hat{\boldsymbol{X}}\rangle_{\mathrm{D}\times\mathrm{T}}=\langle A^*\hat{\boldsymbol{X}}, \boldsymbol{X}\rangle_{\mathrm{D}\times\mathrm{T}}, \tag{2.10}$$

where $\langle\ \rangle$ represents the general definition of the inner product given by

$$\langle \boldsymbol{X}, \boldsymbol{Y}\rangle_{\mathrm{D}\times\mathrm{T}}=\sum_{i=1}^{3}\int_{t_0}^{t_{\mathrm{R}}}\int_{\mathrm{D}} X_i Y_i \,\mathrm{d}D\,\mathrm{d}t, \tag{2.11}$$

where T is the temporal domain, t_0 and t_{R} are the lower and upper limits of the time integration interval, and D is the spatial domain over which the equations are integrated. We can then derive the following expression for A^* (see LeDimet, 1988, Kontarev, 1980):

$$A^*=\begin{pmatrix} -u\frac{\partial (\cdot)}{\partial x}-\frac{\partial (v(\cdot))}{\partial y} & (\cdot)\frac{\partial v}{\partial x}+f & -\phi\frac{\partial (\cdot)}{\partial x} \\ (\cdot)\frac{\partial u}{\partial y}-f & -v\frac{\partial (\cdot)}{\partial y}-\frac{\partial (u(\cdot))}{\partial x} & -\phi\frac{\partial (\cdot)}{\partial y} \\ -\frac{\partial (\cdot)}{\partial x} & -\frac{\partial (\cdot)}{\partial y} & -u\frac{\partial (\cdot)}{\partial x}-v\frac{\partial (\cdot)}{\partial y} \end{pmatrix}. \tag{2.12}$$

2.2. *The objective function and its gradient*

In all the present experiments, the objective function was defined as a simple weighted sum of

squared differences between the observations and the corresponding model values, namely

$$J(X_0) = \langle W(X(X_0) - X_{\text{obs}}),\, X(X_0) - X_{\text{obs}} \rangle$$
$$= W_\phi \sum_{n=1}^{N_\phi} (\phi_n - \phi_{\text{obs}})^2$$
$$+ W_{\text{V}} \sum_{n=1}^{N_{\text{V}}} [(u_n - u_{\text{obs}})^2 + (v_n - v_{\text{obs}})^2], \quad (2.13)$$

where N_ϕ is the total number of geopotential observations available over the assimilation period (t_0, t_{R}), and N_{V} is the total number of wind vector observations. The quantities u_{obs}, v_{obs}, and ϕ_{obs} are the observed values for northward wind component, eastward wind component and geopotential field respectively, while the quantities u, v, and ϕ are the corresponding computed model values. W_ϕ and W_{V} are weighting factors which are taken to be the inverse of estimates of the statistical root-mean-square observational errors on geopotential and wind components respectively. In our test problem, values of $W_\phi = 10^{-4}\,\text{m}^{-2}\,\text{s}^2$ and $W_{\text{V}} = 10^{-2}\,\text{m}^{-2}\,\text{s}^2$ are used.

The gradient of the objective function with respect to the initial condition X_0:

$$X(0) = X_0, \quad (2.14)$$

is given by

$$\nabla_{X_0} J = \hat{X}(0), \quad (2.15)$$

where $\hat{X}(0)$ is obtained by integrating the adjoint system (2.9) backward in time with a zero initial condition at time t_{R}. In addition, a forcing term:

$$2W(X - X_{\text{obs}}) \quad (2.16)$$

is added to the adjoint variable at times when observational data are available.

2.3. *Description of test problems*

To carry out the numerical experiments of the incomplete observations, the initial conditions of Grammeltvedt (1969) were used, given by

$$h = H_0 + H_1 \tanh \frac{9(y - y_0)}{2D}$$
$$+ H_2 \operatorname{sech}^2 \frac{9(y - y_0)}{D} \sin \frac{2\pi x}{L}, \quad (2.17)$$

where $H_0 = 2000$ m, $H_1 = -220$ m, $H_2 = 133$ m, $L = 6000$ km, $D = 4400$ km. Here L is the length of the channel on the β plane, D is the width of the channel, and $y_0 = D/2$ is the middle latitude of the channel. The initial velocity fields were derived from the initial height field via the geostrophic relationship and are given by

$$u = -\frac{g}{f}\frac{\partial h}{\partial y}$$
$$v = \frac{g}{f}\frac{\partial h}{\partial x}. \quad (2.18)$$

where $g = 10\,\text{m}\,\text{s}^{-2}$ and $f = 10^{-4}\,\text{s}^{-1}$. The time and space increments used in the model were

$$\Delta x = 300\text{ km}, \qquad \Delta y = 220\text{ km}, \qquad \Delta t = 600\text{ s}. \quad (2.19)$$

The experiment is devised as follows. The observational data consist of the model-integrated values for wind and geopotential at each time step starting from the Grammeltvedt initial conditions defind by eqs. (2.17) and (2.18). Either random perturbations of these fields, performed using a standard library randomizer RANF on the CRAY-YMP or complete flat fields are then used as the initial guess for the solution. The time length of the assimilation window is 10 hours which in turn involves 60 time steps of 10 min each.

In order to assess and compare the performance of the quadratic penalty method in damping the small scale gravity waves present in the initial fields, we add some artificial perturbations to the geopotential field of the test problem described above. Case 1 was obtained by perturbing the initial height field by the addition of a small perturbation term in the following way

$$\phi'(i, j) = \phi(i, j) + \alpha_{i,j} \times \phi(i, j),$$
$$i, j = 8, 12 \quad (2.20)$$

where $\alpha_{10,10} = 0.9\,\%$, $\alpha_{i,j} = 0.45\,\%$ when i or j equal 9 or 11, and $\alpha_{i,j} = 0.3\,\%$ when i or j equal 8 or 12, where i and j relate to grid locations in the x and y directions, respectively.

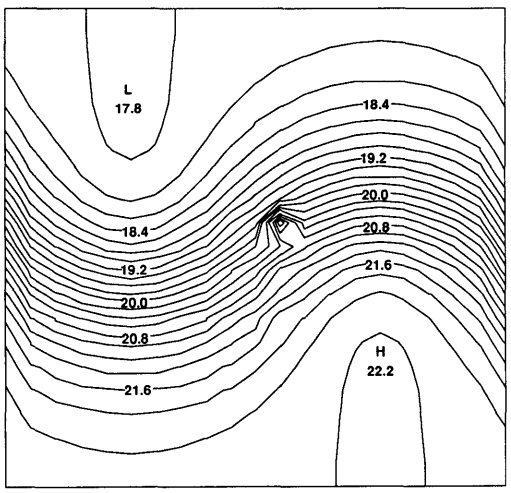

Fig. 1. Case no. 1 initial geopotential field produced by one grid point local perturbation on the Grammeltvedt initial geopotential field condition no. 1. Contour interval is 200 $m^2\ s^{-2}$ and values of the isolines are scaled by 10^3.

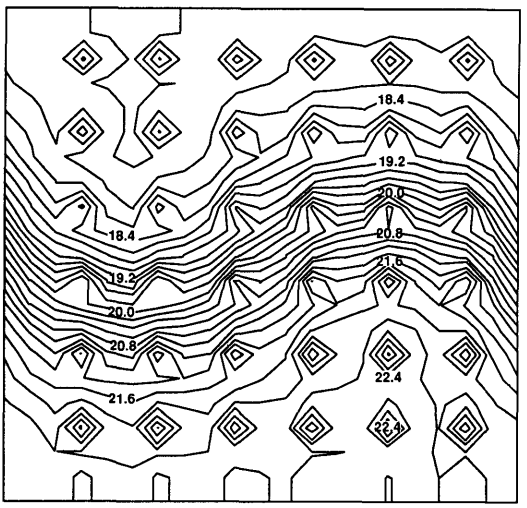

Fig. 2. Case no. 2 initial geopotential field produced by a 6×6 grid point local perturbations on the Grammeltvedt initial geopotential field condition no. 1. Contour intervals and scaling factor for the values of isolines are the same as in Fig. 1.

In case 2, the initial height field was perturbed by

$$\phi''(i, j) = \phi(i, j) + (0.9\,\%) \times \phi(i, j),$$
$$i, j = 3k, \qquad k = 1, 2, ..., 6, \tag{2.21}$$
$$\phi''(i', j') = \phi(i', j') + (0.45\,\%) \times \phi(i', j'),$$

where (i', j') represent all the grid points adjacent to the grid point (i, j).

The isoline graphs of the perturbed height fields ϕ' and ϕ'' are plotted in Figs. 1 and 2, respectively. For both cases the velocity fields were not changed. We see that the height fields thus obtained contain either one local grid point perturbation in Case 1 or a 6×6 local grid point perturbation of the geopotential field for Case 2, respectively.

These two perturbed initial conditions were used to produce the model-generated observations. The variational data assimilation without (with) the penalty term is implemented by starting from a model atmosphere at rest and minimizing the cost function defined in eq. (2.13) (or eqs. (4.3)–(4.5)). When a prescribed convergence criterion is met, the minimization process was stopped. In all the following experiments, we employed L-BFGS, a limited memory quasi-Newton method (Nocedal, 1980, Liu and Nocedal, 1989). The convergence criterion applied for all the numerical experiments was

$$\|\boldsymbol{g}_k\| \leqslant \varepsilon \max\{1, \|\boldsymbol{X}_k\|\}, \tag{2.22}$$

where $\boldsymbol{g}_k$ is the gradient of the cost function at the kth iteration and $\varepsilon = 10^{-14}$ was chosen based on obtaining a perfect solution, i.e., one in which no visible difference with the initial condition used to produce the observations to be assimilated is observed. However the weights used to define the cost function can be interpreted as variances of the data. Errors in the retrieved fields with this convergence criteria are, however, much smaller than the error implied by these weights. The cost function is reduced more than the value of the sum of mean square data errors and it is an overfitting. This is quite acceptable in an "identical twin" approach with lots of data available which corresponds to our cases. In a real data experiment where model and data might be inconsistent (biased), this would lead to the appearance of spurious large-amplitude high-frequency gravity-wave noise in the solution.

3. Penalty and augmented Lagrangian methods—motivation, theory, and numerical algorithms

3.1. Penalty and augmented Lagrangian methods

The basic idea of the penalty method is to eliminate the constraints in constrained optimization problems by adding to the cost function a penalty term which prescribes a high cost to infeasible points. Associated with this method is a parameter r, called the penalty parameter, which determines the severity of the penalty and as a consequence the extent to which the resulting unconstrained problem approximates the original constrained problem.

Consider the problem

$$\text{minimize } J(\boldsymbol{X})$$

$$\text{subject to } h = 0. \tag{3.1}$$

where $\boldsymbol{X}$ is an n-dimensional vector and h is an m-dimensional vector.

Instead of solving the above constrained minimization problem we can solve an unconstrained minimization problem by defining the following augmented Lagrangian function L_r

$$L_r(\boldsymbol{X}, \lambda) = J(\boldsymbol{X}) + \lambda^{\mathrm{T}} h + \tfrac{1}{2} r\,|h|^2 \tag{3.2}$$

for any scalar r, where L_r is called the augmented Lagrangian function and λ is the multiplier vector (Bertsekas, 1982).

The quadratic penalty method consists of solving a sequence of unconstrained minimization problems of the form

$$\text{minimize } L_{r_k}(\boldsymbol{X}, 0), \tag{3.3}$$

where the multipliers are set to zero and $\{r_k\}$ is a penalty parameter sequence satisfying

$$0 < r_k < r_{k+1} \qquad \forall k,\ r_k \to \infty. \tag{3.4}$$

The sequential penalty method iterative procedure can then be algorithmically summarized as follows:

(i) Start with an initial point $\boldsymbol{X}_1$ and an initial value of $r_1 > 0$. Set $l = 1$.

(ii) Minimize $J'(\boldsymbol{X}, r_l)$ by using an unconstrained minimization method and obtain the solution $\boldsymbol{X}_l^*$.

(iii) Test whether $\boldsymbol{X}_l^*$ is a solution of the problem (i.e., satisfying the constraints $h = 0$ within some prescribed accuracy criteria). If this is true, then terminate the process. Otherwise, go to the next step.

(iv) Increase the value of the next penalty parameter, r_{l+1}, using the relationship

$$r_{l+1} = c r_l \qquad (c > 1), \tag{3.5}$$

i.e., $r_{l+1} > r_l > 0$, where $c \in [4, 10]$ is suggested. (See Bertsekas (1982))

(v) Set the new value of $l = l + 1$, use as a new starting point the point $\boldsymbol{X}_1 = \boldsymbol{X}_l^*$ and go to step (ii).

The method depends for its success on sequentially increasing the penalty parameter to infinity. As r takes higher values, the approximation becomes increasingly accurate whilst the problem becomes extremely ill-conditioned (namely the Hessian of the cost function has a very high condition number).

It is, however, possible to improve considerably the performance of the penalty method by employing a nonzero vector of multipliers λ_k and by updating them in an intelligent manner after each minimization of the form

$$\text{minimize } L_{r_k}(\boldsymbol{X}, \lambda_k) \tag{3.6}$$

(see Hestenes, 1969; Powell, 1977; Navon and De Villiers, 1983). This is the augmented Lagrangian approach, also called the method of multipliers.

It is argued that in the penalty method where $\lambda_k = \text{const}$ (in our case λ_k is 0), it is ordinarily necessary to increase r_k to infinity. It is known theoretically (Bertsekas, 1982) that it is not necessary to increase r_k to infinity in order to induce convergence in the augmented Lagrangian method. This is an important advantage, since it results in the elimination or at least moderation of the ill-conditioning problem. A second important advantage of the augmented Lagrangian method is that its convergence rate is considerably better than that of the penalty method. For the method of multipliers, the rate of convergence is linear or superlinear, while for the penalty method, the rate of convergence is much slower and depends essentially on the rate at which the penalty parameter is increased. This advantage of the augmented Lagrangian method in the speed of convergence

has been verified in many computational studies, where a consistent reduction in the computation time ranging from 30 to 80% has been reported.

A formal description of the typical step of the original version of the augmented Lagrangian method is as follows:

Given a multiplier vector λ_k and a penalty parameter r_k, we minimize the augmented Lagrangian function $L_{r_k}(X, \lambda_k)$ using an unconstrained minimization method to obtain a solution vector X_k. We then update the multipliers vector using the following formula

$$\lambda_{k+1} = \lambda_k + r_k h, \tag{3.7}$$

where the penalty parameter is chosen such that $r_{k+1} \geq r_k$ (say for instance $r_{k+1} = cr_k$, $c \in [4, 10]$) and we repeat the entire process (see Bertsekas, 1982).

The choice of the initial vector of multipliers λ_1 is based either on prior knowledge or a choice of either the zero or unit vector is made in the absence of such knowledge.

3.2. *Definition of the penalized cost function*

Throughout this paper we augment (or penalize) the cost functional by adding a quadratic penalty term

$$r \left\| \frac{\partial \phi}{\partial t} \right\|^2 \tag{3.8}$$

to the cost function in order to reduce the amount of gravity waves present in the variationally assimilated fields, and to impose an appropriate balance between the mass and velocity fields, where $\|\cdot\|$ is an L_2 norm. The method follows the same basic premise as used in the nonlinear normal mode initialization approach of Machenhauer (1977) and seeks to reduce the time derivative of the gravity wave component of the flow. The philosophy here is that we employ the penalty method in a weak sense which is different from that of solving a constrained minimization problem. We try to reduce the absolute value of the time derivative until it reaches a prescribed small value instead of trying to attain a perfect steady state solution ($\partial\phi/\partial t \equiv 0$). This is done by stopping the increase of the penalty parameter and not letting it tend to infinity, which is effectively equivalent to imposing an inequality constraint of the form

$$\left\| \frac{\partial \phi}{\partial t} \right\| \leq H, \tag{3.9}$$

where H is a constant related to the frequency gap for the exterior equivalent depth separating the gravity waves from the Rossby waves. This approach is called penalty regularization.

This method is also equivalent to the now classical terminology introduced by Sasaki (1970) and it is intended to allow only a slowly varying evolution of the field, i.e., it will result in a projection on the slow manifold of the motion.

The cost function assumes thus the following form

$$J' = J + r \left\| \frac{\partial \phi}{\partial t} \right\|^2 \tag{3.10}$$

for the penalty method and

$$J' = J + r\psi \left(\left\| \frac{\partial \phi}{\partial t} \right\| \right) \tag{3.11}$$

where

$$\psi(z) = \begin{cases} 0; & \text{if } z \leq H; \\ (z - H)^2 & \text{otherwise.} \end{cases}$$

for the penalty regularization approach.

The augmented Lagrangian function L_r assumes the form

$$L_r(X, \lambda) = J(X) + \lambda^{\mathrm{T}} \frac{\partial \phi}{\partial t} + \frac{1}{2} r \left\| \frac{\partial \phi}{\partial t} \right\|^2 \tag{3.12}$$

3.3. *Gradient with penalty and Lagrangian terms included in the cost function*

To assess the performance of the quadratic penalty method in damping the small scale gravity waves present in the initial fields, one needs to provide the gradient of the cost function to be minimized for any unconstrained optimization routine. From the basic derivation of adjoint model we found out that the gradient of the penalized cost functions can be easily calculated by integrating the same adjoint model with different forcing terms added to the adjoint model, that is, besides adding the forcing terms $2W(t_r)$

$(X(t_r) - X^{obs}(t_r))$ in the adjoint model whenever an analysis time t_r $(r = 0, 1, ..., R)$ is reached, we also add at every time step the following additional terms

$$2rA^* \frac{\partial \phi}{\partial t}, \tag{3.13a}$$

$$rA^* \psi' \left(\frac{\partial \phi}{\partial t} \right), \tag{3.13b}$$

$$2rA^* \frac{\partial \phi}{\partial t} + \lambda \tag{3.13c}$$

to the right-hand side of the adjoint model for calculating the gradients of the cost functions defined in equations (3.10)–(3.12), respectively.

4. Application of the penalty function method

4.1. *Basic experiments: results of penalty and penalty regularization*

Let us first discuss the results obtained for Case 1. Having model produced observations, the minimization (started from 0) of the cost function which does not include a penalty term terminated successfully after 104 iterations and 153 function calls. The cost function decreased by 10 orders of magnitude while the norm of gradient decreased by 6 orders of magnitude. The variational assimilation was thus able to perfectly retrieve the original initial conditions (not shown). This result served as our base experiment.

After that, we carried out several sequential minimizations of the cost function, augmented with the penalty term (3.8) with the intention of damping out the local gravity wave perturbation illustrated in Fig. 1. One might have thought that the gravity wave can be damped to any desired accuracy by simply choosing a very large value of the penalty parameter r, and then carrying out a single unconstrained minimization. However from our experiments, we found out that for smaller values of the penalty parameter r (say for instance $r < 10^5$) the retrieved geopotential field was not smooth enough and for larger values of r $(> 10^5)$ the minimization failed. For a value of $r = 10^5$, a single minimization cycle performed very well at first but after 105 iterations with 141 function calls, the minimization failed again before the same convergence criteria (2.22) was reached. However upon examination of the geopotential field after 105 iterations, we found out that a smooth initial geopotential field has already been obtained. A measure of the efficiency with which the constraint imposed by (3.8) is enforced, is given by the following corresponding variation of the quantity $\|\partial\phi/\partial t\|^2$ which decreases from 0.77×10^2 m^4 s^{-6} to 0.70×10^{-1} m^4 s^{-6} in the course of minimization (see Fig. 3), i.e., a decrease of three orders of magnitude in the squared time tendency of the geopotential field pointing out to a successful filtering of high-frequency gravity-waves.

We also carried out the experiment using the penalty regularization method with $H = 1.0 \times 10^{-3}$. Similar results are obtained.

4.2. *Results using the exterior penalty function method*

The sensitivity of the performance of the minimization for different values of penalty parameter r_k is not surprising since it is well-known that the condition number of the Hessian matrix evaluated at the minimum increases as r becomes larger (the so-called ill-conditioning effect of the penalty method) (Gill et al., 1981, see also Appendix A). The conditioning of the Hessian matrix has a special significance in this case. If the initial value of r is "too large", even a robust unconstrained minimization algorithm (as the one used in our case) will typically experience great difficulty in its attempt to compute the minima due to the slow convergence induced by the increasingly larger condition number of the Hessian of the penalized

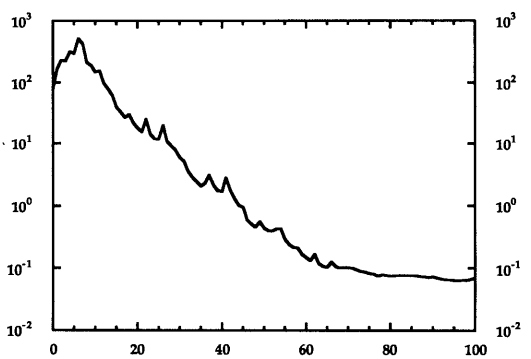

Fig. 3. Variation of the value of $\|\partial\phi/\partial t\|^2$ with the number of iterations during the minimization of the penalized cost function with a penalty parameter $r = 10^5$.

cost function. Therefore, in order to solve the problem by a penalty function method, a sequence of unconstrained minimization problems has to be solved, with moderately increasing values of the penalty parameter (called the exterior quadratic penalty method, see Fiacco and MrCormick, 1968 and Rao, 1983). In the exterior penalty function method, each successive $x^*(r_l)$ is used as the starting point for solving a minimization problem with the next increased value of the penalty parameter, until an acceptable convergence criterion (i.e., a satisfactory solution) is attained for the first time (where $x^*(r_l)$ is the minimum point obtained using the preceding penalty parameter r_l).

Table 1 presents the results of the application of the exterior penalty function method with $r_1 = 1$, $c = 10$, $l = 1, 6$. We see that with a penalty parameter equal to r_1, the minimization terminates with 106 iterations and 157 function calls. If we examine the assimilated initial height field we find that all the characteristics of the large scale features have been retrieved. The following sequential minimization of the cost function augmented by a quadratic penalty term with penalty parameters equal to r_2, r_3, and r_4, respectively, had a rather small impact on the smoothing of the geopotential field (i.e., the elimination of the gravity waves). Much of the damping of the gravity waves occurs however when larger penalty coefficients are used in stages 5 and 6 of the sequential minimization procedure. After stage 6, no sizable amount of high-frequency gravity waves remains to be damped. This can also be inferred by the small decrease in the norm of gradient with a larger penalty parameter r_7 (see Table 1). It seems that the exterior penalty function method is very efficient in damping the small scale gravity wave perturbations when a threshold value of the penalty term has been attained. However, if we consider the total number of iterations (188) and total number of function calls (282) required to reach stage 6, we find out that the required computational effort is too expensive.

Table 1. *Numerical results of the variational data assimilation with application of the exterior penalty function method to assimilate observations with one local perturbation (7 cycle)*

K	r_k	ITER	NFUN	J/J_0	$\lvert G\rvert/\lvert G_0\rvert$	ε
1	10^0	106	157	$4.05 \cdot 10^{-7}$	$1.41 \cdot 10^{-6}$	10^{-14}
2	10^1	5	8	$2.24 \cdot 10^{0}$	$2.94 \cdot 10^{-1}$	10^{-14}
3	10^2	14	21	$9.69 \cdot 10^{-1}$	$2.28 \cdot 10^{-1}$	10^{-14}
4	10^3	16	34	$1.97 \cdot 10^{0}$	$1.64 \cdot 10^{-1}$	10^{-14}
5	10^4	17	27	$4.21 \cdot 10^{-1}$	$8.35 \cdot 10^{-2}$	10^{-14}
6	10^5	30	45	$3.97 \cdot 10^{-1}$	$4.47 \cdot 10^{-2}$	10^{-14}
7	10^6	22	44	$7.89 \cdot 10^{-1}$	$1.51 \cdot 10^{-1}$	10^{-14}

To reduce the computational cost, we carried out a second exterior penalty function numerical experiment with $r_1 = 10^2$, $c = 10$, $l = 1, 4$, in which each subsequent minimization cycle uses an increasingly stringent convergence criteria as shown in Table 2 (also see Navon and deVilliers, 1983). Again most of the large scale characteristic features are retrieved after the first minimization cycle (Figs. 4a–b). The gravity wave oscillations were damped out gradually till the fourth minimization cycle ended with $r_4 = 10^5$. The solution for the geopotential field is practically the same as that in previous exterior experiment. The total number of iterations and function evaluations are 102 and 165, respectively. The computational cost of this procedure is now comparable with the computational cost of minimization without the penalty term being included in the cost function.

4.3. Results using the short cut penalty method

If we examine the performance of the above experiments, one can easily find out that the minimization with a small penalty term or without a penalty term at all but with a looser convergence criteria can retrieve most of the large-scale features of the initial condition fields, while a minimization with larger penalty terms can damp out almost all the small scale gravity wave features present in the meteorological fields. This implies that only two successive minimizations are required to obtain a smooth solution devoid of gravity wave oscillations.

Table 2. *Numerical results of the variational data assimilation with application of the exterior penalty function method to assimilate observations with one local perturbation (4 cycle)*

K	r_k	ITER	NFUN	J/J_0	$\lvert G\rvert/\lvert G_0\rvert$	ε
1	10^2	48	74	$4.46 \cdot 10^{-6}$	$1.41 \cdot 10^{-4}$	10^{-12}
2	10^3	16	26	$7.85 \cdot 10^{-1}$	$1.76 \cdot 10^{-1}$	10^{-13}
3	10^4	19	28	$4.20 \cdot 10^{-1}$	$1.02 \cdot 10^{-1}$	$5 \cdot 10^{-13}$
4	10^5	26	37	$3.96 \cdot 10^{-1}$	$3.13 \cdot 10^{-2}$	10^{-14}

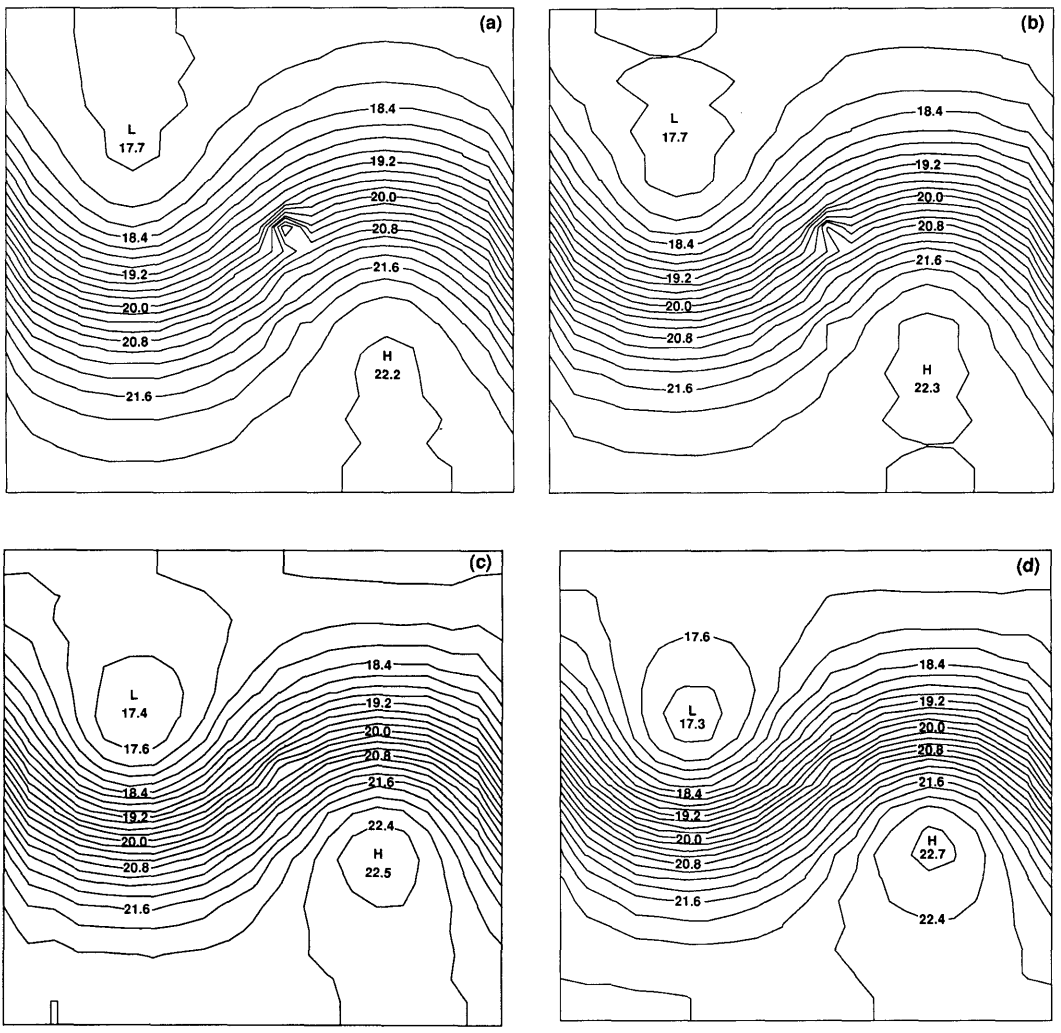

Fig. 4. Retrieved initial geopotential height after each cycle of the exterior penalty function method with a penalty parameter set to (a) $r_1 = 10^2$, (b) $r_2 = 10^3$, (c) $r_3 = 10^4$ and (d) $r_4 = 10^5$, respectively. The Case no. 1 observations were assimilated, where the initial guess for the first minimization cycle is set to zero, the solution of the ith penalized minimization was used as the initial guess for the $(i+1)$th minimization, $i = 1, ..., 3$. Contour intervals and scaling factor for the values of isolines are the same as in Fig. 1.

We conducted therefore, another experiment in which the first unconstrained minimization was carried out without a penalty term being included in the cost function and which was terminated when a convergence criteria of 10^{-12} (instead of a stringent one of 10^{-14}) was reached. The minimization solution of the case where no penalty term was included is then taken as the first guess for the next minimization cycle with a large

Table 3. *Numerical results of the variational data assimilation with combination of the non-penalized assimilation and penalized assimilation to assimilate observations with one local perturbation*

K	r_k	ITER	NFUN	J/J_0	$\lvert G\rvert/\lvert G_0\rvert$	ε
1	0	52	75	$4.46 \cdot 10^{-6}$	$1.41 \cdot 10^{-4}$	10^{-12}
2	10^5	34	45	$7.85 \cdot 10^{-1}$	$1.76 \cdot 10^{-1}$	10^{-14}

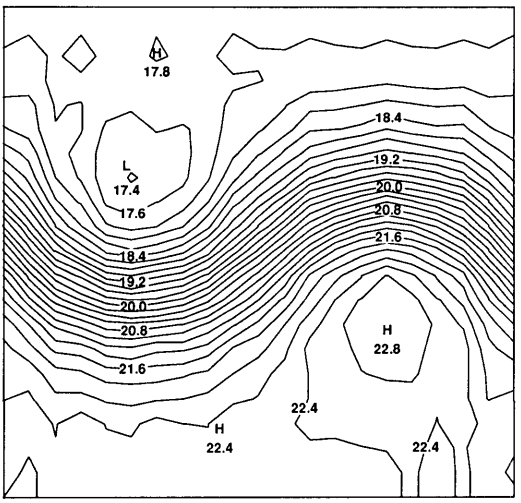

Fig. 5. Retrieved initial geopotential field after the second stage of the short cut minimizations when the Case no. 2 observations are assimilated, where the initial guess for the first minimization is set to zero and the solution of the first minimization without a penalty term was used as the initial guess for the second minimization. Contour intervals and scaling factor for the values of isolines are the same as in Fig. 1.

penalty term being included in the cost function ($r = 10^5$). The minimization of the penalized cost function was then terminated when a convergence criterion of $\varepsilon = 10^{-14}$ was reached. We found that the distribution of the retrieved initial height fields of the two minimizations is the same as in Fig. 4(d). As far as the computational cost is concerned (see Table 3), this experiment required a total of only 86 iterations and 120 function calls for the whole procedure, i.e., it was computationally cheaper than the base experiment.

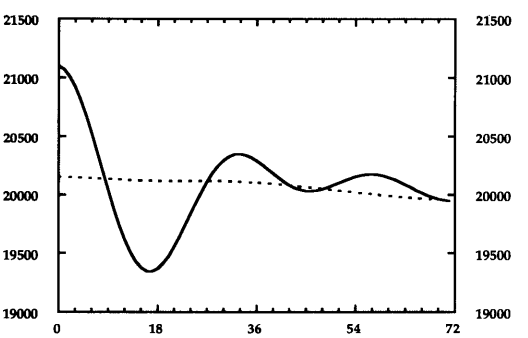

Fig. 6. Time variation of the geopotential fields for 24 h at grid point (10, 10) after variational data assimilation with (dotted line) and without (solid line) a penalty term being included in the cost function when the Case no. 1 observations are assimilated.

Table 4. *Numerical results of the variational data assimilation with combination of the non-penalized assimilation and penalized assimilation to assimilate observations with* 6×6 *local perturbation*

K	r_k	ITER	NFUN	J/J_0	$\lvert G\rvert/\lvert G_0\rvert$	ε
1	0	55	75	$2.12 \cdot 10^{-7}$	$1.27 \cdot 10^{-4}$	10^{-12}
2	10^5	33	45	$2.33 \cdot 10^{-2}$	$1.44 \cdot 10^{-2}$	10^{-14}

To illustrate the validity and effectiveness of this two stage minimization procedure, we implemented it for the test Case No. 2. Table 4 and Fig. 5 present the numerical results obtained for this experiment. It is clear that this procedure (the short-cut penalty approach) also performs very well on Case No. 2, which contains a much larger amount of small scale gravity wave features than Case No. 1, except that near the north and south boundaries there are still some small scale noises. The total computational cost turned out to be rather cheap (86 iterations and 120 function calls) compared to the cost of a single minimization without a penalty term which leaves the small scale gravity wave features undamped (107 iterations and 155 function calls).

4.4. *Impact of the VDA with a penalty term included on short-range forecast (24 h)*

A number of 24 h forecasts (14 h real forecast due to a length of 10 h assimilation window) starting from the variationally data assimilated initial state with and without the inclusion of a quadratic penalty term in the cost function were

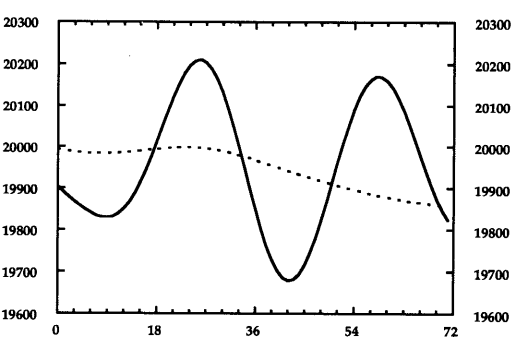

Fig. 7. Same as Fig. 6 except for Case no. 2.

carried out using a finite difference shallow water equations model. The results of these forecasts were then compared to the results of a 24 h integration starting from a non-variationally assimilated state.

In Fig. 6 we plot the time variation of the geopotential field at a given grid point for a variationally assimilated integration without the inclusion of a penalty term as well as the results from the integration with a penalty term being included in the cost function. Fig. 7 is the same as Fig. 6 but for case 2.

From the hourly evolution of the geopotential field for the short-range forecast (24 h) with the penalized variational data assimilation and a run without the inclusion of a penalty term, we observe a major impact for the first 6–10 hours of the numerical integration (see Fig. 6) for one localized grid point perturbation case. In the case of multiple-localized perturbations the impact of the penalty approach remains visible after 24 h.

A perfect damping of the short gravity waves is observed, matching in every respect similar results obtained by the use of nonlinear normal mode initialization. However the present method has several advantages, the first being the fact that no explicit knowledge of the normal modes of the model is required, the second being that using the penalty function approach it is possible to take into account data distribution in time. However, the variational data assimilation procedure itself (without the inclusion of any penalty term) remains computationally expensive. The initialization here comes as a by-product and involves little additional cost beyond the computational cost required by the variational data assimilation.

The results obtained compare favorably with those obtained by other researchers using implicit normal mode initialization of the shallow-water equations model over a limited area domain (See Juvanon du Vachat, 1986; Semazzi and Navon, 1986; and Temperton, 1988).

5. Application of the augmented Lagrangian method

To compare the numerical performance and computational cost of the method of multipliers, we carried out an experiment minimizing the augmented Lagrangian function, identical to the second exterior penalty function experiment ($r_1 = 10^2$, $c = 10$, $l = 1, 4$, where the initial Lagrange multiplier is $\lambda_1 = 0$ and λ_l is updated by equation (3.7), the first order multiplier update formula of the augmented Lagrangian method). The number of iterations and the number of function evaluations for each run is given in Table 5. Comparing Table 5 with Table 2, we can not see any improvement due to the method of multipliers. However by considering the retrieved initial geopotential field after each run (Fig. 8), we found out that most of the large-scale characteristic features were retrieved after the first run. The gravity wave damping effect occurs mainly in the next two runs. A comparison with the Fig. 4 shows that the solution from the third iteration of the method of multipliers is better than the fourth iteration of the penalty method, whilst the total numbers of iterations and function evaluation are 96 and 158, respectively.

Table 5. *Numerical results of the variational data assimilation with application of the method of multiplier to assimilate observations with one local perturbation (4 cycle)*

K	r_k	ITER	NFUN	J/J_0	$\lvert G\rvert/\lvert G_0\rvert$	ε
1	10^2	54	78	$2.07 \cdot 10^{-6}$	$1.07 \cdot 10^{-4}$	10^{-12}
2	10^3	16	32	$5.49 \cdot 10^{-1}$	$6.37 \cdot 10^{-2}$	10^{-13}
3	10^4	24	41	$4.35 \cdot 10^{-1}$	$8.14 \cdot 10^{-2}$	$5 \cdot 10^{-13}$
4	10^5	23	44	$3.50 \cdot 10^{-1}$	$3.32 \cdot 10^{-1}$	10^{-14}

Therefore, with the application of the augmented Lagrangian method we effectively obtain an additional order of decrease in the value of penalty parameter while at the same time obtaining a more satisfactory solution. However the improvement obtained in this case is not impressive in as far as the computational cost is concerned.

Table 6. *Numerical results of the variational data assimilation with application of the method of multiplier to assimilate observations with one local perturbation (short cut)*

K	r_k	ITER	NFUN	J/J_0	$\lvert G\rvert/\lvert G_0\rvert$	ε
1	10^1	4	14	$1.24 \cdot 10^{-2}$	$7.75 \cdot 10^{-2}$	10^{-9}
2	10^2	29	42	$4.05 \cdot 10^{-4}$	$4.63 \cdot 10^{-3}$	10^{-13}
3	10^5	18	24	$3.70 \cdot 10^{-2}$	$3.10 \cdot 10^{-2}$	10^{-14}

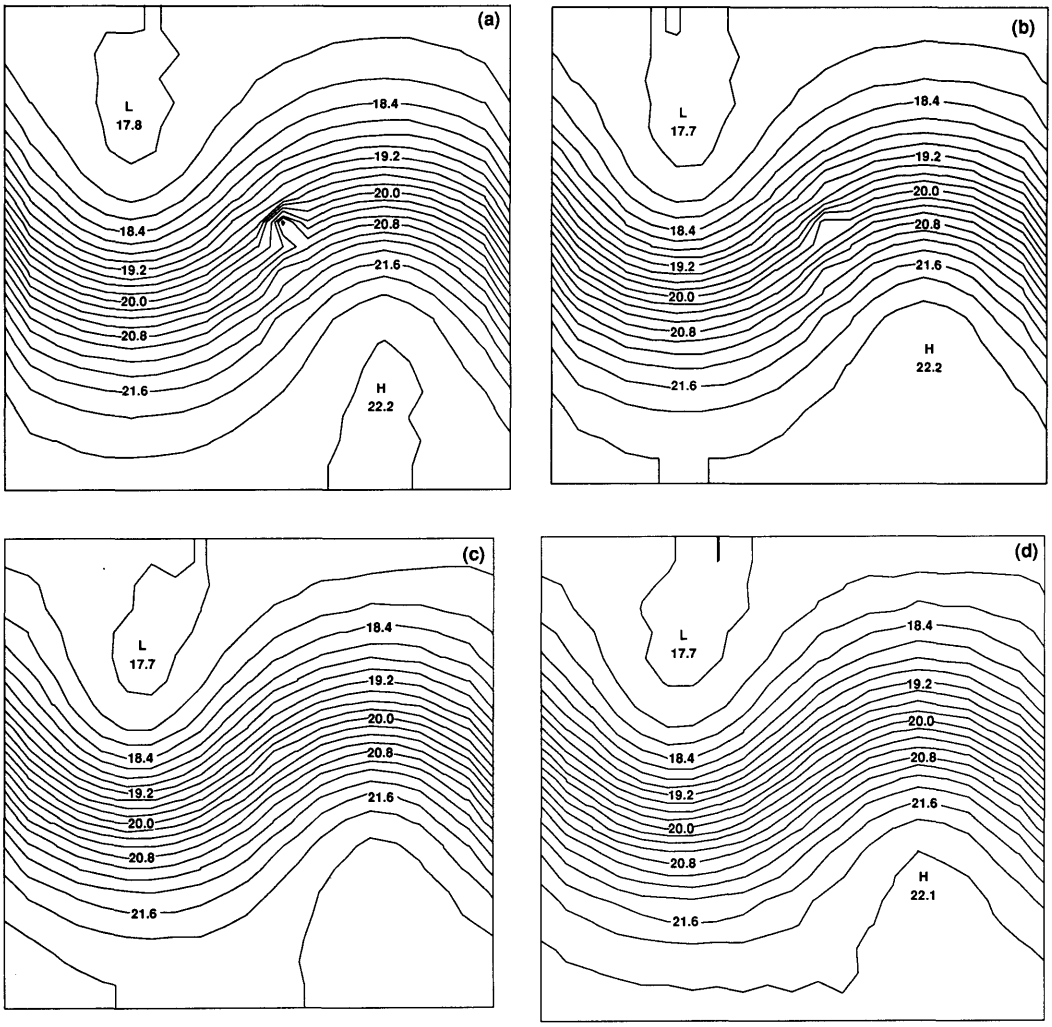

Fig. 8. Same as Fig. 4 except using the augmented Lagrangian method.

And similarly, the short cut augmented Lagrangian method is much cheaper (see Table 6) and performs better than the short cut penalty method in the sense of better quality of the retrieval.

6. The case of incomplete information

6.1. Uniqueness of the solution

In more realistic situations, meteorological observations are temporally and spatially distributed, and are inhomogeneous both in quality as well as in quantity. The question arises as to what will happen if we decrease the number of the observations in space. This is mathematically equivalent to determining the necessary condition for a unique solution of the problem of optimization.

For this case, the objective function may be written as

$$\min J(\boldsymbol{X}_0) = \min \langle W(C\boldsymbol{X}(\boldsymbol{X}_0) - \boldsymbol{X}_{\text{obs}}), C\boldsymbol{X}(\boldsymbol{X}_0) - \boldsymbol{X}_{\text{obs}} \rangle, \quad (6.1)$$

where X is an n-dimensional vector defined in the space R_n, X_{obs} is an m-dimensional vector ($m \leqslant n$) defined in space R_m, and C is a projection operator from the space R_n to the space R_m. In this case the gradient of the objective function with respect to the initial conditions is obtained by the same method described in Subsection 2.1 except that the forcing term added to the adjoint variable during the integration of the adjoint model is changed from (2.16) to the following form

$$2C^{\mathrm{T}}W(CX - X_{obs}). \tag{6.2}$$

The uniqueness of the solution of minimizing problem (6.1) depends on the structure of the Hessian matrix H. (6.1) has a unique solution U^* if the Hessian H is positive definite. In the linear case, the forecast model can be written as

$$\frac{dX}{dt} = AX, \tag{6.3}$$

where A is a constant matrix independent of X. The cost function, J, being a summed up semi-norm, is always positive and consequently the matrix H is positive semi-definite. Thus for problem (6.1) to have a unique solution, H must be positive definite.

An explicit expression for the Hessian matrix H was derived in Appendix B (B8). The uniqueness condition of the solution, i.e., H is positive definite, is transformed into an algebraic condition on A and C. It is stated that if the rank of

$$\mathscr{A} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}, \tag{6.4}$$

is equal to n, then the Hessian matrix H will be positive definite.

It is seen that the Hessian H is independent of the observation itself. It depends on the model (operator A) and on a projection (operator C) mapping the model variables into the space of observations.

In the nonlinear case, however, J is no longer quadratic with respect to the control variable U, where U represents model initial state. Due to the nonlinearity of the forecast model, no global results may be obtained. However since J is bounded from below we know that there exists at least one local minimum U^* which is characterized by $\nabla J(U^*) = 0$. A sufficient condition for U^* to be a unique local minimum is that $H(U^*)$ be positive definite. In Appendix C, we present a particular method to calculate the Hessian matrix using a second order adjoint model integration. One time integration of the second order adjoint (at a cost around 1–2 times integration of the direct foreward model) can provide the value of one column of the Hessian matrix, or the value of the product of the Hessian and a vector. Due to the large dimension of the control variable (10^4 to 10^6) in variational data assimilation, the computation of a full Hessian may be prohibitive for real applications. However it is possible to obtain an estimate of the spectral radius and the condition number of the Hessian by using an iterative power method or a Rayleigh quotient method, both of which require only the value of the Hessian multiplied by a vector. One can also obtain several large and small eigenvalues of the Hessian by a deflation method (Golub and Van Loan, 1989).

At the end of Appendix B an analysis of the relationship between the uniqueness of the solution and the observations was carried out for a simple linearized shallow water equations. In the next section, we will carry out some numerical experiments to estimate the influence of the distribution of available observations on the uniqueness and on the convergence rate to the local minimum.

6.2. Results with incomplete observations

We use the L-BFGS (Liu and Nocedal, 1989) unconstrained minimization algorithm, which seems to be a most efficient and robust large-scale minimization routine (Zou et al., 1991). First we decreased the number of the observational fields. Suppose that only the u and v fields are observed and there are no observations of the geopotential field. The numerical results show that the minimization of the cost function defined in (2.13) without the first summation was able to retrieve a unique minimum, i.e., not only perfect initial fields for the u and v components but also a smooth balanced ϕ field. If only the geopotential field is observed, the minimization of the cost function was also able to retrieve a unique minimum which is the same as the Grammeltvedt initial condition. These results are consistent with the theoretical results of Appendix B, which determine the

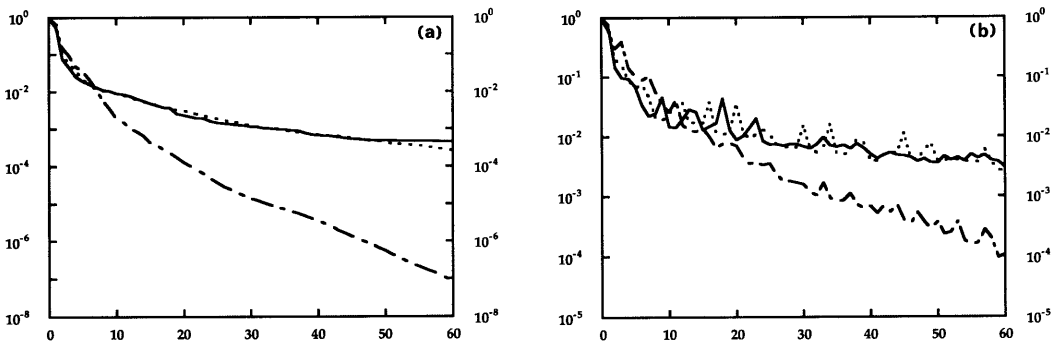

Fig. 9. Variations of (a) the scaled value of the cost function and (b) the scaled norm of the gradient of the cost function with the number of iterations when the observations for the three fields are available at every 1 (dash dot line), 2 (dotted line), and 4 (solid line) grid points in both x and y directions.

necessary conditions for a unique minimum of the cost function.

Next we decreased the number of grid points where observations are available. If we reduce the observational data in grid space (i.e., in the horizontal space dimension) by alternating the available observational data in both the x and y directions, for instance by assuming that data was available only every two grid points in each direction (i.e., the number of observations was decreased from 21×21 to 11×11), the minimization process fails. Namely after 71 iterations, the minimization stops due to rounding errors before a smooth solution was obtained. No reasonable retrieval is obtained and the objective function and the norm of gradient decrease by only about 2–3 orders of magnitude, which is roughly half of the orders of magnitude known to be necessary from our experiment to reach a smooth solution. If we continue to further decrease the number of observations in both space directions to every 4 grid points for instance, no significant difference from the case of every 2 grid points is observed (see Fig. 9) and also no satisfactory convergence and retrieval are obtained. It seems that the performance of the minimization depends really on the number of available observations at the grid points. At least, this is the case for the limited-area shallow-water equations random perturbation test, where the control variables are only the initial conditions.

If we continue further to decrease the number of observations on the grid points (in the space dimension) to every 14 or 19 grid points for instance, there is no significant difference from the case of every 2 grid points, i.e., no good convergence and no satisfactory retrieval are obtained. It seems that the performance of the minimization is more sensitive to a decrease in the number of observations available in space than in time. At least, this is the case for the limited-area shallow-water equations random perturbation test, where the control variables are only the initial conditions. If we calculate the maximum and minimum eigenvalues and the condition number of the Hessian at the initial guess of the minimization (Table 7), we

Table 7. *Values of the maximum eigenvalue (λ_{max}), the minimum eigenvalue (λ_{min}) and the condition number (Cond. no.) of the Hessian of the objective function at the initial guess point using the power method and the shifted power method, when observations are available either at every time step and on each grid point (Case 1), or on each grid point but only at every 30 time steps (Case 2), or at every time step but only every two grid points (Case 3)*

Problems	λ_{max}	λ_{min}	Cond. no.	Convergence criteria
Case 1	$2.1981 \cdot 10^{-4}$	$4.8934 \cdot 10^{-7}$	449.2	$\lvert\lambda_{k+1} - \lambda_k\rvert < 10^{-8}$ (k the iteration number)
Case 2	$1.8362 \cdot 10^{-4}$	$1.0830 \cdot 10^{-7}$	1695.5	
Case 3	$2.9044 \cdot 10^{-4}$	$-1.6569 \cdot 10^{-5}$	17.53	

find that the reduction of observations in the grid space produces a negative minimum eigenvalue of the Hessian. Thus, we obtain an indefinite Hessian, indicating to multiple minima or saddle points.

One could wonder whether the results from experiment in which data is given every two grid points could be partially improved if instead of a random perturbed field, a first guess field lacking small scale structure, e.g., a complete flat field, was used to start the minimization. Minimization starting from a flat initial guess yields similar results with the above experiment. This shows that the failure of the minimization with observations available only every two grid points is independent of the fact that we used a randomly perturbed noisy initial guess.

If the number of observation was diminished only in one space dimension (in this case the x direction), say for instance that observations are available at every two grid points, the minimization was able to retrieve a smooth solution but the convergence rate became rather slow. After 200 function evaluation the retrieved geopotential fields still contain small noises. It seems that having data available at every two grid points in the x direction can still allow the existence of a unique solution but will probably make the shortest waves part of the null space solution. The model still resorts to these degrees of freedom to better fit the data. Imposing a penalty term (3.8) which has the effect of damping the fast waves in the solution may eliminate or reduce the use of these degrees of freedom in the fitting process. Numerical experiments show that with the penalty term included in the cost function the minimization in the case of incomplete data in the x direction was able to reach a faster convergence than without a penalty term. The retrieved geopotential field after 200 function call is rather smooth. However the convergence rate of the incomplete observation with a penalty term included is still slower than the case when observations are available at all the grid points, in which case 83 function evaluations were required to satisfy the convergence criteria in (2.22).

One may ask what will happen if observational data are reduced in time dimension? Suppose that observations are available, say for instance, every 2, 10, 30 or 60 time steps instead of at each time step. Fig. 10 shows the variation of the scaled value of the objective function (J/J_0) and the scaled norm of its gradient ($\|g\|/\|g_0\|$), when the number of time levels at which observations are available decreases. It is evident that if observations are available every 2, 10 or 30 time steps, the performance of the minimization is very similar to the case where observations were available at every time step. But as the time interval between two subsequent observations increases to 60 steps, i.e., observations are available only at the first and last time steps of the assimilation window, the convergence rate becomes slower after 10 iterations, and the minimization reaches the local minimum after 80 iterations instead of 50–60 iterations. A satisfactory retrieval of initial wind and the geopotential height fields is obtained even for the 60 time step interval of observations case. The difference fields between the retrieved and unper-

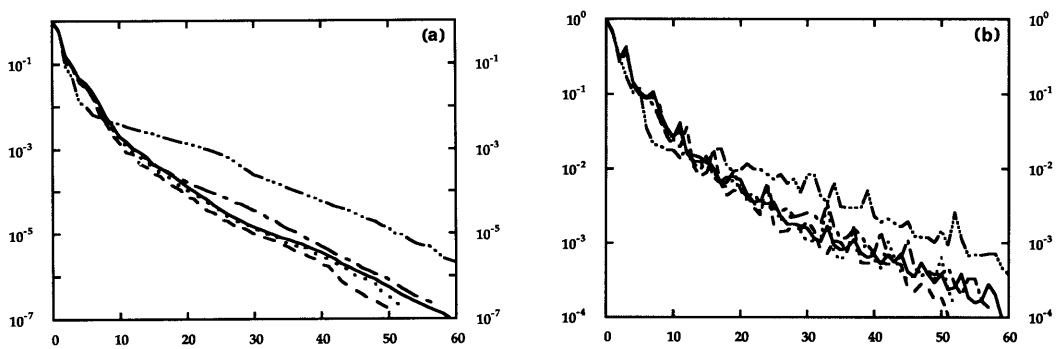

Fig. 10. Variations of (a) the scaled value of the cost function and (b) the scaled norm of the gradient of the cost function with the number of iterations when the observations for the three fields are available at every 1 (solid line), 2 (dotted line), 10 (dash line), 30 (dash dotted line) and 60 (dash dots line) time step, respectively.

turbed initial wind and geopotential are several orders of magnitude smaller than the initial fields.

7. Summary and conclusions

In this paper, we have used optimal control theory of partial differential equations and applied it to a variational 4-D data assimilation in a shallow-water equations model where only initial conditions, or both the initial and the boundary conditions, serve as control variables. The limited-memory quasi-Newton method of Nocedal (Nocedal, 1980; Liu and Nocedal, 1989) (L-BFGS was applied to the problem of minimizing objective functions consisting of the weighted sum of squares between the model-computed quantities and known observations plus some penalty terms. The observations were created from the model integration starting from the Grammeltvedt (1969) initial conditions and the initial guess for the minimization was taken as either a random perturbation of the Grammeltvedt initial conditions or a complete flat field. Two issues related to variational data assimilation, i.e., gravity wave control, and incomplete observations, were addressed.

The minimization was able to retrieve the perfect initial conditions with proper scaling. This occurred both when only the initial conditions or when the initial plus boundary conditions served as the control variables. However, the convergence rate and accuracy of the retrieval are different.

The experiments carried out with diminishing observational data indicate that the case of insufficient data can lead to rather large changes in the pattern of behavior of minimizing the objective function. As a consequence, it may be necessary to include an additional term of penalization in the objective function in order to improve the conditioning of the Hessian of the objective function. Information from second order derivatives of the objective function gives more information on the spectral structure of our problem.

Application of the exterior penalty method shows how the small scale gravity wave noises were reduced with increasing values of the penalty parameter. Results from the short-cut penalty method are very encouraging and show that gravity wave noises can be effectively eliminated without applying the NMI algorithm. Application of the augmented Lagrangian method leads to a better quality of meteorological fields in the retrieved solution.

8. Acknowledgments

This research was funded by NSF Grant ATM-9102851. Additional support was provided by National Meteorological Center and the Supercomputer Computations Research Institute at Florida State University, which is partially funded by the Department of Energy through contract No. DE-FC0583ER250000. The helpful comments of the two reviewers of this paper are gratefully acknowledged.

9. Appendix A

The penalty method theory

The basic idea in penalty methods is to eliminate the equality constraints by adding to the objective function a penalty term which prescribes a high cost to infeasible points.

Associated with the penalty methods is a penalty parameter which determines the severity of the penalty and as a consequence the extent to which the resulting unconstrained problem approximates the original equality constrained problem.

Consider the problem

$$\min f(x) \tag{A.1}$$

$$\text{subject to } h(x) = 0 \tag{A.2}$$

where $f: \mathscr{R}^n \to \mathscr{R}$ and $h: \mathscr{R}^n \to \mathscr{R}^m$ are given functions. We assume throughout that (1) has at least one feasible solution.

Let us define, for any scalar, the augmented Lagrangian function $L_r: \mathscr{R}^n \times \mathscr{R}^m \to R$ by

$$L_r(x, \lambda) = f(x) + \lambda^{\mathrm{T}} h(x) + \tfrac{1}{2} r\, |h(x)|^2, \tag{A.3}$$

where r is penalty parameter and λ is vector of multipliers or simply multiplier.

The quadratic penalty method consists in solving a sequence of unconstrained minimization problems of the form

$$\min L_{r_k}(x, \lambda_k), \tag{A.4}$$

where $\{\lambda_k\}$ is a bounded sequence in $\mathscr{R}^m$ and $\{r_k\}$ the penalty parameter sequence satisfies that

$$r_{k+1} > r_k > 0 \quad \forall k, \quad \text{and} \quad r_k \to \infty \quad \text{as} \quad k \to \infty. \tag{A.5}$$

In the original version of the penalty method we take the multipliers λ_k to be all equal to 0 and the method proceeds by sequentially increasing the penalty parameter to infinity.

The ill-conditioning of the penalty method can be described as follows: Let us denote by

$$P(x, r) = f(x) + \tfrac{1}{2} r h^{\mathrm{T}}(x)\, h(x) = f(x) + \tfrac{1}{2} r \, \|h(x)\|_2^2. \tag{A.6}$$

The Hessian matrices $\nabla^2 P(x^*, r)$ become increasingly ill-conditioned as r increases (Murray, 1969; Lootsma, 1969). To observe why, take the Hessian of P at an arbitrary point x

$$\nabla^2 P(x, r) = H + \sum_{i=1}^{m} r h_i H_i + r A^{\mathrm{T}} A, \tag{A.7}$$

where $A(x)$ is an $m \times n$ matrix whose ith row transpose $a_i^{\mathrm{T}}(x)$ is the gradient vector of the constraint function $h_i(x)$.

At $x^*(r)$ for large r, it follows from

$$\lim_{r \to \infty} \lambda_i(r) = \lambda_i^* \tag{A.8}$$

that the first two terms form an approximation to the bounded matrix $L(x^*, \lambda^*)$, where $L_k(x_k) = H(x_k) + \sum_{i=1}^{m} r h_i H_i$. The eigenvalues of $L_k(x_k)$ are restricted to the tangent space M of the active constraints (see Luenberger, 1984, p375). Thus if $m < n$, the matrix evaluated at $x^*(r)$ is dominated by the unbounded rank deficient matrix $r A^{\mathrm{T}} A$.

As $r \to \infty$, Murray (1971) shows that $n - m$ eigenvalues of $\nabla^2 P(x^*, r)$ are bounded with associated eigenvectors that in the limit lie in the null space of $A(x^*)$ and tend to the eigenvalues of L_M, the restriction or projection of L to the tangent space M. However, the remaining m eigenvalues are of order r. These m eigenvalues tend to infinity as $r \to \infty$. Ill-conditioning can be overcome only by using for each kth iteration a starting point for the unconstrained minimization procedure, which is close to a minimizing point of

$$L_{r_k}(\cdot, \lambda_k)$$

usually taken to be the last point x_{k-1} obtained from the previous unconstrained minimization. This in turn requires that the rate of increase of the penalty parameter be relatively small. One has to balance the benefit of fast convergence with the evil of ill-conditioning. One usually chooses $r_{k+1} = \beta r_k$, $\beta \in [4, 10]$ (see Bertsekas, 1982).

10. Appendix B

Uniqueness of the solution

The solution U^* for the problem (6.1) will be unique if the functional J is strictly convex, i.e., if

$$J(\lambda U + (1 - \lambda) V) < \lambda J(U) + (1 - \lambda) J(V), \quad \lambda \in [0, 1] \tag{B.1}$$

A geometric interpretation of eq. (B.1) is that the graph of J is under any chord between U and V.

If J is only convex, i.e., the inequality in (B.1) is not strict, and then the minimum is not necessarily unique and the functional J may have no minima, one or several minima or a continuum of minima. J may also have stationary points which are not minima, but rather maxima or saddle points.

A difficulty encountered with non-convex functionals is that there is no global characterization of the absolute minimum and only a local analysis can be carried out.

If J has a second derivative, then J will be convex if its Hessian matrix H is positive semi-definite and J will be strictly convex if H is positive definite.

By definition, the functional J depends upon the observations, therefore, computing the Hessian of J where J is defined by (6.1) permits one to estimate the link between the observations and the retrieved fields.

B.1 The linear case:

In the linear case, the model equation may be written as:

$$\frac{dX}{dt} = AX \tag{B.1}$$

$$X(0) = U, \tag{B.2}$$

where A is a $n \times n$ constant matrix.

An explicit solution of (B.1)–(B.2) is

$$X(t) = \mathrm{e}^{At} U, \tag{B.3}$$

where the exponential matrix e^{A} is defined by

$$\mathrm{e}^{A} = I + A + \frac{A^2}{2!} + \cdots + \frac{A^n}{n!} + \cdots$$

The adjoint system is

$$\frac{\mathrm{d}P}{\mathrm{d}t} + A^t \cdot P = C^t(CU\mathrm{e}^{At} - X_{\mathrm{obs}}), \tag{B.4}$$

$$P(T) = 0. \tag{B.5}$$

The explicit solution of (B.4)–(B.5) is

$$P(t) = \int_T^t \mathrm{e}^{-A^t(t-s)} C^t (C\mathrm{e}^{As} U - X_{\mathrm{obs}})\, \mathrm{d}s. \tag{B.6}$$

Using (B.6) and (2.15) we obtain

$$\nabla J(U) = P(0)$$

$$= \int_T^0 e^{A^t s} C^t (C\mathrm{e}^{As} \cdot U - X_{\mathrm{obs}})\, \mathrm{d}s. \tag{B.7}$$

From (B.7) the Hessian matrix H is seen by inspection to be

$$H = \int_T^0 e^{A^t s} C^t C \mathrm{e}^{As}\, \mathrm{d}s, \tag{B.8}$$

where H is a symmetric matrix ($H^t = H$). We see from (B.8) that the Hessian H is independent of the observation itself, depending only on the model (represented by A) and on the operator C mapping the meteorological variables into the space of observations.

Because J is quadratic with respect to U it may be written as

$$J(U) = \tfrac{1}{2} U^t H U.$$

J, being a summed up semi-norm, is always positive and consequently the matrix H is positive. For eq. (6.1) to have a unique solution, one needs to verify that H is positive definite.

Since the condition of positive definiteness is not very convenient to verify, another more convenient formalism may be the following:

Theorem: Let $\mathscr{A}$ be the matrix $[C^t, A^t C^t, \ldots, (A^t)^{n-1} C^t]$. If rank $\mathscr{A} = n$, then H is positive definite and there exists a unique optimal initial condition.

Proof: H is positive definite if and only if

$$X^t H X = 0, \qquad \text{i.e., if and only if} \quad X = 0,$$

but

$$X^t H X = \int_T^0 X^t \mathrm{e}^{A^t s} C^t C \mathrm{e}^{As} X\, \mathrm{d}s$$

$$= \int_T^0 \| C\mathrm{e}^{As} X \|^2\, \mathrm{d}s. \tag{B.9}$$

This expression is equal to zero if and only if

$$C\mathrm{e}^{As} X = 0 \qquad \forall s \in [0, T] \tag{B.10}$$

and by differentiating k times with respect to s, we get

$$CA^k \mathrm{e}^{As} X = 0 \qquad \forall s \in [0, T]. \tag{B.11}$$

For $s = 0$ we get

$$CA^k X = 0 \qquad \forall k > 0. \tag{B.12}$$

By the Cayley-Hamilton theorem, there exist constants $k_0, k_1, \ldots, k_{n-1}$, such that

$$A^n = k_0 I + k_1 A + k_2 A^2 + \cdots + k_{n-1} A^{n-1},$$

and therefore if (B.12) is verified for $k \leqslant n-1$, it will be verified for $k \geqslant n$, i.e.,

$$\operatorname{rank}[C^t, A^t C^t, \ldots, (A^t)^{n-1} C^t]$$

$$= \operatorname{rank} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = r.$$

If $r = n$, then the null space of the linear mapping associated with $\mathscr{A}$ is reduced to 0 and the Hessian H will be positive definite.

The condition of uniqueness of the initial condition is thus transformed into an algebraic condition on the model and the transformation from the

variables of the model to the observed variables. The condition stated above related to the rank of $\mathscr{A}$ is very similar to a condition of observability for the equations (B.1)–(B.2) and a condition of controllability for the adjoint system (B.4)–(B.5) in the framework of optimal control theory.

If only the meteorological variables are observed, the rank of H will depend on the number of observations. The quality of the retrieved fields will depend not only on the number of observations but also on their location.

Similar results may also be obtained if we consider a temporally discretized model.

Example 1: Let us consider a continuous linear advection-diffusion model defined on [0, 1] by

$$\frac{\partial X}{\partial t}+\alpha\frac{\partial X}{\partial x}+\beta\frac{\partial^2 X}{\partial x^2}=0 \qquad \alpha,\beta>0 \tag{B.13}$$

$$X(0,t)=a(t) \tag{B.14}$$

$$X(1,t)=b(t) \qquad t>0 \tag{B.15}$$

$$X(x,0)=U(x), \tag{B.16}$$

where a and b are given.

After a standard spatial discretization on a regular grid with n grid points, the equation (B.13) may be written as

$$\frac{\mathrm{d}X}{\mathrm{d}t}=AX$$

$$X(0)=U,$$

where X is also the discretized variable and the matrix A is given by

$$A=\begin{pmatrix} -2\alpha & \alpha+\beta & & & \\ \alpha-\beta & -2\alpha & \alpha+\beta & & \\ & & & \alpha-\beta & -2\alpha \end{pmatrix}$$

and rank $A=n$.

If there is an observation of X at each grid point then C is the identity matrix and rank $C=n$, therefore rank $\mathscr{A}=n$ and there is a unique optimal initial condition.

If only one out of every two grid points is observed then the dimension of matrix C is $(n/2, n)$ (n even)

$$C=\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & & 0 \\ 0 & 0 & 1 & 0 & 0 & & 0 \\ 0 & 0 & 0 & 0 & 1 & & 0 \\ & & & & & & \\ 0 & & & & & & 1 \end{pmatrix}$$

and rank $C=n/2$.

The matrix $A'C'$ assumes the following form

$$A'C'=\begin{pmatrix} -2\alpha & 0 & 0 & 0 & 0 \\ \alpha+\beta & \alpha-\beta & 0 & 0 & 0 \\ 0 & -2\alpha & 0 & 0 & 0 \\ & \alpha+\beta & \alpha-\beta & 0 & 0 \\ & 0 & -2\alpha & 0 & 0 \\ & & \alpha+\beta & \alpha-\beta & 0 \\ & & & -2\alpha & 0 \\ & & & \alpha+\beta & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The first $2k-3$ elements of the kth column of $A'C'$ are null, while the first $2k-2$ elements of the kth column of C' are 0. Therefore the matrix $[C', A'C']$ is an $n\times n$ matrix and its columns may be reordered in such a way that we obtain a lower triangular matrix whose diagonal elements are nonzero, therefore its rank is equal to n and there is a unique optimal initial condition with respect to the observations.

If there are less than $n/2$ points of observations then it will be necessary to compute the successive powers of the matrix A. This can be done if one observes that A may be written as

$$A=-2\alpha I+(\alpha-\beta)L+(\alpha+\beta)L',$$

where

$$L=\begin{pmatrix} 0 & & & & & \\ 1 & 0 & & & & \\ & 1 & 0 & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & \cdot & \cdot \\ & & & & 1 & 0 \end{pmatrix}.$$

Example 2: In this example, we will consider the linearized shallow water equations discretized in space using a standard centered finite-difference scheme on a square domain using a step size $\Delta x = \Delta y = h$ with N^2 grid points, i.e.,

$$\frac{\partial u_{i,j}}{\partial t} + \frac{\bar{u}_{i,j}}{h}(u_{i+1,j} - u_{i-1,j}) + \frac{\bar{v}_{i,j}}{h}(u_{i,j+1} - u_{i,j-1}) - f v_{i,j} + \frac{1}{h}(\phi_{i+1,j} - \phi_{i-1,j}) = 0, \quad \text{(B.17)}$$

$$\frac{\partial v_{i,j}}{\partial t} + \frac{\bar{u}_{i,j}}{h}(v_{i+1,j} - v_{i-1,j}) + \frac{\bar{v}_{i,j}}{h}(v_{i,j+1} - v_{i,j-1}) + f u_{i,j} + \frac{1}{h}(\phi_{i,j+1} - \phi_{i,j-1}) = 0 \quad \text{(B.18)}$$

$$\frac{\partial \phi_{i,j}}{\partial t} + \frac{\bar{u}_{i,j}}{h}(\phi_{i+1,j} - \phi_{i-1,j}) + \frac{\bar{v}_{i,j}}{h}(\phi_{i,j+1} - \phi_{i,j-1}) + \frac{\bar{\phi}_{i,j}}{h}(u_{i+1,j} - u_{i-1,j} + v_{i,j+1} - v_{i,j-1}) = 0, \quad \text{(B.19)}$$

where $\bar{u}$, $\bar{v}$ and $\bar{\phi}$ are given and $u_{i,j}$ is the approximate value of u at grid point $M = (ih, jh)$.

Let $X = (u, v, \phi)$ be the vector of grid point values, then the system (B.17)–(B.19) can be written as

$$\frac{dX}{dt} = AX,$$

where A is the $3N \times 3N$ matrix assuming the form written as

$$A = \begin{pmatrix} A_{uu} & A_{vv} & A_{u\phi} \\ A_{vu} & A_{vv} & A_{v\phi} \\ A_{\phi u} & A_{\phi v} & A_{\phi\phi} \end{pmatrix}.$$

A is divided into 9 $N \times N$ square submatrices, where A_{uu}, A_{vv} and $A_{\phi\phi}$ are tetradiagonal, A_{uv} and A_{vu} are diagonal with the Coriolis parameter on the diagonal. Other matrices are bidiagonal. According to Theorem 2, in order to obtain a unique solution, the rank of the matrix

$$\mathcal{A} = [C^t, A^t C^t, ..., (A^t)^{n-1} C^t]$$

must be equal to $3N$. As above if u, v and ϕ are observed at each grid point then $C = I$ and we will have a unique solution.

If only two components of the fields, u and v for instance, are observed at all grid points, then C is a $2N \times 3N$ matrix assuming the form

$$C = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \end{pmatrix},$$

where I is the $N \times N$ identity matrix. In this case $\text{rank}(C^t) = 2N$, therefore the matrix $A^t C^t$ must be computed giving

$$A^t C^t = \begin{pmatrix} A_{uu} & A_{vu} \\ A_{uv} & A_{vv} \\ A_{u\phi} & A_{v\phi} \end{pmatrix}.$$

Due to the bidiagonal structure of $A_{u\phi}$ and $A_{v\phi}$, it is possible to extract N columns of $A^t C^t$ independent of the $2N$ columns of C^t, and therefore the uniqueness of the optimal initial condition with respect to the observations is guaranteed.

A similar result may be obtained in the case where only one of the three components of X is observed. But this result may be obtained only after computing the matrix $(A^t)^2 C^t$.

This analysis has been carried out only for the simplest linear version of the shallow water equations model. Using formal calculus a similar analysis might be carried out with a more complicated version of the shallow water model and a more realistic distribution of observation points.

In summary, the results given in this appendix can be used as tools to derive, for particular cases, sufficient conditions required to retrieve dynamical fields from observations when optimal control methods are carried out. Furthermore, if a criteria on the quality of the retrieval is given by the condition number of the Hessian matrix, this approach can be used to carry out an analysis on the optimal location of sensors.

B.2 The nonlinear case

In the nonlinear case J is no longer quadratic with respect to the control variable U. If the model is nonlinear, no global results may be obtained. Because J is bounded from below we know that there exists at least one local minimum U^* which is characterized by $\nabla J(U^*) = 0$.

A sufficient condition for U^* to be a unique

local minimum is that $H(U^*)$, the Hessian at U^*, be positive definite. In Appendix C we presented a way to calculate the Hessian matrix by integrating a second order adjoint model. One time integration of such a model will yield one column of the Hessian, or values of the Hessian multiplied by a vector. Since the dimension of the control variable in variational data assimilation is very large (up to 10^6 or more), it is still too expensive to calculate the full Hessian and then to examine the properties of that Hessian matrix. However it is possible to obtain an estimate of the spectral radius and the corresponding condition number of the Hessian by using an iterative power method or a Rayleigh quotient method which only requires the value of Hessian multiplied by a vector (Golub and Van Loan, 1989).

For example the sequence defined by

$$X^{(k+1)} = \frac{HX^{(k)}}{\|X^{(k)}\|}$$

converges to the maximum eigenvector $X_{\max}$ associated with the largest eigenvalue $\lambda_{\max}$ of the matrix H if this eigenvalue is simple. Furthermore

$$\frac{\|X^{(k+1)}\|}{\|X^{(k)}\|} \to |\lambda_{\max}|.$$

To estimate the smallest eigenvalue of the matrix H, it is sufficient to make a shift and to consider the matrix

$$K = \lambda I - H, \qquad \text{where} \quad \lambda > \lambda_{\max};$$

the largest eigenvalue of K is $\lambda - \lambda_{\min}$ (i.e., to use the shifted power iteration).

One will also be able to calculate several large and small eigenvalues of the Hessian by integrating the second order adjoint model by a deflation method (see Golub and Van Loan, 1989).

11. Appendix C

Second-order adjoint model

For the sake of simplicity, we will assume that the model has been discretized with respect to the space variable and that the state of the atmosphere may be described in the time interval from 0 to T, as the solution of an ordinary differential equation

$$\frac{dX}{dt} = F(X) \tag{C.1}$$

$$X(0) = U, \tag{C.2}$$

where X represents the state of the atmosphere and belongs to the n-dimensional space $\mathscr{X}$. F is a nonlinear mapping from $\mathscr{X}$ to $\mathscr{X}$. We assume that F is such that with the initial condition (C.2) at time 0, the model described by equation (2.1) has a unique solution between times 0 and T.

The adjoint equations corresponding to (C.1)–(C.2) are written as

$$\frac{dP}{dt} + \left[\frac{\partial F}{\partial X}\right]^t P = C^t(CX - X_{\text{obs}}), \tag{C.3}$$

$$P(T) = 0, \tag{C.4}$$

where P represents the adjoint variable. Integrations of the adjoint system (C.3)–(C.4) results in the gradient of the cost function

$$\nabla J(U) = -P(0). \tag{C.5}$$

Applying a perturbation K on U we obtain from (C.1)–(C.2)

$$\frac{d\hat{X}}{dt} = \left[\frac{\partial F}{\partial X}\right] \cdot \hat{X} \tag{C.6}$$

$$\hat{X}(0) = K, \tag{C.7}$$

where $\hat{X}$ is the perturbation of X. For the adjoint system we obtain from (C.3)–(C.5)

$$\frac{d\hat{P}}{dt} + \left[\frac{\partial^2 F}{\partial X^2} \cdot \hat{X}\right]^t \cdot P + \left[\frac{\partial F}{\partial X}\right]^t \cdot \hat{P} = C^t C\hat{X}, \tag{C.8}$$

$$\hat{P}(T) = 0, \tag{C.9}$$

$$\hat{\nabla} J(U, K) = -\hat{P}(0). \tag{C.10}$$

The Hessian is obtained by exhibiting the linear dependence of $P(0)$ with respect to K.

Let us introduce Q and R, the second order adjoint variables, which will be defined explicitly in the following.

We take the inner product of eq. (C.6) with Q

and of eq. (C.8) with R, integrate from time 0 to T, and add the two results together to get

$$\int_0^T \left\{ \left\langle \frac{d\hat{X}}{dt}, Q \right\rangle + \left\langle \frac{d\hat{P}}{dt}, R \right\rangle \right\} dt$$
$$= \int_0^T \left\{ \left\langle \left[\frac{\partial F}{\partial X}\right] \cdot \hat{X}, Q \right\rangle - \left\langle \left[\frac{\partial^2 F}{\partial X^2} \cdot \hat{X}\right]^t \cdot P, R \right\rangle \right.$$
$$\left. - \left\langle \left[\frac{\partial F}{\partial X}\right]^t \cdot \hat{P}, R \right\rangle + \langle C^t C \hat{X}, R \rangle \right\} dt \quad \text{(C.11)}$$

The left-hand side of (C.11) is integrated by parts while the operators in the right-hand side are transposed giving:

$$\langle \hat{X}(T), Q(T) \rangle - \langle \hat{X}(0), Q(0) \rangle$$
$$+ \langle \hat{P}(T), R(T) \rangle - \langle \hat{P}(0), R(0) \rangle$$
$$- \int_0^T \left\langle \hat{X}, \frac{dQ}{dt} \right\rangle dt - \int_0^T \left\langle \hat{P}, \frac{dR}{dt} \right\rangle dt$$
$$= \int_0^T \left\langle \hat{X}, \left[\frac{\partial F}{\partial X}\right]^t \cdot Q \right\rangle dt$$
$$- \int_0^T \left\langle \hat{X}, \left[\frac{\partial^2 F}{\partial X^2} \cdot R\right]^t \cdot P \right\rangle dt$$
$$- \int_0^T \left\langle \hat{P}, \left[\frac{\partial F}{\partial X}\right] \cdot R \right\rangle dt$$
$$+ \int_0^T \langle \hat{X}, C^t C R \rangle \, dt. \quad \text{(C.12)}$$

Therefore if the variables Q and R are defined as the solution of the following differential systems:

$$\frac{dQ}{dt} + \left[\frac{\partial F}{\partial X}\right]^t \cdot Q - \left[\frac{\partial^2 F}{\partial X^2} \cdot R\right]^t \cdot P$$
$$+ C^t C R = 0, \quad \text{(C.13)}$$

$$\frac{dR}{dt} = \left[\frac{\partial F}{\partial X}\right] \cdot R, \quad \text{(C.14)}$$

$$Q(T) = 0, \quad \text{(C.15)}$$

then using (C.7) and (C.9) from (C.12) one obtains

$$-\langle \hat{P}(0), R(0) \rangle = \langle K, Q(0) \rangle \quad \text{(C.16)}$$

Now, using (C.10) and the definition of the Gateaux derivative, we have

$$\langle H \cdot K, R(0) \rangle = \langle K, Q(0) \rangle, \quad \text{(C.17)}$$

and therefore, since H is symmetric, we have

$$H \cdot R(0) = Q(0). \quad \text{(C.18)}$$

To obtain the full Hessian matrix H, we have to integrate n times the differential system (C.13)–(C.15) with the condition on $R(0)$ given as:

$$R(0) = e_i, \qquad i = 1, \ldots, n, \quad \text{(C.19)}$$

where e_i is the vector of the canonical basis of $\mathscr{R}^n$, then from (C.17) the n vectors $Q(0)$ obtained from the integration are the rows of the matrix H.

The condition for the existence of a unique optimal initial condition is that H is positive definite. A necessary and sufficient condition for H to be positive definite is that its null space is reduced to zero. This will be the case if the n vectors $Q(0)$ obtained by solving (C.13)–(C.15), (C.18) are linearly independent. This is a problem of controllability for the system (C.13)–(C.14).

REFERENCES

Axelsson, O. and Barker, V. A. 1984. *Finite element solution of boundary value problems: theory and computation*, in Computer Science and Applied Mathematics Series. Academic Press, Inc., 111 Fifth Avenue, New York, New York 10003, 437 pp.

Ballish, B. 1981. A simple test of the initialization of gravity modes. *Mon. Wea. Rev. 109*, 1318–1321.

Bertsekas, D. P. 1982. *Constrained optimization and Lagrange multiplier methods*, in Computer Science and Applied Mathematics, Series (ed. Werner Rheinboldt). Academic Press, Inc., 111 Fifth Avenue, New York, New York 10003, 395 pp.

Browning, G. and Kreiss, H. 1982. Initialization of the shallow water equations with open boundaries by the bounded derivative method. *Tellus 34*, 334–351.

Courtier, P. and Talagrand, O. 1987. Variational assimilation of meteorological observations with the adjoint equation, Part II. Numerical results. *Q. J. R. Meteorol. Soc. 113*, 1329–1347.

Courtier, P. and Talagrand, O. 1990. Variational

assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus 42A*, 531–549.

Derber, J. C. 1985. The variational 4-D assimilation of analyses using filtered models as constraints. Ph. D. Thesis, University of Wisconsin-Madison, 141 pp.

Fiacco, A. V. and McCormick, G. P. 1968. *Nonlinear programming: sequential unconstrained minimization techniques.* John Wiley and Sons, New York and Toronto, 210 pp.

Gill, P. E., Murray, W. and Wright, M. H. 1981. *Practical optimization.* United States Edition published by Academic Press, INC., Orlando, Florida 32887, 401 pp.

Golub, H. G. and Van Loan, C. 1989. *Matrix computation*, 2nd Edition, Johns Hopkins Series in the Mathematical Sciences, The Johns Hopkins University Press, 642 pp.

Grammeltvedt, A. 1969. A survey of finite-difference schemes for the primitive equations for a barotropic fluid. *Mon. Wea. Rev. 97*, 387–404.

Hestenes, M. R. 1969. Multiplier and gradient methods. *J. Opt. Th. Applics. 4*, 303–320.

Juvanon du Vachat, R. 1986. A general formulation of normal modes for limited-area models: Applications to initialization. *Mon. Wea. Rev. 114*, 2478–2487.

Kontarev, G. 1980. *The adjoint equation technique applied to meteorological problems.* ECMWF Tech. Rep. 21, Reading, UK, 1–21.

Kreiss, H. O. 1980. Problems with different time scales for ordinary differential equations. *Commun. Pure Appl. Math. 33*, 399–437.

LeDimet, F. X. 1988. *Determination of the adjoint of a numerical weather prediction model.* Tech. Rep. FSU-SCRI-88-79, Florida State University, Tallahassee, Florida 32306–4052, 22 pp.

LeDimet, F. X. and Talagrand, O. 1986. Variational algorithms for analysis and assimilation of Meteorological Observations: Theoretical Aspects. *Tellus 38A*, 97–110.

Liu, D. C. and Nocedal, J. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming 45*, 503–528.

Lootsma, F. A. 1969. Hessian matrices of penalty functions for solving constrained optimization problems. *Philips Res. Repts 24*, 322–331.

Luenberger, D. G. 1984. *Linear and nonlinear programming.* Addison-Wesley Publishing Company, Second Edition, 491 pp.

Machenhauer, B. 1977. On the dynamics of gravity oscillations in a shallow water model, with applications to normal mode initialization. *Contrib. Atmos. Phys. 50*, 253–271.

Murray, W. 1969. An algorithm for constrained minimization. *Symposium of the Institute of Mathematics and Its Applications.* (ed. R. Fletcher). London, Newrok, Academic Press, 111 Fifth Avenue, New York, New York 10003, 354 pp.

Murray, W. 1971. Analytical expressions for the eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions. *J. Opt. Th. Applics. 7*, 189–196.

Navon, I. M. and De Villiers, R. 1983. Combined penalty multiplier optimization methods to enforce integral invariant conservation. *Mon. Wea. Rev. 111*, 1228–1243.

Nocedal, J. 1980. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation 35*, 773–782.

Powell, M. J. D. 1977. Restart procedures for the conjugate gradient method. *Mathematical Programming 12*, 241–254.

Rao, S. S. 1983. *Optimization: theory and applications.* Published in the Western Hemisphere by Halsted Press, a Division of John Wiley & Sons, INC., New York, 747 pp.

Sasaki, Y. 1970. Some basic formalisms in numerical variational analysis. *Mon. Wea. Rev. 98*, 875–883.

Semazzi, F. and Navon, I. M. 1986. A comparison of the bounded derivative and the normal mode initialization methods using real data. *Mon. Wea. Rev. 114*, 2106–2121.

Talagrand, O. and Courtier, P. 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation, Part I. Theory. *Q. J. R. Meteorol. Soc. 113*, 1311–1328.

Temperton, C. 1988. Implicit normal mode initialization. *Mon. Wea. Rev. 116*, 1013–1031.

Temperton, C. 1989. Implicit normal mode initialization for spectral models. *Mon. Wea. Rev. 117*, 436–451.

Thacker, W. C. 1988. Fitting models to inadequate data by enforcing spacial and temporal smoothness. *J. Geophys. Res. 93*, 10655–10665.

Thacker, W. C. and Long, R. 1988. Fitting dynamics to data. *J. Geophys. Res. 93*, 1227–1240.

Zou, X., Navon, I. M., Berger, M., Phua, Paul K. H. and Schlick, T. 1991. *Numerical experience with limited-memory quasi-Newton and truncated Newton methods.* Tech. Rep., FSU-SCRI-90-167, available at SCRI, Florida State University, Tallahassee, FL 32306–4052, 44 pp.