Chapter

Variational Data Assimilation: Theory and Applications[1]

Xiaolei Zou

Supercomputer Computations Research Institute
Florida State University
Tallahassee, FL 32306-4052, USA

I. M. Navon

Dept. of Mathematics and Supercomputer Computations
Research Institute, Florida State University
Tallahassee,FL 32306, USA

Abstract

Variational data assimilation is performed using either a T40 18 layer version of the National Meteorological Center spectral model or a limited-area shallow-water equations model, with operationally analyzed fields as well as simulated datasets. Issues of incomplete observations, control of gravitational oscillations, and inclusion of "on-off" physical processes are addressed in the framework of Variational data assimilation.

Keywords

Variational data assimilation, incomplete observations, gravitational oscillations, Adjoint of "on-off" physical processes.

[1]To appear in *Environmental Modeling—Volume II*, P. Zannetti, editor.

1 Introduction

Numerical modeling has now become a major tool for the study of atmospheric dynamics. These models are constructed based on the physical laws that govern the temporal evolution of the flow and that express the mass, momentum, and energy of the fluid. In numerical form, they appear as a computer code that uses as independent variables (or input), the initial condition from which the model must be integrated and many parameters that define the physical and numerical conditions of the time integration. The dependent variables (or output), of models consist of the temporal sequence of meteorological fields produced by the integration, and also of the many derived quantities that one may wish to compute from that sequence of meteorological fields.

In many real-life situations, meteorologists are led to consider the following problems:

a) To determine values of part of the input parameters, such as initial conditions and/or model parameters, corresponding to given values of output parameters (i.e., observations). Such problems will normally be solved by optimal control of partial differential equations as unconstrained optimization problems. A prescribed scalar function, the cost function, measuring the misfit between the output parameters and the observed quantities will be defined and the optimal values of the input parameters will be found out using various large-scale unconstrained minimization algorithms which require information of function and gradient with respect to the input parameters. Two typical examples of such problems are variational data assimilation and parameter estimation.

b) To determine to which input parameters are some quantities calculated from the output parameters most sensitive. In some situations, one will be interested in the sensitivities of one forecast aspect (called the response) with respect to a large number of input parameters.

In all these situations one is required to determine the gradient of a function or a response with respect to a large number ($\sim 10^6$) of input parameters. The method of adjoint model equations provides an efficient way of solving such problems, which would of course be totally prohibitive if one were to use the only other method that could be considered for numerically determining the gradient: namely the finite-difference scheme.

In this chapter we will discuss variational 4-dimensional data assimilation. For parameter estimation see Smedstad and O'Brien (1991) and Zou, *et al.* (1992b) and for sensitivity analysis using the adjoint method, see Cacuci (1981a,b), Hall, *et al* (1982), Hall and Cacuci (1983) and Zou, *et al.* (1993d) and the references therein.

The complete and accurate specification of the 4-dimensional structure of the atmospheric state is a basic problem in meteorology. The conventional observational network alone is not capable of providing this complete description of the atmospheric state, since the observational density is not high enough in either space or time. For this reason it is desirable to include additional information (i.e., statistical or dynamical information) into the data assimilation system. There are many techniques used in data assimilation. Detailed reviews of data assimilation are given in Navon (1986), LeDimet and Navon (1989),

Ghil and Mananotte-Rizzoli (1991) and Daley (1991). Recently, considerable attention has been focused on variational data assimilation, which fits the data as well as possible in a least-square sense and satisfies the dynamics of the forecast model by adjusting only the initial fields. It originates in the control theory of partial differential equations and produces high-quality analyses of the atmosphere for the production of reliable forecasts. The basis of variational data assimilation is to use all the available information, observations which are distributed more or less regularly in both time and space and vary greatly in their nature and accuracy as well as numerical prediction model itself which describes physical conservation laws, in order to produce the best possible initial state in a least-squares norm optimal sense. The goal aimed at can be described as optimizing a cost function which measures the distance between model states and observations plus some prior knowledge. For the underlying optimal control theory see Lions (1971) and for a general description of its application to meteorology, see LeDimet and Talagrand (1986). Since then, a sizable amount of research has been carried out in this area in meteorology such as by Derber (1985), Lewis and Derber (1987), Courtier and Talagrand (1987), Talagrand and Courtier (1987), Zou, *et al.* (1992a), Thépaut and Courtier (1992), Navon, *et al.* (1992), Zou, *et al.* (1993a,c) and Zupanski (1992).

Apart from its conceptual simplicity and its adaptability, variational data assimilation possesses many advantages. First, it can be shown that it is equivalent to the optimal Kalman Filter in the linear context and under the hypothesis of a perfect model (Daley, 1991; Lorenc, 1986; Rabier *et al.*, 1992). In the nonlinear case and if the model is still assumed to be perfect, results from variational data assimilation or the extended Kalman Filter are expected to be very close if the tangent-linear hypothesis is valid for the duration of the assimilation period. Comparison of variational data assimilation with simplified classic sequential assimilation by Rabier, *et al.* (1992) shows a significantly better performance of variational data assimilation scheme, as it yields consistently better analyses than its sequential counterpart and as the error grows in the subsequent forecasts at a lesser rate.

The model forecast error can also be considered in the context of variational data assimilation by introducing an additional term in the model equation (Derber, 1989, Zupanski, 1992). The initialization problem, aiming at establishing balanced initial conditions without high frequency gravity wave noise, can also be incorporated into variational data assimilation by adding a penalty term into the cost function (Thépaut and Courtier, 1992; Zou, *et al.*, 1992a; Zou, *et al.*, 1993a).

The main limitation of variational data assimilation is that it remains rather computationally expensive and extensive testing has to be carried out before its possible operational implementation.

This chapter presents the main aspects underlying a rigorous derivation of variational data assimilation. As this chapter is intended to be self-contained, the basic formalism of the variational problem including the linkage between the nonlinear model, tangent linear model, adjoint model, cost function and the gradient of the cost function with respect to the initial conditions is briefly reviewed in Section 2. This section is divided into four main parts. Section 2.1 presents variational data assimilation theory for nonlinear systems with a standard cost function. The system of nonlinear operator equations and the associated

cost function, gradient calculation, and adjoint are introduced and described in Section 2.1; altogether, they are intended to be sufficiently general to include the mathematical representation of a large number of problems arising in a wide variety of fields. Section 2.2 centers on practical coding of adjoint models and their verification. State-of-the-art large-scale unconstrained minimization schemes are described and summarized in Section 2.3, where different limited-memory quasi-Newton methods and truncated Newton methods as well as their relative performance are briefly reviewed. As an illustrative example, the formalisms in Sections 2.1-2.3 are applied in Section 2.4 to perform a twin experiment of variational data assimilation using a shallow-water equations model. Theory and numerical results concerning the impact of incomplete observations on the uniqueness of the solution and the convergence rate of the minimization processes are presented in Section 3. Methods aimed at the specification of suitably balanced optimal initial fields, i.e., applying different penalty terms in the framework of variational data assimilation to control gravitational oscillations, are provided in Section 4. A sequential penalty method is applied to a shallow-water equations model in Section 4.1 and different simple penalty terms constraining directly the time tendency of model variables are applied to a 3-dimensional adiabatic version of the operational primitive equations model at the National Meteorological Center (NMC) in Section 4.2. The computational aspects and the difference in the retrieved initial fields due to the inclusion of "threshold" processes, i.e., large-scale precipitation and deep cumulus convection, are presented in Section 5. These "threshold" processes represent locally nondifferentiable phenomena. Finally, Section 6 summarizes and highlights the main points underlying variational data assimilation and discusses the potential of using adjoint techniques to further extend the scope of this theory for meteorological applications.

2 Basic Framework of Variational Data Assimilation

2.1 Variational Data Assimilation Formalism

Cost function

The objective of variational 4-D data assimilation is to find the solution to a numerical forecast model which will best fit a series of observational fields distributed over some space and time interval. One possible measure of the lack of fit, the cost function J, consists of a weighted least square fit of the model forecast to the observations:

$$J(\mathbf{x}(t_0)) = \frac{1}{2}\sum_{r=0}^{R}\Big(\mathbf{C}\mathbf{x}(t_r) - \mathbf{x}^{obs}(t_r)\Big)^T \mathbf{W}(t_r)\Big(\mathbf{C}\mathbf{x}(t_r) - \mathbf{x}^{obs}(t_r)\Big) \tag{1}$$

where t_r represents the time when an observation occurs in the assimilation interval $[t_0, t_a]$, R is the total number of time levels in the assimilation interval when observations are available, $\mathbf{x}(t_r)$ is the N-component vector in space R_N containing values of model variables at time t_r, $\mathbf{x}^{obs}(t_r)$ is the M-component ($M \leq N$) vector in space R_M containing

values of observations at time t_r, $\mathbf{C}$ is a projection operator from the space R_N to the space R_M, and $\mathbf{W}(t_r)$ is an $M \times M$ weighting matrix. If $\mathbf{W}(t_r)$ are taken as unit matrices, J is defined as an $\mathcal{L}_2$ norm of $\mathbf{Cx} - \mathbf{x}^{obs}$. One can also choose an energy norm to define J, which will provide a natural weight for different model variables.

The values of $\mathbf{x}(t_r)$, $r = 0, 1, \ldots, R$ are obtained by integrating a numerical model which can be written in a general form

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(t, \mathbf{x}) \tag{2}$$

from the initial state $\mathbf{x}(t_0)$. In time discretized form, (2) will assume the following general form

$$\mathbf{x}(t_0 + \Delta t) = \quad \mathbf{F}_1(\mathbf{x}(t_0)) + \mathbf{L}_1\mathbf{x}(t_0) \tag{3}$$

$$\mathbf{x}(t_r + \Delta t) = \quad \mathbf{F}(\mathbf{x}(t_r)) + \mathbf{L}(t_r)\mathbf{x}(t_r - \Delta t) + \mathbf{m}(t_r)\mathbf{x}(t_r), \tag{4}$$

$$\text{for} \quad r = 1, 2, \ldots,$$

where $\mathbf{F}_1(\mathbf{x})$ and $\mathbf{F}(\mathbf{x})$ are nonlinear operators and $\mathbf{L}_1$, $\mathbf{L}$ and $\mathbf{m}$ are linear operators.

Weighting

Weights in the cost function serve the following purposes: (i) scale the cost function J to become a non-dimensional quantity; (ii) reflect the confidence we have in the quality of the observed data, and (iii) provide also initial conditioning essential in efficient minimization. Weights are usually chosen as the inverses of the covariance matrices of the observation errors. However, these variances are difficult to specify properly and much more research work is necessary in this area. Thépaut and Courtier (1992) used an energy norm to define the cost function. The choice of an energy form defines a physical measure of atmospheric fields as well as a natural weighting for the different model variables. In the variational data assimilation experiments of Navon *et al.* (1992) with the adiabatic version of the NMC spectral model, a diagonal weighting matrix is used with $\mathbf{W}_\chi, \mathbf{W}_\psi, \mathbf{W}_T, \mathbf{W}_{\ln ps}$ and $\mathbf{W}_q$ as $\mathbf{W}$'s diagonal submatrices. These submatrices represent the weighting factors for the potential velocity (χ), stream function (ψ), temperature (T), surface pressure (p_s) and moisture (q) fields, respectively. The submatrix $\mathbf{W}_\psi$, for instance, is calculated by the following formula

$$\mathbf{W}_\psi = \frac{1}{\max_{i,j,k} |\psi^{obs}_{i,j,k}(t_0) - \psi^{obs}_{i,j,k}(t_R)|^2} \tag{5}$$

with similar expressions for the velocity potential, temperature, surface pressure, and moisture fields, i.e., the inverse of the maximum difference between the two observational fields at times t_0 and t_R. In the experiments of Zou, *et al.* (1992a) using a shallow-water equations model, a constant diagonal weight is used with values of $10^{-4}m^{-2}s^2$ and $10^{-2}m^{-2}s^2$ as the weighting coefficients for geopotential and wind fields respectively. Courtier and Talagrand (1990) used a temporal weighting in which weights given to individual observations varied linearly with time, the total sum of the weights assuming

the same value as the reciprocal of squared estimates of the statistical root-mean-square observational errors. The idea is that the model not being perfect, it cannot adjust uniformly to the whole set of observations at intermediate times. In the case of variational data assimilation, where one performs a forecast starting from the final time of the assimilation period, a better adjustment to later observations is obviously preferable, and larger weights are assigned to more recent observations in the definition of the cost function.

Gradient

To find the minimum of the cost function J, most of large-scale efficient minimization algorithms require the user to supply values of the gradient of the cost function with respect to control variables which in our case are the initial conditions.

To calculate the gradient of the cost function with respect to the initial condition, we will define a quantity J':

$$\begin{aligned} J'(\mathbf{x}(t_0)) \quad &\equiv J\left(\mathbf{x}(t_0)+\mathbf{x}'(t_0)\right)-J\left(\mathbf{x}(t_0)\right) \\ &= \sum_{r=0}^{R}\Big(\mathbf{W}(t_r)\Big(\mathbf{C}\mathbf{x}(t_r)-\mathbf{x}^{obs}(t_r)\Big)\Big)^T\mathbf{C}\mathbf{x}'(t_r)+\sum_{r=0}^{R}O(\mathbf{x}'^2(t_r)). \end{aligned} \tag{6}$$

i.e., the change in the cost function resulting from a small perturbation $\mathbf{x}'(t_0)$ about the initial conditions $\mathbf{x}(t_0)$. In limit as $\|\mathbf{x}'\|\to 0$, J' is the directional derivative in the $\mathbf{x}'(t_0)$ direction and is given by

$$J'\left(\mathbf{x}(t_0)\right)=\left(\nabla J\left(\mathbf{x}(t_0)\right)\right)^T\mathbf{x}'(t_0). \tag{7}$$

Equating (6) and (7) results in

$$\left(\nabla J\left(\mathbf{x}(t_0)\right)\right)^T\mathbf{x}'(t_0)=\sum_{r=0}^{R}\Big(\mathbf{W}(t_r)\Big(\mathbf{C}\mathbf{x}(t_r)-\mathbf{x}^{obs}(t_r)\Big)\Big)^T\mathbf{x}'(t_r). \tag{8}$$

It is clear that if $\mathbf{x}'(t_r)$ can be expressed as a function of $\mathbf{x}'(t_0)$, then the gradient of the cost function with respect to the initial conditions can be found.

Since $\mathbf{x}'(t_r)$ is the perturbation at time t_r in the forecast resulting from the initial perturbation $\mathbf{x}'(t_0)$, it can be obtained by integrating the tangent linear model, which in turn can be obtained by linearizing the nonlinear model (3)-(4):

$$\mathbf{x}'(t_0+\Delta t)= \quad \frac{\partial\mathbf{F}_1\left(\mathbf{x}(t_0)\right)\mathbf{x}'(t_0)}{\partial\mathbf{x}}+\mathbf{L}_1\mathbf{x}'(t_0) \tag{9}$$

$$\mathbf{x}'(t_r+\Delta t)= \quad \frac{\partial\mathbf{F}\left(\mathbf{x}(t_r)\right)\mathbf{x}'(t_r)}{\partial\mathbf{x}}+\mathbf{L}\mathbf{x}'(t_r-\Delta t)+\mathbf{m}(t_r)\mathbf{x}'(t_r), \tag{10}$$

$$\text{for } r=1,2,\ldots,$$

which may then be rewritten symbolically as

$$\mathbf{x}'(t_r)=\mathbf{P}_r\mathbf{x}'(t_0) \tag{11}$$

where $\mathbf{P}_r$ represents the result of applying all the operator matrices in the linear model to obtain $\mathbf{x}'(t_r)$ from $\mathbf{x}'(t_0)$.

Using the tangent linear model (11), (8) becomes

$$\left(\nabla J\left(\mathbf{x}(t_0)\right)\right)^T \mathbf{x}'(t_0) = \sum_{r=0}^{R} \left(\mathbf{W}(t_r)\left(\mathbf{C}\mathbf{x}(t_r) - \mathbf{x}^{obs}(t_r)\right)\right)^T \mathbf{C}\mathbf{P}_r\mathbf{x}'(t_0). \tag{12}$$

This implies

$$\nabla J\left(\mathbf{x}(t_0)\right) = \sum_{r=0}^{R} \mathbf{P}_r^T\mathbf{C}^T\mathbf{W}(t_r)\left(\mathbf{C}\mathbf{x}(t_r) - \mathbf{x}^{obs}(t_r)\right), \tag{13}$$

where $\mathbf{P}_r^T$, $r = 1, 2, \ldots, R$ are the corresponding adjoint operators of the linear operators $\mathbf{P}_r$, $r = 1, 2, \ldots, R$ in the tangent linear model. Therefore, the gradient of the cost function may be obtained by a single integration of the adjoint model from final time t_a to initial time t_0 of the assimilation window with zero initial conditions for the adjoint variables at time t_a while the weighted differences

$$\mathbf{C}^T\mathbf{W}(t_r)\left(\mathbf{C}\mathbf{x}(t_r) - \mathbf{x}^{obs}(t_r)\right), \quad r = R, R-1, \ldots, 0 \tag{14}$$

are inserted on the right-hand-side of the following adjoint model

$$\hat{\mathbf{x}}(t_0) = \mathbf{P}^T\hat{\mathbf{x}}(t_a), \tag{15}$$

whenever an observational time $t_r (r = R, R-1, \ldots, 0)$ is reached, where $\hat{\mathbf{x}}$ represents the adjoint variables.

2.2 Adjoint models

The discretized atmospheric models are implemented by rather long and complicated codes. The models are in permanent evolution and are constantly subject to modifications. The possibility of routinely performing work involving adjoint model integration requires that, whenever a modification is made on the direct model, a corresponding modification should also be made on the adjoint. Experience shows that even minor mistakes in the adjoint model code can render it totally useless. This requires that the adjoint code must be developed directly from the basic direct code (and not, for instance, from the partial differential equations from which the direct code is built). It also implies that the components of the adjoint code will be in one-to-one correspondence with the components of the direct code: to each subroutine of the direct code, there is a mirror subroutine of the adjoint performing the corresponding adjoint task. Since the adjoint operator $\mathbf{P}^T$ is based on the tangent linear operator $\mathbf{P}$, the above discussion implies that discrete operations in the forward nonlinear model have unique corresponding discrete operations in the tangent linear model with only one exception: all the nonlinear terms in the original model are linearized locally.

For an adiabatic primitive equations forecast model, the only nonlinear terms are those due to the quadratic terms introduced by the advection processes. Linearization of

this type of nonlinear terms is quite simple. For example, a term $u * T$ in the nonlinear model becomes two terms $u * T' + T * u'$ in the tangent linear model. However, when the nonlinear model includes physical processes, the linearization of these processes is not at all evident due to the high degree of nonlinearity and the appearance of either table lookups or iterative procedures which were employed in the direct code. For example, in the case of table lookups, the values in the tables are built for the nonlinear calculation. The original analytical formulas used to build the lookup tables are then required in order to derive the tangent linear version of the nonlinear model. When an iterative method is used to solve a highly nonlinear equation in the nonlinear model, the tangent linear version of this part can not be derived directly from the nonlinear code, i.e., by linearizing all the procedure of the iterative process. One is required to derive the analytical tangent linear equation of the relevant equation which is solved iteratively in the direct code and then write the independent linear code of this part using the same variable notations as in the direct code. These tasks are rather tedious and complicated since normally one has at his disposal the direct code but does not always possess the corresponding updated complete documentation of the direct code. A check of the correctness of the tangent linear model is crucial in these situations.

To date, adjoint models have been used, in the context of meteorology, only at the experimental level. Meteorologists write adjoint models with few basic underlying principles in mind. The first and foremost basic rule for writing the adjoint code is the matrix method, i.e., each DO loop in the direct code corresponds to a matrix/vector multiplication: $\mathbf{y}^{\text{output}} = \mathbf{A}\mathbf{x}^{\text{input}}$. The adjoint code can be developed based on the equality: $\hat{x}^{\text{output}} = \mathbf{A}^T \hat{y}^{\text{input}}$, where $\hat{x}$ represents the adjoint variables corresponding to $\mathbf{x}$. A simple example of the construction of the discrete adjoint code from the discrete linear code based on this matrix method is provided in the Appendix A of the paper by Navon, *et al.* (1992).

If we view the linear model as the result of the multiplication of a number of operator matrices:

$$\mathbf{P} = \mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_N, \tag{16}$$

where each matrix $A_i (i = 1, \cdots, N)$ represents either a subroutine or a single DO loop, then the adjoint model can be viewed as being a product of adjoint subproblems

$$\mathbf{P}^T = \mathbf{A}_N^T \mathbf{A}_{N-1}^T \cdots \mathbf{A}_1^T, \tag{17}$$

i.e., the adjoint model has to be written backward from the tangent linear model. The second rule is the rule of locality: whenever a local modification is made in the direct model, a corresponding modification must be made locally in the adjoint code. This requires transparent notations, between the direct and adjoint codes, in as far as subroutine and variable names, instruction labeling, etc. are concerned. Some examples may be found in the paper by Talagrand (1991). The third principle of building the adjoint code is the trajectory management when an IF statement is encountered in the direct code. The adjoint model will have to follow the same route of the direct model but backward in time.

The above formulated rules must of course be applied with some discernment, since there are situations where there is simply no need for them. A typical example is the

Fourier transform, which being unitary, has as adjoint its own inverse. There is therefore no need to develop the adjoint of a Fourier transform and it is sufficient in the adjoint code to invoke the inverse transform at the appropriate location.

In other words, the discrete adjoint model can be obtained directly from the discrete linear model, which in turn is obtained from the original nonlinear model by linearization around a model state. The linearization is made locally in the code and the structure of the linear model is exactly the same as that of the nonlinear model. These rules simplify not only the complexity of constructing the adjoint model but also avoid the inconsistency generally arising from the derivation of the adjoint equations in analytic form followed by the discrete approximation since there is no commutativity between adjoint and discretization.

The correctness of the adjoint can be checked at any level of the code of the development of the discrete adjoint model by applying the following identity

$$(\mathbf{Ax})^T(\mathbf{Ax}) = \mathbf{x}^T\left(\mathbf{A}^T(\mathbf{Ax})\right), \tag{18}$$

where $\mathbf{A}$ represents either a single DO loop or one/several subroutines and $\mathbf{x}$ represents the input of $\mathbf{A}$.

2.3 Large-scale minimization algorithms

Limited-memory quasi-Newton and truncated Newton methods

Once the values of the cost function J (obtained by integrating the nonlinear model forward) and its gradient with respect to the control variables (obtained by integrating the adjoint model backward in time) are found, the optimal initial condition $\mathbf{x}^*(t_0)$ of minimizing J can be obtained by any unconstrained minimization algorithms which require the user to supply the values of the cost function and its gradient. Since the model already taxes the capability of the largest available supercomputers and the optimization of the initial conditions must conform to the operational requirements of timeliness, the choice of a robust and efficient minimization algorithm is crucial.

Limited-memory quasi-Newton (Shanno and Phua, 1980; Buckley, 1978; Gill and Murray, 1979; Buckley and Lenir, 1983, 1985; Buckley, 1989; Liu and Nocedal, 1989; Nocedal, 1980) and truncated Newton (Nash, 1984a,b, 1985; Schlick and Fogelson, 1992a,b) methods represent two classes of algorithms which are attractive for large-scale problems because of their modest storage requirements. They use a low and adjustable amount of storage and require the function and gradient values at each iteration. Limited-memory quasi-Newton methods can be viewed as extensions of conjugate-gradient methods in which the addition of some modest storage serves to accelerate the convergence rate. Truncated Newton methods attempt to retain the rapid convergence rate of classical Newton methods while economizing storage and computational requirements so as to become feasible for large-scale applications.

Limited-memory quasi-Newton algorithms have the following basic structure for minimizing $J(\mathbf{x})$, $\mathbf{x} \in \mathcal{R}^N$:

1) Choose an initial guess $\mathbf{x}_0$, and a positive definite initial approximation to the inverse Hessian matrix $\mathbf{H}_0$ (which may be chosen as the identity matrix).

2) Compute

$$\mathbf{g}_0 = \mathbf{g}(\mathbf{x}_0) = \nabla J(\mathbf{x}_0), \tag{19}$$

and set

$$\mathbf{d}_0 = -\mathbf{H}_0\mathbf{g}_0. \tag{20}$$

3) For $k = 0, 1, \ldots$, set

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k, \tag{21}$$

where α_k is the step-size obtained by a safeguarded procedure.

4) Compute

$$\mathbf{g}_{k+1} = \nabla J(\mathbf{x}_{k+1}) \tag{22}$$

5) Generate a new search direction, $\mathbf{d}_{k+1}$, by setting

$$\mathbf{d}_{k+1} = -\mathbf{H}_{k+1}\mathbf{g}_{k+1}, \tag{23}$$

6) Check for convergence: If

$$\|\mathbf{g}_{k+1}\| \leq \epsilon \max\{1, \|\mathbf{x}_{k+1}\|\}, \tag{24}$$

stop, where $\epsilon = 10^{-5}$. Otherwise continue from step 3.

In practice, limited-memory quasi-Newton methods update formula (Liu and Nocedal, 1989) forms an approximate inverse Hessian from $\mathbf{H}_0$ and k pairs of vectors $(\mathbf{q}_i, \mathbf{p}_i)$, where $\mathbf{q}_i = \mathbf{g}_{i+1} - \mathbf{g}_i$ and $\mathbf{p}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$ for $i \geq 0$. Since $\mathbf{H}_0$ is generally taken to be the identity matrix or some other diagonal matrix, the pairs $(\mathbf{q}_i, \mathbf{p}_i)$ are stored instead of $\mathbf{H}_k$ (which requires a large memory), and $\mathbf{H}_k\mathbf{g}_k$ is computed by a recursive algorithm. All the limited-memory quasi-Newton methods presented below fit into this conceptual framework. They differ only in the selection of the vector couples $(\mathbf{q}_i, \mathbf{p}_i)$, the choice of $\mathbf{H}_0$, the method for computing $\mathbf{H}_k\mathbf{g}_k$, the line search implementation, and handling of restarts.

Just as limited-memory quasi-Newton methods attempt to combine modest storage and computational requirements of conjugate-gradient methods with the convergence properties of the standard quasi-Newton methods, truncated Newton methods attempt to retain the rapid (quadratic) convergence rate of classic Newton methods while making storage and computational requirements feasible for large-scale applications. Recall that Newton methods for minimizing a multivariate function $J(\mathbf{x}_k)$ are iterative techniques based on minimizing a local quadratic approximation to J at every step. The quadratic model of J at a point $\mathbf{x}_k$ along the direction of a vector $\mathbf{d}_k$ can be written as

$$J(\mathbf{x}_k + \mathbf{d}_k) \approx J(\mathbf{x}_k) + \mathbf{g}_k^T\mathbf{d}_k + \frac{1}{2}\mathbf{d}_k^T\mathbf{H}_k\mathbf{d}_k \tag{25}$$

where $\mathbf{g}_k$ and $\mathbf{H}_k$ denote the gradient and Hessian of J, respectively, at point $\mathbf{x}_k$. Minimization of this quadratic approximation results in a linear system of equations for the search vector $\mathbf{d}_k$ known as the Newton equations:

$$\mathbf{H}_k \mathbf{d}_k = -\mathbf{g}_k. \tag{26}$$

In the modified Newton framework, a sequence of iterates is generated from $\mathbf{x}_0$ by the rule $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$. The vector $\mathbf{d}_k$ is obtained as the solution (or approximate solution) of the system (26) or, possibly, a modified version of it, where some positive definite approximation to $\mathbf{H}_k$, $\tilde{\mathbf{H}}_k$, replaces $\mathbf{H}_k$.

When an approximate solution is used, the method is referred to as a "truncated" Newton method because the solution process of (26) is not carried out to completion. In this case, $\mathbf{d}_k$ may be considered satisfactory when the residual vector $\mathbf{r}_k = \mathbf{H}_k \mathbf{d}_k + \mathbf{g}_k$ is sufficiently small. Truncation may be justified since accurate search directions are not essential in regions far away from local minima. For such regions, any descent direction suffices, so the effort expended in solving the system accurately is often unwarranted. However, as a solution of the optimization problem is approached, the quadratic approximation of (25) is likely to become more accurate and a smaller residual may be more important. Thus, the truncation criterion should be chosen to enforce a smaller residual systematically as minimization proceeds. One such effective strategy requires

$$\|\mathbf{r}_k\| \leq \eta_k \|\mathbf{g}_k\| \tag{27}$$

where

$$\eta_k = \min\{\frac{c}{k}, \|\mathbf{g}_k\|\}, \quad c \leq 1. \tag{28}$$

Other truncation criteria have also been discussed (Nash, 1984a,b; Schlick and Fogelson, 1992a,b).

The quadratic subproblem of computing an approximate search direction at each step is accomplished through some iterative scheme. This produces a nested iteration structure: an "outer" loop for updating $\mathbf{x}_k$, and an "inner" loop for computing $\mathbf{d}_k$. The *linear* conjugate-gradient method is attractive for large-scale problems because of its modest computational requirements and theoretical convergence in at most N iterations (Golub and van Loan, 1989). However, since conjugate-gradient methods were developed for positive definite systems, adaptations must be made in the present context where the Hessian may be indefinite. Typically, this is handled by terminating the inner loop (at iteration q) when a direction of negative curvature is detected ($\mathbf{d}_q^T \mathbf{H}_k \mathbf{d}_q < \xi$ where ξ is a small positive tolerance such as 10^{-10}); an exit direction that is guaranteed to be a descent direction is then chosen (Dembo and Steihaug, 1983; Schlick and Fogelson, 1992a,b). An alternative procedure to linear conjugate-gradient for the inner loop is based on the Lanczos factorization (Golub and van Loan, 1989), which works for symmetric but not necessarily positive definite systems. It is important to note that different procedures for the inner loop can lead to a very different overall performance in the minimization.

Several comparisons have been made on different large-scale unconstrained minimization algorithms. Navon and Legler (1987) compared a number of different conjugate-gradient and limited-memory quasi-Newton methods for problems in meteorology and

concluded that the Shanno-Phua (1980) limited-memory quasi-Newton algorithm was the most adequate for their test problems. The studies of Gilbert and Lemaréchal (1989) and Liu and Nocedal (1989) indicate that the L-BFGS method is among the best limited-memory quasi-Newton methods available to date. Nash and Nocedal (1989) compared the L-BFGS method with the truncated Newton method of Nash (1984a,b, 1985) on 53 problems of dimensions ranging between 10^2 - 10^4. Their results suggest that performance is correlated with the degree of nonlinearity of the objective function: for quadratic and approximately quadratic problems, the truncated Newton algorithm outperformed L-BFGS, while for most of the highly nonlinear problems L-BFGS performed better.

Zou, *et al.* (1993b) compared four of the state-of-the-art limited memory quasi-Newton methods and two truncated Newton methods on several test problems including problems in meteorology and oceanography. Their results confirm that the L-BFGS algorithm of Liu and Nocedal (1989) seems to be the most efficient, particularly robust and user-friendly. It deals with the critical issue of storage in large-scale problems. The L-BFGS update formula generates matrices using information from the last m Quasi-Newton iterations, where m is the number of quasi-Newton updates determined by the user(generally $3 \leq m \leq 7$). After having used the m vector storage locations for m quasi-Newton updates, the quasi-Newton approximation of the Hessian matrix is updated by dropping the oldest information and replacing it by the newest information. A new search direction, which is an estimate of the relative change to the current variables vector that produces the maximum reduction in the cost function, is then computed. It employs a cubic line search required to satisfy a Wolfe (1969) condition and a unit stepsize is always tried first. This algorithm uses a limited amount of storage and the quasi-Newton approximation of the Hessian matrix is updated continuously. The general algorithm for L-BFGS and other limited-memory quasi-Newton algorithms and truncated Newton methods can be found in the paper of Zou, *et al.* (1993b).

Scaling

The term "scaling" is invariably used in many meteorological applications in a vague sense to discuss numerical difficulties whose existence is universally acknowledged, but cannot be described precisely in general terms. Therefore, it is not surprising that much confusion exists about scaling. One normally considers the effect on problem scaling of replacing the original variables $\mathbf{x}$ of the problem by a set of linearly transformed variables $\mathbf{y}$: $\mathbf{x} = \mathbf{L}\mathbf{y}$, where $\mathbf{L}$ is a fixed non-single matrix. The choice of the value of $\mathbf{L}$ depends upon what does one wishes to scale. There are four scaling methods: by variable, by constraint, by gradient, and by Hessian (Gill, *et al.*, 1981). In theory, such a transformation does not affect the optimal solution; however, the scaling of variables may have an enormous effect on the behaviour of finite-precision calculations. One example of poor scaling is an imbalance between the values of the function and changes in $\mathbf{x}$ in the sense that the function values may change very little even though $\mathbf{x}$ changes significantly, or vice versa. The effect of poor scaling is to invalidate the criteria of the line search used to accept the step length. This will lead to the failure of a step-length algorithm to find an acceptable point along the search direction. Let $F(\mathbf{y})$ denote the function of the

transformed variables:

$$F(\mathbf{y}) = J(\mathbf{L}\mathbf{y}). \tag{29}$$

The derivatives of the function F with respect to the transformed variables $\mathbf{y}$ are

$$\nabla_y F(\mathbf{y}) = \mathbf{L}\nabla_x J(\mathbf{x}), \qquad \nabla_y^2 F(\mathbf{y}) = \mathbf{L}^T \nabla_x^2 J(\mathbf{x})\mathbf{L}. \tag{30}$$

One instance of bad scaling can occur when the variables $\mathbf{x}$ have widely differing orders of magnitude. A basic rule for scaling is that the variables of the scaled problem should be of similar magnitude and of order unity because within optimization routines convergence tolerances and other criteria are necessarily based upon an implicit definition of "small" and "large" and thus variables with widely varying orders of magnitude may cause serious difficulties for some minimization algorithms (Gill, *et al.*, 1981). One simple direct way to determine the scaling factor is to use the typical values for different fields (for instance 10^{-5} can be used as the scaling factor of vorticity), i.e.,

$$\mathbf{x} = \mathbf{D}\mathbf{y} \tag{31}$$

where d_j is set to a typical value of the j-th variable, causing each variable to be of similar "weight" during the optimization. Another form of bad scaling occurs when the partial derivatives of a function with respect to a particular variable are not "balanced". A scaling based on the first derivative can be used where $d_j = (1 + |J(\mathbf{x})|)/(2|\nabla_{x_j} J|)$ (see Gill, *et al.*, 1981).

Therefore, scaling is a crucial issue in the success of nonlinear unconstrained optimization problems and some research has been carried out on scaling nonlinear programming problems. It is well known that a badly scaled nonlinear programming problem can be almost impossible to solve (see also Navon and de Villiers, 1983; Courtier and Talagrand, 1990). An effective automatic scaling procedure would ease these difficulties, and could also render problems which are well scaled easier to solve by improving the condition number of their Hessian matrix which controls the rate of convergence of the conjugate-gradient and quasi-Newton algorithms (Thacker, 1989).

In meteorological problems, the variables in the control vector have also enormously different magnitudes varying over a range of 8 orders of magnitude. 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 unconstrained minimization process. In our experiment with the NMC spectral model, the scaling constants for the different fields are calculated by

$$\mathbf{L}_{\psi ii} = \max_{i,j,k} |\psi^{obs}_{i,j,k}(t_0) - \psi^{obs}_{i,j,k}(t_R)|/2, \tag{32}$$

where ψ represents the stream function field, and similarly for χ (velocity potential), T (temperature), $\ln ps$ (log of the surface pressure), and q (specific humidity).

For complicated functions, difficulties may be encountered in choosing suitable scaling factors. There is no general rule to determine the best scaling factors for all minimization problems, and a good scaling is problem dependent. Further improvement of the condition number can be obtained by a more sophisticated scaling (Gill, *et al.*, 1981; Bertsekas, 1982; Luenberger, 1984).

Preconditioning

Preconditioning is a way to speed-up the convergence and to "relax" ill-conditioning of the Hessian matrix by altering both the condition number as well as the distribution of eigenvalues "clustering" within the spectrum (Axelsson and Barker, 1984). The conditioning of the Hessian matrix at the solution of an unconstrained problem determines the accuracy with which the solution can be computed in different directions. In particular, when the Hessian is ill-conditioned, the cost function will vary much more rapidly along some directions in R_N than along others. An ill-conditioned Hessian at the solution is thus a symptom of bad scaling, in the sense that similar changes in the control variables $\|\mathbf{x}\|$ do not lead to similar changes in the cost function J.

We can briefly describe preconditioning as follows:

The solution of

$$\mathbf{H}\mathbf{x} = -\mathbf{b} \tag{33}$$

can be found by minimizing

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} - \mathbf{b}^T\mathbf{x}. \tag{34}$$

The transformation

$$\mathbf{y} = \mathbf{E}^T\mathbf{x} \tag{35}$$

leads to the problem of minimizing

$$\tilde{f}(\mathbf{y}) = f(\mathbf{E}^{-T}\mathbf{y}) = \frac{1}{2}\mathbf{y}^T\tilde{\mathbf{H}}\mathbf{y} - \tilde{\mathbf{b}}^T\mathbf{y} \tag{36}$$

where

$$\tilde{\mathbf{H}} = \mathbf{E}^{-1}\mathbf{H}\mathbf{E}^{-T}, \tilde{\mathbf{b}} = \mathbf{E}^{-1}\mathbf{b}. \tag{37}$$

If the spectral condition number of the preconditioned matrix $\tilde{\mathbf{H}}$ is such that

$$\kappa(\tilde{\mathbf{H}}) < \kappa(\mathbf{H}) \equiv \frac{\lambda_{\max}}{\lambda_{\min}}, \tag{38}$$

the rate of convergence is improved (see Axelsson and Baker, 1984).

The matrix $\tilde{\mathbf{H}}$ has the same eigenvalues as the matrix $\left(\mathbf{E}\mathbf{E}^T\right)^{-1}\mathbf{H}$. The matrix $\mathbf{Q} = \mathbf{E}\mathbf{E}^T$ is called the preconditioning matrix.

One can prove that

(i) $\mathbf{Q}$ is a good approximation to $\mathbf{H}$ if and only if

$$\mathbf{Q} = \alpha\mathbf{H}, (\alpha \text{ is a scalar}), \tag{39}$$

since in this case, $\kappa(\tilde{\mathbf{H}}) = 1$.

(ii) $\mathbf{Q}$ needs not to be a particular good approximation to $\mathbf{H}$ since if $\mathbf{Q} = \mathbf{I}$, $\kappa(\tilde{\mathbf{H}}) = \kappa(\mathbf{H})$.

For the preconditioning of the conjugate-gradient method, the preconditioning matrix $\mathbf{Q}$ can be taken as $\mathbf{Q} = \mathbf{M}^{-1}$, where $\mathbf{M}$ is obtained by performing r steps of a limited-memory quasi-Newton method from the one-parameter family of quasi-Newton updates. The matrix $\mathbf{MH}$ has r unit eigenvalues. This is roughly equivalent to choosing $\mathbf{Q}$ so that the condition number of $\tilde{\mathbf{H}}$ is as small as possible. In nonlinear problems, the matrix $\mathbf{Q}$ will vary from iteration to iteration. For example a limited-memory approximate Hessian matrix may be used as preconditioning matrix. A preconditioning based on a diagonal scaling is often used in practice, which is a less sophisticated technique that has been applied to general problems. Recently, a new preconditioning method was proposed by Zupanski (1993), which can be used for any minimization algorithm if the preconditioned steepest descent is used in the first iteration. His preconditioning is based on the requirements that the control variable adjustment is balanced by the calculated gradient norm and that a second order Taylor series expansion approximation is valid in the vicinity of the first guess. A significant improvement was found in his experiments. For recent research on preconditioning methods applied to discretized shallow-water equations model see Navon And Cai (1993) as well as the code NSPCG (Holter, *et al.*, 1991).

2.4 A simple example of variational data assimilation

A simple 2-D limited-area shallow water equations model is used to evaluate a quadratic objective function. The equations in Cartesian coordinates may be written as

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - fv + \frac{\partial \phi}{\partial x} = 0 \tag{40}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + fu + \frac{\partial \phi}{\partial y} = 0 \tag{41}$$

$$\frac{\partial \phi}{\partial t} + u\frac{\partial \phi}{\partial x} + v\frac{\partial \phi}{\partial y} + \phi(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}) = 0, \tag{42}$$

where f is the Coriolis parameter, u, v,and ϕ are the two components of the velocity field and the geopotential field, respectively; both are spatially discretized with a centered finite-difference scheme in space and an explicit leapfrog integration scheme in time. A rectangular domain of size $L = 6000km, D = 4400km$ is used along with space and time mesh increments of $\Delta x = 300km, \Delta y = 220km$ and $\Delta t = 600s$, respectively.

This model is widely used in meteorology and oceanography since it contains most of the physical degrees of freedom (including gravity waves) present in the more sophisticated 3-D primitive equation models. It is computationally less expensive to implement and results using this model can be expected to be similar to those obtained from a more complicated primitive equation model.

The objective function is defined as a simple weighted sum of squared differences between the observations and the corresponding prediction model values:

$$J = W_\phi \sum_{n=1}^{N_\phi} (\phi_n - \phi_n^{obs})^2 + W_V \sum_{n=1}^{N_V} [(u_n - u_n^{obs})^2 + (v_n - v_n^{obs})^2], \tag{43}$$

where N_ϕ is the total number of geopotential observations available over the 10 hour assimilation window (t_0, t_R), and N_V is the total number of wind vector observations. The quantities u_n^{obs}, v_n^{obs}, and ϕ_n^{obs} are the observed values for the northward wind component, the eastward wind component and the geopotential field respectively, while the quantities u_n, v_n, and ϕ_n are the corresponding computed model values. W_ϕ and W_V are weighting factors, taken to be the inverse of estimates of the statistical root-mean-square observational errors on geopotential and wind components respectively. Values of $W_\phi = 10^{-4}m^{-4}s^4$ and $W_V = 10^{-2}m^{-2}s^2$ are used. The objective function J is viewed as a function of $\mathbf{x}_0 = (u(t_0), v(t_0), \phi(t_0))^T$.

For the numerical experiments, the observational data consist of the model-integrated values for wind and geopotential at each time step starting from the Grammeltvedt initial conditions (Figure 1, Grammeltvedt (1969)). Random perturbations of these fields, performed using a standard library randomizer RANF on the CRAY-YMP (shown in Figure 2) are then used as the initial guess for the solution. A grid of a resolution 21×21 points in space results in a control variables vector of dimension 1323.

Two different scaling procedures were considered: "gradient" and "consistent". The first scales the gradient of the objective function. The second makes the shallow water equations model nondimensional.

The limited-memory quasi-Newton method of Liu and Nocedal (1989) —L-BFGS— was successful only with gradient scaling. The limited-memory quasi-Newton method of Gill and Murray (1979) —E04DGF — (see NAG, 1990) worked only with the nondimensional shallow water equations model. It appears that additional scaling is crucial for the success of the minimization algorithms.

Due to the different scaling procedures employed in the different minimization methods the minimization is stopped when the following convergence criterion

$$\|\mathbf{g}_k\| \leq 10^{-4} \times \|\mathbf{g}_0\| \tag{44}$$

is satisfied.

The performance of the truncated Newton often depends on the specified maximum number of permitted inner iterations per outer iteration (MXITCG). Our experience suggests that different settings for MXITCG have small impact on the performance of the truncated Newton method of Nash but a rather large impact on that of the truncated Newton method of Schlick and Fogelson (see Table 1). This results from our current unpreconditioned implementation for the truncated Newton method of Schlick and Fogelson, since the inner conjugate gradient linear equation solver will require more iterations to find a search direction.

To clarify this idea and to see which differences in performance between the two truncated Newton methods were due to the different truncation criteria, conjugate gradient versus Lanczos, and preconditioning, we performed minimization for the truncated Newton method of Nash without diagonal preconditioning. The results are presented in Table 1. Similar trends are identified for both the truncated Newton method of Nash and the truncated Newton method of Schlick and Fogelson in this case: the cost for large MXITCG is much lower than that for small MXITCG. However, the truncated Newton method of Schlick and Fogelson with MXITCG=50 performs much better than

Figure 1: (a) Geopotential fields based on the Grammeltvedt initial condition and (b) the wind field calculated from the geopotential fields in Fig. (a) by the geostrophic approximation. Contour interval is 200 $m^2 s^{-2}$ and the value of maximum vector is 29.9 ms^{-1}.

Figure 2: Random perturbation of (a) the geopotential and (b) the wind fields in Fig. 1. Contour interval is 500 m^2s^{-2} and the value of maximum vector is 54.4 ms^{-1}.

Table 1: Initial condition control problam in meteorology

algorithm	MXTITCG	Iter	Nfun	NCG	MTM (total CPU time)	FTM (function calls' CPU time)
TN1	3	19	20	50	12.20	11.31
TN1	50	20	26	54	13.79	12.89
TN1	3	63	64	170	38.82	37.15
(no prec.)	50	39	40	165	32.78	31.53
TN2	3	81	82	242	68.68	67.21
	50	4	5	91	16.41	16.30

Table 2: Meteorological problem with the limited memory quasi-Newton and truncated Newton methods

Algorithm	Iter	Nfun	MTM (total CPU time)	FTM (Function calls' CPU time)
E04DGF	72	203	36.89	33.56
L–BFGS	66	89	15.53	14.76
TN1	19	70	12.20	11.31
TN2	4	96	16.41	16.30

the truncated Newton method of Nash with MXITCG=50 in terms of Newton iterations, conjugate-gradient iterations, function evaluations, and CPU time. This strongly suggests that with a suitable preconditioner for the problem in meteorology, the truncated Newton method of Schlick and Fogelson might perform best. Thus, the use of preconditioning in the truncated Newton method of Nash accelerates performance as expected.

Table 2 presents the performance of the E04DGF, L-BFGS and the two truncated Newton methods. We observe from Table 2 that most of the CPU time is spent on function calls rather than in the minimization iteration. By comparing the number of function calls and CPU time, we find that the computational cost of L-BFGS is much lower than that of E04DGF. In 66 iterations with 89 function calls, L-BFGS converged. In contrast, E04DGF required 72 iterations and 203 function calls to reach the same convergence criterion. This produces rather large differences in the CPU time spent in minimization. L-BFGS uses less than half of the total CPU time required for E04DGF. The truncated Newton methods are competitive with L-BFGS and the truncated Newton method of Nash is slightly better than L-BFGS.

The differences between the figures showing the retrieved initial wind and geopotential and the one in Figure 1 are imperceptible (figures omitted). Table 3 gives the maximum differences between the retrieval and the unperturbed initial conditions. An accuracy of at least 10^{-3} is reached for both the wind and geopotential fields using any of the aforementioned four minimization codes This clearly shows the capability of the unconstrained limited-memory quasi-Newton and the truncated Newton methods to adjust a numerical

Table 3: Maximum absolute differences between the retrieval and the unperturbaed initial wind and geopotential fields using the limited memory quasi-Newton and truncated Newton methods

Algorithm	$(u'^2 + v'^2)^{1/2}$	$\lvert\phi'\rvert$
E04DGF	7.5×10^{-3}	1.2×10^{1}
L–BFGS	3.8×10^{-2}	9.0×10^{-1}
TN1	8.9×10^{-3}	5.4×10^{1}
TN2	5.8×10^{-3}	4.1×10^{1}

weather prediction model to a set of "observations" distributed in both time and space.

Recently, Wang, *et al.* (1993) indicated that the performance of the truncated Newton method of Nash can be significantly improved by using a more accurate Hessian/vector product which is required when solving the Newton equation (26) to obtain the search direction. A second order adjoint model was integrated backwards in time instead of using a finite-difference approximation to obtain the value of the Hessian/vector product.

3 Incomplete observations

3.1 Uniqueness of the solution

In real 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 have only a limited number of the observations in space and time. This is mathematically equivalent to determining the necessary condition for a unique solution of the problem of optimal control of distributed parameters.

The solution U^* for the problem (1) will be unique if the functional J is strictly convex., i.e. if

$$J(\lambda U + (1-\lambda)V) < \lambda J(U) + (1-\lambda)J(V), \quad \lambda \in [0,1],\ U, V \in R_N. \tag{45}$$

A geometric interpretation of (45) is that the graph of J is under any chord between U and V.

If J is only convex, i.e., the inequality in (45) 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.

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.

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.

Therefore, the uniqueness of the solution of minimizing problem (1) depends on the structure of the Hessian matrix H. (1) has a unique solution U^* if the Hessian matrix H is positive definite.

By definition, the functional J depends upon the observations, therefore, computing the Hessian of J, where J is defined by (1), permits one to estimate the link between the observations and the retrieved fields.

In the linear case, the forecast model can be written as

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}, \tag{46}$$

where $\mathbf{A}$ is an $N \times N$ constant matrix independent of $\mathbf{x}$.

An explicit expression for the gradient $\nabla J(U)$ and the Hessian matrix $\mathbf{H}$ can be derived by

$$\nabla J(U) = P(0) = \sum_{r=0}^{R} e^{\mathbf{A}^T t_r} \mathbf{C}^T (\mathbf{C} e^{\mathbf{A} t_r} \cdot U - \mathbf{x}^{obs}), \tag{47}$$

$$\mathbf{H} = \sum_{r=0}^{R} e^{\mathbf{A}^T t_r} \mathbf{C}^T \mathbf{C} e^{\mathbf{A} t_r}, \tag{48}$$

where $\mathbf{H}$ is a symmetric matrix ($\mathbf{H}^T = \mathbf{H}$).

It was proved (Zou, *et al.*, 1992a) that $\mathbf{H}$ is positive definite if and only if the rank of

$$\mathcal{A} = \begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{n-1} \end{bmatrix} \tag{49}$$

is equal to N.

We see from (48) or (49) that the Hessian $\mathbf{H}$ is independent of the observation itself, depending only on the model (represented by the operator $\mathbf{A}$) and on the operator $\mathbf{C}$ mapping the meteorological variables into the space of observations. Therefore, the uniqueness of the solution is dependent on the model and the mapping operator.

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. Since U^* is not known in practice, one can only estimate uniqueness of the solution by calculating the value of the Hessian at the initial guess point assuming that the initial guess is not far from the solution. 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, one can estimate the minimum eigenvalues $\lambda_{\min}(\mathbf{H})$ of the Hessian of the cost function by using an iterative power method or a Rayleigh quotient method(Golub and Van Loan, 1989), both of which require only the value of the Hessian/vector product, which can be obtained by integrating the so-called second order adjoint model (Zou, *et al.*, 1992a, Wang, *et al.*, 1992). If $\lambda_{\min}(\mathbf{H}) > 0$, $\mathbf{H}$ is positive definite and the solution of J is unique.

3.2 Results with incomplete observations

In this subsection, some numerical experiments, similar to the one in section 2.4, are carried out to estimate the influence of the distribution of available observations on the uniqueness of the solution and on the rate of convergence to the local minimum.

First we decreased the number of the observational fields. Suppose that only the wind fields (u and v) are observed and there are no observations of the geopotential field (ϕ). The numerical results show that the minimization of the cost function defined in (43) 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 for u and v which is consistent (i.e., in geostrophic balance) with the Grammeltvedt (1969) initial ϕ field. Using a linearized shallow-water equations, one can derive that the rank of $\mathcal{A}$ (the linearized model operator) is equal to N for both cases. This shows that our numerical results are consistent with the theoretical conclusion of (49), which determines the sufficient conditions for the existence of 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 and also no satisfactory convergence and retrieval are obtained. It seems that the performance of the minimization depends 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.

One could wonder whether the results from the 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. 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.

Finally 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. We observe that the performance of the minimization is very similar to the case where observations were available at every time step, with the convergence rate of the

Table 4: Values of the maximum and minimum eigenvalues ($\lambda_{\max}$, $\lambda_{\min}$) and the condition number ($\kappa(\mathbf{H})$) of the Hessian of the cost function at the initial guess point using the power method and the shifted power method (with the convergence criteria: $|\lambda_{k+1} - \lambda_k| < 10^{-8}$)

Problems	$\lambda_{\max}$	$\lambda_{\max}$	$\kappa(\mathbf{H})$
complete observations	2.20×10^{-4}	4.89×10^{-7}	449.2
less observations in time space (30 steps)	1.84×10^{-4}	1.08×10^{-7}	1695.5
less observations in grid space (2 grids)	2.90×10^{-4}	-1.66×10^{-5}	17.53

former being slightly slower than that of the later. A satisfactory retrieval of initial wind and the geopotential height fields is obtained even for all the cases. The difference fields between the retrieved and unperturbed initial wind and geopotential are several orders of magnitude smaller than the initial fields.

In Table 4 we calculated the maximum and minimum eigenvalues and the condition number of the Hessian at the initial guess of the minimization and we find that the reduction of observations in the grid space produces a negative minimum eigenvalue of the Hessian. Thus, we obtain an indefinite Hessian, pointing to the existence of either multiple minima or saddle points. The condition number in the case that observations are available at each grid point but only every 30 time steps (case 2) is larger that in the case of complete observation, which explains that the convergence rate in the case of less observations in time space is slightly slower than that in the case of complete observations in time space.

4 Control of gravitational oscillations

Gravitational oscillations with unrealistically large amplitudes may be generated in numerical models due to the insertion of erroneous data or due to errors in models. Methods aimed at the specification of suitably balanced initial fields, which will not give rise to spurious gravity oscillations, are known as initialization methods. 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 carrying out 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. 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 normal mode initialization (NMI) procedure. Other penalty terms controlling the total time tendency of the flow were proposed by Zou *et al.* (1992a,

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

1993a). The proposed method requires less memory and computational effort than the method of combined penalty and NMI proposed by Courtier and Talagrand (1990).

In the following, both the shallow-water equations model and the primitive equations model are used to test the performance of the penalty method on controlling gravitational oscillations.

4.1 A 2-dimensional problem

In this subsection experiments similar to the one described in section 2.4 are carried out using the shallow-water equations model. 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, a small perturbation was added to the initial height field defined in section 2.4 in the following way

$$\phi'(i,j) = \phi(i,j) + \alpha_{i,j} \times \phi(i,j), \quad i,j = 8,12 \tag{50}$$

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.

The isoline graphs of the perturbed height fields ϕ' are plotted in Figure 3. We see that the height field thus obtained contains one local grid point perturbation. This perturbed initial condition was used to produce model-generated observations. The variational data

assimilation with or without the penalty term is implemented by starting from a model atmosphere at rest and minimizing the cost function defined in (43). When the prescribed convergence criterion

$$\|\vec{g}_k\| \leq \epsilon \max\{1, \|\mathbf{x}_k\|\}, \tag{51}$$

is met, the minimization process was stopped, where $\vec{g}_k$ is the gradient of the cost function at the k-th iteration and $\epsilon = 10^{-14}$, a number close to the machine accuracy of the CRAY-YMP, was chosen based on expecting to obtain 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 may also 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 corresponds to an overfitting. This is quite acceptable in an "identical twin" approach with an abundance of data available which corresponds to our case. 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.

Now, instead of minimizing the cost function defined in (43), a penalized cost function J_r is minimized, where J_r is defined as

$$J_r = J + r \left\| \frac{\partial \phi}{\partial t} \right\|^2 . \tag{52}$$

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 usual forcing terms $2W(t_r)(\mathbf{x}(t_r) - \mathbf{x}^{obs}(t_r))$ in the adjoint model whenever an analysis time t_r $(r = 0, 1, \cdots, R)$ is reached, we also add at every time step the following additional terms

$$2r\mathbf{A}^* \frac{\partial \phi}{\partial t}, \tag{53}$$

to the right hand side of the adjoint model.

For comparison, a minimization of the cost function J_0, which does not include a penalty term, was carried out to serve as a benchmark. The minimization 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).

Then a number of minimizations are carried out which minimize J_r, $(r \neq 0)$ to damp the perturbation shown in Figure 3. One might have thought at the beginning that the gravity waves may 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 while 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

Figure 4: 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 $r = 10^5$.

105 iterations with 141 function calls, the minimization failed again before the same convergence criteria (51) was attained. 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 of the penalty method is given by the following corresponding variation of the quantity $\|\partial\phi/\partial t\|^2$ which decreased from $0.77 \times 10^2 m^4 s^{-6}$ to $0.70 \times 10^{-1} m^4 s^{-6}$ in the course of the minimization (see Figure 4), 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.

The sensitivity of the performance of the minimization for different values of penalty parameter r 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 of the paper by Zou *et al.*, 1992a). 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 attempts to compute the minima due to the slow convergence induced by the increasingly larger condition number of the Hessian of the penalized 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 (a method called the exterior quadratic penalty method, see Fiacco and McCormick, 1968 and Rao, 1983). In the exterior penalty function method, each successive $\vec{x}^*(r_l)$ 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 $\vec{x}^*(r_l)$ is the minimum point obtained using the preceding penalty parameter r_l). Another elegant method avoiding the ill-conditioning of penalty methods is the augmented Lagrangian

Table 5: 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	$\lvert\nabla J^*/\nabla J_0\rvert$	ϵ
1	10^2	48	74	4.46×10^{-6}	1.41×10^{-4}	10^{-12}
2	10^3	16	26	7.85×10^{-1}	1.76×10^{-1}	10^{-13}
3	10^4	19	28	4.20×10^{-1}	1.02×10^{-1}	$?5 \times 10^{-13}$
4	10^5	26	37	3.96×10^{-1}	3.13×10^{-2}	10^{-14}

method (Navon and de Villiers, 1983).

An exterior penalty function numerical experiment with $r_1 = 10^2$, $c = 10$, $l = 1, 4$ was carried out, in which each subsequent minimization cycle uses an increasingly stringent convergence criteria as shown in Table 5 (also see Navon and de Villiers, 1983). Again most of the large scale characteristic features are retrieved after the first minimization cycle (Figs. 5a-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 comparable with the computational cost of the base minimization.

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.

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 the stringent one of 10^{-14}) was reached. The minimization solution of the case where no penalty term was included is then taken to serve as the first guess for the next minimization cycle with a large 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 $\epsilon = 10^{-14}$ was reached. We call this two- stage minimization the "short-cut" penalty method. We found that the distribution of the retrieved initial height fields after the two minimizations is the same as in Figure 5(d). As far as the computational cost is concerned (see Table 6), 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.

To illustrate the effectiveness of this two stage minimization procedure on damping the gravitational oscillations, 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 carried out. In Figure 6 we plot the time variation of the geopotential field at a given

Figure 5: 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 i^{th} penalized minimization was used as the initial guess for the $(i+1)^{\text{th}}$ minimization, $i = 1, \ldots, 3$. Contour intervals and scaling factor for the values on isolines are the same as in Figure 3.

Figure 5: (Continued)

Table 6: 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	$\lvert\nabla J^*/\nabla J_0\rvert$	ϵ
1	0	52	75	4.46×10^{-6}	1.41×10^{-4}	10^{-12}
2	10^5	34	45	7.85×10^{-1}	1.76×10^{-1}	10^{-14}

Figure 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.

grid point for a variationally assimilated integration without the inclusion of a penalty term as well as the results from the integration with the penalty term being included in the cost function. From the 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 the occurrence of a major impact for the first 6-10 hours of the numerical integration. 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 by 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 results as a by-product of the adjoint model integration and involves little additional computational 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).

4.2 A 3-dimensional problem

Variational 4-D data assimilation combined with a penalty method constraining time derivatives of the surface pressure and divergence is implemented on an adiabatic version of the NMC 18 level primitive equations spectral model (NMC, 1988) with surface drag and horizontal diffusion.

High-frequency fluctuations in the surface pressure tendency reflect the presence of external gravity waves. A way to suppress these external gravity waves consists in imposing as a constraint the vanishing of the time tendency of the surface pressure: $\frac{\partial \ln p_s}{\partial t}$. A quadratic penalty term of the form

$$J_{r_1} = r_1 \sum_{i=0}^{R-1} \left(\frac{\partial \ln \vec{p}_s(t_i)}{\partial t} \right)^T \left(\frac{\partial \ln \vec{p}_s(t_i)}{\partial t} \right), \tag{54}$$

was thus augmented the cost function J, where $\ln \vec{p}_s$ is an M dimensional vector of the logarithm of the surface pressure, M is the total number of Gaussian grid points used in the definition of the cost function, and R is the number of time steps in the time interval spanning the window of assimilation. Variational data assimilation with a penalty term J_{p1} included in the cost function means that while minimizing the distance between the model solution and data, the high frequency oscillations presented in the surface pressure field are also controlled.

We may also impose a constraint on the time tendency of the divergence to damp high-frequency oscillations of internal gravity waves in a manner similar to (54). We define a penalty term

$$J_{r_2} = r_2 \sum_{i=0}^{R-1} \left(\frac{\partial \vec{D}(t_i)}{\partial t} \right)^T \left(\frac{\partial \vec{D}(t_i)}{\partial t} \right), \tag{55}$$

where $\vec{D}$ is the $M \times K$ dimensional vector of divergence and K is the total number of vertical levels.

The penalized cost function thus assumes the following form

$$J_{r_1,r_2} = J + J_{r_1} + J_{r_2}. \tag{56}$$

The gradient of the penalized cost function can be obtained by integrating the same adjoint model as the one used in obtaining the gradient of the non-penalized cost function. The only difference is that more forcing terms are added to the right-hand-side of the adjoint equations model besides the weighted differences between the model solution and data (Zou, *et al.*, 1993a).

Numerical experiments are devised as follows: initial conditions for the model and analyzed data were obtained from the Global Data Assimilation scheme at NMC. The variational data assimilation was carried out adiabatically, as well as with a surface drag and horizontal diffusion terms included in the model. Two uninitialized state vectors,

separated by a 6 hour time interval, were taken as analyzed data to be assimilated. The cost function was defined as

$$J(\mathbf{x}(t_0)) = \frac{1}{2}\left(\mathbf{x}(t_0)-\mathbf{x}^{obs}(t_0)\right)^T \mathbf{W}\left(\mathbf{x}(t_0)-\mathbf{x}^{obs}(t_0)\right) + \frac{1}{2}\left(\vec{X}(t_R)-\vec{X}^{obs}(t_R)\right)^T \mathbf{W}\left(\vec{X}(t_R)-\vec{X}^{obs}(t_R)\right) \quad (57)$$

where $\mathbf{x}$ is the model state vector, $\mathbf{W}$ is a diagonal weighting matrix whose values are calculated as the inverse of the maximum squared difference between the two analyzed data fields at times t_0 and t_R. The time difference between the two analyzed times is 6 hours. A limited-memory quasi-Newton method (Navon and Legler, 1987) due to Liu and Nocedal (1989) was employed throughout the minimization experiments.

In order to illustrate the performance of the penalty method we calculated the amount of residual gravity wave component in the retrieved initial state, which was 0.076. The amount of residual gravity wave component before and after the application of nonlinear NMI are 2.56 and 0.15 respectively, which provides a benchmark here for assessing the efficiency of filtering gravitational oscillations by the simple penalty method developed within the framework of variational data assimilation. We see that the solution of the minimization with both penalty terms J_{r_1} and J_{r_2} being added to the cost function contains a lesser amount of residual gravity component than the one contained in the nonlinear NMI initialized data.

The minimization of J_{r_1,r_2} with $r_1 = 10^{13}$ and $r_2 = 10^{19}$ performed very well with both J_{r_1}/r_1 and J_{r_2}/r_2 decreasing by about 2 orders of magnitudes. To show that the gravitational oscillations were damped out after minimizing the penalized cost function J_{r_1,r_2} and to confirm that after the high-frequency gravity wave oscillations have been completely eliminated by the penalty procedure while the Rossby modes are not degraded, and furthermore that once these oscillations have been eliminated, they do not seem to re-occur, a closer inspection of the integration results was carried out.

A 3 days forecast was carried out from the nonlinear NMI data and the retrieved initial states without a penalty term and with the two penalty terms, using the adiabatic version of the NMC spectral model with surface drag and horizontal diffusion. Figure 7 shows the time variation of the surface pressure at Point 1 (a point in the Indian Ocean) and Point 2 (a point near the Rocky Mountains) during this period. It is seen that the value from the retrieved initial state without a penalty term oscillates around a slowly varying value from the retrieved initial state with penalty terms at both points (solid and dotted lines in Figure 7). This fact indicates that the gravity oscillations present in the integration with the non-penalized retrieval neither grow nor influence significantly the slowly varying part of the fields. This fact is confirmed by a comparison between the penalization result and the nonlinear NMI result (dashed line in Figure 7). The time variation of the surface pressure is nearly the same at Point 2 for the results obtained from the retrieved initial state with penalty terms and from that obtained from the the nonlinear normal mode initialized state, with the former being even smoother than the latter. At Point 1 the time variation of the penalty result is smooth and exhibits a 12 hours oscillation. However, after nonlinear NMI both the large amplitude 12 hours oscillations

Figure 7: Time variation of 3 days forecasts at Point 1 and Point 2. Solid line: no penalty term. Dotted line: with penalty terms $J_{r_1} + J_{r_2}$. Dashed line: nonlinear NMI.

and the low amplitude high-frequency oscillations are reduced. After two and half days the time variation is nearly same.

In view of these experimental results, we are led to the conclusion that adequate penalization of the cost function is able to suppress undesirable high frequency gravity wave oscillations. The constraints of requiring the vanishing of the surface pressure and divergence at each time step were implemented weakly by the penalty method. The Rossby waves are not affected by the penalty procedure. It is also worth noting that very little additional computational cost is required to calculate the values of the penalized cost function and its gradient, and that the number of iterations and function evaluations is about the same as that required by the minimization without a penalty term. The advantage of using of simple penalty terms over penalty terms including the model normal modes results in a simplification of the procedure, allowing a more direct control over the model variables, and the possibility of using weak constraints to eliminate high frequency gravity wave oscillations. This approach does not require direct information about the model normal modes. One of the encouraging results obtained is that the introduction of the penalty terms does not slow down the convergence rate of the minimization process.

5 Inclusion of "on-off" processes in variational data assimilation

In the last several years most attention on variational methods for 4-D data assimilation has been focused on 2-D models (LeDimet and Talagrand, 1986; Derber, 1985; Lewis and Derber, 1985; Hoffman, 1986; Talagrand and Courtier, 1987; Courtier and Talagrand, 1987; Zou *et al*, 1992a) or 3-D adiabatic models (Thépaut and Courtier, 1992, Navon *et al.*, 1992 , and Chao and Chang, 1992). The issue of the impact of "on-off" threshold processes on variational data assimilation was addressed only recently (Douady and Talagrand, (1990); Zou, *et al.*, 1993c; Vukićecić (1993); Zupanski, 1993), when the inclusion of the adjoint of diabatic processes is considered.

The specific problem related to the presence of "threshold", or "on-off" processes, triggered when a given parameter attains a prescribed critical value, is how to code the adjoint of these processes and how will the minimization process behaves for locally nondifferentiable terms. The mathematical modeling of these physical processes defies a convenient quantification at the level of the resolution presently used in numerical forecast models. Different parameterization schemes are used and some involves special computational treatment due to the complicated nonlinear relationship existing between variables and the high cost of the computer time. Examples of these are, in the deep cumulus convection, the calculation of the dew point temperature from the vapor pressure, the calculation of the equivalent potential temperature (θ) from temperature and pressure fields at the lifting condensation level (LCL), and the derivation of the temperature and specific humidity of an ascending parcel from its potential temperature and pressure. In order to alleviate the computational load, the operational model employs several pre-calculated tables along with linear or bilinear interpolations. For the tangent linear and adjoint models, one can no longer use the original tables which were employed in the direct nonlinear

model and linearize the nonlinear model technically and locally in the code as one does in developing the tangent linear model of the adiabatic model. The values in the tables are built for the nonlinear calculation. The original (analytical) formulas used to build the lookup tables are required to compute the first derivatives. This is quite different from the development of the adjoint of the adiabatic model, where the only nonlinear terms present are those due to the quadratic terms. The "on-off" processes appearing in the code as IF statements depend on the model variables' value which produce a finite number of first-order discontinuous points. These points represent the "on-off" switches in time and space domains. The adjoint and gradient calculations assumes that the initial small perturbation does not change when and where the "on-off" switch occurs. Therefore, the adjoint model integration follows the same path backward as the nonlinear model's integration, i.e., the switches are determined by the basic state.

Once these problems of building the adjoint and calculating the value of the gradient of the cost function defined by a model constraint involving "on-off" physical processes are solved, the following questions arise: what will the computational load due to these processes be, and whether the minimization will still work for first order discontinuous points and moreover what will be the influence on the solution of the optimized initial conditions?

As an initial estimation to these problems, some preliminary experiments of variational data assimilation were carried out using a version of 18 vertical levels, triangularly truncated at wavenumber T40, of the NMC operational spectral model. In the experiments, the "observations" consisted of two complete states of vorticity, divergence, temperature, surface pressure field and moisture from the NMC global operational data assimilation system, 6-hour apart, which are denoted by $x^{\text{obs}}(t_0)$ and $x^{\text{obs}}(t_R)$. The adjustment was performed on the 6-hour interval $[t_0, t_R]$ preceding t_R. The problem is to find an optimal initial state $x^*(t_0)$ which minimizes the cost function J given by (57).

The computational cost of calculating the gradient of the cost function using the adjoint technique is independent of the dimension of the problem, meaning

$$\frac{\mathcal{T}(J, \nabla J)}{\mathcal{T}(J)} \leq c \tag{58}$$

where $\mathcal{T}(J)$ and $\mathcal{T}(J, \nabla J)$ measure the CPU time to compute the cost function J and the pair of $(J, \nabla J)$, respectively, and c is a constant independent of n. This constant is small, between 2 and 4, depending on the number of operations performed in the direct model. Table 7 displays the CPU time used by the NMC T40, 18 layer model, its tangent linear model and its adjoint model for a six hour time integration, with or without the large-scale precipitation and deep cumulus convection, respectively. The time step is 1800s. All the "on-off" routes and the nonlinear model solutions at each time step were stored in memory during the direct integration to avoid major recomputation costs in the adjoint. We observe that the required CPU time increased from $16.65s$ for an adiabatic model to $29.15s$ for the partially diabatic model due to the "on-off" physical processes. The tangent linear model integration is slightly more expensive than that of the nonlinear model integration, while the integration of the adjoint model is most expensive but still less than twice the CPU time required for the integration of the nonlinear model.

Table 7: CPU time used by the NMC T40, 18 layer primitive equation model, its tangent linear model and its adjoint model for a six hour time integration with a time step of 1800 seconds

CPU time (second)	nonlinear model	linear model	adjoint model
adiabatic	16.65	21.79	25.14
with "on-off" processes	29.15	31.10	34.49

Table 8: CPU time used for calculating the cost function J and its gradient with respect to the initial state ∇J

CPU time (second)	$\mathcal{T}(J)$	$\mathcal{T}(\nabla J)$	$\mathcal{T}(\nabla J)/\mathcal{T}(J)$
adiabatic	17.31	27.35	2.6
with "on-off" processes	30.01	37.73	2.3

Table 8 displays the values of c for the gradient calculation. For the adiabatic version of the model, c is equal to 2.6. When the "on-off" physical processes are included in the direct model and the corresponding adjoint model, $c = 2.3$. In addition, we note that the cost function calculation adds 1 second of CPU time, in both cases, to the model's integration. An additional 3 seconds are required for including the forcing term in the adjoint model. Therefore, the direct access of data in the nonlinear model and adjoint model, for both the model solution and the "on-off" route, is computationally very efficient.

The first test consisted of two parallel assimilations with and without the large-scale precipitation and cumulus convection, both including the horizontal diffusion and a simplified surface drag (see Navon et al., 1992). The minimization was found to be similar in both cases with the value of the cost function decreasing to about 15% of its original value. The norm of gradient decreased about two orders of magnitude. The total CPU times required for 60 iterations of the minimization with and without the physical processes were 76 and 44 minutes, respectively. The maximum memory used was 7.4MW and 6.4MW respectively.

From this experiment, we conclude that "on-off" threshold processes when included in the model do not hinder the minimization process. Theoretically, if the gradient of the cost function calculated by integrating the adjoint model is always a one-sided gradient, i.e., the model integration follows the same "on-off" route at each iteration, the large-scale unconstrained minimization will still work. We note however that this is not the case in our minimization experiment, and despite this fact, the minimization process performed successfully. Figures 8(a–c) show three examples of the large-scale saturation, evaporation, and the cumulus convection during the first 3 iterations of minimization, respectively. In

Figure 8: (a) Large scale supersaturation amounts at the point $(90^oE, 35^oN, 225mb)$ as a function of time. (b) Cumulus convective heating of the environment at the point $(25^oW, 38^oN, 920mb)$ as a function of time. (c) Cumulus convective moistening of the environment at the point $(25^oW, 38^oN, 920mb)$ as a function of time.

Figure 8(a) we plotted the amount of δq_c (the increment of specific humidity due to the large-scale precipitation process) when large-scale saturation occurs ($q > q_s$) at a model grid point located at $(90^o E, 35^o N, 225mb)$. We observe that the saturation process was turned on from the fourth leap-frog time step during the first and second iterations. During the third iteration, this process was also turned on at the fourth time step but it was turned off at the last time step of the assimilation window. We can draw a similar conclusion from Figures 8(b,c) for the cumulus convection process.

In the following, we compare the differences between the retrieved initial states from the adiabatic and partially adiabatic runs. The root-mean- square of the retrieved divergence field was found to increase slightly when the large-scale precipitation and the cumulus convection processes were included in the variational data assimilation (Figure omitted).

In order to observe the differences in the retrieved initial conditions with or without the inclusion of the precipitation processes, we now consider the horizontal distribution of the model's variables displayed over an area encompassing North America and plotted in Figures 9(a,b) the difference of the divergence, vorticity, temperature and the moisture fields at high and low levels respectively. In the area where the divergence in the higher layers and the convergence in lower layers increases due to the inclusion of the "on-off" processes in the minimization, (we shall call it the active region), the temperature and the vorticity were also found to have increased. At the lower levels, as the convergence increased, the temperature decreased and the moisture increased. In addition, a northwest- southeast wave train is observed in the vorticity field which propagates toward the equator.

Figure 10 displays the retrieved initial stream functions near 500mb (level 9). We observe that there is a trough upstream of the active region, and that the differences in the minimization solution with or without the physical processes are strongly correlated with the weather system.

We also carried out 6 h forecasts of the precipitation from the two initial conditions (Figure 11). We find that the initial condition obtained by the variational data assimilation including the "on-off" processes produces larger amounts of precipitation and greater area coverage over most of the map. However, the amount of precipitation near Lake Erie is smaller than the amount forecasted from the initial condition obtained without the inclusion of the physical processes. Since we do not have verifying data, we cannot conclude which result is better. We have however demonstrated that differences near meteorologically active areas do result from the two respective minimizations.

Figure 9: The differences of divergence, temperature, vorticity and the moisture fields at (a-d) the 856mb level and (e-h) the 225mb level between the two retrieved initial states. Contour intervals for divergence, vorticity, temperature and moisture are $8.0 \times^{-7} s^{-1}$, $2.0 \times 10^{-6} s^{-1}$, $0.07K$ and $8.0 \times 10^{-5} gkg^{-1}$ in the upper layer and $1.5^{-6} s^{-1}$, $5.0 \times 10^{-6} s^{-1}$, $0.2K$ and $8.0 \times 10^{-5} gkg^{-1}$ in the lower layer respectively. Values on the isolines of divergence and moisture fields are scaled by 10^8 and 10^6 respectively. Values on the isolines of vorticity in the upper and lower layers are scaled by 10^8 and 10^7 respectively.

Figure 9: (Continued)

Figure 9: (Continued)

Figure 9: (Continued)

Figure 10: The distribution of the retrieved initial stream function at 500mb obtained by the minimization including the "on-off" processes. Contour interval is $6.0 \times 10^6\ m^2 s^{-1}$ and values on the isolines are scaled by 10^{-5}.

Figure 11: The total 6 hour precipitation forecasted from the initial state with(a) or without(b) the "on-off" processes. Contour interval is 2.0×10^{-2} m and values on the isolines are scaled by 10^3.

6 Summary and Conclusions

The methods and concepts of optimal control theory have been employed to formulate a variational 4-dimensional data assimilation for meteorological problems described by systems of coupled nonlinear equations and the minimization of nonlinear cost functions. The adjoint method is the most economical method to use, if the physical problem involves a large data base.

It is of practical interest to mention that, in particular applications, additional constraints may be required to be imposed on the cost function. 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 high frequency gravity waves. Other constraints such as energy and enstrophy conservation, model bias, and observational errors can be included in the cost function in different situations.

Inclusion of the "on-off" physical processes in the nonlinear and adjoint models does not hinder the minimization processes. Along this line, further research is needed related to the inclusion of additional physical processes in the adjoint model, such as the planetary boundary layer processes, vertical diffusion, topographic variations, shallow convection and radiation. Once the full adjoint of the operational model including the physics package is obtained, it can be used to tune the forecast model by carrying out an adjoint sensitivity study or a parameter estimation. As such the adjoint model constitutes a useful diagnostic tool and helps explore the physical details of the numerical model.

References

Axelsson, O. & Barker, V. A. (1984) *Finite Element Solution of Boundary Value Problems: Theory and Computation*, Computer Science and Applied Mathematics, Academic Press, 111 Fifth Avenue, New York, New York 10003, 437pp.

Bertsekas, D. P. (1982) *Constrained Optimization and Lagrange Multiplier Methods*, Computer Science and Applied Mathematics, Edited by Werner Rheinboldt, Academic Press, 111 Fifth Avenue, New York, New York 10003, 395pp.

Buckley, A. G. (1978) A combined conjugate-gradient quasi-Newton minimization algorithm, *Mathematical Programming*, **15**, 200-210.

Buckley, A. G. (1989) Remark on algorithm 630. *ACM Transactions of Mathematical software*, **15**, 262-274.

Buckley, A. G. & Lenir, A. (1983) QN-like variable storage conjugate gradients, *Mathematical Programming*, **27**, 155-175.

Buckley, A. G. & Lenir, A. (1985) Algorithm 630-BBVSCG. A variable storage algorithm for function minimization. *ACM. Trans. Math. Software*, **11**, 103-119.

Cacuci, D. G. (1981a) Sensitivity theory for nonlinear systems: I. Nonlinear functional analysis approach. *J. Math. Phys.*, **22**, 2794-2802.

Cacuci, D. G. (1981b) Sensitivity theory for nonlinear systems: II. Extensions to additional classes of responses. *J. Math. Phys.*, **22**, 2803-2812.

Chao W. C. & Chang, L. P. (1992) Development of a 4-dimensional variational analysis system using the adjoint method at GLA. Part I: Dynamics. *Mon. Wea. Rev.*, **120**, 1661-1673.

Courtier, P. & Talagrand O. (1987) Variational assimilation of meteorological observations with the adjoint equation - Part I. Numerical results. *Q. J. R. Meteorol. Soc.*, **113**, 1329-1347.

Courtier, P. & Talagrand O. (1990) Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus*, **42A**, 531-549.

Daley, R. (1991) Atmospheric Data Analysis. *Cambridge Atmospheric and Space Science Series, eds. Cambridge University Press*, 457pp.

Dembo, R. S. & T. Steihaug (1983) Truncated-Newton algorithms for large-scale unconstrained optimization. *Mathematical Programming*, **26**, 190-212.

Derber J. C. (1985) The variational 4-D assimilation of analysis using filtered models as constraints. *Ph. D. Thesis, University of Wisconsin-Madison, Madison, Wisconsin 53706*, 142pp.

Derber J. C. (1989) A variational continuous assimilation technique. *Mon. Wea. Rev.*, **117**, 2437-2446.

Douady, D. & Talagrand, O. (1990) The impact of threshold processes on variational assimilation. World Meteorological Organisation, Proceedings from International Symposium on Assimilation of Observations in Meteorology and Oceanography, Clermont-Ferrand, France, July 9-13, 486-487.

Fiacco, A. V. & McCormick, G. P. (1968) *Nonlinear Programming: Sequential unconstrained minimization techniques*, John Wiley and Sons, New York and Toronto, 210pp.

Ghil M. & Malanotte-Rizzoli, P. (1990) Data assimilation in meteorology and oceanography. *Adv. Geophys.*, **33**, 141-266.

Gilbert, J. C. & Lemaréchal, C. (1989) Some numerical experiments with variable-storage quasi-Newton algorithms. *Mathematical Programming*, **45**, 407-435.

Gill, P. E. & Murray, W. (1979) Conjugate-gradient methods for large-scale nonlinear optimization. *Technical Report SOL 79-15*, Systems Optimization Laboratory, Department of Operations Research, Stanford University, 61pp.

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

Golub, H. G. & van Loan, C. (1989) Matrix Computation. Second Edition, Johns Hopkins Series in the Mathematical Sciences, The Johns Hopkins University Press, 642pp.

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

Hall, M. C. G., Cacuci, D. G. (1983) Physical interpretation of the adjoint functions for sensitivity analysis of atmospheric models. *J. Atmos. Sci.* , **40**, 2537-2546.

Hall, M. C. G., Cacuci, D. G. & Schlesinger, M. E. (1982) Sensitivity analysis of a radiative-convective model by the adjoint method. *J. Atmos. Sci.*, **39**, 2038-2050.

Hoffmann, R.N. (1986) A four-dimensional analysis exactly satisfying equations of motion. *Mon.Wea. Rev.*. **114**, 388-397.

Holter, W. H., Navon I. M. & Oppe, T. C. (1993) Parallelizable preconditioned conjugate-gradient methods for the Cray Y-MP and the TMC CM-2, *submitted to International J. of High Speed Computing.* Tech. Rep., FSU-SCRI-92-128, available at SCRI, Florida State University, Tallahassee, FL 32306-4052, 49pp.

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

LeDimet, F. X. & Navon, I. M. (1989) Variational and optimization methods in meteorology - A review, Research Report, Supercomputer Computations Research Institute, Florida State University, Tallahassee.

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

Lewis, J. M. & Derber, J. C. (1985) The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus* **37A**, Vol.4, pp. 309-322.

Lions, J. L. (1971) Optimal control of systems governed by partial differential equations. *Translated by S. K. Mitter, Springer-Verlag, Berlin-Heidelberg*, pp396.

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

Lorenc, A. C. (1986) Analysis methods for numerical weather prediction, *Q. J. R. Meteorol. Soc.*, **112**, 1177-1194.

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

NAG (1990) Fortran library reference manual (Mark 14), Numerical Algorithms Group. **3**. [Available from NAG Limited, Oxford OX2-8DR, United Kingdom, or NAG Inc., Downers Grove, IL 60515-5702.]

Nash, S. G. (1984a) Newton-type minimization via the Lanczos method, *SIAM J. Numer. Anal.*, **21**, 770-788.

Nash, S. G. (1984b) Solving nonlinear programming problems using truncated-Newton techniques. *Numerical Optimization, Proceedings of the SIAM Conference on Numerical Optimization*, Boulder, Colorado, June 12-14, 1984. 119-136.

Nash, S. G. (1985) Preconditioning of truncated-Newton methods. *SIAM J. Sci. Stat. Comput*, **6**, 599-616.

Nash, S. G. & Nocedal, J. (1989) A numerical study of the limited memory BFGS method and the truncated-Newton method for large scale optimization. *Tech. Rep.*, Northwestern University, **NAM 02**, 1-19.

Navon, I. M. (1986) A review of variational and optimization methods in meteorology, in: Y. Sasaki, ed., Variational Methods in Geosciences (Elsevier, New York), 29-34.

Navon, I. M. & Cai, Y. (1993) Domain decomposition and parallel processing of a finite element model of the shallow water equations, *Computer Methods in Applied Mechanics and Engineering*, **Vol 106,No 1–2**,179- 212.

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

Navon, I. M. & Legler, D. M. (1987) Conjugate-gradient methods for large-scale minimization in meteorology. *Mon. Wea. Rev.*, **115**, 1479-1502.

Navon, I. M., Zou, X., Derber J. & Sela, J. (1992) Variational data assimilation with an adiabatic version of the NMC spectral model. Mon. Wea. Rev., **120**, 1433-1446.

NMC (1988) Documentation of the Research Version of the NMC medium range forecasting model. *National Meteorological Center, W/NMC2, Room 204, Development Division, World Weather Building, 5200 Auth Road, Camp Springs, Maryland 20746, U.S.A.*

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

Rabier, F. & Courtier, P. (1992) Four-dimensional assimilation in the presence of baroclinic instability. *Quart. J. Roy. Meteor. Soc.*, **118**, 649-672.

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, 747pp.

Schlick, T. & Fogelson, A. (1992a) TNPACK — A truncated Newton minimization Package for large-scale problems. I. Algorithm and usage, *ACM Trans. Math. Softw.*, **18**, 46-70.

Schlick, T. & Fogelson, A. (1992b) TNPACK — A truncated Newton minimization Package for large-scale problems. II. Implementation Examples. *ACM Trans. Math. Softw.*, **18**, 71-111.

Semazzi, F. & 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.

Shanno, D. F. & Phua, K. H. (1980) Remark on algorithm 500 - a variable method subroutine for unconstrained nonlinear minimization. *ACM Trans. on Mathematical Software*, **6**, 618-622.

Smedstad, O. M. & O'Brien, J. J. (1991) Variational data assimilation and parameter estimation in an equatorial Pacific Ocean model. *Prog. Oceanog.*, **26**, 179-241

Talagrand, O. (1992) The use of adjoint equations in numerical modeling of the atmospheric circulation, in: A. Griewank and G. F. Corliss, ed., Automatic Differentiation of Algorithms: Theory, Implementation, and Application (Society for Industrial and Applied Mathematics), 169-180.

Talagrand, O. & 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.

Thépaut J. N. & Courtier, P. (1992) Four-dimensional variational data assimilation using the adjoint of a multilevel primitive equation model. Q. J. R. M. S., **117**, 1225-1254.

Thacker, W. C. (1989) The role of the Hessian matrix in fitting models to measurements. *J. Geophy. Res.*, **94**, 6177-6196.

Vukićecić, T. (1993) Linearization and adjoint of parameterized moist diabatic processes, *accepted for publication in Tellus*

Wang, Z., Navon, I. M., LeDimet, F. X. & Zou, X. (1992) The Second Order Adjoint Analysis: Theory and Application. *Meteorol. Atmos. Phys.*, **50**, 3-20.

Wang, Z., Navon, I. M. & Zou, X. (1993) The adjoint truncated Newton algorithm for large-scale unconstrained optimization. *submitted to Optimization Methods and Software*

Zou, X., Navon, I. M. & LeDimet, F. X. (1992a) Incomplete observations and control of gravity waves in variational data assimilation. *Tellus*, bf 44A, 273-296.

Zou, X., Navon, I. M. & LeDimet, F. X. (1992b) An optimal nudging data assimilation scheme using parameter estimation. *Q. J. of Roy. Met. Soc.*, **118**, 1163-1186.

Zou, X., Navon, I. M. & Sela, J. (1993a) Control of gravitational oscillations in variational data assimilation, *Mon. Wea. Rev.*, **121**, 272-289.

Zou, X., Navon, I. M., Berger, M., Phua, Paul K. H., Schlick, T. & LeDimet, F. X. (1993b) Numerical experience with limited-memory quasi-Newton and truncated Newton methods. *SIAM J. on Optimization*, **Vol 3**, 582-608. Also Tech. Rep., FSU-SCRI-90-167, available at SCRI, Florida State University, Tallahassee, FL 32306-4052, 44pp.

Zou, X., Navon, I. M. & Sela, J. G. (1993c) Variational data assimilation with moist threshold processes using the NMC spectral model. *accepted for publication in Tellus*

Zou, X., Barcilon, A. , Navon, I. M., Whitaker, J. & Cacuci, D. G. (1993d) An adjoint sensitivity study of blocking in a two-layer isentropic model, *Mon. Wea. Rev.*, **Vol 121**, 2833-2857.

Zupanski, D. (1993) The Effects of Discontinuities in the Betts-Miller Cumulus Convection Scheme on Four-Dimensional Variational Data Assimilation, *accepted for publication in Tellus*

Zupanski, M. (1993) A preconditioning algorithm for large scale minimization problems. *Accepted for publication in Mon. Wea. Rev.*