National Oceanic & Atmospheric Administation
under Contracts NA37RJ0202 and NA67RJ0152

Air Force Office of Scientific Research
under Grant F49620-95-1-0132

A METHOD OF CONTINUOUS DATA ASSIMILATION USING SHORT-TERM 4D-VAR ANALYSIS

by Shuowen Yang and William R. Cotton

William R. Cotton, P.I.

Colorado State University

DEPARTMENT OF ATMOSPHERIC SCIENCE

PAPER NO. 653

A METHOD OF CONTINUOUS DATA ASSIMILATION USING SHORT-TERM 4D-VAR ANALYSIS

Shuowen Yang and William R. Cotton
Department of Atmospheric Science
Colorado State University
Fort Collins, CO 80523

Research Supported by

National Oceanic & Atmospheric Administration
under Contracts NA37RJ0202 and NA67RJ0152

Air Force Office of Scientific Research
under Grant F49620-95-1-0132

June 30, 1998

Atmospheric Science Paper No. 653

Table of Contents

Abstract

A continuous data assimilation method based on short-term four-dimensional variational data assimilation (4D-Var) is proposed. This method consists of forecast and analysis steps. The analysis increment (analyzed value minus forecasted value) is assumed to be proportional to the gradient of a cost function, which measures the misfit between model prediction and observations over a period of time. The gradient of the cost function is calculated with the adjoint method and is updated cyclically. This technique is a kind of retrospective analysis and can continuously assimilate data in an infinite time period. Different forecast model versions (or models) can be used in the forecast and analysis steps.

A two-dimensional shallow-water system with horizontal diffusion, Rayleigh friction and external forcing is used to test the proposed method through identical-twin numerical experiments. The control run represents a typical mesoscale case with energy cascaded in two ways (upscale and downscale). The influence of model error and resolution of the analysis grid on the assimilated results is examined. Results show that when model error is small or moderate, the assimilated wind and geopotential fields correlate well with the true fields. When model error is large, the proposed method can still recover a large portion of small-scale motions which are not resolved by observations. Model error can lead to the generation of spurious small-scale gravity waves because of the inconsistency between model and observations. Numerical experiments show that bounding wind divergence and its time tendency can considerably suppress high-frequency spurious gravity waves and improve the assimilated results.

1. Introduction

With the advent of new observing systems, such as NEXRAD network (Telesetsky 1995) and GOES-NEXT satellites (Menzel and Purdom 1994), the atmosphere can be consistently monitored with very high temporal and spatial resolution. For example, GOES-I snapshots the United Sates continent every hour and every 15 minutes during severe weather episodes in a mesoscale area. Unfortunately, data provided by such observing systems are incomplete in terms of atmospheric wind and thermodynamic parameters. Analysis is needed to reconstruct four-dimensional fields of atmospheric state variables. To better understand synoptic and climatic phenomena, analysis has to be performed continuously over a long period of time from hours to years. Several data assimilation methods have been developed and are still undergoing intensive investigations, some of these are the method of four-dimensional variational data assimilation (4D-Var) (e.g., Talagrand 1981; Navon et al. 1992; Courtier et al. 1993; Verlinde and Cotton 1993; Zupanski 1993; Courtier et al. 1994; Sun and Crook 1994; Zou et al. 1995; Järvinen et al. 1996; Xu 1996a,b; Yang and Xu 1996), nudging assimilation (e.g., Anthes 1974; Walko et al. 1989; Staufer and Seaman 1990), Kalman (KF) and extended Kalman filters (EKF) (e.g., Kalman 1960; Ghil et al. 1982; Cohn and Parrish 1991; Daley and Menard 1993; Cohn et al. 1994), and intermittent assimilation (e.g., Mahfouf 1991; Ruggiero et al. 1996).

Among these data assimilation methods, 4D-Var technique is the most promising method as it obtains an optimal model initial condition by minimizing a cost function that measures the misfit between model forecasts and observations over a period of time. However, 4D-Var is not suitable for assimilating data continuously over a long period (say, one week) because the assimilation period of 4D-Var is controlled by many factors such as accuracy of linearizations, model error, weather predictability, and the huge computational cost involved. For example, using a barotropic β-plane model and without introducing model and observational errors, Tanguay et al. (1995) showed that for a given model resolution, exact initial conditions cannot be recovered if the assimilation period is much larger than the validity timescale (of an upper limit of about 3 days) of the tangent linear model. Motivated by the capability of 4D-Var method for assimilating data from diverse sources, we describe an assimilation algorithm [referred to as gradient-descent data assimilation (GDDA)] that can continuously assimilate data in an infinite time period using the concepts of 4D-Var technique and Kalman filter or nudging assimilation.

The computation of a cost function and its gradient is very computationally expensive, this is the major factor that makes operational application of 4D-Var method very difficult. To reduce total computational cost, Courtier et al. (1994) proposed an incremental approach as an approximation to the full 4D-Var problem. This strategy obtains the analyzed increment of initial value for a model through a 4D-Var analysis performed on a coarse grid with a linearized prediction model. To reduce the peak computation task, Järvinen et al. (1996) proposed a scheme of quasi-continuous variational data assimilation. This scheme divides the 4D-Var assimilation task into smaller parts of which only the last one is time and memory critical, and can reduce about 40% of peak computational work for a 24-h assimilation using primitive equations and real observations. While maintaining the capability of assimilating data from a variety of observation systems like Doppler radar and GOES-next series, the proposed GDDA scheme is also designed to reduce both total and peak computational tasks when compared with the standard 4D-Var method.

This paper is organized as follows. Section 2 shows how the GDDA scheme is formulated by using a cost functional and its gradient (calculated with an adjoint method). To examine the efficiency of the GDDA scheme, identical-twin numerical experiments were performed using a two-dimensional shallow water model, which is described in Section 3. The proposed method includes forecast and analysis steps. Different forecast models (or model grid) can be used in the forecast and analysis steps. Subsection 4b reports the results assimilated with different analysis grids. The proposed method does not explicitly handle model error, which is one of the challenging issues faced by the 4D-Var technique (Thépaut et al. 1993; Ménard and Daley 1996). The influence of model error on the assimilated results is discussed in Subsection 4c. In Section 5, we investigate methods to damp spurious gravity waves in the assimilated fields, which can be caused by the inconsistency between model and data. In Section 6, we summarize the findings and results of this paper.

2. The algorithm of gradient-descent data assimilation

a. *Basic formulation*

Without loss of generality and for the sake of simplicity, we assume the observation grid is the same as the model grid, and the observed variables are model state variables. Like the KF (or EKF) and nudging method, the proposed method can be written as:

Forecast step:

$$\mathbf{X}_n^f = \mathbf{F}_{n-1}(\mathbf{X}_{n-1}^a), \tag{1}$$

Analysis step:

$$\mathbf{X}_n^a = \mathbf{X}_n^f + \mathbf{d}_n, \tag{2}$$

where $\mathbf{X}$ represents model state variable, $\mathbf{F}$ is the forecast model, superscripts f and a indicate forecasted and analyzed values respectively, subscript n indicates time step, and d is the analysis increment (i.e., analyzed value minus forecasted value). In the KF and nudging methods, analysis increment $\mathbf{d}_n$ is constructed as a linear function of the misfit between observation (denoted with superscript o) and forecast:

$$\mathbf{d}_n = -\mathsf{G}_n(\mathbf{X}_n^f - \mathbf{X}_n^o), \tag{3}$$

where G_n is a weighting matrix (gain matrix in KF). The analysis increment in (3) can generally be written as follows

$$\mathbf{d}_n = -\mathsf{G}_n \nabla J(\mathbf{X}_n^f), \tag{4}$$

where ∇ is a gradient operator with respect to $\mathbf{X}_n^f$, and J is a cost function measuring the discrepancy between the model forecast and observations at time t_n,

$$J(\mathbf{X}_n^f) = \frac{1}{2}(\mathbf{X}_n^f - \mathbf{X}_n^o)^T(\mathbf{X}_n^f - \mathbf{X}_n^o), \tag{5}$$

where T stands for transpose. In (4), the analysis increment is proportional to the gradient of a cost function. Since a cost function can be constructed with considerable freedom depending upon the problems of interest, this makes it feasible to assimilate a variety of observations that are complicated functions of model state, such as satellite-derived radiances, and to use future data in the analysis steps.

An analysis that uses future data to analyze a current model state is termed retrospective analysis (e.g., Cohn et al. 1994). Using a fixed-lag Kalman smoother and a shallow-water model, Cohn et al. (1994) demonstrated that retrospective analysis is very efficient in reconstructing model states. To make use of future data in the analysis steps, the cost function J is redefined in a form commonly used in 4D-Var,

$$J(\mathbf{X}_n^f) = \sum_{t'=t_n}^{t_n+\tau} [\tilde{\mathbf{X}}^f(t') - \mathbf{X}^o(t')]^T W^{-1} [\tilde{\mathbf{X}}^f(t') - \mathbf{X}^o(t')], \tag{6}$$

where τ is the length of a time period after t_n, W is the covariance matrix of observational and model errors, and $\tilde{\mathbf{X}}^f$ is the model state predicted with

$$\tilde{\mathbf{X}}_r^f = \tilde{\mathbf{F}}_{r-1}(\tilde{\mathbf{X}}_{r-1}^f), \tag{7}$$

starting from an initial condition

$$\tilde{\mathbf{X}}_r^f = \mathbf{X}_n^f, \qquad \text{when } r = 0, \tag{8}$$

where $\tilde{\mathbf{F}}$ is the forecast model used in the analysis steps. It can be different from the one used in the forecast steps. The gradient of J in (6) with respect to $\mathbf{X}_n^f$ can be calculated with the adjoint method used in 4D-Var. For 4D-Var, τ is termed as assimilation period (or time window). Since the proposed scheme can assimilate data indefinitely, to avoid confusion, in this paper τ is referred to as the assimilation period (time window) of future data. Also, in the remainder of this paper, the quoted cost function J is the one defined by (6). In this paper, a data assimilation algorithm formulated with (1), (2), (4) and (6) is referred to as a scheme of gradient-descent data assimilation (GDDA).

b. *Determination of the weighting matrix*

In this paper, G_n is assumed to be

$$\mathsf{G}_n = \alpha_n \mathsf{I}, \tag{9}$$

where I is an identity matrix, and α_n is a parameter which changes with time. As a necessary condition of a good analysis, G_n should make J decrease after an analysis step, i.e.,

$$J(\mathbf{X}_n^a) \leq J(\mathbf{X}_n^f). \tag{10}$$

According to (4), this condition requires α_n be positive and bounds α_n to a finite value. From (4) and (9), one can estimate α_n as

$$\alpha_n \approx [J(\mathbf{X}_n^f) - J(\mathbf{X}_n^a)] / \parallel \nabla J \parallel^2 \approx C[J(\mathbf{X}_n^f) - J_{min}] / \parallel \nabla J \parallel^2, \tag{11}$$

where C is a proportional parameter, and J_{min} is the minimum value of J. Since (11) is an approximate formula, there is no need to find the exact minimum of J. In practical applications, J_{min} can be estimated based on observational errors in data. In this paper, J_{min} is set to zero.

Theoretically speaking, if the observed data contain sufficient information to give an optimal solution (by minimizing J) that are very close to the "true" state, then the optimal solution can be directly taken as the analyzed value — $\mathbf{X}_n^a$. However, this is not the case for the real situation. For this reason we require the cost function decreases only a fraction after an analysis step. In general, one may expect that $0 < C < 1$.

For a four-dimensional assimilation problem, calculating the gradient of J is very time-consuming. It is not practical to compute the gradient of J at every analysis step. In this paper, the gradient of J is updated periodically with a time period of t_h. According to (4), (9), and (11), the analysis increment is thus modified to

$$\mathbf{d}_n = -C[J(\mathbf{X}_n^f) - J_{min}]\nabla J / \parallel \nabla J \parallel^2, \qquad \text{for } n = 0, M, 2M, ..., \tag{12}$$

$$\mathbf{d}_{n+m} = \mathbf{d}_n \Big(1 - \beta \frac{t_{n+m} - t_n}{t_h}\Big), \qquad \text{for } m = 1, ..., M-1, \tag{13}$$

where $M = [t_h/\Delta t]$, β is a positive parameter. In (13), the term in the braces is introduced to reduce the norm of the analysis increment because cost function J decreases in the updating cycles. Cyclical updating of the analysis increment acts like a low-pass filter (in the time domain) for the analysis increment (Bloom et al. 1996) with oscillations of periods smaller than t_h being heavily damped. However, if small-scale features exist in the analysis increment, high-frequency oscillation can still be generated in the forecast fields through the spatial forcing of the analysis increment. For C, it is assigned a value so that J may approach to J_{min} after a period of t_h, i.e.,

$$C = C_o \Delta t / t_h, \tag{14}$$

where C_o is a proportional parameter near unity in magnitude.

To give a clear picture of the GDDA method, the GDDA method is schematically illustrated in Fig. 1. In Fig. 1, an analysis is performed at time $t = t_0$. The analysis is done through two steps: 1)

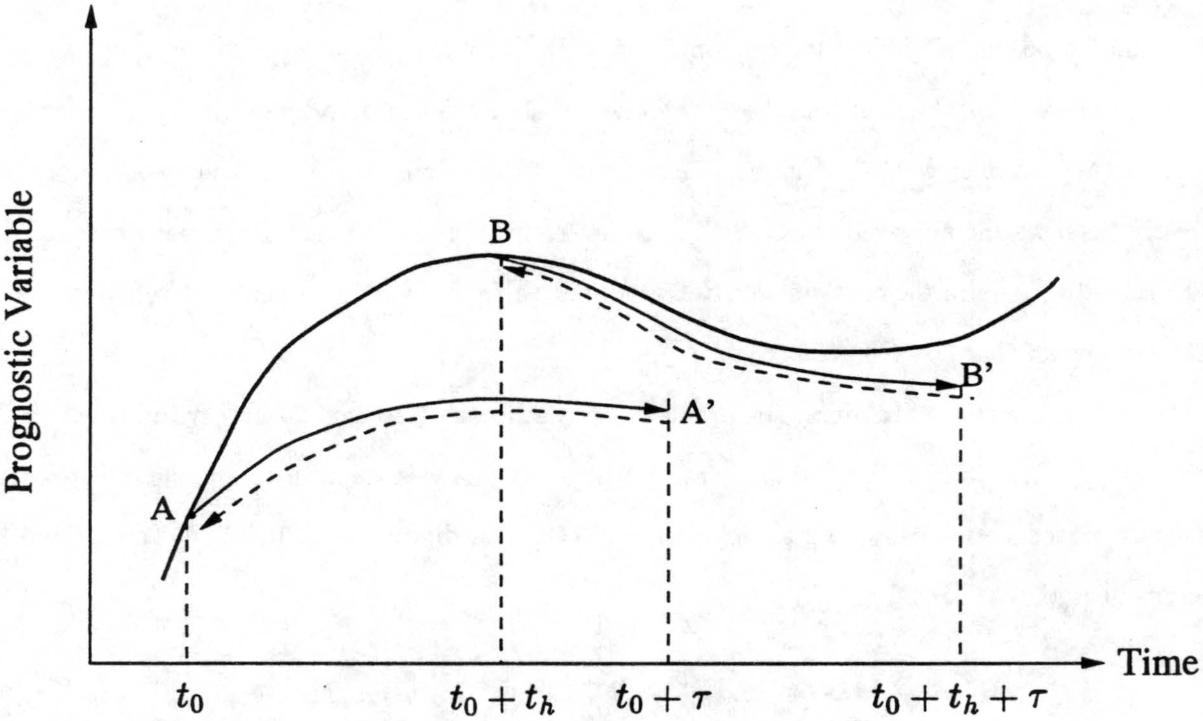


Figure 1: Schematic diagram of the GDDA scheme. At time t_0, analysis increment is updated by performing 4D-Var analysis (Forward integration – thin solid line AA', backward adjoint integration – dotted line under AA'). Forecast (thick solid line) continues until time $t_0 + t_h$. At time $t_0 + t_h$, analysis increment is updated again through 4D-Var analysis (BB') and data assimilation continues.

integrate the forecast model used in the analysis step for a period of time τ (solid line AA′), and 2) integrate the adjoint model backward (dash line under AA′). After these two steps, the gradient of J, and therefore the analysis increment can be obtained. Then data assimilation continues without updating the gradient of J until $t = t_0 + t_h$. In the time interval $(t_0,\ t_0 + t_h)$, the forecast model state follows curve AB due to the modification of analysis, otherwise it will evolve along AA′. At time $t_0 + t_h$ a new analysis begins at point B and ∇J is updated. Data assimilation is thus executed continuously by updating ∇J cyclically at time $t = t_0 + \ell t_h$ $(\ell = 0, 1, 2, ...)$.

The computation efficiency of the GDDA method can be estimated as follows. Assume that adjoint integration costs as about 2 times of CPU time as the forward integration, it can be proved that the total CPU time (CPU_t) cost by the proposed method can be approximated as

$$CPU_t \approx CPU_s + 3\tau/t_h CPU_s', \tag{15}$$

where CPU_s is the CPU time cost by the forecast model used in the forecast steps, and CPU_s' is the CPU time required by a single integration of the forecast model used in the analysis steps. CPU_s' equals CPU_s if the forecast models used in the forecast and analysis steps are the same. CPU_s' will be much smaller than CPU_s if the forecast model used the analysis steps has a lower grid/temporal resolution (or simpler parameterization schemes of microphysics) than the one used in the forecast steps.

It is also worth pointing out that model variables should be properly scaled when calculating the gradient of J as is done in many 4D-Var schemes. Otherwise, the analysis increment will seriously deviate from the optimal descent direction of J and the efficiency of the proposed algorithm will be low.

3. The shallow-water model and statistics

a. *Experimental design*

The prototype test bed model is a two-dimensional shallow-water model with horizontal diffusion, Rayleigh friction and external forcing:

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} + fv - \frac{\partial \phi}{\partial x} + K_m D_m u - c_r u + F_x \tag{16}$$

$$\frac{\partial v}{\partial t} = -u\frac{\partial v}{\partial x} - v\frac{\partial v}{\partial y} - fu - \frac{\partial \phi}{\partial y} + K_m D_m v - c_r v + F_y \tag{17}$$

$$\frac{\partial \phi}{\partial t} = -u\frac{\partial \phi}{\partial x} - v\frac{\partial \phi}{\partial y} - \phi\Big(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\Big) \tag{18}$$

where u, v are eastward and northward wind components respectively; ϕ is the geopotential height; f is the Coriolis parameter ($f = 1.0^{-4}s^{-1}$); K_m is the horizontal diffusion coefficient, D_m is an iterated Laplacian (e.g., Tanguay et al. 1995), $D_m = (\nabla^2)^\mu$ (unless otherwise specified, $\mu = 1$); c_r is the Rayleigh friction coefficient ($c_r = 1/14$ day^{-1}), F_x and F_y represent external forces along the x and y directions, respectively.

Double periodic boundary conditions are assumed for the model domain. The model is discretized with a potential-enstrophy conserving method (Sadourny 1975; Washington and Parkinson 1986) and is initialized with

$$u = -\frac{\partial \psi}{\partial y} + u_o, \tag{19}$$

$$v = \frac{\partial \psi}{\partial x} - A\cos(1 + 4\pi x/L)\cos(-1 + 4\pi y/L), \tag{20}$$

where u_o is the mean (basic) flow, A=2ms^{-1}, L is the model domain dimension, $0 \leq x \leq L$ and $0 \leq y \leq L$, and ψ is defined by

$$\psi = B\sin(2\pi k_x^\psi x/L)\sin(2\pi k_y^\psi y/L), \tag{21}$$

where k_x^ψ and k_y^ψ are dimensionless wave-numbers of ψ, $k_x^\psi = k_y^\psi = 2$, and $B = 4 \times 10^6$m^{2}s^{-1}. The initial value of geopotential height is given by

$$\phi = \phi_o + f\psi, \quad \text{at } t = 0, \tag{22}$$

where ϕ_o is the mean geopotential height; $\phi_o = 7500m^2s^{-2}$. The model domain is set to a mesoscale area, $1600 \times 1600km^2$. The number of grid points ($N_x \times N_y$) is set to 32×32. Accordingly, the grid spacing (d) is 50km. For this grid spacing, K is set to $4 \times 10^4 m^2s^{-1}$, and Δt is set to 100s. In this 2D shallow water system, energy cascades in two ways in wavenumber space. Small-scale features with wavelength smaller than or equaling $4d$ are fairly quickly (with dissipation time scales less than 45h) dissipated by horizontal diffusion. In order for the flow to maintain a wide energy spectrum in the assimilation period, the external forcing is set as

$$F_x = \rho A_x \cos[2\pi k_x(x - u_o t)/L + 2\pi k_y y/L], \tag{23}$$

$$F_y = f u_o + \rho A_y \cos[2\pi k_x(x - u_o t)/L + 2\pi k_y y/L], \tag{24}$$

where ρ is a multiplier used to specify the strength of external forcing, $A_x = 8.10 \times 10^{-4}ms^{-2}$, $A_y = 6.94 \times 10^{-4}ms^{-2}$, and k_x and k_y are the dimensionless wave-numbers, $k_x = 4$, and $k_y = 3$. Unless otherwise stated, ρ is set to 1.0. The first term in (24) balances the Coriolis force of the mean flow. In dimensionless wavenumber space, the external forcing has two components with wave-numbers of zero and 5. Note that initial u, v and ϕ are not in a balanced state. Consequently, gravity waves exist in the model solution. This is to let the model to represent asynoptic processes occurring on mesoscales (Holton 1992, p.266).

Although this shallow-water system is driven by external forcing (23) and (24), numerical integrations (not shown) demonstrate that the model solution is very sensitive to initial conditions. Therefore, the above-described modeling system is appropriate for testing data assimilation schemes.

We assume that the observed quantities are u and v. The observed u and v are simulated by adding a random noise to the "true" values of u and v, which is generated by running the shallow-water model from a given initial condition (control run). The random noise has a Gaussian distribution with a standard deviation $\sigma = 1.0\text{ms}^{-1}$. Data noise is assumed uncorrelated in space.

Data coverage (denoted as R) is expressed a fraction of total model grid points and observation stations are randomly distributed on the model grid. R is assumed to be 6.2%. This corresponds to an average spatial resolution of 200km for the observation stations. In the wavenumber space associated with the model grid, these data can only resolve spatial structures with wave-numbers of 0-3 with reasonable accuracy.

Data temporal resolution (denoted as t_d) is set to 3h. The assimilation time window of future data (τ) is set to 4h. Because t_d=3h, the assimilation time interval $(t_n, t_n + \tau)$ in (6) may cover either one or two levels of observations, depending upon t_n. Accordingly, the actual assimilation time window of future data may vary between 1h and 4h, with an average of 2.5h.

The initial values of analyzed u, v, and ϕ are the domain averages of the true initial values. The gradient of J is updated every 1200s; $t_h = 1200s$. Because of data errors and low data coverage, the gradient field of J contains short-wave components that cannot properly be resolved by the model grid and may cause wave-aliasing problems. In 4D-Var numerical experiments without data noise, Tanguay *et al.* (1995) showed that during early iterations, spurious short-wave components appear in retrieved initial fields, implying spurious short-wave components exist in the gradient field of the cost function when the cost function is not very close to its minimum. In this paper, when the analysis increment is updated, it is smoothed with a five-point filter, which can be written as

$$\tilde{q}_{i,j} = q_{i,j} + \gamma(q_{i+1,j} + q_{i-1,j} + q_{i,j+1} + q_{i,j-1} - 4q_{i,j}), \tag{25}$$

where q can represent any one of the analysis increment fields of wind and geopotential height, $\tilde{q}$ is the filtered value of q, γ is the filter coefficient. In order to filter out only those components having very short wavelength, γ is set to a small value of 0.05, and (25) is iterated 3 times.

b. *Definitions of statistics*

In order to assess the quality of assimilated products, the following statistics are introduced for a variable q, which can represent any one of u, v and ϕ:

$$\mathrm{RMS}_q = \overline{(q^a - q^t)^2}^{1/2}, \tag{26}$$

$$\mathrm{CR}_q = \frac{\overline{(q^a - \overline{q^a})(q^t - \overline{q^t})}}{\overline{(q^a - \overline{q^a})^2}^{1/2}\,\overline{(q^t - \overline{q^t})^2}^{1/2}}, \tag{27}$$

where RMS_q is the root-mean square error (RMS) of q, CR_q is the correlation coefficient (CR) between assimilated and true values, q^t is the true value of q. The overbar represents an average taken over the model domain. For velocity field, we also define the rms error of wind (denoted as RMS_w) and the correlation coefficient (denoted as CR_w) between the assimilated and true winds as follows:

$$\mathrm{RMS}_w = \overline{\parallel \mathbf{V^a} - \mathbf{V^t} \parallel^2}^{1/2} = \left(\mathrm{RMS}_u^2 + \mathrm{RMS}_v^2\right)^{1/2}, \tag{28}$$

$$\mathrm{CR}_w = \frac{\overline{(\mathbf{V^a} - \overline{\mathbf{V^a}}) \cdot (\mathbf{V^t} - \overline{\mathbf{V^t}})}}{\overline{\parallel \mathbf{V^a} - \overline{\mathbf{V^a}} \parallel^2}^{1/2}\,\overline{\parallel \mathbf{V^t} - \overline{\mathbf{V^t}} \parallel^2}^{1/2}}, \tag{29}$$

where $\mathbf{V}$ is the wind vector, $\mathbf{V} = (\mathbf{u}, \mathbf{v})$.

More detailed comparison can be made in the dimensionless wavenumber (k) space associated with the model grid. The correlation coefficient between the assimilated and true fields at wave number k is defined as

$$r_k \equiv \frac{\overline{q^a(k)[q^t(k)]^*}}{\overline{|q^a(k)|^2}^{1/2}\,\overline{|q^t(k)|^2}^{1/2}} = \frac{\sum\limits_{k-1/2 \le |\mathbf{k}'| \le k+1/2} \hat{q}^a(\mathbf{k}')[\hat{q}^t(\mathbf{k}')]^*}{[Q^a(k)]^{1/2}[Q^t(k)]^{1/2}}, \tag{30}$$

where $\mathbf{k}$ is the wave vector, $k = |\mathbf{k}|$; $\hat{q}(\mathbf{k})$ indicates the Fourier coefficients of $q - \bar{q}$, a star indicates the complex conjugate, $q(k)$ is the sum of $\hat{q}(\mathbf{k}')$ in k-space rings of unit thickness centered on wavenumber k,

$$q(k) = \sum_{k-1/2 \le |\mathbf{k}'| \le k+1/2} \hat{q}(\mathbf{k}') \exp\left(\frac{2\pi}{L}\hat{\imath}\mathbf{k}' \cdot \mathbf{x}\right), \tag{31}$$

where $\hat{\imath} = \sqrt{-1}$. In (30), $Q(k)$ is the power (energy) spectrum of $q - \bar{q}$,

$$Q(k) = \overline{q(k)q^*(k)} = \sum_{k-1/2 \le |\mathbf{k}'| \le k+1/2} \hat{q}(\mathbf{k}')\hat{q}^*(\mathbf{k}'). \tag{32}$$

It can be proven that CR_q is related to r_k through

$$\mathrm{CR}_q = \sum_k \tilde{r}_k, \tag{33}$$

where $\tilde{r}_k$ is defined as

$$\tilde{r}_k = r_k[\tilde{Q}^a(k)\tilde{Q}^t(k)]^{1/2}, \tag{34}$$

and $\tilde{Q}(k)$ is the normalized energy spectrum of $q - \bar{q}$,

$$\tilde{Q}(k) = \frac{Q(k)}{\sum\limits_k Q(k)}. \tag{35}$$

We refer to $\tilde{r}_k$ as the spectrum (spectral density) of the correlation coefficient CR_q. Equations (33)–(34) show that the contribution of a component of wavenumber k to the total correlation coefficient CR_q is weighted by the square-root of normalized energy spectra of the two fields. If the spatial pattern of the assimilated field is identical to that of the true field, both r_k and CR_q equal unity and $\tilde{Q}^a(k)$ equals $\tilde{Q}^t(k)$. In that case, (34) becomes

$$\tilde{r}_k = \tilde{Q}^t. \tag{36}$$

Equation (36) indicates that for two perfectly correlated fields, the spectrum of correlation coefficient is the same as the normalized power spectrum (of true field). In this paper, we refer to an assimilation as a perfect assimilation when the assimilated fields are the same as the true ones.

4. Numerical results

In this section, identical-twin numerical experiments are performed to test the proposed method. Sensitivities of the assimilated results to C_o and β are presented and the affect of model error on assimilated results is examined. Results using different grids in the forecast and analysis steps are also presented. In the remainder of the paper, the model grid used the forecast steps is referred to as forecast grid and that used in the analysis steps is referred to as analysis grid (denoted as $N_x^a \times N_y^a$). The forecast grid is the same as the model grid used to produce true fields and has 32×32 grid points. If not specifically stated, the analysis grid is the same as the forecast grid.

a. *Control run and assimilation without model error*

The normalized power spectra of true fields are shown in Fig. 2. For the case studied here, the power spectral density of u, v and ϕ are negligible for wave-numbers greater than 10. Therefore, we plot the spectrum of a quantity for $k \leq 10$, although the maximum resolvable wavenumber is $16\sqrt{2}$ for a forecast grid of 32×32 grid points. Fig. 2 shows that energy cascades in two directions: downscale and upscale. At the initial time, the spectra of perturbation winds (wind minus the basic flow) and perturbation geopotential height $(\phi - \phi_o)$ concentrate at wavenumber 3 [not shown, see (19)–(22)]. With the evolution of time, the power spectra of true fields broaden. At 24h, the power spectra of $u - u_o$, v and $\phi - \phi_o$ cover more than one wavenumber, especially the spectrum of $\phi - \phi_o$, which has a significant part distributed at wave-numbers different from wavenumber 3. At 120h, the power spectra of the true fields spread in the range $k \in [1, 7]$, and the spectral density of $\phi - \phi_o$ is fairly large at $k = 1$ and 2, indicating that part of the small-scale kinetic energy is transferred to large-scale geopotential energy.

The upscale cascading of energy raises a challenging issue for data assimilation (Tanguay et al. 1995) since small-scale flow features cannot be retrieved from observations on large-scales. It is possible, however, that small-scale features can be inferred through the inherent nonlinear dynamic relationship between small-scale and large-scale flows and/or (horizontal) advection of information. The following results show that this is achievable for the shallow-water system investigated in this paper.

Fig. 3 shows the RMS and CR of the assimilated results obtained with $C_o = 1$ and $\beta = 0.8$. It can be seen in Fig. 3 that the root-mean square errors (RMS) of wind and geopotential height decrease considerably with the evolution of time. RMS_u and RMS_v are larger than 10ms^{-1} at the initial time. After about a 40h assimilation period, they drop below 1.0ms^{-1}, which equals the standard deviation of the observed u and v. At 120h, RMS_u and RMS_v are 0.3m/s, indicating that the assimilated fields can be more accurate than the observed values. Consistently, the improvement of the assimilated results with time is reflected in the improved correlation between the assimilated and true fields. At 120h, CR reaches 0.99 or larger for all assimilated fields. As one can be seen in Fig. 3, ϕ is not

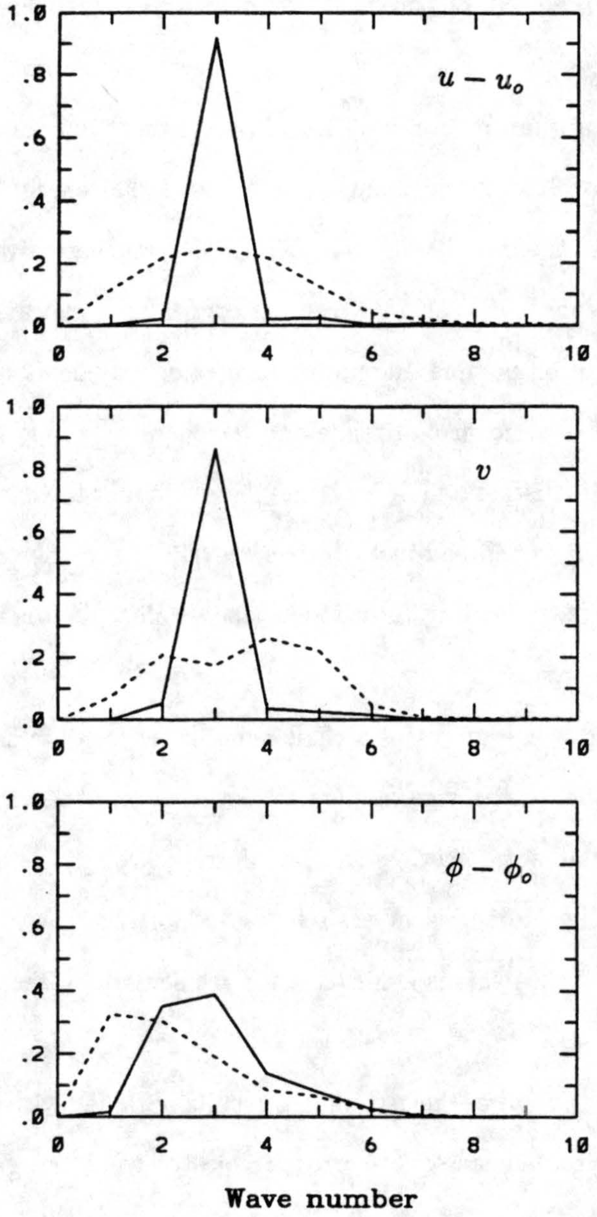
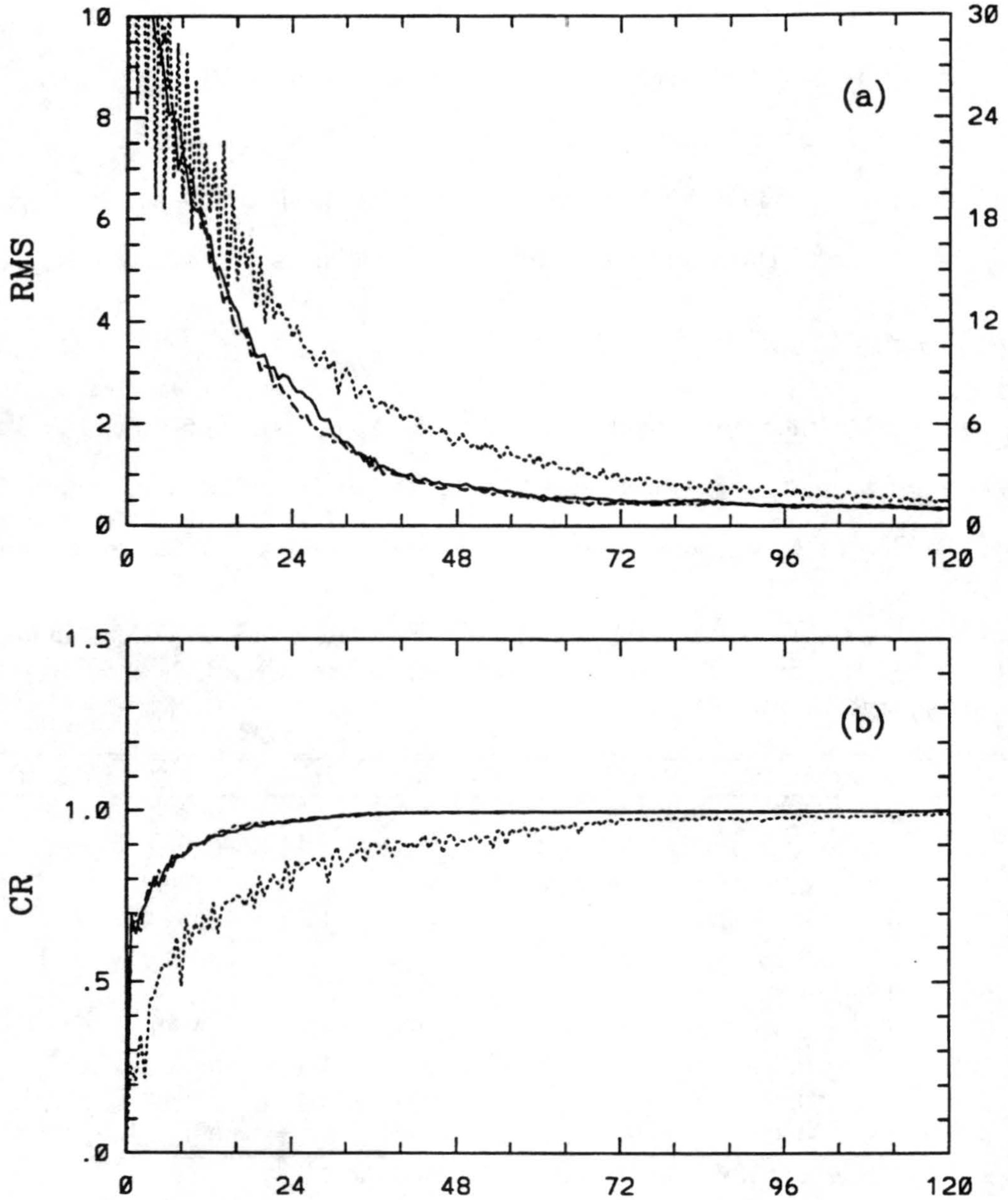


Figure 2: Normalized power spectra of perturbations of true wind ($u - u_o$, v) and geopotential height ($\phi - \phi_o$) at time t=24h (solid line) and 120h (dotted line).



time (hr)

Figure 3: Root-mean square error (RMS) of the assimilated fields and correlation coefficients (CR) between the assimilated and true fields. Curves are plotted with a time interval of 40min. Solid line — u, dash-dotted line — v, dotted line — ϕ. In panel (a), the unit of RMS is ms^{-1} for the wind (left labels), and $10m^2s^{-2}$ for the geopotential height (right labels). $C_o = 1$, and $\beta = 0.8$.

recovered as quickly as the wind, this is because of the high-frequency oscillation (with a period as small as about 1h) of ϕ.

Results (not shown) show that when $t > 72$h, the spectra of correlation coefficients between the assimilated and true fields are very close to those for a perfect assimilation. The above results demonstrate that the full spectra of wind and geopotential height fields can be accurately recovered by assimilating wind data which can only accurately resolve spatial features of wave-numbers 0-3.

b. *Sensitivity to C_o and β*

It is impossible to determine, *a prior*, the optimal values of the parameters C_o and β. Therefore experiments were conducted to check the sensitivity of the assimilated results to C_o and β. Table 1 lists the RMS and CR of the assimilated results at time 120h and the time after which both RMS$_u$ and RMS$_v$ are below the standard deviation of errors in the observed wind (1.0ms^{-1}). As can be seen in Table 1, the assimilated results at t=120h are not very sensitive to C_o and β. When β is fixed at 0.0 or 0.8, RMS$_u$ and RMS$_v$ vary in the range of 0.3–0.5ms^{-1} when C_o is changed from 0.5 to 5. When $\beta = 0$, the assimilated geopotential height is a little more sensitive to C_o than the assimilated wind. CR$_\phi$ changes by 0.01 when C_o changes from 0.5 to 5.0, whereas CR$_u$ and CR$_v$

Table 1: Root-mean square error of the assimilated wind components (RMS$_u$ and RMS$_v$, in ms^{-1}) and geopotential height (RMS$_\phi$, in $10m^2s^{-2}$) and correlation coefficient (CR$_u$, CR$_v$ and CR$_\phi$) between the assimilated and true fields at time t=120h.

Experiment	C_o	β	RMS$_u$	CR$_u$	RMS$_v$	CR$_v$	RMS$_\phi$	CR$_\phi$	t^*_{uv} (h)
ES1	0.2	0.0	0.6	0.9947	0.6	0.9945	2.0	0.9789	112.0
ES2	0.5	0.0	0.3	0.9985	0.4	0.9982	1.4	0.9890	42.0
ES3	1.0	0.0	0.3	0.9989	0.3	0.9988	1.4	0.9892	37.0
ES4	2.0	0.0	0.3	0.9987	0.3	0.9986	1.6	0.9859	33.7
ES5	5.0	0.0	0.4	0.9980	0.4	0.9978	2.3	0.9729	36.3
ES6	0.2	0.8	1.8	0.9595	1.7	0.9600	3.2	0.9429	—
ES7	0.5	0.8	0.5	0.9972	0.5	0.9968	1.5	0.9878	60.0
ES8	1.0	0.8	0.3	0.9987	0.3	0.9985	1.4	0.9894	40.3
ES9	2.0	0.8	0.3	0.9989	0.3	0.9988	1.4	0.9896	36.7
ES10	5.0	0.8	0.3	0.9985	0.3	0.9985	1.7	0.9845	35.7

* t_{uv} — time after which both RMS$_u$ and RMS$_v$ are below 1.0ms^{-1}.

change by less than 0.05. Although the accuracy of the assimilated fields at 120h are not sensitive to C_o and β, the results in Table 1 do indicate that the convergence rate of the assimilated fields to the true ones is affected by C_o and β. Overall, the results in Table 1 show that the assimilated products are more sensitive to the factor of $C_o[1-\beta(t_{n+m}-t_n)/t_h]$ derived from (12)–(14). When this factor does not significantly deviate from unity, the assimilated fields converge quickly to the true ones and have high accuracy.

Because the cost function decreases in each updating cycle of its gradient, one may therefore expect the analysis increment also decreases. Based on this consideration and the results in Table 1, we set $C_o = 1$ and $\beta = 0.8$ in the remainder of this paper.

c. *Mixed use of models with different grid resolution*

Although the adjoint method is efficient in calculating the gradient of a cost function, the computational cost is very large when compared with a single forecast integration. Courtier et al. (1994) proposed to reduce the grid resolution of gradient fields so that computational cost can be reduced. This approach can be used in the proposed method. From (1), (2), (4), and (6)–(8), one can see that the forecast models used in the forecast and analysis steps can have different grids if the analysis increment (forecasted fields) is properly interpolated into the forecast (analysis) grid. Linear interpolation is used in mapping data from one grid to another.

Fig. 4 shows the RMS and CR of the assimilated fields obtained with $N_x^a \times N_y^a = 20 \times 20$. As can be seen in Fig. 4, the assimilated results are still very good. Over the period from 72h to 120h, RMS_u and RMS_v are close to the standard deviation (1.0ms^{-1}) of errors in the observed u and v, and CR_u and CR_v are larger than 0.98. Comparing Fig. 3 and 4, one can find that owing to the reduction of the resolution of the analysis grid, the results assimilated with $N_x^a \times N_y^a = 20 \times 20$ is degraded, especially the assimilated geopotential height. This degradation is caused by the truncation error in the forecast model used in the analysis steps. For a grid of 20×20 points, it cannot accurately resolve spatial patterns with wave-numbers greater than about 5, which exist in the true fields (see Fig. 2).

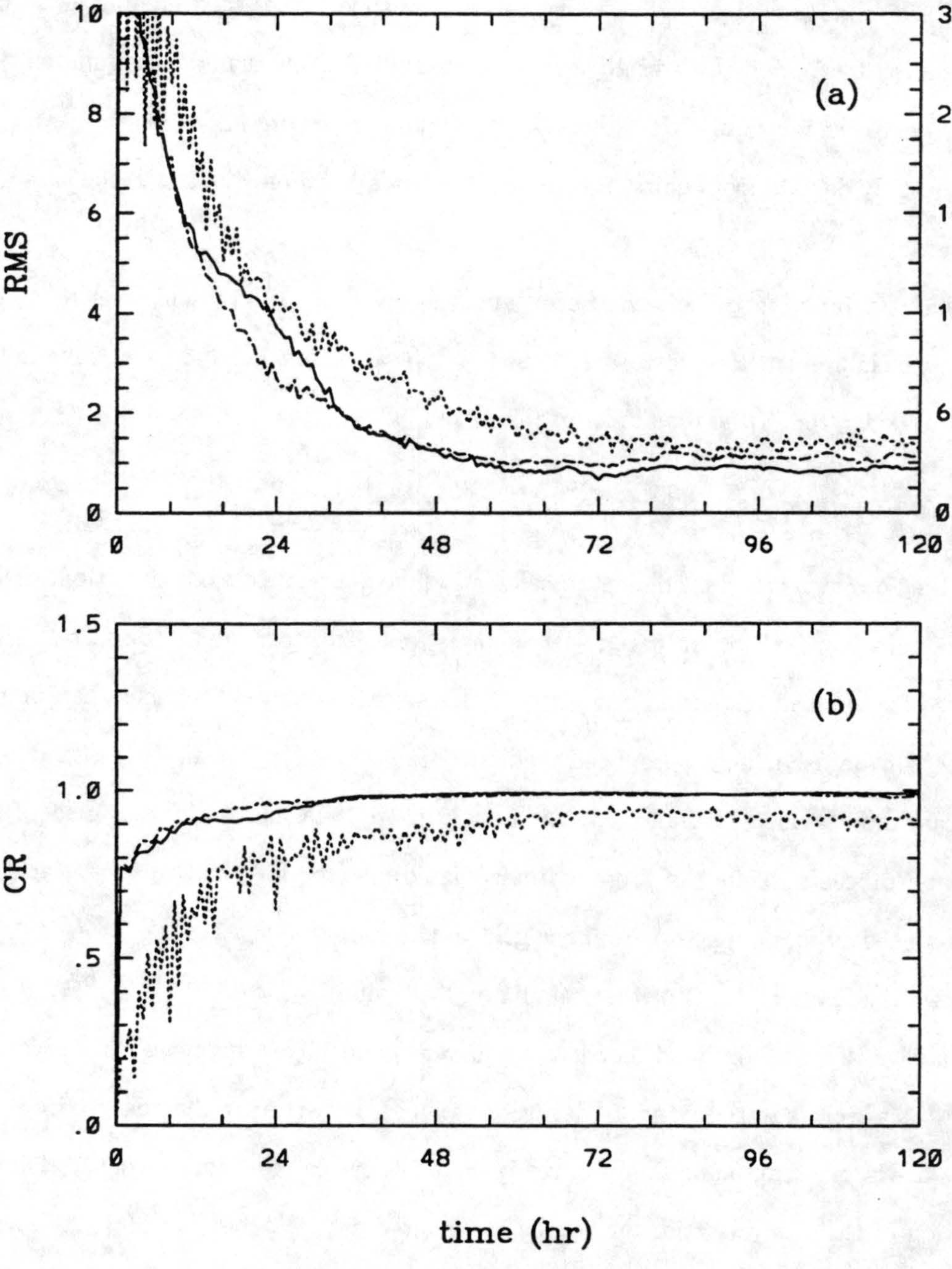


Figure 4: As in Fig. 3, but for $N_x^a \times N_y^a = 20 \times 20$.

When the resolution of the analysis grid is further reduced, truncation error is more severe and the quality of the assimilated results degrades further. Results show that when $N_x^a \times N_y^a = 16 \times 16$, CR$_\phi$ oscillates around 0.83 when $t > 72$h. In the next subsection, the influence of model error on the assimilated results is investigated in association with different $N_x^a \times N_y^a$.

d. *Affect of model error*

The atmosphere and other fluids (say, oceans) are continuums, the scales of motions in these fluid systems may range over several or tens of orders of magnitude. For example, the atmosphere may simultaneously contain eddies of sizes of 10^{-2}m and planetary waves of horizontal scale of 10^7m (Holton 1992, p.5). For the simulation of motions over such a wide range of scales, numerical models inevitably have model errors due to limited spatial resolution (truncation error) and inaccurate parameterizations of various subgrid processes. Dealing with model errors is still a challenging issue faced by 4D-Var techniques. In this subsection, the influence of model error on the assimilated fields with the proposed method was investigated by introducing an error in the external forcing in (16)–(17).

The shallow-water system defined by (16)–(24) with $\rho = 1$ is treated as an actual or a perfect system, and used to generate simulated observations of wind according to the method described in Section 3. With the other terms being the same as those in (16)–(18), (23) and (24), an imperfect forecast model used in the forecast and analysis steps is formulated by changing ρ to a value different from unity. The model error of an imperfect model can be assessed by RMS and CR of the forecasts produced by the imperfect model from the same initial condition as used in the perfect system. Fig. 5 shows the RMS and CR of a forecast made with $\rho = 0.5$. As one can see in Fig. 5, the wind and geopotential field simulated with the imperfect model exhibit very large error after 48h integration; RMS of u and v can be larger than about 8ms^{-1}. After 72h, the CRs of wind and geopotential height averaged over the period from 72h to 120h are only 0.02 and -0.03, respectively, indicating that u, v and ϕ forecasted by the imperfect model becomes totally uncorrelated with the true fields.

Fig. 6 displays the error statistics of the results assimilated with the imperfect model having

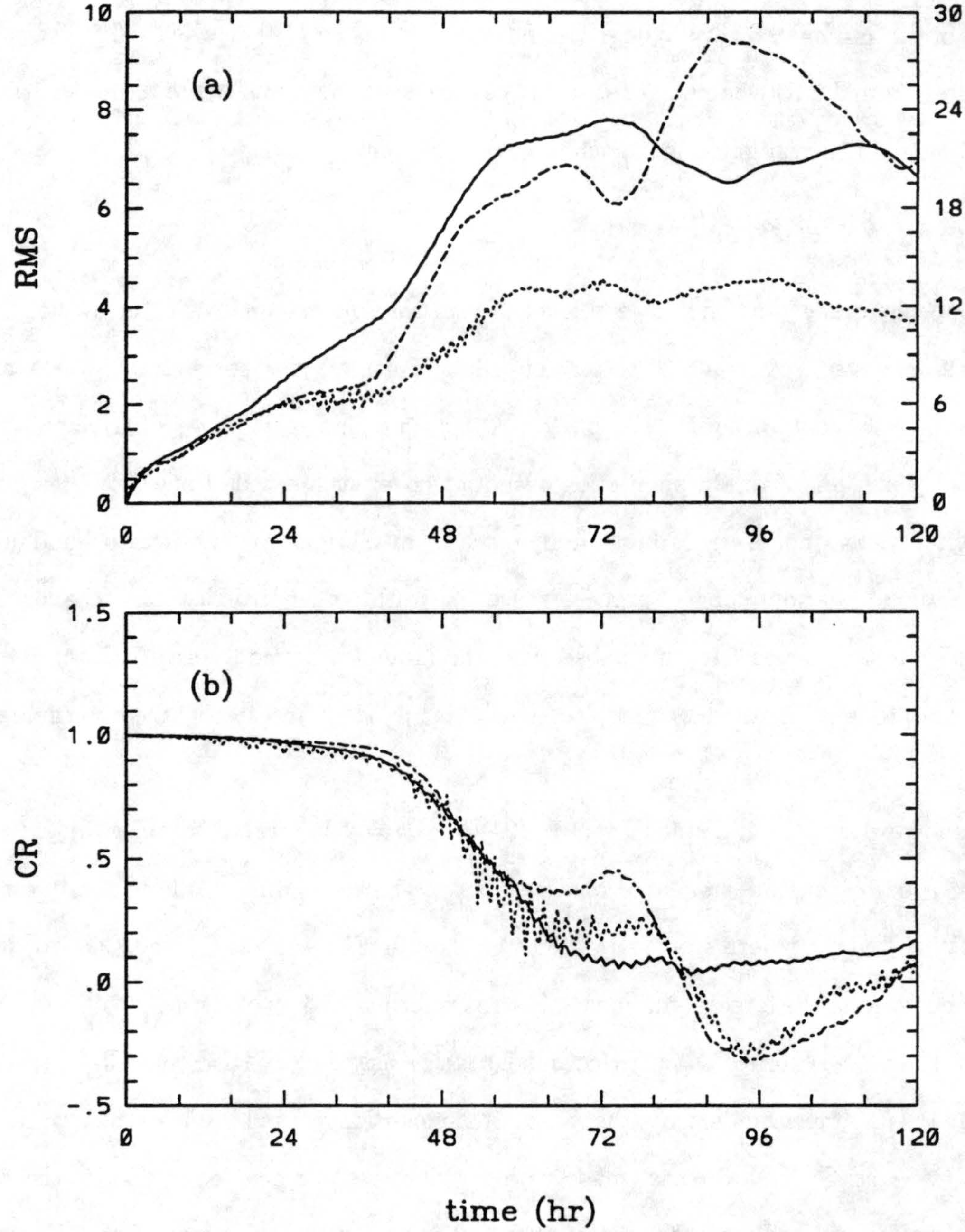


Figure 5: Root-mean square error (RMS) of a forecast made with an imperfect model having $\rho = 0.5$ and correlation coefficients (CR) between the forecast and true fields. Otherwise the same as in Fig. 3.

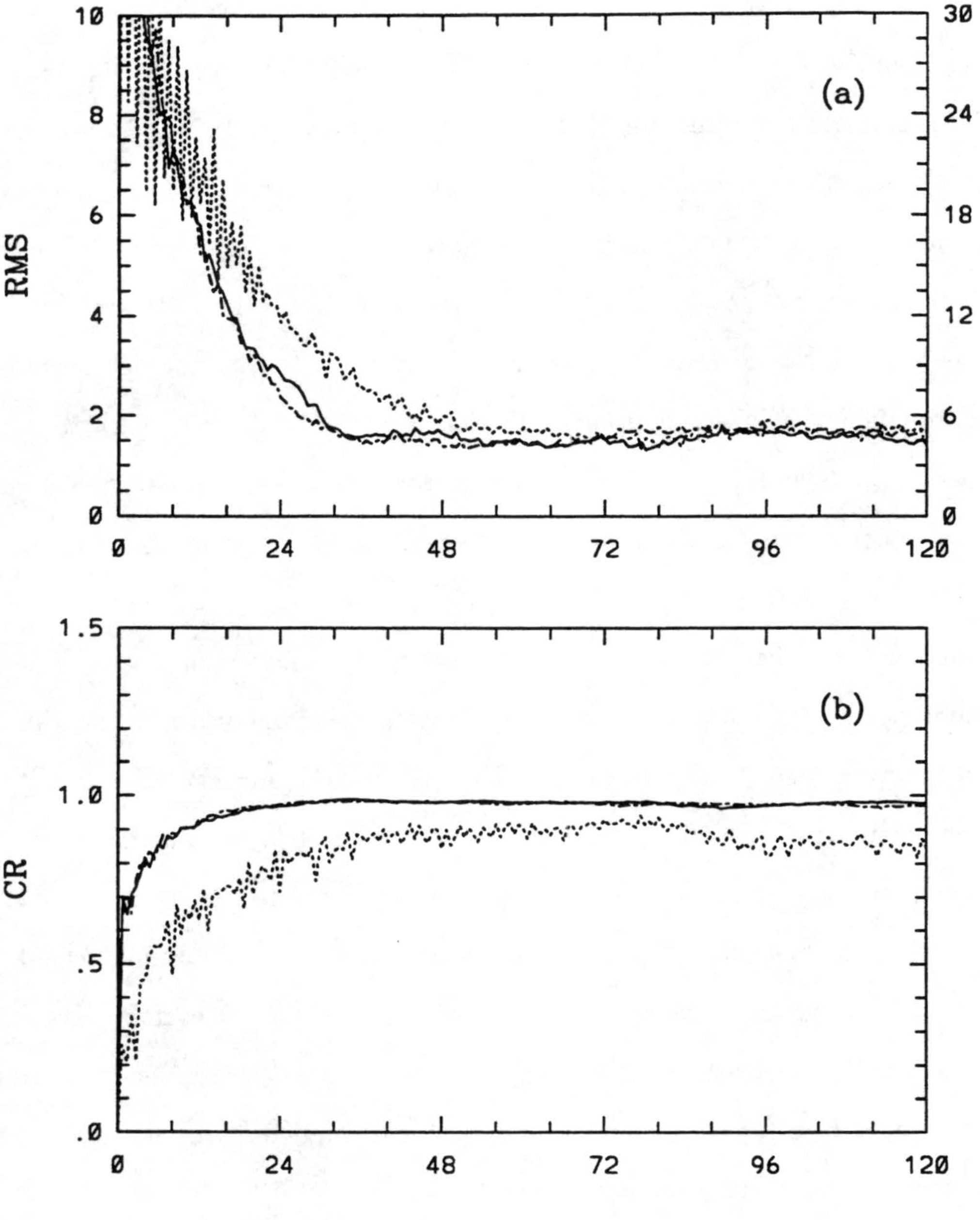


Figure 6: As in Fig. 3, but the assimilation is performed with an imperfect model having $\rho = 0.5$.

$\rho = 0.5$. As shown in Fig. 6, the assimilated wind and geopotential fields still correlate well with the true fields although the forecast model used in the assimilation has very large error. Comparing Fig. 3 and 6, one can find that CR$_\phi$ is affected more severely by model error than that of wind. In Fig. 6, CR$_\phi$ is smaller and fluctuates more violently with time than those of wind. This is caused by high-frequency oscillation of inertia-gravity waves, which result from the imbalance between the observed wind and forecasts made by the imperfect model. Since the imperfect model has error, a balanced flow in the perfect model (which is used to generate observations) is unbalanced in the imperfect model and gravity waves are triggered in the imperfect model after each analysis step. The prediction equation of ϕ has no horizontal diffusion to damp small-scale and high-frequency (with a time period of about 1h) components of ϕ and observations are made every 3h, therefore the assimilated ϕ is not as accurate as the assimilated wind. In Fig. 7, the assimilated ϕ has more small-scale features than the true ϕ, especially at t=120h.

Fig. 6 also demonstrates that unlike the results obtained with the perfect model (see Fig. 3), the results assimilated with an imperfect model cannot be progressively improved with time. In Fig. 6, the accuracy of the fields assimilated with the imperfect model does not increase and remains the same (statistically) with time when $t > 48$h. This was verified by a 30-day assimilation (not shown).

Fig. 8 shows the true wind field and that assimilated with an imperfect model having $\rho = 0$. As shown in Fig. 8, although the external forcing is zero in the imperfect model, small-scale features of the true velocity field are largely recovered in the assimilated field. This example demonstrates that even if large model error exists and energy cascades in two ways, the proposed method can still partially recover the small-scale (wavenumber $k = 4 - 6$) motions from data that can only approximately resolve motions with wave-numbers of 0-3.

Fig. 8 also shows that the recovered small-scale motions are not as strong as the true ones, therefore the contribution of small-scale motions to CR is lower than that of perfect assimilation. This can be seen in Fig. 9, in which the spectral densities of CR$_u$ and CR$_v$ are smaller than those of a perfect assimilation when wave-numbers are larger than 3. For the geopotential height, due to spurious high-frequency oscillation of small-scale gravity waves the spectral density of the correlation

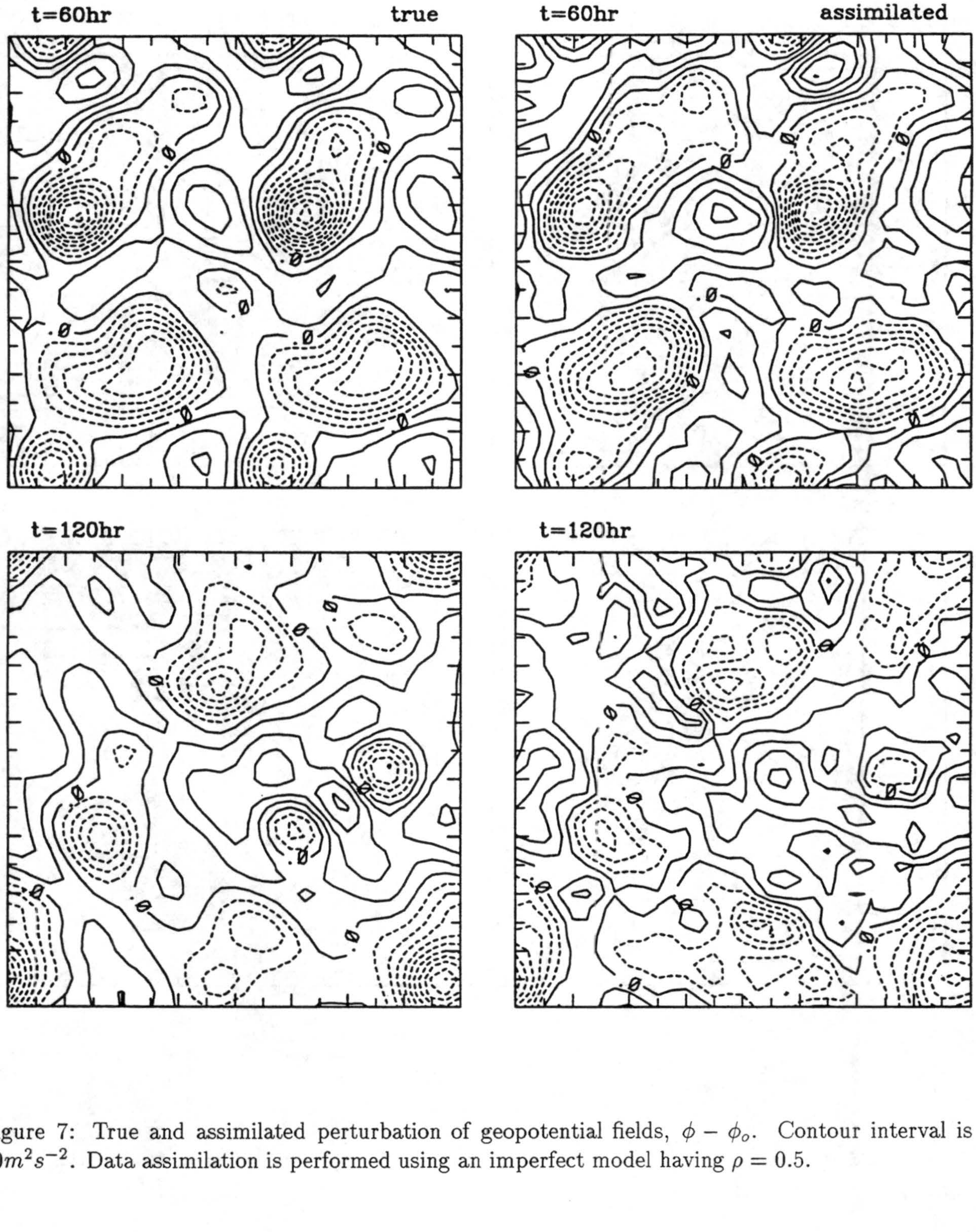


Figure 7: True and assimilated perturbation of geopotential fields, $\phi - \phi_o$. Contour interval is $50 m^2 s^{-2}$. Data assimilation is performed using an imperfect model having $\rho = 0.5$.

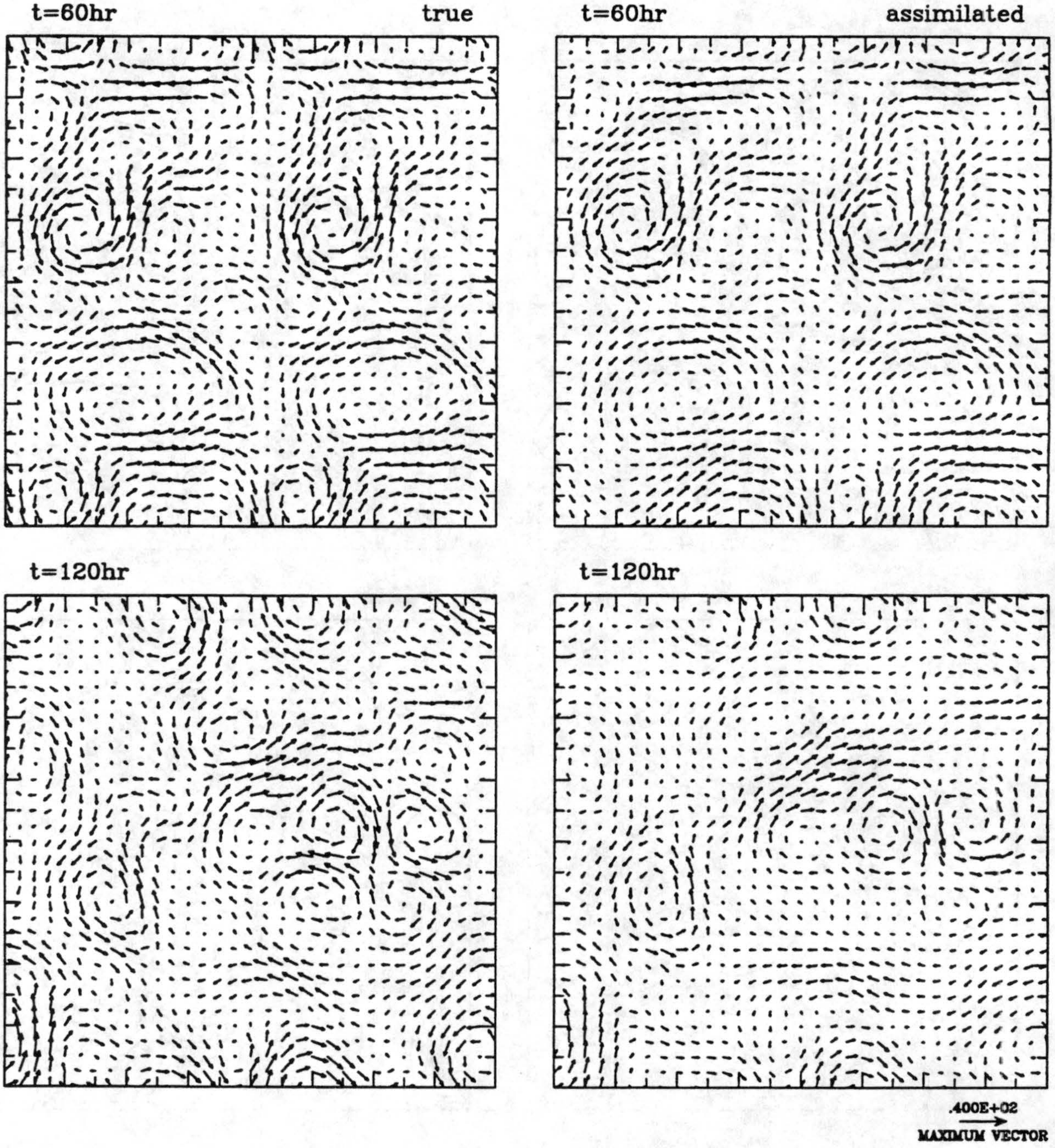


Figure 8: Vectors of true and assimilated perturbation wind $(u - u_o, v)$. Data assimilation is performed using an imperfect model having $\rho = 0$.

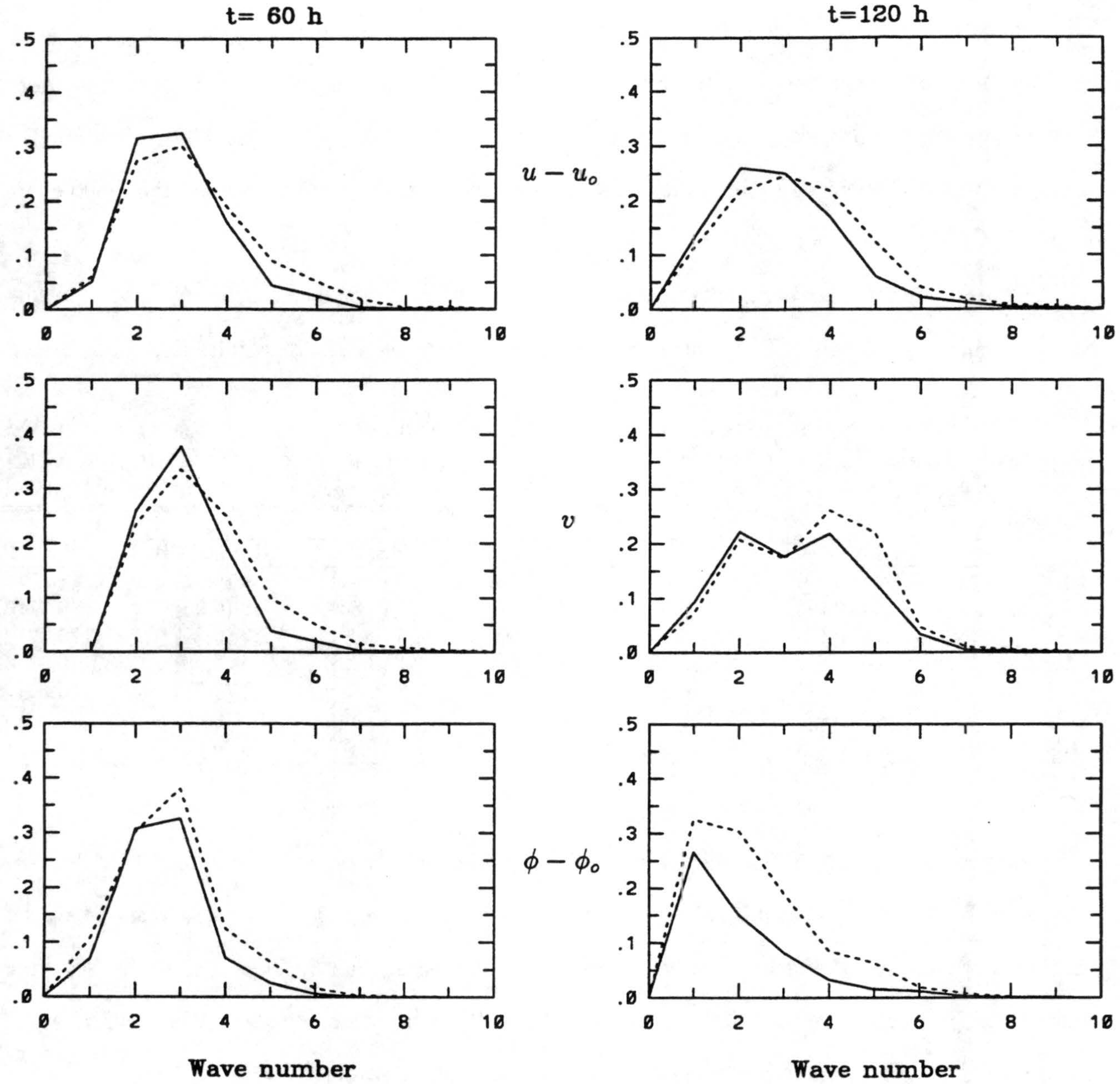


Figure 9: Spectra of correlation coefficients between the assimilated and true fields (solid line). The dotted line is the spectra of correlation coefficients for a perfect assimilation. Data assimilation is performed using an imperfect model having $\rho = 0$.

coefficient is about one half or one third of that of perfect assimilation when $t = 120$h and $k \geq 2$. This indicates that small-scale features of geopotential fields cannot be recovered when model error is too large. However, the assimilated results could be significantly improved (see the next section) if gravity waves are properly controlled. The good retrieval of ϕ at $t = 60$h results from a smaller model error as the external forcing is relatively smaller in the early stage of forecast than in the late stage of forecast (see also Fig.5).

A series of numerical experiments were conducted for different model errors and $N_x^a \times N_y^a$. Results are listed in Table 2. Since RMS and CR change with time (see Fig. 6), results listed in Table 2 are averaged from 72h to 120h. As truncation error can generally be considered as a model error, results in Table 2 show that in general, the assimilated results degrade with an increase in model errors. When the relative error of external forcing is not greater than 50% (i.e, $0.5 \leq \rho \leq 1.5$), the assimilated products have very high accuracy if $N_x^a \times N_y^a$ is not below 20×20. When $N_x^a \times N_y^a$ is 16×16, the analysis grid can not accurately resolve the external forcing ($k = 5$) and the small-scale features ($k = 4-7$) embedded in the true flow. Therefore the assimilated fields have a lower accuracy than those obtained with a finer analysis grid. In Table 2, the quality of assimilated products by a perfect model with $N_x^a \times N_y^a = 16 \times 16$ (in Experiment ER4) is close to that obtained with an imperfect model with $N_x^a \times N_y^a = 32 \times 32$ and $\rho = 0.5$ or 1.5 (in Experiments ER2 and ER6).

Table 2: Time-averaged root-mean square error of the assimilated wind (RMS$_w$, in ms^{-1}) and geopotential height (RMS$_\phi$, in $10m^2s^{-2}$) and time-averaged correlation coefficient (CR$_w$ and CR$_\phi$) between the assimilated and true fields. The time period used for averaging is from 72h to 120h.

Experiment	ρ	32×32*		20 × 20*		16 × 16*	
		RMS$_w$ (CR$_w$)	RMS$_\phi$ (CR$_\phi$)	RMS$_w$ (CR$_w$)	RMS$_\phi$ (CR$_\phi$)	RMS$_w$ (CR$_w$)	RMS$_\phi$ (CR$_\phi$)
ER1	0.0	4.3 (0.89)	9.1 (0.65)	4.4 (0.88)	9.7 (0.51)	4.6 (0.87)	13.2 (0.51)
ER2	0.5	2.3 (0.97)	4.9 (0.87)	2.6 (0.96)	6.1 (0.83)	2.8 (0.95)	8.7 (0.71)
ER3	0.75	1.4 (0.99)	3.0 (0.95)	1.8 (0.98)	4.5 (0.90)	2.2 (0.97)	7.3 (0.79)
ER4	1.0	0.6 (1.00)	2.0 (0.98)	1.4 (0.99)	4.1 (0.92)	1.9 (0.98)	6.6 (0.83)
ER5	1.25	1.4 (0.99)	3.3 (0.96)	1.9 (0.98)	5.3 (0.89)	2.3 (0.97)	7.4 (0.81)
ER6	1.5	2.3 (0.98)	5.2 (0.91)	2.7 (0.97)	7.3 (0.83)	3.1 (0.96)	9.3 (0.76)
ER7	2.0	4.3 (0.95)	9.5 (0.80)	4.8 (0.93)	11.8 (0.70)	5.5 (0.89)	14.8 (0.62)

* — Number of grid points of analysis grid ($N_x^a \times N_y^a$).

5. Treatment of spurious gravity waves and effect of data configuration

As seen in the above section, model error can lead to the generation of spurious gravity waves and thus reduces the accuracy of the assimilated variables. Generally speaking, inconsistency between model and data can cause the generation of spurious gravity waves in the assimilated fields. In addition, even if there is no model and data errors, spurious gravity waves can also be generated if the analysis increment, which acts like a forcing term for the model used in the forecast steps, is not smooth in space. This can arise from low temporal and spatial resolution of the observed data. To control spurious high-frequency small-scale gravity waves, one has to use certain methods to reduce the source strength of small-scale gravity waves and to damp these waves in the forecast steps, or increase data temporal and spatial resolution so that high-frequency oscillation can be identified.

a. *Controlling gravity waves*

A method to control gravity waves is to damp the divergence rate ($D = \partial u/\partial x + \partial v/\partial y$) of the wind. Through scaling analysis of (18) (e.g., Von Hinkelmann 1969; Browning et al. 1980), one can find that the temporal rate of gravity mode of ϕ is controlled by the divergence term (the last term) in (18). For high-frequency gravity waves, the time derivatives of D — $\partial D^l/\partial t^l$ ($l = 0, 1, 2, ..$) have large amplitudes. In this paper, the analyzed value is adjusted by bounding the amplitudes of D and its time tendency at each analysis step. We define a penalty function J_d as follows:

$$J_d = \sum_{|D_{i,j}|>E} \left(\frac{D_{i,j}}{E}\right)^2 + \sum_{|\frac{\partial D_{i,j}}{\partial t}|>E_t} \frac{1}{E_t^2}\left(\frac{\partial D_{i,j}}{\partial t}\right)^2, \tag{37}$$

where $D_{i,j}$ is the divergence at grid point (i, j) at an analysis step, and E and E_t are the upper bounds set for D and $\partial D/\partial t$, respectively; $E_t = E/T_D$ where T_D is the minimum time scale set for D. The analyzed value is modified into

$$\mathbf{x}^a \leftarrow \mathbf{x}^a - \alpha_d \nabla J_d, \tag{38}$$

where α_d is a weighting coefficient which can be determined like α_n used in (9). Writing (37) in a continuous form, it can be shown that if α_d is a constant and $E = 0$ and $E_t \to \infty$, the modification made in (38) is equivalent to adding a diffusion term in the momentum equations, i.e.,

$$\frac{\partial u}{\partial t} = ... + K_d \frac{\partial D}{\partial x}, \tag{39}$$

$$\frac{\partial v}{\partial t} = ... + K_d \frac{\partial D}{\partial y}, \tag{40}$$

where $K_d = \alpha_d / \Delta t$. Equations (39) and (40) are the linear dissipation form for divergence proposed by Sadourny (1975). Note that other diagnostic weak constraints can be similarly transformed into linear dissipation terms in the prognostic equations.

To demonstrate the efficacy of the above method for controlling gravity waves, data assimilation was performed for a case in which the true fields are generated using $c_r = 0$, $F_x = 0$, $F_y = fu_o$, $K_m = 300m^2s^{-2}$, and $D_m = (\nabla^2)^8$ (Tanguay et al. 1995). The high-order iterated Laplacian is used to damp motions of scales of about two and three grid spacings. These motions result from the equilibrium distribution mechanism of enstrophy in the wavenumber space, which is determined by the dynamic nature of a two-dimensional shallow water model (Sadourny 1975). For this set of model parameters, high-frequency gravity waves, once generated, can last a very long time. To see clearly the spurious high-frequency oscillations generated in the assimilated fields, the true fields are generated in such a way that no high-frequency oscillations (with period of about 1-2h) exist. This is achieved using the bounded derivative method (e.g., Browning et al. 1980). A variational approach to the bounded derivative method is presented in the Appendix. The true flow contains a substantial amount of small-scale motions (see Fig. 10 and the Appendix) which cannot be resolved by observations. The observed u and v are simulated in the same way as described in Section 4b, except that the temporal resolution of the observed data is assumed to be 1h. Data assimilation is performed with a perfect model.

Numerical experiments showed that for the case described above, strong spurious gravity waves are generated in the assimilated fields and the forecast is numerically unstable if the penalty for D and $\partial D/\partial t$ is not imposed. Controlling D can ensure the numerical stability of the model solution,

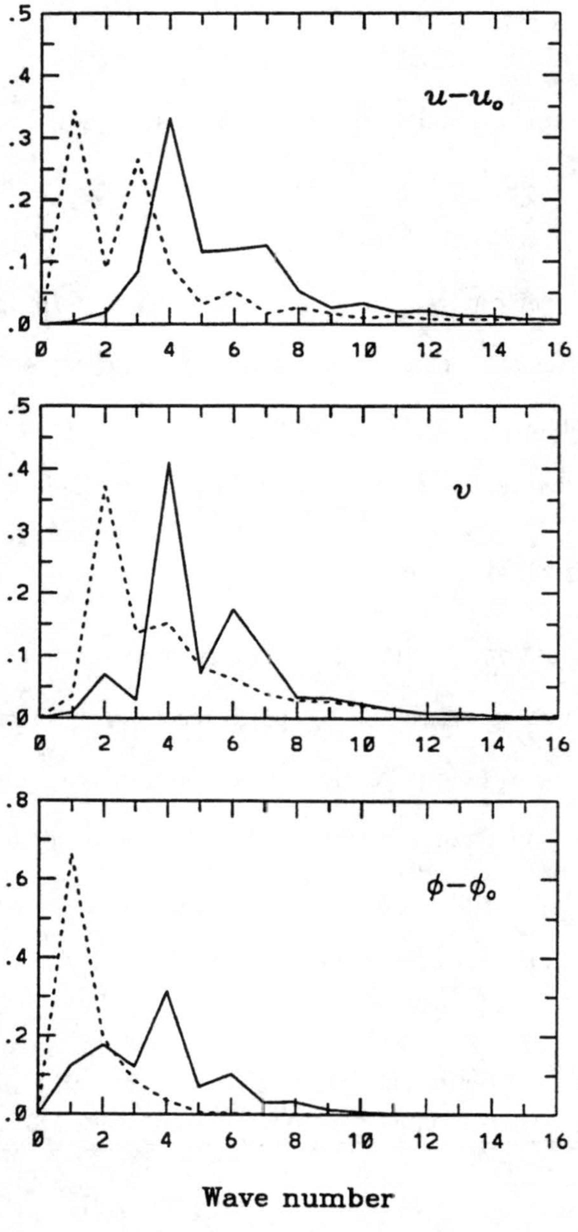


Figure 10: Normalized power spectra of perturbations of true wind and geopotential at t=60h (solid line) and t=120h (dotted line). Initial fields are generated with the bounded derivative method (see the Appendix).

but controlling D only cannot very efficiently eliminate spurious gravity waves, controlling both D and $\partial D/\partial t$ gives better results (Fig. 11).

The assimilated results obtained using $E = 1.0 \times 10^{-5}s^{-1}$ and $T_D = 6h$ are very accurate. At t=120h, RMS$_u$ and RMS$_v$ are 1.1ms^{-1} and 1.1ms^{-1}, respectively; RMS$_\phi$ is $10m^2s^{-2}$, and (CR$_u$, CR$_v$, CR$_\phi$) = (0.981, 0.971, 0.996). Spectral analyses for correlation coefficients and rms error show that errors in the assimilated fields mainly arise from the underestimation of the strength of small-scale features with wave-numbers $k \geq 10$.

Results (not shown) demonstrate that adding a diffusion term into the prognostic equation of ϕ and increasing the diffusion coefficient K_m can ensure the numerical stability of the assimilated solution and increase the accuracy of the assimilated results, but this method is not as effective as controlling the divergence rate of the wind; a large diffusion coefficient ($K_m \sim 10^6m^2s^{-2}$) is required.

b. *Influence of data configuration*

The amount of information contained in a data set is determined by its coverage and temporal resolution. Increasing data coverage and temporal resolution will certainly increase the ability of assimilation algorithms to detect small-scale features. We repeated the assimilation experiment ER1 in Table 2 with different data configurations. Results are listed in Table 3. From Table 3, one can see that results obtained with high spatial and temporal resolutions are much better than those with lower data coverage and temporal resolution. With gravity waves being controlled, the results

Table 3: As in Table 2, but for different data coverage (R) and temporal resolution (t_d) and different observed quantities. Data assimilations are performed using an imperfect model with $\rho = 0$ and $N_x^a \times N_y^a = 33 \times 33$. Gravity waves are controlled with equations (37) and (38).

Experiment	Observations	R (%)	t_d (h)	RMS$_w$ (CR$_w$)	RMS$_\phi$ (CR$_\phi$)
ED1	u, v	6.2	3	4.0 (0.91)	4.9 (0.87)
ED2	u, v	6.2	1	3.1 (0.95)	4.2 (0.91)
ED3	u, v	25	3	1.7 (0.98)	3.2 (0.95)
ED4	v	25	1	2.3 (0.97)	4.0 (0.92)
ED5	v	50	3	2.1 (0.98)	3.7 (0.93)

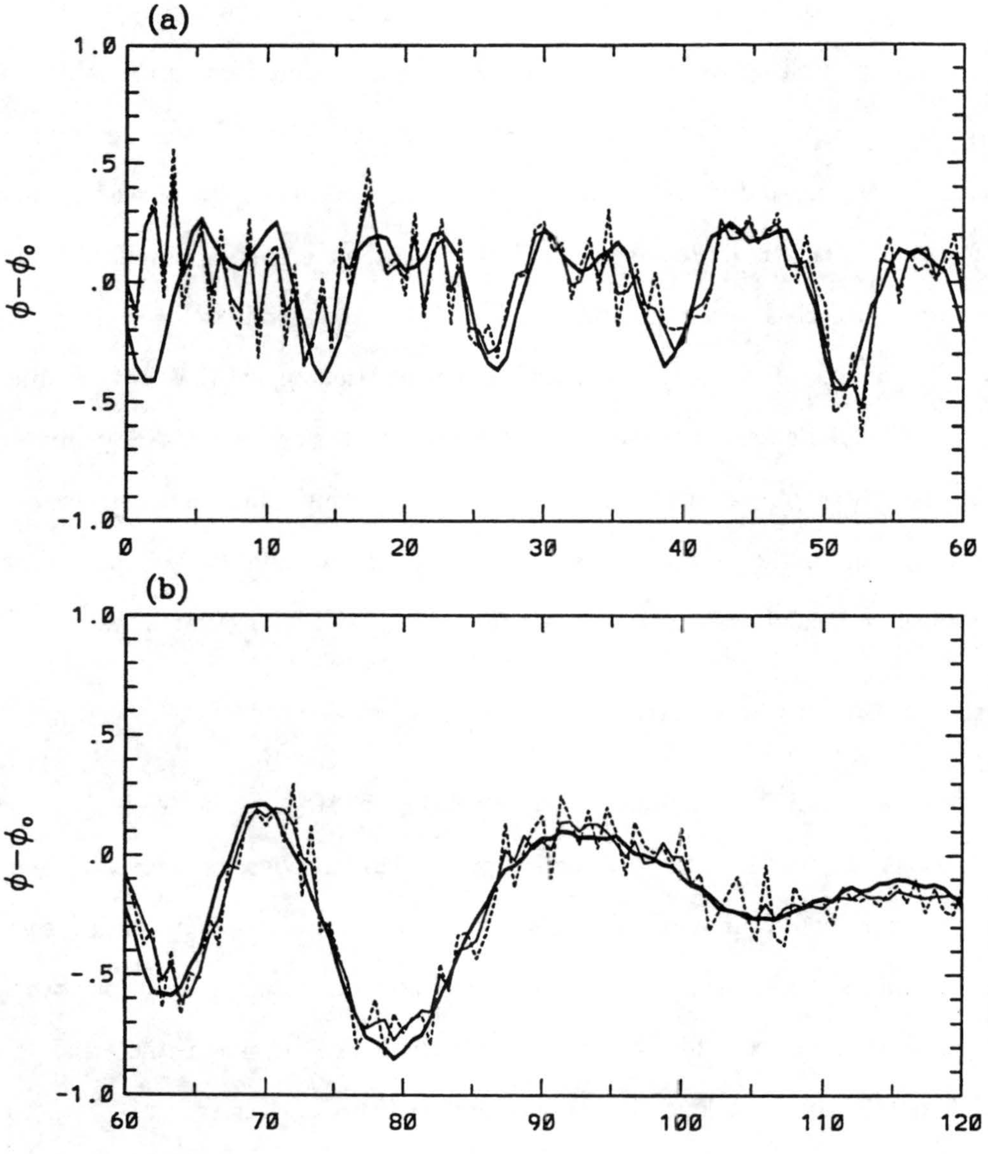


Figure 11: Time series of perturbation geopotential $(\phi - \phi_o)$ at $(x, y) = (500km, 1000km)$. Gravity waves are controlled with equations (36) and (37). Thick solid line — true solution, thin solid line — $E = 1.0 \times 10^{-5}s^{-1}$ and $T_D = 6h$, dotted line — $E = 1.0 \times 10^{-5}s^{-1}$ and $T_D = 0$.

of experiment ED1 are much better than those of experiment ER1 in Table 1, especially for the geopotential field. CR$_\phi$ is 0.87 in ED1, while it is only 0.63 in ER1 for the case $N_x^a \times N_y^a = 33 \times 33$.

Experiments ED4 and ED5 have only one observed velocity component v. This is designed to simulate the radial velocity observed by Doppler radars. Although only v is observed and model error is large ($\rho = 0$), the assimilated fields are very accurate. Experiment ED4 and ED5 also indicate that increasing data coverage is more effective in improving the assimilated results than increasing data temporal resolution. ED4 uses more data than ED5, but the assimilated results of ED4 is not as accurate as those of ED5. Other experiments (not shown) also support the above finding.

Experiments in Table 3 demonstrate that model error is not a big problem for the proposed GDDA method if data spatial and temporal resolutions are sufficiently high and inertia-gravity waves are properly controlled. The data coverage and temporal resolution used in Table 3 have already been achieved in the observations of WRS-88D Doppler radars and GOES-Next satellites.

6. Summary and conclusions

A continuous data assimilation method based on short-term 4D-Var analysis is described. This method is termed as gradient-descent data assimilation (GDDA). It consists of forecast and analysis steps. The analysis increment (analyzed value minus forecast value) is proportional to the gradient of a cost function, which measures the misfit between model prediction and observations over a period of time. The gradient of the cost function is calculated with the adjoint method and is updated periodically. This technique is a kind of retrospective analysis as it uses future data to analyze current model state. Unlike standard 4D-Var algorithms (e.g., Järvinen et al. 1996) that can only assimilate data in a limited time window, the proposed method can continuously assimilate data in an infinite time period and is less computer intensive than the standard 4D-Var technique. Moreover, the proposed method can use different forecast models (or different grids) in the forecast and analysis steps.

A two-dimensional shallow-water model with horizontal diffusion, Rayleigh friction and external forcing is used to test the GDDA method through identical-twin numerical experiments. The model domain has a mesoscale size of 1600km×1600km and 32×32 grid cells. Energy in this system cascades two-ways (upscale and downscale). Simulated observations of wind with a coverage of 6.2% are made every 3h. The observed wind components have random observational errors with a standard deviation of 1m/s and observation stations are randomly distributed on the model grid. The influence of model error and resolution of the analysis grid on the assimilated results is examined. Results show that when the relative error of external forcing is not greater than 50% and the analysis grid is not coarser than 20×20, the correlation coefficients between the assimilated and true fields are above 0.96 for the wind and above 0.83 for the geopotential height. When the relative error of external forcing is 100%, the proposed method can still recover a large portion of the small-scale motions.

Model error causes the generation of spurious small-scale inertia-gravity waves because of the inconsistency between the model and data. A penalty function is constructed to bound the time derivatives of the divergence rate of wind [see (37)] and thus to control inertia-gravity waves. Such a controlling procedure for gravity waves ensures the numerical stability of the assimilated fields (see subsection 5a) when horizontal diffusion is very small. When model error is larger, the assimilated results obtained with such a penalty function are improved considerably, especially for the geopotential field (see subsection 5b). Results also show that the influence of model error can be further reduced by increasing data spatial and temporal resolution.

When the GDDA method is applied to three dimensional (3D) mesoscale and large scale problems, spurious small-scale inertia-gravity waves will also exist in the assimilated fields. These spurious waves can be partially eliminated by controlling the divergence field of the wind and using a balance constraint (Parrish and Derber 1992). In addition, spurious small-scale gravity waves will be heavily dissipated by diffusion because mesoscale and large scale models use large horizontal diffusion coefficients. For example, the Regional Atmospheric Modeling System (RAMS) developed at Colorado State University uses a horizontal eddy diffusion coefficient of $K_m \geq 0.075 d^{4/3} m^2 s^{-1}$ (Robert Walko, private communication), where d is the horizontal grid spacing in meters. Preliminary studies show that artificial gravity waves are negligible in the assimilated fields when the GDDA scheme and the adjoint of RAMS are used to assimilate operational weather data and Doppler radar winds. These results will be reported elsewhere.

Aknowledgements

The authors would like to acknowledge Drs. Tomas Vonder Haar, Wayne Schubert, Gerald Browning, and Robert Walko for many critic comments and helpful suggestions. This work was supported by the National Oceanic and Atmospheric Administration under contracts NA37RJ0202 (S.Y, W.C.) and NA67RJ0152 (S.Y.), and the Air Force Office of Scientific Research under grant #F9620-95-01-0132 (S.Y., W.C.).

Appendix A
A variational approach to the bounded derivative method

The bounded derivative (BD) method (Browning et al. 1980; Kreiss 1980) is based on the observation that a solution of a hyperbolic system which varies slowly with respect to time must have a number of time derivatives on the order of the slow time scale. In the BD method, temporal derivatives of dependent variables at the initial time are constrained to the order of the slow time scale so that the amplitudes of the ensuing high-frequency motions remain small in a fixed time interval (~ 1 day). Alternatively, to achieve the same results one can bound the temporal derivatives of dependent variables to the order of the slow time scale in a short period of time. This can be fulfilled with a 4D-Var algorithm. For this particular problem, the cost function should be defined as the difference between observation and dependent variables at the initial time,

$$J_b = (\mathbf{X}(0) - \mathbf{X}^o(0))^T W^{-1} (\mathbf{X}(0) - \mathbf{X}^o(0)). \tag{A1}$$

An optimal solution of $\mathbf{X}$ can be found through minimizing J_b subject to the constraints of

$$\frac{\partial \mathbf{X}}{\partial t} = \mathbf{F}(\mathbf{X}), \tag{A2}$$

$$\frac{\partial^l x_i}{\partial t^l} \leq g_i^l, \qquad l = 1, 2, ..., K, \qquad \text{for } 0 \leq t \leq t_b, \tag{A3}$$

where (A2) is the prediction model of $\mathbf{X}$, g_i^l is the upper bound set for the lth derivative of the ith element (x_i) of $\mathbf{X}$, and t_b is the time period in which time derivatives are bounded. J_b can be minimized using penalty or augmented Lagrangian methods (e.g., Zou et al. 1993).

Using a two-dimensional shallow-water model, Zou et al. (1993) showed that constraining the first-order temporal derivative of geopotential can effectively remove spurious gravity waves in a 4D-Var numerical experiment for recovering large scale flows. They also found that treating the first-order time derivative of geopotential as a penalty term in the cost function is suffucient to remove spurious gravity waves. In this paper, (A3) is imposed as a weak constraint, and J_b is modified into

$$J_b' = J_b + \sum_{0\leq t\leq t_b} \sum_{l=1}^{K} \left(\frac{\partial^l \mathbf{X}}{\partial t^l}\right)^T G^l \frac{\partial^l \mathbf{X}}{\partial t^l}, \tag{A4}$$

where G^l is a weighting matrix for the lth order of time derivative

$$G_{ij}^l = (g_i^l)^{-2}\delta_{ij} \tag{A5}$$

where δ_{ij} is the Kroneckle symbole.

For the case described in section 5b, the initial values of the true solution is generated using the above 4D-Var approach to the BD method. The observed initial geopotential field is produced with (21) and (22) except that $k_x^\psi = 3$ and $k_y^\psi = 4$ and the observed initial wind is assumed to be in geostrophic balance. An optimal initial condition is then obtained by minimizing J_b'. The upper bounds of time derivatives are given as

$$g_i^l = \begin{cases} U/T_c^l, & \text{for wind,} \\ \Delta\phi/T_c^l, & \text{for geopotential,} \end{cases} \tag{A6}$$

where U (=10ms^{-1}) and $\Delta\phi$ (=300m^2s^{-2}) are the characteristic scales of wind and geopotential deviation from mean geopotential, respectively; and T_c is the minimum time scale set for the problem. In order to remove only those high-frequency oscillations with a period less than about two hours, T_c is set to 2h. The shallow-water model is integrated 3 time steps (i.e., $t_b = 3\Delta t$). Results show that if K=1, oscillations with a period of one or two hours are heavily damped, but still visible. When $K = 2$, 1-2h oscillations are eliminated. This result is consistent with that obtained by Browning et al. (1980). We also found that the difference between the results obtained with $K = 2$ and those with $K = 3$ is negligible for $t < 60h$. The true solution for the case in section 5b is obtained with $K = 3$.

References

Anthes, R. A., 1974: Data assimilation and initialization of hurricane prediction models. *J. Atmos. Sci.,* **31**, 702–719.

Bloom, S.C., L.L. Takacs, A.M. Da Silva, and D. Ledvina, 1996: Data assimilation using incremental analysis updates. *Mon. Wea. Rev.*, **124**, 1256–1271.

Browning G., A. Kasahara and H.-O. Kreiss, 1980: Initialization of the primitive equations by the bounded derivative method. *J. Atmos. Sci.*, **37**, 1424–1436.

Cohn, S. E., N. S. Sivakumaran, and R. Todling, 1994: A fixed-lag Kalman smoother for retrospective data assimilation. *Mon. Wea. Rev.*, **122**, 2838–2864.

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

Courtier, P., J. Derber, R. Errico, J. F. Louis, and T. Vukicevic, 1993: Important literature on the use of adjoint, variational methods and the Kalman filter in meteorology. *Tellus*, **45A**, 342–357.

Courtier, P., J.-N. Thepaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.*, **120**, 1367–1387.

Daley, R., and R. Menard, 1993: Spectral characteristics of Kalman filter systems for atmospheric data assimilation. *Mon. Wea. Rev.*, **121**, 1554–1565.

Ghil, M., S. E. Cohn and A. Dalcher, 1982: Sequential estimation, data assimilation, and initialization. D. Williamson (ed.), The interaction between objective analysis and initialization. *Publ. Meteorol.*, **127**, McGill University, Montreal, 83–97.

Holton, J. R., 1992: *An introduction to dynamic meteorology.* 3rd ed. Academic Press, Inc., San Diego, 511 pp.

Järvinen, H., J.-N. Thepaut and P. Courtier, 1996: Quasi-continuous variational data assimilation. *Q. J. R. Meteorol. Soc.*, **122**, 515–534.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. Trans. ASME, *J. Basic Eng.*, **82D**, 35–45.

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

Mahfouf, J.-F., 1991: Analysis of soil moisture from near-surface parameters: A feasibility study. *J. Appl. Meteor.*, **30**, 1534–1547.

Ménard, R., and R. Daley, 1996: The application of Kalman smoother theory to the estimation of 4D-Var error statistics. *Tellus*, **48A**, 221–237.

Menzel, W. P. and J. F. W. Purdom, 1994: Introducing GOES-I: The first of a new generation of geostationary operational environmental satellites. *Bull. Amer. Meteor. Soc.*, **75**, 757–781.

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

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral-interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747–1763.

Ruggiero F.H., K.D. Sashegyi, R.V. Madala, and S. Raman, 1996: The use of surface observations in four-dimensional data assimilation using a mesoscale model. *Mon. Wea. Rev.*, **124**, 1018–1033.

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

Staufer, D. R., and M. L. Seaman, 1990: Use of four-dimensional data assimilation on a limited area mesoscale model. Part I: Experiments with synoptic-scale data. *Mon. Wea. Rev.*, **118**, 1250–1277.

Sun, J., and A. Crook, 1994: Wind and thermodynamic retrieval from single–Doppler measurements of a gust front observed during Phoenix II. *Mon. Wea. Rev.* , **122**, 1075–1091.

Talagrand, O., 1981: A study of the dynamics of four–dimensional data assimilation. *Tellus,* **33**, 43–60.

Tanguay, M., P. Bartello and P. Gauthier, 1995: Four-dimensional data assimilation with a wide range of scales. *Tellus*, **47A**, 974–997.

Telesetsky, W., 1995: Current status and issues of the weather surveillance radar-88 Doppler program. Preprints, *27th Conference on Radar Meteorology*. 9-13 Oct., 1995, Vail, CO., Amer. Meteor. Soc., xxxvii–xlii.

Thépaut, J.-N., D. Vasiljevic, P. Courtier and J. Pailleux, 1993: Variational assimilation of conventional observations with a multi-level primitive equations model. *Q. J. R. Meteorol. Soc.*, **119**, 153–186.

Verlinde, J., and W. R. Cotton, 1993: Fitting microphysical observations of non-steady convective clouds to a numerical model: An application of the adjoint technique of data assimilation to a kinematic model. *Mon. Wea. Rev.*, **121**, 2776–2793.

Von Hinkelmann, K., 1969: Primitive equations. *Lectures on Numerical Weather Prediction*, Hydrometerizdat, Leningrad, 306–375.

Walko, R.L., C.J. Tremback, and W.R. Cotton, 1989: Assimilation of Doppler radar wind data into a numerical prediction model: A demonstration of certain hazards. Preprints, *24th Conf. on Radar Meteor.*, Tallahassee, FL, Amer. Meteor. Soc., 248–250.

Washington W. M., and C. L. Parkinson, 1986: *An Introduction to Three-dimensional Climate Modelling*. Univ. Sci. Books, 422 pp.

Xu, Q., 1996a: Generalized adjoint for physical processes with parameterized discontinuities. Part I: Basic issues and heuristic examples. *J. Atmos. Sci.*, **53**, 1123-1142.

Xu, Q., 1996b: Generalized adjoint for physical processes with parameterized discontinuities. Part II: Vector formulations and matching conditions. *J. Atmos. Sci.*, **53**, 1143-1155.

Yang, S., and Q. Xu, 1996: Statistical errors in variational data assimilation – a theoretical one-dimensional analysis applied to Doppler wind retrieval. *J. Atmos. Sci.*, **53**, 2563-2577.

Zou, X., I. M. Navon, and F. X. Le Dimet, 1993: Incomplete observations and control of gravity waves in variational data assimilation. *Tellus*, **44A**, 273–296.

Zou, X. , Y.-H. Kou, and Y.-R. Guo, 1995: Assimilation of atmospheric radio refractivity using a nonhydrostatic adjoint model. *Mon. Wea. Rev.* , **123**, 2229–2249.

Zupanski, M., 1993: Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Mon. Wea. Rev.*, **121**, 2396–2408.