Q. J. R. Meteorol. Soc. (2001), **127**, pp. 1095–1122

Four-dimensional variational data assimilation with a diabatic version of the NCEP global spectral model: System development and preliminary results

By X. ZOU[1]*, H. LIU[1,2], J. DERBER[3], J. G. SELA[3], R. TREADON[3], I. M. NAVON[1] and B. WANG[2]

[1]*Florida State University, USA*
[2]*Institute of Atmospheric Physics, Beijing, China*
[3]*National Centers for Environmental Prediction, USA*

(Received 30 December 1999; revised 29 August 2000)

SUMMARY

The formulation of the National Centers for Environmental Prediction four-dimensional variational data-assimilation (4D-Var) system is described. Results of applying 4D-Var over a one-week assimilation period, with a full set of physical parametrizations, are presented and compared with those of 3D-Var. The linearization has been performed without simplifications and, therefore, the tangent-linear and adjoint codes are consistent with the nonlinear physical parametrizations. The 4D-Var assimilation is similar in formulation to the 3D-Var analysis, except that observations are used at the appropriate time in 4D-Var. Compared with the 3D-Var runs, the 4D-Var results showed good convergences, smaller analysis increments, and a comparable fit of analyses and short-range forecasts to observations. A consistent improvement with the 4D-Var system is observed in short-range (six-hour) forecasts of all model variables except the specific humidity. The temperature analyses from 4D-Var were found to be better in most of the areas where the analysis errors from 3D-Var were largest, although the globally averaged root-mean-square difference in the 4D-Var temperature analysis was larger due to a very small degradation in some parts of the globe that include data-rich areas. The globally averaged root-mean-square difference in the 4D-Var specific-humidity analysis, compared with that of 3D-Var, was larger and was found to result from slightly increased analysis-error maxima in the 4D-Var results over data-sparse tropical regions. The 3–4 day forecasts from 4D-Var analyses compared more favourably than forecasts from the 3D-Var analyses with the targeted mid-Pacific dropwindsonde observations available from the 1998 North Pacific Experiment. Compared with conventional observations, a consistent improvement in the 1–5 day forecasts of wind and temperature was shown in the tropics and the southern hemisphere.

KEYWORDS: Adjoint of physical parametrizations Numerical weather prediction 4D-Var

1. INTRODUCTION

The atmosphere, whose past, present and future changes in weather are of great interest to meteorologists and to society, is governed by complex mechanical and thermodynamical laws. These laws can be formulated into a set of partial differential equations (the governing equations) which can be solved numerically, but not analytically. The set that is solved numerically is usually a simplified, approximate, and discretized form of the governing equations. They are solved as an initial-value problem in which atmospheric states at future times are predicted from the current atmospheric state (initial condition). Besides theoretical and arithmetical errors, uncertainties in the initial conditions are a major source of errors in numerical weather prediction. Four-dimensional variational data assimilation (4D-Var) is a logical and rigorous mathematical method to obtain the 'best' estimate of the model initial condition—a three-dimensional discrete representation of the atmosphere—from observational measurements and a priori knowledge of the atmospheric state. A cost function that involves a model trajectory (an additional time dimension as compared with 3D-Var) has to be minimized, which requires the adjoint model in order to solve this problem at a reasonable computing cost.

Originally, 4D-Var was applied to different theoretical or adiabatic models (Le Dimet and Talagrand 1986; Derber 1985; Lewis and Derber 1985; Talagrand and Courtier 1987; Courtier and Talagrand 1987; Derber 1989; Thépaut and Courtier 1991; Navon *et al.* 1992; Chao and Chang 1992). Theoretical analyses suggest that a model

* Corresponding author: Florida State University, Department of Meteorology, 404 Love Building, Tallahassee, FL32306-4520, USA. e-mail: zou@bamboo.met.fsu.edu

that includes diabatic physical processes can provide a better data-assimilation vehicle, and reduces the negative impact of the perfect-model assumption in the current 4D-Var formulation. A 4D-Var system with physical processes is more adequate than an adiabatic version for using observations of quantities derived from physical processes (such as rainfall, radar reflectivity, cloud water, and rain water). Direct assimilation of non-conventional data, such as rainfall data, requires that the moist physics and its adjoint be included in the nonlinear forecast model and in the adjoint model (Zupanski and Mesinger 1995; Zou and Kuo 1996; Xiao *et al.* 2000). Recently, a 4D-Var system, with physics ranging from simplified to improved and complex, has been developed at ECMWF* (Rabier *et al.* 1998; Mahfouf and Rabier 2000; Rabier *et al.* 2000; Klinker *et al.* 2000), at NCEP† using a mesoscale limited-area model (Zupanski and Zupanski 1995), and at NCAR‡ and Florida State University using a non-hydrostatic mesoscale model (Zou and Kuo 1996). Given the many unique features, the encouraging results and the future potential of 4D-Var, a major effort has been made towards the development of a 4D-Var system using the NCEP global spectral model and its 'full-physics' adjoint model. This paper reports and describes the recently completed NCEP 4D-Var system and presents some preliminary numerical results.

Development of the NCEP 4D-Var system started with an adiabatic adjoint version of the NCEP global spectral model (Navon *et al.* 1992). Later, some penalty terms were added to the cost function to test the control of gravity-wave oscillations in the 4D-Var framework (Zou *et al.* 1993a). The adjoints of two moist physical parametrization schemes (grid-scale precipitation and cumulus convection) were then developed and included in the 4D-Var assimilation model (Zou *et al.* 1993b). All these experiments were carried out in the absence of a background term, and used NCEP analyses as 'observations'. The current paper summarizes the continuing development of a 'full-physics' NCEP 4D-Var system which combines an operational 3D-Var system (Parrish and Derber 1992) with a diabatic version of the NCEP global medium-range forecast model and its tangent linear and adjoint models. We compare the numerical results obtained using 4D-Var with those of 3D-Var when the same number of conventional observations (mainly radiosonde, surface, aircraft, satellite, and dropsonde data) were used, and examine the influence of the use of 4D-Var on the analyses and subsequent forecasts.

Note that there are differences between the global forecast model used in this study (a version that was operational before 1995) and the current NCEP medium-range forecast model, as well as between the 3D-Var system used in this study and the current NCEP 3D-Var system. The current operational system uses a simplified Arakawa–Schubert scheme for cumulus parametrization, three- six- and nine-hour forecasts from the guess fields to interpolate the guess fields to the observation time, a constraint on supersaturation and negative specific-humidity values, a new formulation of background error defining the spectral statistics as a function of the total wave number and allowing a specification of a spatial variance field, a three-dimensional ozone analysis, and the use of TOVS-1B§ and GOES¶ radiances.

The paper is arranged as follows. In section 2, we briefly describe the physical processes and their adjoint operators which were not included in our previous studies.

* European Centre for Medium-Range Weather Forecasts.
† National Centers for Environmental Prediction.
‡ National Center for Atmospheric Research.
§ TIROS (Television InfraRed Operational Satellite) Operational Vertical Sounder-1B.
¶ Geostationary Operational Environmental Satellite.

These include surface processes, vertical diffusion, shallow convection and gravity-wave drag parametrizations. An interesting example illustrating challenges related to the development of adjoint physics is reported in the same section. Some test results showing the validity of the tangent linear model (TLM) with complex physics, as well as the correctness of both the tangent linear and adjoint models, are also included in section 2. Experimental design and formulations of the 3D-Var and 4D-Var problems (the cost functions) are described in section 3. Results of analyses using both 3D-Var and 4D-Var are presented in section 4. Forecast differences resulting from 3D-Var and 4D-Var analyses are shown in section 5. The misfit of the 3D-Var and 4D-Var analyses to conventional data are examined and their geographical dependences with respect to observations are discussed. Conclusions are presented in section 6.

2. PHYSICAL PROCESSES AND THEIR TANGENT LINEAR AND ADJOINT OPERATORS

A set of linear and adjoint operators for the physical processes in NCEP's global forecast model has been developed, without any simplification or modification to the original schemes. Thus, the main feedback loops between the processes are the same as in the original nonlinear forecast model except for radiation. The adjoint of radiation processes was developed (Li and Navon 1998), but was not included in these experiments. In the following we briefly describe the six physical parametrization schemes that, along with their tangent linear and adjoint operators, were included in the 4D-Var system.

(*a*) *Surface processes*

The main purpose of surface processes in the NCEP model is to predict the surface (air–ground interface) temperature and humidity, and to estimate the fluxes of momentum, heat and humidity in the surface layer of the atmosphere (Miyakoda and Sirutis 1977). The surface-process scheme calculates the surface temperature (T_s), the subsurface soil temperatures ($T_{g,1}$ and $T_{g,2}$), the transfer coefficients for momentum and heat (C_D and C_H), the snow melt, the surface specific humidity, and the roughness length (z_0). The calculation of the above variables requires the iterative solution of a series of implicit nonlinear equations.

A full linearization of the parametrization scheme for the surface processes was performed. A complete adjoint operator for the surface parametrization scheme was developed. However, for the experiments that were conducted in this study, the perturbations for the surface temperature, the subsurface soil temperatures, the snow melt, the surface specific humidity, and the roughness length were set to zero, resulting an effective partial linearization for variables associated with the calculations of the transfer coefficients in the surface boundary layer.

(*b*) *Vertical diffusion*

The effects of vertical turbulent eddy transfer of momentum, heat, and moisture throughout the atmosphere are represented by a local-K approach. The diffusivity coefficients are parametrized as functions of the local Richardson number. The vertical diffusion calculation is carried out after the nonlinear dynamics and is added to the adiabatic nonlinear tendencies (splitting method). Numerical calculations of the vertical diffusion involve the formulation of the diffusion coefficients, the discretization of the vertical differentiations, and the set-up of the upper and lower boundary conditions. The lower boundary condition is determined by the previously mentioned surface processes.

We have done a complete linearization of the original vertical diffusion scheme without simplification. When a new variable-dependent denominator is created during the course of linearizing the original NLM, we add a small number (determined by machine accuracy) to that denominator to avoid abnormal growth of the tangent linear perturbation solution. We did not experience a serious convergence problem of minimization due to the spurious noise that may still be present in the tangent linear perturbation solution of vertical diffusion (Janiskova *et al.* 1999).

(*c*) *Shallow convection*

Shallow convection simulates the effect of shallow non-precipitating cumulus clouds by carrying out an enhanced vertical diffusion of specific humidity and temperature in the model columns that contain a conditionally unstable layer near the surface and in which no deep convection has been performed. This allows a vigorous vertical mixing of water vapour that would otherwise tend to accumulate near the surface in synoptically inactive regions.

The shallow-convection process is treated in the same way as the model's basic vertical diffusion, except that the values of the diffusion coefficient are prescribed. Cloud base is determined from the values of the lifting condensation level, which has been calculated from the deep-convection cumulus parametrization scheme. The cloud top is determined as the minimum between the highest unstable layer and the sixth σ layer from the model top.

(*d*) *Gravity-wave drag parametrization*

The gravity-wave drag parametrization was used as a momentum-damping mechanism to improve the medium-range performance of the NCEP model (Pierrehumbert 1986)*.

(*e*) *Moist processes*

Moist processes in the NCEP global model include (i) large-scale precipitation and (ii) deep-cumulus convection. The large-scale precipitation simulates the condensation of excess water vapour when supersaturation is reached, and turns it into large-scale precipitation. Some of the precipitation formed in the upper layers is allowed to re-evaporate into the unsaturated lower layers when the condensed water falls through them. The cumulus parametrization simulates deep convection. It is modelled by a Kuo–Anthes type scheme. A certain amount of moisture convergence, a deep conditional instability, a warm low-level temperature, and the absence of a low-level inversion, are required to trigger the convection. Details of these two moist processes have been described by Zou *et al.* (1993b) and the references therein. We mention that the current NCEP medium-range forecast model uses a simplified Arakawa–Schubert cumulus parametrization (Pan and Wu 1995).

(*f*) *Linear and adjoint operators of the physical processes*

Given the discretized version of a particular physical parametrization scheme in the form of a sequence of computer codes, both the tangent linear and the adjoint models were constructed by differentiation and transposition of these codes. This process is done in two steps: (a) linearize the forward discretized nonlinear model (NLM), with respect to the NLM state, to obtain the discretized TLM as a sequence of computer

* An example of adjoint coding for part of the codes in the gravity-wave drag parametrization is given at http://www.met.fsu.edu/adjoint.

codes in which the 'on–off' switches are kept the same as in the NLM (the basic-state variables are used for the 'IF' statements in the computer code); and (b) view the operator (a matrix) of the tangent linear physics, $\mathbf{M}$, as a consequence of multiple matrices $\mathbf{M}_N \ldots \mathbf{M}_1$, and develop the computer codes that represent the transposition of these matrices, i.e. $\mathbf{M}_1^{\mathrm{T}}, \mathbf{M}_2^{\mathrm{T}}, \ldots, \mathbf{M}_N^{\mathrm{T}}$. The matrix $\mathbf{M}^{\mathrm{T}} = \mathbf{M}_1^{\mathrm{T}} \ldots \mathbf{M}_N^{\mathrm{T}}$ constitutes the operator of the adjoint physics.

Notice that, in the entire adjoint development procedure, we never store a full matrix. We are only concerned with the following task: given the input $\mathbf{x}_r$ to $\mathbf{M}_r$ or the input $\mathbf{y}_r$ to $\mathbf{M}_r^{\mathrm{T}}$, obtain an output vector of $\mathbf{M}_r\mathbf{x}_r$ or $\mathbf{M}_r^{\mathrm{T}}\mathbf{y}_r$. In other words, we do not need the explicit form of the matrices, but are only interested in the result of each matrix (the tangent linear operator or adjoint operator) multiplied by a vector (the input to the tangent linear operator or to the adjoint operator). Examples have been illustrated by Navon *et al.* (1992), Zou *et al.* (1993b) and Zou (1996) for the practical adjoint coding using the adjoint of finite-difference method (for example, see Sirkes and Tziperman 1997). For an adiabatic model, or for simple physics, the logic of the numerical calculation in the original nonlinear code is simple and straightforward. For other physical processes, it can be very difficult and the adjoint coding can become very tricky*.

(*g*) *Perturbation solutions and their linear approximations*

The correctness of a TLM may be checked by examining how well its solution approximates the difference between the two NLMs' solutions as the size of the initial perturbation tends toward zero. The ratio of the two solutions approaches unity linearly as the magnitude of the initial perturbation is reduced. For a finite-amplitude initial perturbation, the degree to which the adiabatic TLM solution approximates the NLM perturbation solution depends on the degree of nonlinearity, which is related to the usefulness of the TLM.

The difference between the solution and the perturbation solution of a nonlinear diabatic model depends, not only on the nonlinearity, but also on discontinuities caused by 'on–off' switches in the model physical processes. Figure 1 shows the time evolution of the globally averaged root-mean-square (rms) temperature difference between the perturbed and unperturbed (the basic state) NLM solutions:

$$\mathbf{x}^{\text{perturbed NLM}}(t, \alpha) = M_t\{\mathbf{x}_0 + \alpha(\mathbf{x}_0^{(2)} - \mathbf{x}_0)\} \tag{1}$$

$$\mathbf{x}^{\text{basic state}}(t) = M_t(\mathbf{x}_0), \tag{2}$$

where M_t is the operator representing the operations performed in the NLM to obtain the model forecast at time t from an initial condition at time t_0 ($t > t_0$), $\mathbf{x}_0$ is the NCEP analysis at 00 UTC 21 February 1998, $\mathbf{x}_0^{(2)}$ is the NCEP analysis at 00 UTC 15 February 1998, and α is a real number with its values ranging from 10^{-1} to 10^{-12} controlling the magnitude of the initial perturbation. Therefore, the initial perturbation $\alpha\Delta\mathbf{x}_0$ ($= \alpha(\mathbf{x}_0^{(2)} - \mathbf{x}_0)$) is the difference between the two analyses multiplied by a factor of α.

We observe from Fig. 1 that the nonlinear perturbation solutions are close to the tangent linear solution only when the initial perturbations are very small ($\alpha \leqslant 10^{-5}$). However, a large discrepancy is observed between the NLM perturbation and its TLM

* One such example that was encountered in developing the adjoint of NCEP's gravity-wave drag parametrization scheme is shown at http://www.met.fsu.edu/adjoint—this example may serve as a test for a comprehensive 'automatic' adjoint-code generator.

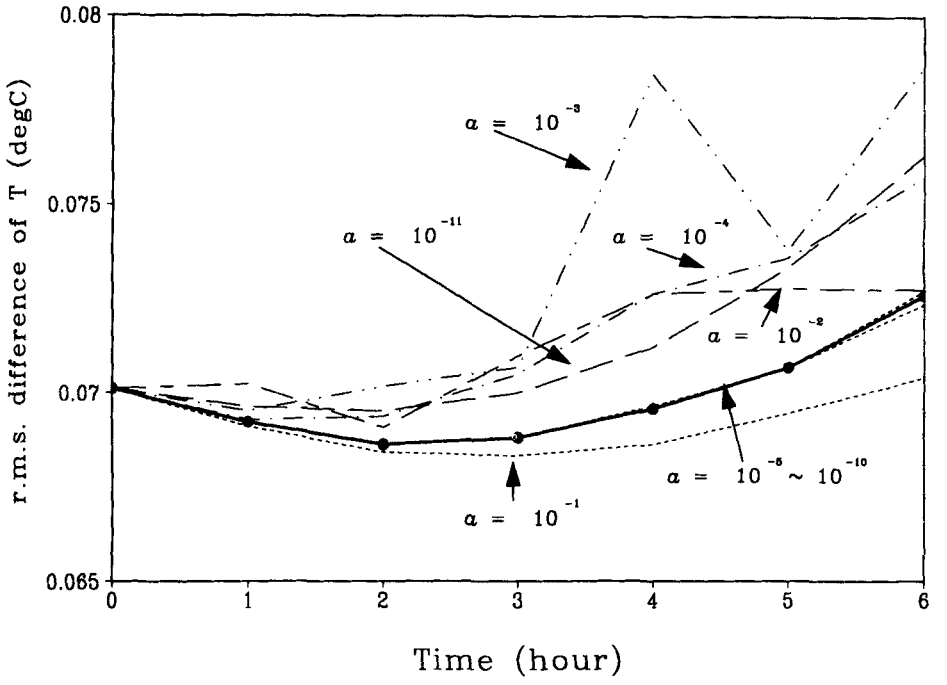

Figure 1. The time evolution of the globally averaged root-mean-square (rms) differences of the temperature perturbation from the tangent linear model (thick solid line with circles) and the nonlinear model (broken lines) with the initial perturbation of $\alpha\delta\mathbf{x}_0$ at 00 UTC 21 February 1998, with α taking various values in the range 10^{-1} to 10^{-11}, $\delta\mathbf{x}_0$ is the difference between the two analyses at 00 UTC 15 February and 00 UTC 21 February 1998. The rms values are normalized by α.

approximation in the course of the time integration for initial perturbations ranging from $\alpha = 10^{-1}$ to $\alpha = 10^{-4}$. This behaviour may be attributed to strong nonlinearities associated with the physical processes and the presence of the 'on–off' switches in various physical parametrization schemes. We notice that, sometimes, the nonlinear perturbation solution with a larger initial perturbation (for example $\alpha = 10^{-1}$ (dotted line in Fig. 1)) can diverge more slowly from the TLM approximation than the solution with a smaller initial perturbation such as $\alpha = 10^{-3}$ (dash-double-dots line), indicating a strong nonlinear effect in the diabatic forecast model when the initial perturbations are not sufficiently small.

The tangent linear solution for the initial perturbation of $\alpha\Delta\mathbf{x}_0$ can be expressed as

$$\mathbf{x}^{\mathrm{TLM}}(t, \alpha) = \alpha\mathbf{M}_t\delta\mathbf{x}_0, \tag{3}$$

where $\mathbf{M}_t = \partial M_t/\partial\mathbf{x}$ is the operator of the TLM. We observe from (3) that, as α decreases, the rms evolution in time changes only in magnitude. Therefore, we can conclude from Fig. 1 that the nonlinear perturbation solutions will not be approximated well by the TLM solution when $10^{-1} \geqslant \alpha \geqslant 10^{-4}$, without even examining the variation of the TLM solution in time, since the shapes of the NLM solutions in this range of α change. This conclusion does not depend on how we code the TLM; i.e. whether the 'on–off' switches are kept the same as in the NLM, or whether the switches are different (Zou 1996). From Fig. 1, we find that the TLM solutions with physics approximate the NLM perturbation solutions very well when $10^{-10} \leqslant \alpha \leqslant 10^{-5}$, but not so well when

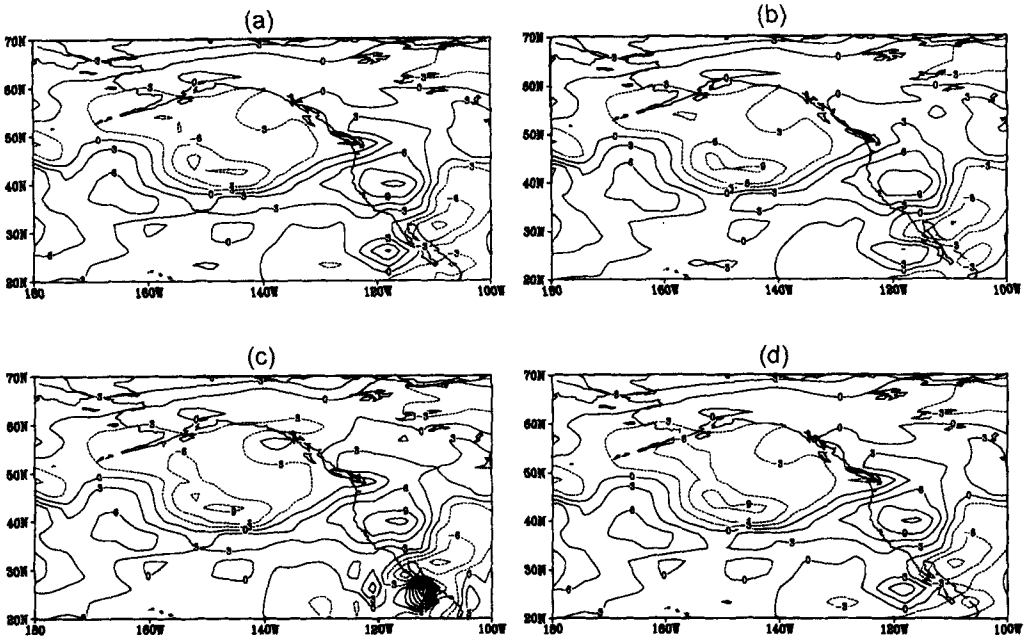

Figure 2. The six-hour forecasts of the temperature perturbations on the 500 hPa surface from (a) the tangent linear model with the initial perturbation of $\delta \mathbf{x}_0$ and from the nonlinear models with the initial perturbations as in Fig. 1 for (b) $\alpha = 10^{-1}$, (c) $\alpha = 10^{-3}$ and (d) $\alpha = 10^{-5}$.

$\alpha \geqslant 10^{-4}$. The differences between the TLM solutions and the nonlinear perturbation solutions when $\alpha \leqslant 10^{-11}$ are due to round-off errors.

For correctness, a diabatic TLM may not compare better with a diabatic NLM than an adiabatic TLM with an adiabatic NLM. But, for usefulness, a TLM that includes the linearized model physics will always compare better with a diabatic NLM than with an adiabatic TLM. In an incremental approach, the TLM is used to advance the analysis increment forward in time, and the adjoint model is used to calculate the gradient of the cost function defined in the TLM. Therefore, the discrepancy between the TLM solution and the nonlinear perturbation solution does not affect the accuracy of gradients calculated at the inner loops of the minimization. For a non-incremental approach, the TLM model is not used and the adjoint model is used to calculate the gradient of the cost function, which is defined using the NLM forecasts. As was indicated by Zou (1996) and Zhang *et al.* (2000), the results of an adjoint integration with discontinuous physics provide useful subgradient information. The discontinuity in the physical processes, however, gives rise to a discontinuous cost function and a discontinuous gradient, which may render a minimization problem more difficult to solve.

In order to compare the TLM forecast results more precisely with those of the NLM, we show in Fig. 2 the six-hour forecasts of the 500 hPa temperature perturbation over a selected area from the TLM (Fig. 2(a)) and NLM, with various sizes of initial perturbation represented by the values of α (Figs. 2(b)–(d)). We find that the TLM solution could be a very good approximation of the corresponding nonlinear perturbation solutions when $10^{-10} \leqslant \alpha \leqslant 10^{s}$, where s varies from -1 to -4 depending on the physical locations of the verification regions. It seems that the initial perturbation should be smaller for the TLM to approximate the NLM with the same accuracy over regions where the diabatic processes have a larger effect on the atmospheric state than elsewhere.

3. Description of the 3D-Var and 4D-Var data assimilation experiments

(*a*) *3D-Var system*

The NCEP operational 3D-Var system has been described by Parrish and Derber (1992), Derber and Wu (1998), and Derber and Bouttier (1999). A brief description is provided below:

An objective function J is defined as a summation of the background term, an observational term, and a penalty term:

$$2J(\mathbf{x}_a) = (\mathbf{x}_a - \mathbf{x}_b)^T \mathbf{B}_0^{-1} (\mathbf{x}_a - \mathbf{x}_b) + \{H(\mathbf{x}_a) - \mathbf{y}_{obs}\}^T \mathbf{R}^{-1} \{H(\mathbf{x}_a) - \mathbf{y}_{obs}\} + 2J_c, \tag{4}$$

where $\mathbf{x}_a$ is an N-component vector of analysis variables (the vorticity, the unbalanced part of the divergence, the unbalanced temperature, the logarithm of the unbalanced surface pressure, and the water-vapour mixing ratio), $\mathbf{x}_b$ is a six-hour forecast from the previous cycle of analysis (also known as the background or first guess), and $\mathbf{y}_{obs}$ is an M-component vector of observations. $\mathbf{B}_0$ is a diagonal $N \times N$ forecast-error covariance matrix, $\mathbf{R}$ is an $M \times M$ observational-error covariance matrix (including both the instrument and the representative errors), H is an observation operator that converts the analysis variables on model grids to the observation quantities and locations, and J_c is a dynamical constraint term used to increase the balance in the analysis increment.

Under the transform of

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{C}\mathbf{z}, \quad \mathbf{C} = (\sqrt{B_{ii}}), \tag{5}$$

the cost function becomes a function of $\mathbf{z}$:

$$2J(\mathbf{z}) = \mathbf{z}^T\mathbf{z} + \{H(\mathbf{x}_b + \mathbf{C}\mathbf{z}) - \mathbf{y}_{obs}\}^T \mathbf{R}^{-1} \{H(\mathbf{x}_b + \mathbf{C}\mathbf{z}) - \mathbf{y}_{obs}\} + 2J_c. \tag{6}$$

Minimization of J with respect to $\mathbf{z}$ is carried out by solving the equation

$$\frac{\partial J}{\partial \mathbf{z}} \equiv \left(\mathbf{z} - \mathbf{C}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}_k + \frac{\partial J_c}{\partial \mathbf{z}}\right)\Bigg|_{\mathbf{z}=\mathbf{z}^*} = 0, \tag{7}$$

using a perturbation method, where $\mathbf{H} = \partial H/\partial \mathbf{x}$ is the tangent linear of the observation operator H, and $\mathbf{d}_k = \mathbf{y}_{obs} - H(\mathbf{x}_b + \mathbf{C}\mathbf{z}_k)$. It consists of an outer loop (k) and an inner loop for each outer loop. Assume that at the kth iteration ($k = 1, 2, \ldots$), $\mathbf{z} = \mathbf{z}_k$ and $\mathbf{x} = \mathbf{x}_k$ (notice that $\mathbf{z}_1 = 0$ and $\mathbf{x}_1 = \mathbf{x}_b$). The value of $\mathbf{z}$ at the $(k+1)$th iteration, $\mathbf{z}_{k+1}$, is found by adding a small perturbation δ_k (to be determined by an inner loop) to $\mathbf{z}_k$; i.e.

$$\mathbf{z}_{k+1} = \mathbf{z}_k + \delta_k. \tag{8}$$

The value of δ_k is found by solving the following linear equation

$$\frac{\partial J}{\partial \mathbf{z}}\bigg|_{\mathbf{z}_k} + \left(\frac{\partial}{\partial \mathbf{z}}\frac{\partial J}{\partial \mathbf{z}}\right)\Bigg|_{\mathbf{z}_k} \delta_k = 0 \tag{9}$$

using a conjugate-gradient algorithm (the inner loop). Equation (9) can be written in an explicit form:

$$\underbrace{\left(\mathbf{I} + \mathbf{C}^T\mathbf{H}_k^T\mathbf{R}^{-1}\mathbf{H}_k\mathbf{C} + \frac{\partial^2 J_c}{\partial \mathbf{z}^2}\right)}_{\left(\frac{\partial}{\partial \mathbf{z}}\frac{\partial J}{\partial \mathbf{z}}\right)\Big|_{\mathbf{z}_k},\ \text{coefficients}} \delta_k = -\underbrace{\left(\mathbf{z}_k - \mathbf{C}^T\mathbf{H}_k^T\mathbf{R}^{-1}\mathbf{d}_k + \frac{\partial J_c}{\partial \mathbf{z}}\right)}_{\frac{\partial J}{\partial \mathbf{z}}\Big|_{\mathbf{z}_k},\ \text{forcings}} \tag{10}$$

The data-assimilation cycle using 3D-Var is carried out at six-hour intervals with analysis times (t_i) centred at 00 UTC, 06 UTC, 12 UTC and 18 UTC. Data within $t_i \pm 3$ h are taken as observations at time t_i in our 3D-Var experiments (in the current operational version, observation increments are calculated at observational times through time interpolation). The six-hour model forecast starting from the 3D analysis at the analysis time t_i is used as the background field $\mathbf{x}_b$ for the 3D analysis at the next analysis time t_{i+1}. All the experiments are carried out at the T62L28 resolution.

(*b*) *4D-Var system*

To facilitate the system validation, 4D-Var experiments were made very similar to those of 3D-Var. The two systems have the same background covariance matrix, the same observational data, the same spectral resolution used during the minimization, and the same number of iterations. The only difference between 3D-Var and 4D-Var was that, in 4D-Var, observations were grouped into a smaller one-hour time interval instead of a six-hour interval, as in 3D-Var. Specifically, a modification is made to the J_o term (the second term in (4) is $2 \times J_o$) of the total cost function in 4D-Var experiments:

$$J_o = \frac{1}{2} \sum_j [H\{\mathbf{A}^{-1} M_{t_j}(\mathbf{A}\mathbf{x}_a)\} - \mathbf{y}_{obs}(t_j)]^T \mathbf{R}^{-1} [H\{\mathbf{A}^{-T} M_{t_j}(\mathbf{A}\mathbf{x}_a)\} - \mathbf{y}_{obs}(t_j)], \quad (11)$$

where $\mathbf{A}$ is the transformation from analysis variable $\mathbf{x}_a$ to model variable $\mathbf{x}$, $t_j = t_0 + j\Delta t$, $\Delta t = 1$ h, and t_0 is the time at the beginning of the six-hour assimilation window (e.g. $\mathbf{x}(t_0) \equiv \mathbf{x}_a$). Therefore, NCEP's global spectral model (represented by the M operator) is incorporated into the NCEP Spectral Statistical Interpolation analysis system, forming the so-called NCEP 4D-Var system. Data within $t_j \pm 0.5$ h are assimilated into the model at time t_j in 4D-Var. The global spectral model and its TLM are used to propagate the background field ($\mathbf{x}_b$) and analysis increments ($\mathbf{x} - \mathbf{x}_b$) in time.

The linear equation solved in 4D-Var, corresponding to (10) solved in 3D-Var, is modified into

$$\left(\mathbf{I} + \sum_j \mathbf{C}^T \mathbf{A}^T \mathbf{M}_{t_j}^T \mathbf{A}^{-T} \mathbf{H}_k^T \mathbf{R}^{-1} \mathbf{H}_k \mathbf{A}^{-1} \mathbf{M}_{t_j} \mathbf{A}\mathbf{C} + \frac{\partial^2 J_c}{\partial \mathbf{z}^2}\right) \delta_k = \mathbf{F}$$

$$\mathbf{F} = -\left(\mathbf{z}_k - \sum_j \mathbf{C}^T \mathbf{A}^T \mathbf{M}_{t_j}^T \mathbf{A}^{-T} \mathbf{H}_k^T \mathbf{R}^{-1} \mathbf{d}_k(t_j) + \frac{\partial J_c}{\partial \mathbf{z}}\right), \quad (12)$$

where $\mathbf{d}_k(t_j) = \mathbf{y}_{obs}(t_j) - H\{\mathbf{A}^{-1} M_{t_j} \mathbf{A}(\mathbf{x}_b + \mathbf{C}\mathbf{z}_k)\}$, M_{t_j} is the NLM operator representing the operations performed in the NLM to obtain the model forecast at t_j from an initial condition at time t_0, $\mathbf{M}_{t_j}$ is the tangent linear operator representing the operations performed in the TLM to obtain the model forecast of perturbation at t_j from an initial perturbation at time t_0, and $\mathbf{M}_{t_j}^T$ is the adjoint model operator representing the operations performed in the adjoint model to obtain the gradient of a forecast aspect J with respect to the initial condition at time t_0, given the forcings $\partial J / \partial \mathbf{x}_{t_j}$ at time t_j.

(*c*) *Experiment design*

A one-week period of data-assimilation cycles from 00 UTC 15 February to 00 UTC 21 February 1998 was chosen for running both the 3D-Var and 4D-Var experiments. This period during February 1998 was selected because of the North Pacific Experiment (NORPEX) (Langland *et al.* 1999), which took place in the north-eastern Pacific Ocean

TABLE 1. THE CPU TIMES ON THE NCEP CRAY C90 COMPUTER

Integration	CPU times (s)
Six-hour integration	
Nonlinear model	140
Tangent linear model	250
Adjoint model	304
20 iterations	
Six-hour 4D-Var	13 850
3D-Var	220

and collected targeted airborne dropwindsonde data. For every 3D-Var and 4D-Var minimization, we carried out 20 iterations in total over two outer loops. The small number of 20 iterations with two outer loops was used for a one-week period of data-assimilation cycles to save computational resources, and was determined by examining the convergence rate during a few minimizations with more than 20 iterations. It was found that the decrease in the values of the cost function during the initial 20 iterations was 75% of the total reduction in 100 iterations. The large number of total 100 iterations with four outer loops was carried out for the first one-day cycle to examine the extra benefit of more iterations.

The first 3D-Var minimization was carried out at 00 UTC 15 February 1998 (t_0), followed by a six-hour model forecast starting from the first analysis. The second 3D-Var was carried out at 06 UTC 15 February 1998 (t_1) with the six-hour forecast from the previous analysis time used as the background field, followed by the next cycle of data assimilation. The 4D-Var assimilation cycles started with the first minimization being carried out in a six-hour time window of $[(t_0 - 3\text{ h}), (t_0 + 3\text{ h})]$ in order to include the same observations as the first 3D-Var experiment. The initial condition at the zeroth iteration at the time $(t_0 - 3\text{ h})$ used the NCEP background field at 21 UTC 14 February 1998. The ith 4D-Var experiment was carried out in the ith six-hour window $[(t_i - 3\text{ h}), (t_i + 3\text{ h})]$, where $(t_i = t_0 + 6 \times i\text{ h})$, using the six-hour forecast starting from the previous 4D-Var analysis at $(t_{i-1} - 3\text{ h})$.

In 4D-Var, the minimization calls the tangent linear model twice and the adjoint model once for each inner loop. The NLM is called for each outer loop. The TLM and the adjoint model used in the 4D-Var assimilation include all the physical processes except radiation, and have the same resolution as the nonlinear full forecast model (T62L28). Computational costs on the NCEP CRAY C90 computer for a six-hour time integration of the NLM, TLM and adjoint model, and for the 3D-Var and 4D-Var with a total of 20 iterations (with two outer loops) are shown in Table 1. The TLM costs twice as much as the NLM, and the adjoint model costs more than twice (a factor of about 2.2) the cost of the NLM. Most of the computational expenses in 4D-Var are due to forward and backward time integrations of the forward models (the NLM and TLM) and the adjoint model. A CPU-time increase by a factor of about 60 is observed in replacing 3D-Var with 4D-Var, implying the need for 'wall-clock' time reduction before the 4D-Var can be considered for operational implementation. We mention that this factor of 60 may be overestimating the 4D-Var CPU increase for several reasons. First, the tangent linear and adjoint models are run at the same resolution as the NLM. If the 3D-Var and 4D-Var basic state were run at a resolution of, say, T212 instead of T62, the ratio between the 4D-Var CPU time and the 3D-Var CPU time could be reduced by a factor of 10. Second, the 3D-Var cost is underestimated due to the use of conventional

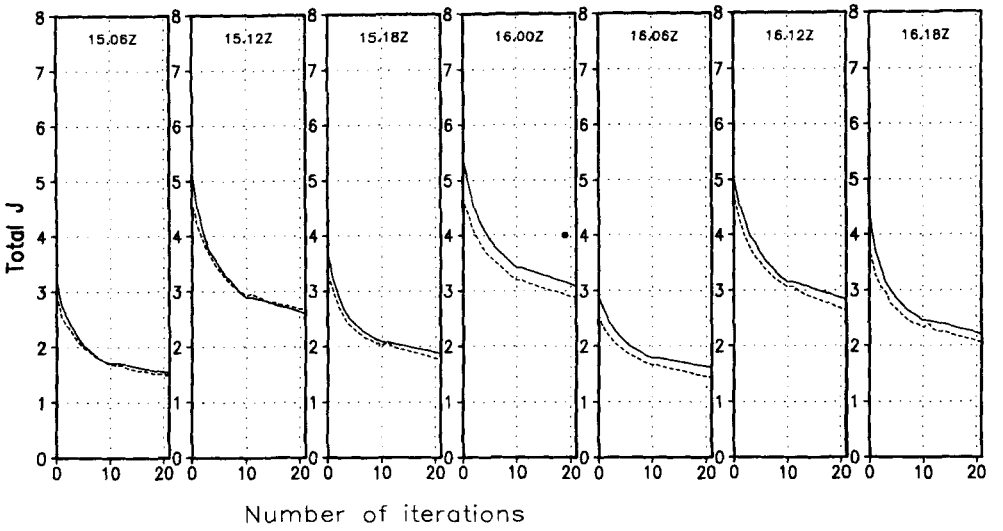

Figure 3. Variations of the values of the cost functions ($J = J_b + J_o + J_c$) during the minimization processes of the first seven assimilation cycles (from 06 UTC 15 February to 18 UTC 17 February 1998) for 3D-Var (solid line) and 4D-Var (dashed line). The 4D-Var was carried out including all the physics processes. The dot on the middle panel represents the value of the cost function calculated using the 3D-Var forecasts (the 3–6 hour forecast starting from 18 UTC 15 February for observations during the period 21 UTC 15 February to 00 UTC 16 February and the 0–3 hour forecast starting from 00 UTC 16 February for observations during the period 00 UTC to 03 UTC 16 February).

data only. If satellite data, such as GPS* occultation measurements of refraction angles (Zou *et al.* 2000), were included, the 3D-Var cost would be much higher and the ratio between 4D-Var and 3D-Var costs would be reduced. For example, if the 3D-Var cost is doubled due to the use of indirect observations, the ratio between 4D-Var and 3D-Var costs will be reduced from about 60 to about 30. Finally, the data communication is carried out by 'writing to' and 'reading from' a hard disk, which takes much more CPU time than keeping the data in memory.

4. THE FIT OF THE BACKGROUND AND ANALYSIS TO THE DATA

Figure 3 shows the evolution of the cost function for the seven continuous six-hour assimilation cycles from 06 UTC 15 to 00 UTC 17 February 1998 in the 3D-Var and 4D-Var experiments. The decrease of J in the 4D-Var results (dashed line) was very similar to those for 3D-Var (solid line). This implies that the nonlinearity and discontinuity associated with the complicated physics do not cause the minimization to fail, which is consistent with the previous work by Zou *et al.* (1993b), Zou (1996), and Zupanski and Mesinger (1995). It is interesting to note that, in the 4D-Var experiments, J at the beginning of all the assimilation cycles (at the point defined by the background field) was always smaller than in the 3D-Var results. The value of the cost function reached at the end of each assimilation cycle in 4D-Var was, in most cases, smaller than in 3D-Var, except at 12 UTC 15 February.

The background term J_b (the first term in (4)) and penalty term J_c were much smaller than the observation term J_o, and do not represent major contributors to the total cost function. The variation of mismatch between the observations and the model fields with respect to the number of iterations thus has a behaviour similar to the total

* Global Positioning System.

TABLE 2. THE ROOT-MEAN-SQUARE DIFFERENCES BETWEEN THE 3D-VAR AND 4D-VAR BACKGROUND AND ANALYSIS FIELDS AND THE OBSERVATIONS (VECTOR WIND, TEMPERATURE, SPECIFIC HUMIDITY AND SURFACE PRESSURE) OVER A ONE-WEEK PERIOD (THE FIRST CYCLE IS OMITTED)

Data assimilation	**v** ($m\ s^{-1}$)	T (degC)	q ($kg\ kg^{-1}$)	p_s (hPa)
		Background		
3D-Var	6.80	1.98	1.28	1.95
4D-Var	6.43	1.89	1.49	1.81
		Analysis		
3D-Var	3.87	1.61	1.13	1.49
4D-Var	3.83	1.64	1.31	1.46

cost function shown in Fig. 3 for both the 3D-Var and 4D-Var runs, and is not shown here. The fact that the fit to the observations is slightly better in 4D-Var than in 3D-Var is surprising, since in 4D-Var the forecast-model constraint is imposed to the time evolution of analysis increments, which is not required in 3D-Var. By examining the assimilation results carefully, we find that the better fit to observations in 4D-Var was due to the fact that the observational times were being considered more precisely in 4D-Var than in 3D-Var. In 4D-Var the fit to observations of model fields means the fit to observations of the fields from the assimilation window, $t_i - 3$ h to $t_i + 3$ h, at the time of the observations as defined by J_o, whereas in 3D-Var the fit as given by J_o is measured using a single 'analysis' time field, t_i, irrespective of the observation times. The nonlinear and linear forecast models are used for evolving the backgrounds and analysis increments in 4D-Var. Using the 3–6 forecasts from the previous 3D-Var analysis at 18 UTC 15 February, and the 0–3 hour forecasts from the current analysis at 00 UTC 16 February to calculate the value of the cost function, we obtained an even larger value of the cost function (the dot in the middle panel of Fig. 3) than in 3D-Var. This implies that the distance to the observations of the short-range forecast from the 3D-Var analysis is larger than that from the 4D-Var analysis. This suggests that 4D-Var may produce better quality time-continuous reanalysis data than 3D-Var, a task that is not limited by the operational constraint on computational time.

The fit of the background and analyses to the observations at each analysis time (3D-Var) or six-hour assimilation window (4D-Var) was computed and averaged over the one-week period of data-assimilation cycles, omitting the first cycle. Results are presented in Table 2. Compared with the 3D-Var analyses, the 4D-Var analyses fit the radiosonde wind and surface pressure data better, but fit the temperature and specific-humidity data less well. The fit of the 4D-Var background fields to the radiosonde temperature, wind and surface pressure observations is better than the fit of the 3D-Var backgrounds to the observations. The 4D-Var background and analysis of the specific humidity do not fit as well as those of the 3D-Var. This may be related either to large model errors in the humidity field, or to the inadequate specification of background-error variances for specific humidity in the 4D-Var, or to both. The weak dynamical link between the moisture and other variables and the very few observations in the tropics, where the specific humidity is largest, are also reasons for a poor moisture analysis. Further efforts are required to improve the global analysis of moisture variables.

TABLE 3. THE ROOT-MEAN-SQUARE DIFFERENCES BETWEEN THE 3D-VAR AND 4D-VAR ANALYSES AND THE CONVENTIONAL DATA (VECTOR WIND, TEMPERATURE, SPECIFIC HUMIDITY AND SURFACE PRESSURE) OVER LAND DURING A ONE-DAY PERIOD

Data assimilation	**v** ($m\ s^{-1}$)	T (degC)	q ($kg\ kg^{-1}$)	$\ln p_s$
		20 iterations		
3D-Var	5.04	1.52	0.81	1.52
4D-Var	4.91	1.59	0.88	1.49
		100 iterations		
3D-Var	4.49	1.29	0.73	1.39
4D-Var	4.31	1.38	0.87	1.33

The units for p_s are hPa.

5. 3D-VAR AND 4D-VAR RUNS WITH INCREASED NUMBER OF ITERATIONS

In order to see how the data-assimilation results change if more iterations (>20) are carried out for each analysis (for 3D-Var) or six-hour assimilation (for 4D-Var), and how the 4D-Var data-assimilation results differ from those of the 3D-Var, we repeated the first one-day assimilation cycles with 100 iterations for each minimization. Both the convergence and the spatial distributions of the rms differences between the analyses and the conventional data are examined.

The fit of analyses with 20 and 100 iterations to conventional data over land are shown in Table 3. Both the 3D-Var and the 4D-Var with 100 iterations improve the analysis fit to the observations in all model fields, compared with analyses with 20 iterations. Compared with 3D-Var, the fit of the 100 iteration 4D-Var analysis to the observations shows a similar result to the 20 iteration run: an improved fit in the wind and surface pressure fields, and a degraded fit in the temperature and specific-humidity fields. We mention, however, that the wind analysis over the ocean in the 100 iteration run of 3D-Var is degraded in comparison with its 20 iteration run. Further study is required to understand what could have caused this to happen.

Figure 4 shows the rms errors of the 3D-Var and 4D-Var analyses of wind for all the observations in the six-hour window centred at 00 UTC 16 February. The 4D-Var fit to observations was much better than that of 3D-Var at all times in the assimilation window. (Notice that the observations at the 3D-Var analysis time are in the middle of the 4D-Var assimilation window). Again, we note that the operational NCEP 3D-Var system uses three-hour, six-hour and nine-hour forecasts from the previous analysis for the calculation of the innovation vectors. This modification noticeably improves the 3D-Var fit to data at times different from the analysis times. This enhancement, however, is not included in these 3D-Var runs. We notice from Fig. 4(a) that the 4D-Var fit to the observations at the beginning of the assimilation window is poorest, which could be caused by the adjustment in the model state at the initial time (21 UTC 15 February) due to model errors when the observations at the future times (t > 21 UTC 15 February) are closely fitted.

To gain some insights into where the misfits come from, we examine the spatial distributions of the rms differences between the analyses and the observations at 00 UTC 16 February 1998 for each field (Figs. 5–9). For the wind analysis (Fig. 5), we find that the rms differences based on 3D-Var are largest over the western Atlantic and the Pacific Oceans (Fig. 5(a)). The 4D-Var procedure significantly reduced these differences

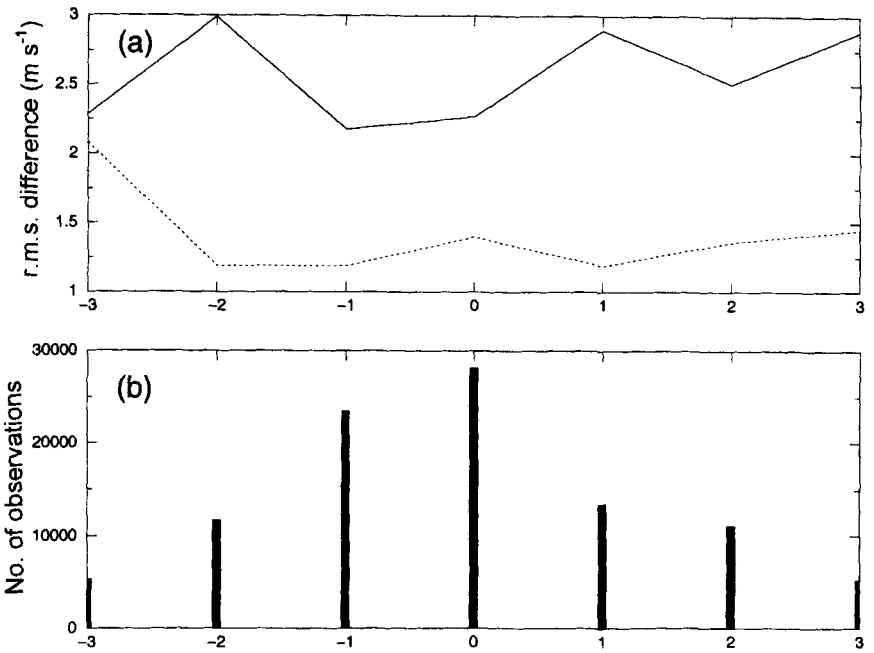

Figure 4. (a) The root-mean-square differences between the observations for the six-hour wind data centred at 00 UTC 16 February 1998 and the 3D-Var (solid line) and 4D-Var (dashed line) analyses after 100 iterations, and (b) the total number of wind observations at different times during the six-hour window. The horizontal axis indicates the time (in hours) before and after the central analysis time.

(Fig. 5(b)). The rms differences of the wind analyses are distributed more homogeneously in 4D-Var than in 3D-Var. The differences between the two rms differences (Fig. 5(c)) show more clearly where the 4D-Var improved the fit to the wind observations (negative areas) and where it fits the wind observations less well (positive areas). The 4D-Var reduced the wind analysis errors mostly over oceans near the eastern coasts in the northern hemisphere. The wind analysis errors are slightly increased over land near the western coasts, and some degradation is also seen in the eastern oceanic areas. The geographical locations where large improvements are observed in the fit to wind observations by 4D-Var are found to be closely correlated with the locations where a high number of wind observations are available (Fig. 5(d)). It seems that 4D-Var is able to fit the wind analysis better than 3D-Var. The maximum error reduction in the wind fields by 4D-Var is found to be over data-rich oceanic regions.

It is less obvious how the differences in the temperature analyses are associated with the distribution of observations (Fig. 6). Comparing the spatial distributions of the rms differences in temperature between 3D-Var and 4D-Var (Figs. 6(a) and (b)), it is surprising to find larger numbers of extrema of the rms differences from 3D-Var than in 4D-Var. For example, there are nine rms error extrema in the 3D-Var analysis where the rms errors of temperature exceed 4 degC, while there are only three such difference extrema in the 4D-Var analysis. The differences between the rms errors of 4D-Var and 3D-Var over areas where 3D-Var fits the observations better are below 1 degC. On the other hand, the differences between rms errors of 4D-Var and 3D-Var over areas where 4D-Var fits the observations better are as large as 5 degC. However, the positive areas of the 4D-Var analysis errors minus the 3D-Var analysis errors (Fig. 6(c)) cover a larger portion of the globe and include data-rich areas such as the USA, the east coast of China,

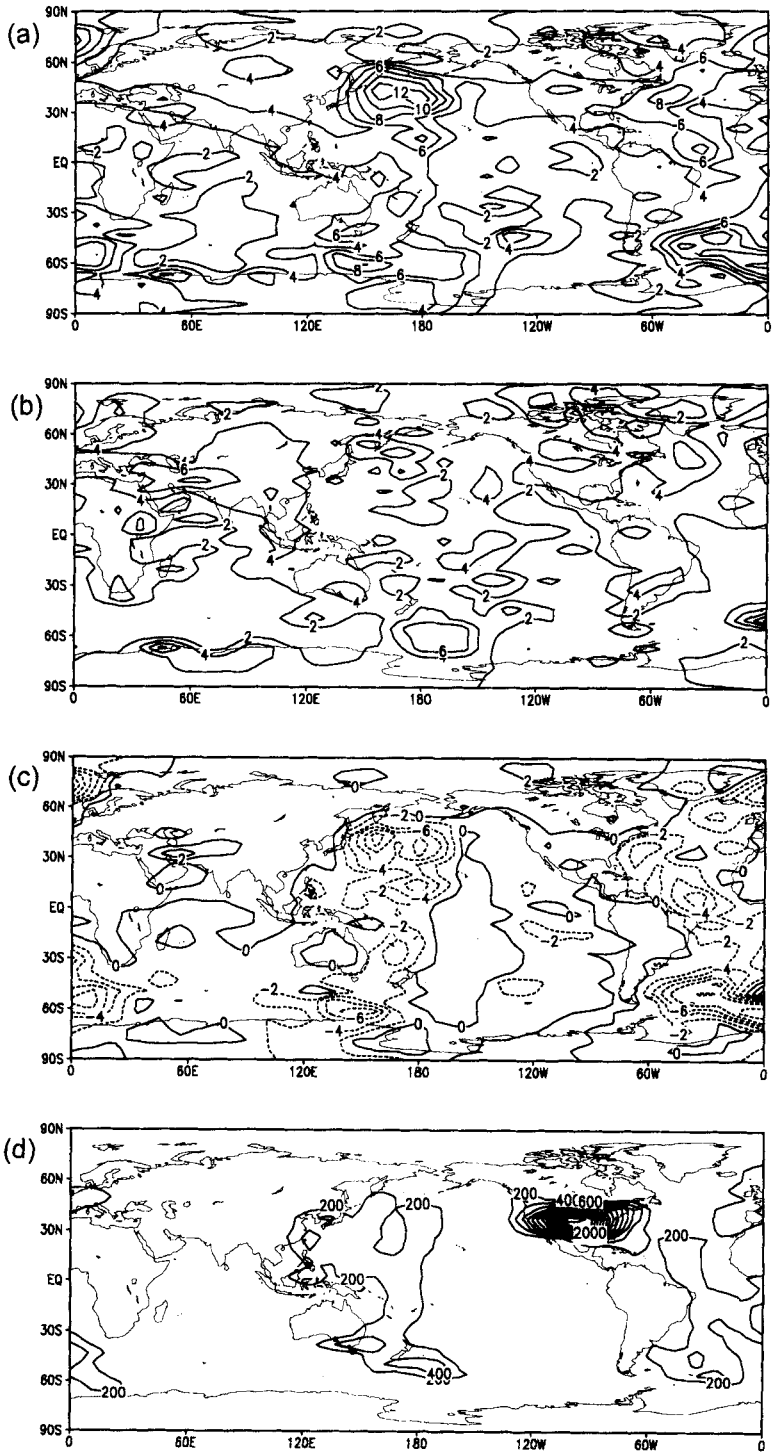

Figure 5. The spatial distributions of the root-mean-square differences of wind between the analyses and observations at 00 UTC 16 February 1998 after one-day cycles of data assimilation: (a) 3D-Var, (b) 4D-Var, (c) differences between 4D-Var and 3D-Var, and (d) the number of wind observations in a 4 × 6 grid box. Contour intervals are 2 m s^{-1} for (a), (b) and (c) and 200 for (d).

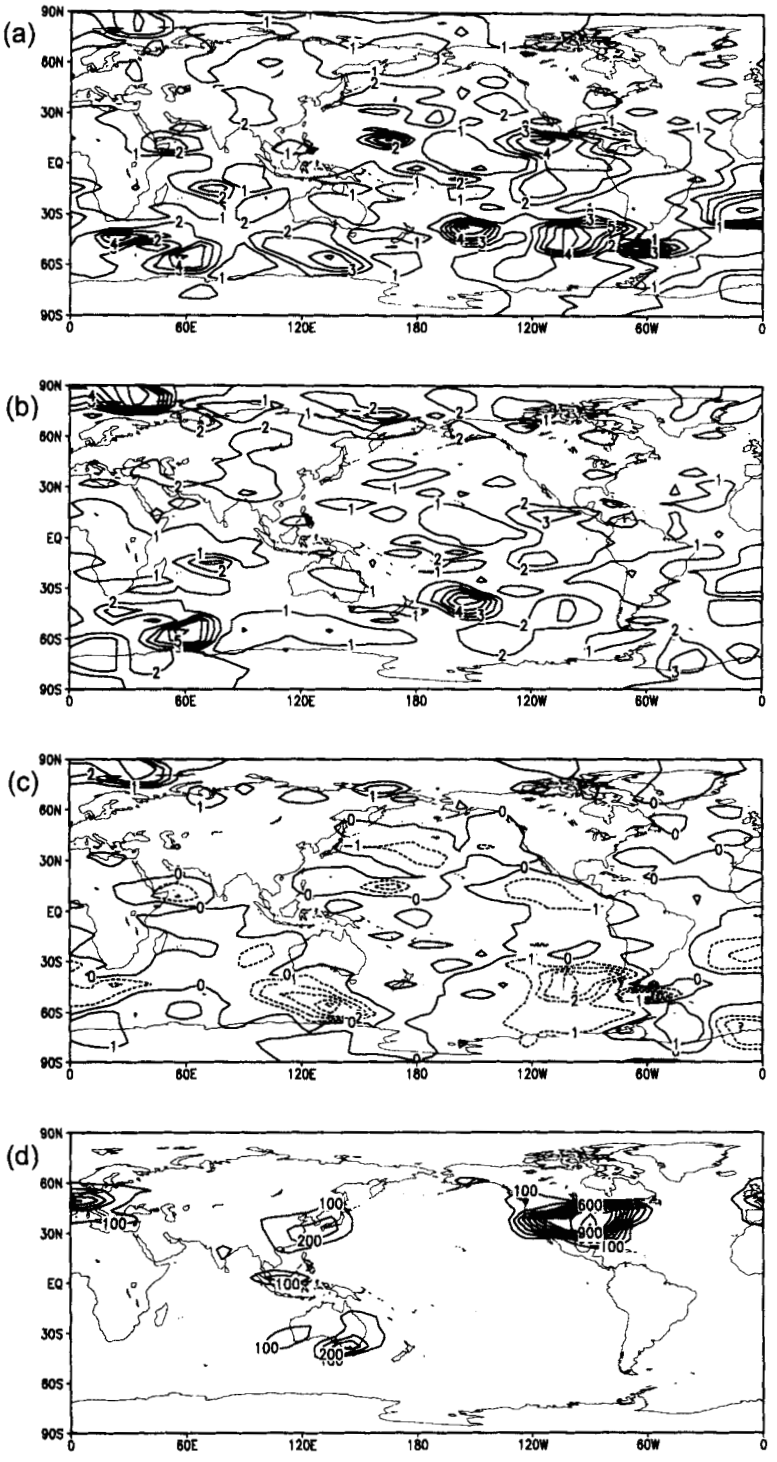

Figure 6. As Fig. 5, but for temperature. Contour intervals are 1 degC for (a), (b) and (c) and 100 for (d).

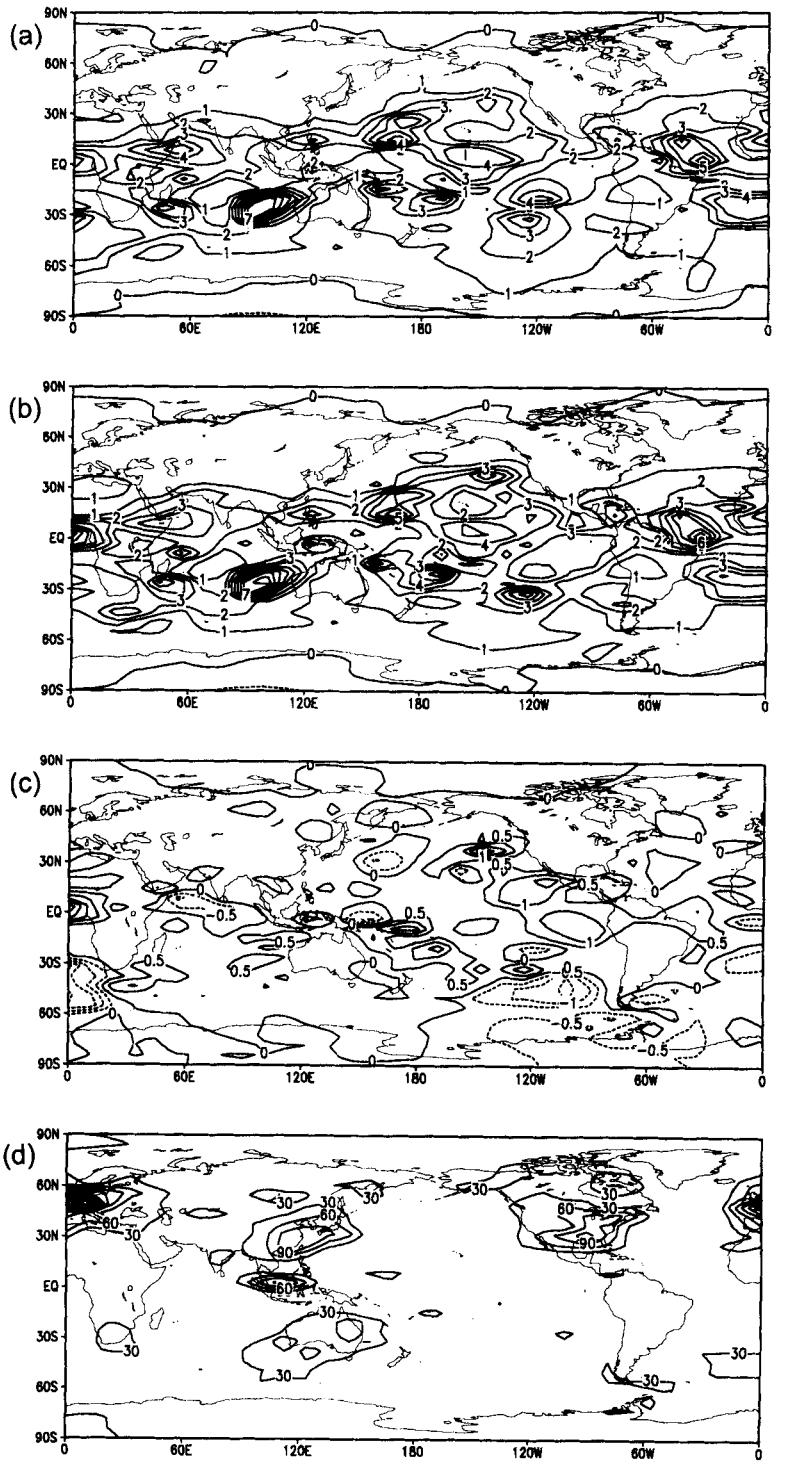

Figure 7. As Fig. 5, but for specific humidity. Contour intervals are 1 g kg^{-1} for (a) and (b), 0.5 g kg^{-1} for (c) and 30 for (d).

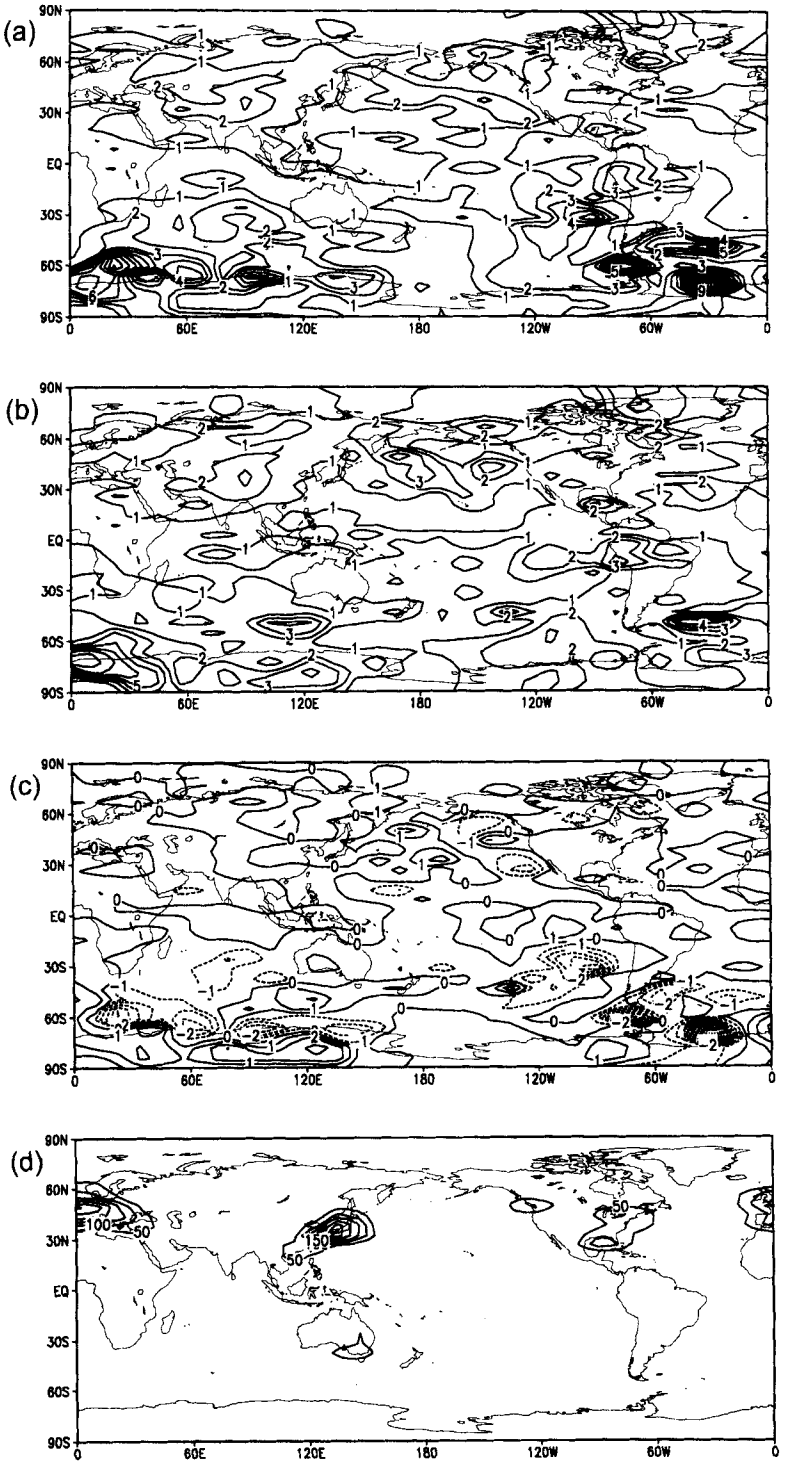

Figure 8. As Fig. 5, but for surface pressure. Contour intervals are 1 hPa for (a), (b) and (c) and 50 for (d).

and Australia (Fig. 6(d)). Globally, this results in larger rms differences in the 4D-Var temperature analysis than 3D-Var. Regionally, 4D-Var fits the observations much better than 3D-Var over several local areas, particularly those where 3D-Var differences are large.

Figures 7 and 8 are similar to Figs. 5 and 6 except that they are for the specific humidity and surface pressure fields, respectively. The rms differences of the moisture analyses in both 3D-Var and 4D-Var are largest over the data-sparse tropical oceanic regions (Fig. 7(a) and (b)). Over data-rich regions (see Fig. 7(d)), the analysis differences are small in both analyses. The difference maxima in 4D-Var correspond to the difference maxima in 3D-Var, except that maximum values are usually slightly higher in 4D-Var than in 3D-Var. However, some improvement in the 4D-Var moisture analysis fit to observations is found in the southern hemisphere (Fig. 7(c)). From these results, we feel that a slight modification to the background difference variance of moisture in 4D-Var may bring the moisture analysis differences in 4D-Var to a level similar to those in 3D-Var. Further change in the moisture analysis can be expected by including more satellite data, which are sensitive to the moisture content in the atmosphere over tropical regions. In general, the analysis differences of the surface pressure compared with the observations are smaller in 4D-Var than in 3D-Var over the globe. Large analysis differences in the surface pressure found in 3D-Var in the southern hemisphere (Fig. 8(a)) are significantly smaller in 4D-Var (Fig. 8(b)).

6. Forecast results from 3D-Var and 4D-Var analyses

An important test of analysis quality is the resultant accuracy of the forecasts. In order to obtain further understanding of the 4D-Var analysis results, 1–5 day forecasts have been conducted starting with the 3D-Var and 4D-Var analyses from 00 UTC 16 February to 00 UTC 21 February 1998. These forecasts are compared with the NORPEX targeted aircraft dropwindsonde data and the conventional radiosonde data.

(a) Forecast verification with NORPEX targeted dropwindsonde data

The NORPEX field program took place over the north-east Pacific Ocean for the period 28 January to 31 February 1998 to study the impact of targeted dropwindsonde and satellite observations on 1–4 day model forecasts of weather. Here, we use the targeted dropwindsonde data deployed by high-altitude jet aircraft as verification data for the 1–5 day forecasts from the 3D-Var and 4D-Var analyses. There were about 33 aircraft dropwindsonde profiles available around 00 UTC 20 February and 40 aircraft dropwindsonde profiles around 00 UTC 22 February 1998 over the mid-Pacific area north of Hawaii. These dropwindsonde data were distributed along the flight track (see dots in Fig. 10) in a baroclinic zone. The fit of the four-day forecasts from the 3D-Var and 4D-Var analyses at 00 UTC 16 February 1998 (after one day of data-assimilation cycles) is shown in Fig. 9. The forecast from the 4D-Var analysis produced, on average, a closer fit to the dropwindsonde data than the forecast from the 3D-Var analysis. The largest improvement of the 4D-Var forecast over the 3D-Var forecast was found near 500 hPa. The large differences between 3D-Var and 4D-Var analyses, when verified with dropsondes over the central Pacific, are partly due to the exclusion of all satellite radiance data from our experiments.

The differences in the fit of the four-day forecasts to the NORPEX dropwindsonde data at 00 UTC 20 February 1998 partially reflect the differences between the 3D-Var and the 4D-Var forecasts near a tropospheric jet region. Figure 10 shows the predicted wind speed distributions at 500 hPa (Fig. 10(a)) and the difference between the two

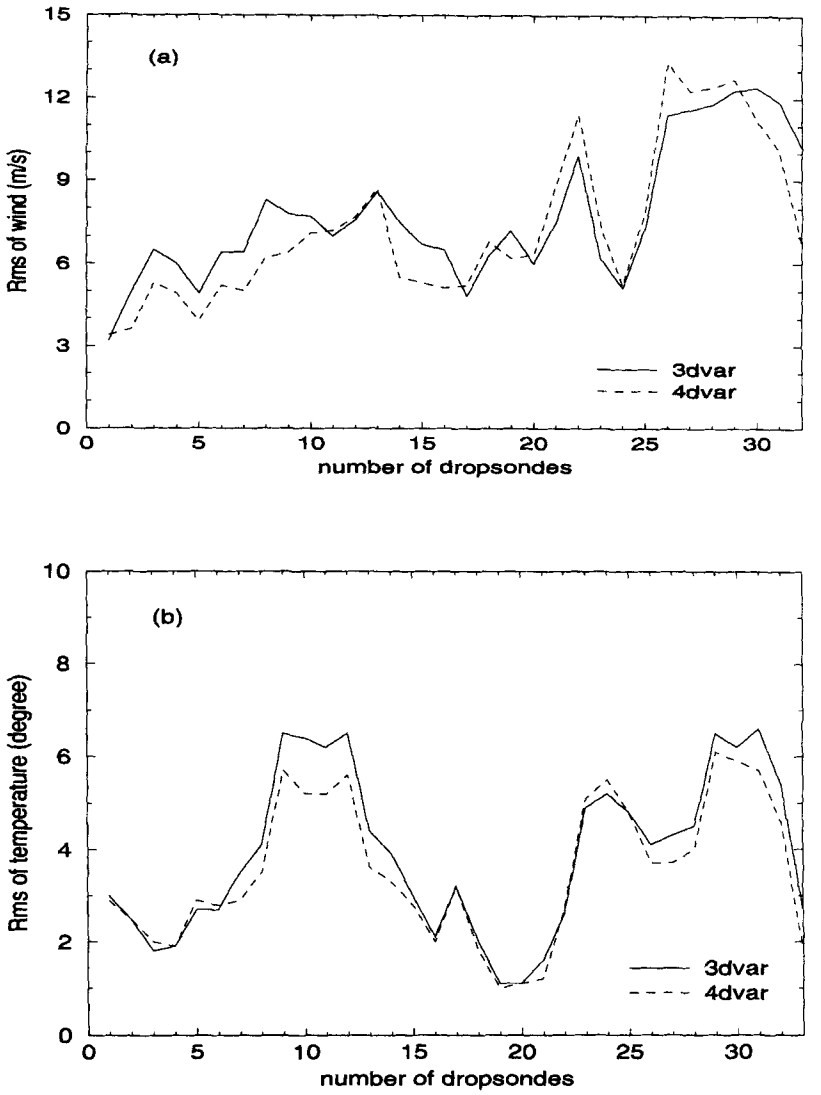

Figure 9. The root-mean-square differences of (a) wind (m s^{-1}) and (b) temperature (degC) between the four-day forecasts and all the dropwindsonde data at 00 UTC 20 February 1998 from 3D-Var (solid lines) and 4D-Var (dashed lines) experiments. The numbers of dropsondes are indicated in Fig. 10.

(Fig. 10(b)). The aircraft dropwindsonde data obtained in the 1998 NORPEX were located mostly in the south of the jet axis and the jet-exit region. The jet in the 4D-Var forecast is narrower and more to the west than the jet in the 3D-Var forecast. The wind-speed difference is negative at the jet-axis region and positive north and south of the jet in mid latitudes. The differences between the 3D-Var and 4D-Var forecasts and the dropwindsonde data could have been greater if these dropwindsonde verification data were located further to the north.

Differences in the temperature fields on the 500 hPa of the four-day forecast (Figs. 10(c) and (d)) are mainly in the distribution of the mid-latitude thermal ridge. The thermal ridge in the 4D-Var forecast is behind that of the 3D-Var forecast. Temperature differences between the 3D-Var and 4D-Var forecasts (Fig. 10(d)) show a positive–negative dipole structure along the thermal ridge, strongly reflecting the phase

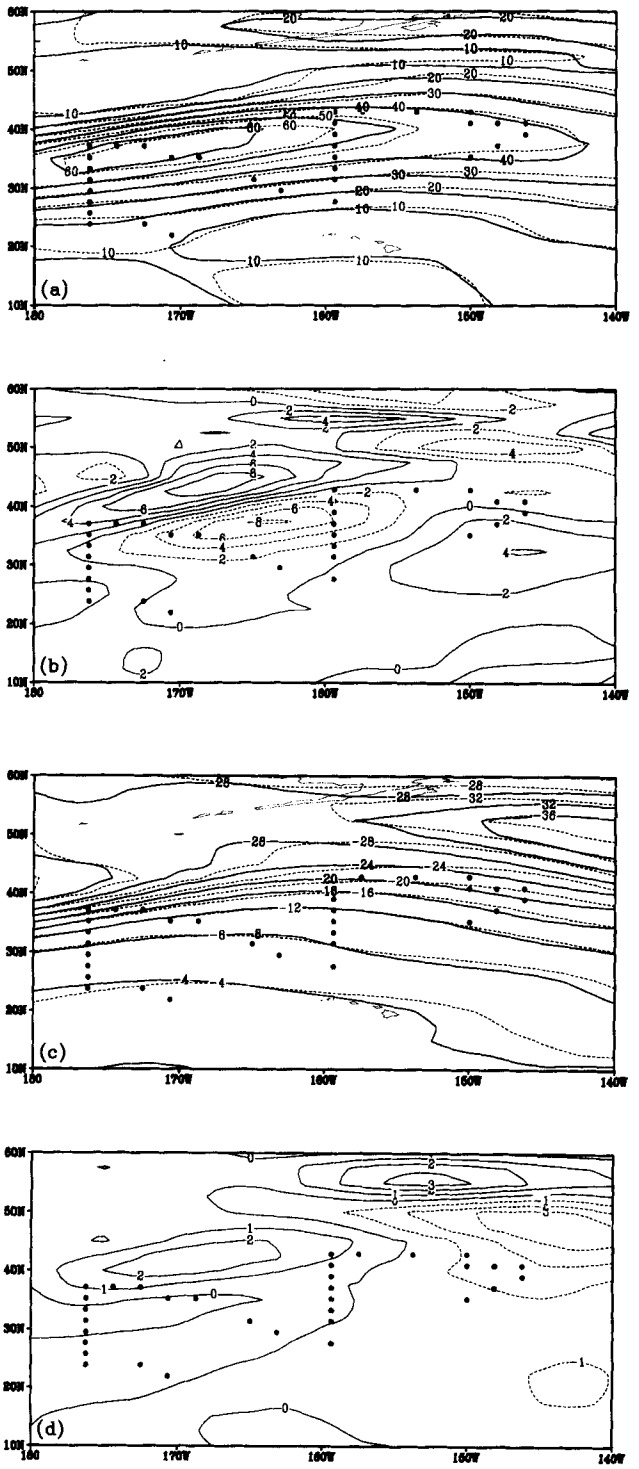

Figure 10. The four-day forecasts of (a) wind and (b) temperature on the 500 hPa surface at 00 UTC 20 February 1998 from the 4D-Var (solid lines) and 3D-Var (dashed lines) analyses at 00 UTC 16 February 1998 after a one-day cycle of assimilation, and the differences between the forecasts from the 4D-Var and 3D-Var analyses of (c) wind and (d) temperature on the 500 hPa surface at 00 UTC 20 February 1998. Contour intervals for (a), (b), (c) and (d) are 10 m s^{-1}, 2 degC, 4 m s^{-1}, and 1 degC, respectively.

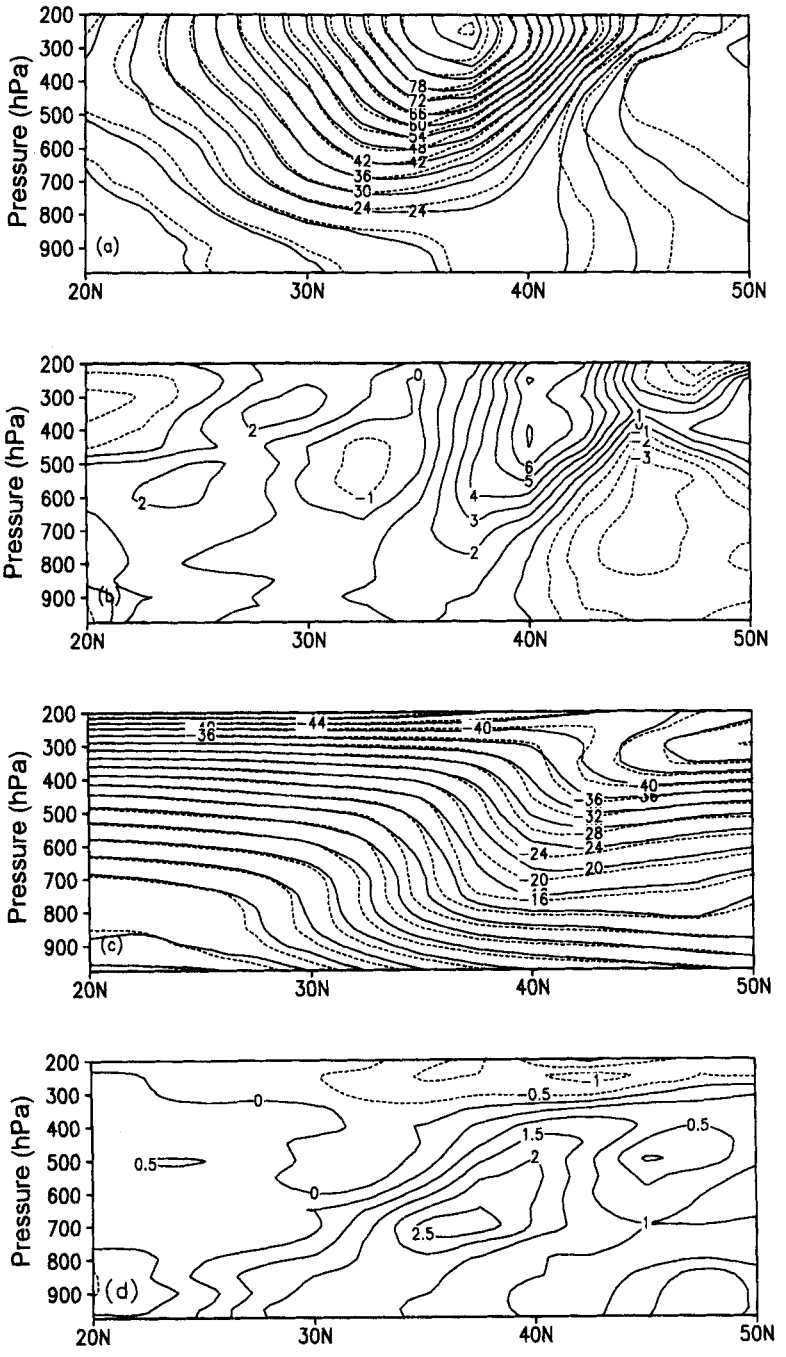

Figure 11. Cross-sections of the four-day forecasts of (a) wind and (c) temperature along the longitude 176°W at 00 UTC 20 February 1998 from the 4D-Var (solid line) and 3D-Var (dashed line) analyses at 00 UTC 16 February 1998 after a one-day cycle of assimilation, and of the differences between the forecasts from the 4D-Var and 3D-Var analyses of (b) wind and (d) temperature. Contour intervals for (a), (b), (c), and (d) are 6 m s^{-1}, 1 m s^{-1}, 4 degC, and 0.5 degC, respectively.

TABLE 4. THE ROOT-MEAN-SQUARE WIND ($m\ s^{-1}$) AND TEMPERATURE (degC) DIFFERENCES BETWEEN THE FORECASTS FROM THE 3D-VAR AND 4D-VAR ANALYSES AND NORPEX DROPWINDSONDE DATA AT 00 UTC FEBRUARY 1998

	1-day forecasts		2-day forecasts		3-day forecasts		4-day forecasts	
	3D-Var	4D-Var	3D-Var	4D-Var	3D-Var	4D-Var	3D-Var	4D-Var
Wind	5.44	5.18	6.48	6.50	10.41	9.34	8.06	7.65
Temperature	2.25	2.23	2.70	3.19	3.93	3.83	4.28	3.84

TABLE 5. THE ROOT-MEAN-SQUARE WIND ($m\ s^{-1}$) AND TEMPERATURE (degC) DIFFERENCES BETWEEN THE FORECASTS FROM THE 3D-VAR AND 4D-VAR ANALYSES AND CONVENTIONAL DATA OVER A ONE-WEEK PERIOD

	1-day forecasts		2-day forecasts		3-day forecasts		4-day forecasts		5-day forecasts	
	3D-Var	4D-Var	3D-Var	4D-Var	3D-Var	4D-Var	3D-Var	4D-Var	3D-Var	4D-Var
					Southern hemisphere					
Wind	7.6	5.41	8.14	6.43	8.34	7.13	8.50	7.90	8.73	8.18
Temperature	3.03	2.95	3.75	3.49	4.17	3.93	4.70	4.67	4.69	4.63
					Tropics					
Wind	6.99	6.24	7.69	7.24	9.32	7.91	8.69	8.36	9.50	9.19
Temperature	2.43	2.42	2.81	2.81	3.28	3.25	3.65	3.64	3.93	3.88
					Northern hemisphere					
Wind	8.39	7.85	10.18	9.87	11.73	11.72	13.36	13.33	15.16	15.24
Temperature	2.89	2.87	3.78	3.84	4.50	4.55	5.17	5.21	5.57	5.53

differences in the two forecasts. The magnitude of the temperature differences for the two four-day forecasts was as large as 2.5 degC at 500 hPa.

Figure 11 shows a cross-section of the temperature and wind fields along the south-to-north flight track at 176°W. We find that the mid-latitude front and the upper-level jet were stronger in the 4D-Var simulation than in the 3D-Var simulation. The maximum differences in wind and temperature between the two simulations were as large as 7 $m\ s^{-1}$ and 2.5 degC.

Since the data assimilation was done in a cycling mode, we can also examine the differences in the 1–4 day forecasts compared with the same dropwindsonde data at 00 UTC 20 February 1998. The overall performances of the 3D-Var and 4D-Var forecasts, ranging from 1–4 days and verified by these dropwindsonde data, are provided in Table 4. We find that, except for the two-day forecast, the one-day, three-day and four-day forecasts from the 4D-Var analyses, as verified by the NORPEX dropwindsonde data, are better than the forecasts from the 3D-Var analyses.

During the NORPEX experiment, targeted dropwindsonde observations were also available at 00 UTC 22 February over the north Pacific ocean. Comparing the 3D-Var and 4D-Var forecasts with these dropwindsonde observations, the same positive impact of 4D-Var is found.

(*b*) *Forecast verification with conventional observations*

In this section, the 1–5 day forecasts are compared with all the conventional data. The model forecasts were initialized with analyses obtained during the one-week period of data-assimilation cycles. Results are presented in Table 5.

Improvements in the wind and temperature forecasts due to the use of 4D-Var are found in the southern hemisphere and tropics. In the northern hemisphere, improvement to the model forecasts is found only in the wind field. The temperature forecasts show mixed results, with slight improvements in the one-day and five-day forecasts, and a degradation in the two- to four-day forecasts. These results are encouraging since the 4D-Var system used a small number of iterations, a static 3D-Var background-error covariance matrix, and conventional data only.

7. CONCLUSIONS AND DISCUSSIONS

This paper presents the first results produced using a newly developed NCEP 4D-Var system. A one-week period of experimentation with a resolution of T62L28, including a comprehensive set of physics parametrizations (with the exception of radiation) and the 1997 operational 3D-Var background formulation, have been studied. The one-week period was chosen from the NORPEX field experiment, for which targeted dropwindsonde data were available for forecast verification. We find that the very-short-range forecasts used as backgrounds were consistently closer to the data in 4D-Var than in 3D-Var. The analysis increments in 4D-Var at each analysis cycle were smaller than those in 3D-Var when observations were fitted to a similar level in both experiments. These results are consistent with the findings in Rabier *et al.* (2000) using ECMWF's 3D-Var and 4D-Var systems. Model forecasts from the 4D-Var analyses compared better than 3D-Var with the targeted dropwindsonde data available during the NORPEX experiment. The forecast verifications by conventional observations showed mixed results. The 1–5 day 4D-Var forecasts are better for both wind and temperature fields over the southern hemisphere and tropics when compared with 3D-Var forecasts. In the northern hemisphere, 4D-Var improved one- to four-day wind forecasts but degraded the five-day forecast. Temperature fields were improved on forecast days one and five, but degraded on forecast days two to four.

The numerical results of 4D-Var shown in this paper are mainly presented for the purpose of system validation. This first version of the 4D-Var system has been made as similar as possible to the 3D-Var system. The two systems used six-hour assimilation windows, the same amount of data, the same number of iterations, the same spectral resolution, and the same quality control for observations. Satellite data have not yet been included in the 4D-Var system. There is room for further improvement of the 4D-Var results by including satellite data, speeding up the convergence rate, and modifying the current use of 3D-Var background information in the 4D-Var system.

The high computational cost of the 4D-Var system has limited us to a one-week period of experimentation. A one-week assimilation is certainly not enough to draw firm conclusions on the performance of a forecast system. Aspects of the 4D-Var analysis, including reduction of rms increments and better fit of the background to observations, are more robust than the forecast scores, given the similarity of these analysis features to those obtained at ECMWF. The potential advantages of using 4D-Var in operations have not been investigated. The development of a diabatic global 4D-Var system, however, will permit the exploration of scientific and practical issues that are related to the operational use of 4D data assimilation. These issues include improving the minimization algorithm, modifying the 4D-Var background term, defining a new 4D-Var configuration with physical processes activated during selective periods of the minimization, optimizing the 4D-Var code, and running it on a multi-processor massively-parallel computer. More comprehensive tests, with longer periods of assimilation, will be conducted when some of these issues are resolved.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation under project No. ATM-9812729, and partially supported by the Integrated Program Office of the National Oceanographic and Atmospheric Administration's National Polar-orbiting Operational Environmental Satellite System under SMC/CIPN Project Order No. Q000C1737600086. The authors would like to thank Drs E. Kalnay and S. Lord for their persistent support for the NCEP 4D-Var system development.

APPENDIX

The correctness check of the 4D-Var system with 'full-physics' linear and adjoint models

First, the correctness of the tangent linear and adjoint operators of all the physical processes is carefully checked in isolation. These operators are then linked to the adiabatic version of the NCEP model and its adjoint. The correctness of the 'full-physics' TLM is tested against the NLM by a Taylor expansion and the correctness of the 'full-physics' adjoint model is tested against the 'full-physics' TLM through an algebraic equality equation (Navon *et al.* 1992).

Table A.1 shows some of the test results, From Table A.1 and many other test results not shown here, we feel that both the tangent linear and the adjoint models of the NCEP global forecast model with 'full-physics' are constructed correctly and are used for the further examination of the linear model with physics (section 2.8) and the 4D-Var experiments in section 4.

Having correctly developed the 'full-physics' tangent linear and adjoint models, we proceed with the linkage of these models to the NCEP 3D-VAR analysis system.

TABLE A.1. CORRECTNESS CHECK OF THE TANGENT LINEAR AND ADJOINT MODELS WITH 'FULL-PHYSICS'

(a) Tangent linear model test

α	$\Phi(\alpha) \equiv \dfrac{\|\ln p_s(\mathbf{x}_0 + \alpha\mathbf{h}) - \ln p_s(\mathbf{x}_0)\|}{\alpha\|\ln p_s'\mathbf{h}\|}$
10^{-1}	1.00000230
10^{-2}	1.00000860
10^{-3}	0.99999999
10^{-4}	1.00000000
10^{-5}	0.99999991
10^{-6}	0.99999843
10^{-7}	0.99999493
10^{-8}	0.99987260
10^{-9}	0.99821241
10^{-10}	1.04891570
10^{-11}	0.87474641

(b) Adjoint model test

$\mathbf{x}'^{\mathrm{T}}(t_R)\mathbf{x}'(t_R)$	$\mathbf{x}'(t_0)^{\mathrm{T}}\hat{\mathbf{x}}(t_0)$
$0.3318034167069\times10^{7}$	$0.3318034167091\times10^{7}$

$\mathbf{x}_0$ is the NCEP guess field at t_0 (18 UTC 20 February 1998), $\mathbf{h}$ is taken as $\mathbf{x}_0$, $\mathbf{x}'(t_R)$ is the two-hour forecast of the TLM, and $\hat{\mathbf{x}}(t_0)$ is the result of the adjoint model integration starting with $\hat{\mathbf{x}}(t_R) = \mathbf{x}'(t_R)$ at the time t_R.

TABLE A.2. GRADIENT CHECK WITH THE 'FULL-PHYSICS' ADJOINT MODEL

α	$\Phi(\alpha) \equiv \dfrac{J(\mathbf{x}_0 + \alpha\nabla_{\mathbf{x}_0} J) - J(\mathbf{x}_0)}{\alpha\|\nabla_{\mathbf{x}_0} J\|}$
10^{-1}	1.12857500
10^{-2}	1.01285740
10^{-3}	1.00128570
10^{-4}	1.00012860
10^{-5}	1.00001290
10^{-6}	1.00000120
10^{-7}	0.99999996
10^{-8}	0.99999936
10^{-9}	1.00000940
10^{-10}	1.00033100
10^{-11}	0.99490424
10^{-12}	1.00495380
10^{-13}	1.00495380
10^{-14}	2.00990760

$\mathbf{x}_0$ is the NCEP guess field at t_0 (18 UTC 20 February 1998).

A gradient check follows after the linkage is done. The cost function J used for the gradient check included observations over a six-hour window from 15 UTC 20 February to 21 UTC 20 February 1998. Test results are presented in Table A.2. We find that the gradient calculation using the 'full-physics' adjoint model has a similar degree of accuracy to that using the adiabatic adjoint model (Navon *et al.* 1992).

REFERENCES

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

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

Derber, J. C. 1985 'The variational 4-D assimilation of analysis using filtered models as constraints'. PhD thesis, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA

1989 A variational continuous assimilation technique. *Mon. Weather Rev.*, **117,** 2437–2446

Derber, J. C. and Bouttier, F. 1999 A reformulation of the background error covariance in the ECMWF global data assimilation scheme. *Tellus*, **51A,** 195–221

Derber, J. C. and Wu, W. S. 1998 The use of TOVS cloud-cleared radiances in the NCEP's SSI analysis system. *Mon. Weather Rev.*, **126,** 2287–2299

Helfand, H. M., Jusem, J. C., Pfaendtner, J., Tenenbaum, J. and Kalnay, E. 1986 'The effect of a gravity wave drag parameterization scheme on GLA fourth order GCM forecasts'. In 'Short- and medium-range numerical weather prediction', a collection of papers presented at the WMO/IUGG NWP Symposium, Tokyo, 4–8 August, 1986

Janiskova, M., Thépaut, J.-N. and Geleyn, J.-F. 1999 Simplified and regular physical parameterizations for incremental four-dimensional variational assimilation. *Mon. Weather Rev.*, **127,** 26–45

Klinker, E., Rabier, F., Kelly, G. and Mahfouf, J.-F. 2000 The ECMWF operational implementation of four-dimensional variational assimilation. III: Experimental results and diagnostics with operational configuration. *Q. J. R. Meteorol. Soc.*, **126,** 1191–1215

Langland, R. H, Toth, Z., Gelaro, R., Szunyogh, I., Shapiro, M. A., Majumdar, S. J., Morss, R. E., Rohaly, G. D., Velden, C., Bond, N. and Bishop, C. H. 1999 The North Pacific Experiment (NORPEX-98): Targeted observations for improved North American weather forecasts. *Bull. Amer. Meteorol. Soc.*, **80,** 1363–1384

Le Dimet, F. X. and Talagrand, O. 1986 Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus*, **38A,** 97–110

Lewis, J. M. and Derber, J. C. 1985 The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus*, **37A,** 309–322

Li, Z. and Navon, I. M. 1998 Adjoint sensitivity of the earth's radiation budget in the NCEP/MRF model. *J. Geophys. Res.*, **103,** 3801–3814

Mahfouf, J.-F. and Rabier, F. 2000 The ECMWF operational implementation of four-dimensional variational assimilation Part II: Experimental results with improved physics. *Q. J. R. Meteorol. Soc.*, **126,** 1171–1190

Miyakoda, K. and Sirutis, J. 1977 Comparative integrations of global models with various parameterized processes of subgrid-scale vertical transports: Description of the parameterization. *Beitr. Atmos. Phys.*, **50.** 445–488

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

Pan, H. L. and Wu, W. S. 1995 'Implementing a mass flux convection parameterization package for the NMC medium-range forecast model'. Technical Report for NMC/NOAA/NWS, No. 409

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

Pierrehumbert, R. T. 1986 'An essay on the parameterization of orographic gravity wave drag'. GFDL/NOAA report, Princeton University, Princeton, NJ08542

Rabier, F., Jarvinen, H., Klinker, E., Mahfouf, J.-F. and Simmons, A. 2000 The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.*, **126,** 1143–1170

Rabier, F., Thépaut, J.-N. and Courtier, P. 1998 Extended assimilation and forecast experiments with a four-dimensional variational assimilation system, *Q. J. R. Meteorol. Soc.*, **124,** 1861–1887

Sirkes, Z. and Tziperman, E. 1997 Finite difference of adjoint or adjoint of finite difference? *Mon. Weather Rev.*, **125,** 3373–3378

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

Thépaut, J. N. and Courtier, P. 1991 Four-dimensional variational data assimilation using the adjoint of a multilevel primitive equation model. *Q. J. R. Meteorol. Soc.*, **117,** 1225–1254

Xiao, Q., Zou, X. and Kuo, Y.-H. 2000 Incorporating the SSM/I derived precipitable water vapor and rain rate into a numerical model: A case study for ERICA IOP-4 cyclone. *Mon. Weather Rev.*, **128,** 87–108

Zhang, S., Zou, X., Alquist, J., Navon, I. M. and Sela, J. G. 2000 Use of differentiable and nondifferentiable optimization algorithms for variational data assimilation with discontinuous cost functions. *Mon. Weather Rev.*, **128,** 4031–4044

Zou, X. 1996 Tangent linear and adjoint of 'on–off' processes and their feasibility for use in four-dimensional variational data assimilation. *Tellus*, **49A,** 3–31

Zou, X. and Kuo, Y.-H. 1996 Rainfall assimilation through an optimal control of initial and boundary conditions in a limited-area mesoscale model. *Mon. Weather Rev.*, **124,** 2859–2882

Zou, X., Navon, I. M. and Sela, J. 1993a Control of gravity oscillations in variational data assimilation. *Mon. Weather Rev.*, **121,** 272–289

1993b Variational data assimilation with moist threshold processes using the NMC spectral model. *Tellus*, **45A,** 370–387

Zou, X., Wang, B., Liu, H., Anthes, R. A., Matsumura, T. and Zhu, Y.-J. 2000 Use of GPS/MET refraction angles in 3D variational analysis. *Q. J. R. Meteorol. Soc.*, **126,** 3013–3040

Zupanski, D. and Mesinger, F. 1995 Four-dimensional variational assimilation of precipitation data. *Mon. Weather Rev.*, **123,** 1112–1127

Zupanski, M. and Zupanski, D. 1995 'Recent development of NCEP's regional four-dimensional variational data assimilation system'. Pp. 367–372 in Proceedings of the second international symposium on assimilation of observation in meteorology and oceanography, Tokyo, Japan. World Meteorological Organization, Geneva, Switzerland