Science in China Series D: Earth Sciences

An explicit four-dimensional variational data assimilation method

QIU ChongJian†, ZHANG Lei & SHAO AiMei

Key Laboratory of Arid Climatic Changing and Reducing Disaster of Gansu Province, College of Atmospheric Sciences, Lanzhou University, Lanzhou 730000, China

A new data assimilation method called the explicit four-dimensional variational (4DVAR) method is proposed. In this method, the singular value decomposition (SVD) is used to construct the orthogonal basis vectors from a forecast ensemble in a 4D space. The basis vectors represent not only the spatial structure of the analysis variables but also the temporal evolution. After the analysis variables are expressed by a truncated expansion of the basis vectors in the 4D space, the control variables in the cost function appear explicitly, so that the adjoint model, which is used to derive the gradient of cost function with respect to the control variables, is no longer needed. The new technique significantly simplifies the data assimilation process. The advantage of the proposed method is demonstrated by several experiments using a shallow water numerical model and the results are compared with those of the conventional 4DVAR. It is shown that when the observation points are very dense, the conventional 4DVAR is better than the proposed method. However, when the observation points are sparse, the proposed method performs better. The sensitivity of the proposed method with respect to errors in the observations and the numerical model is lower than that of the conventional method.

data assimilation, four-dimensional variation, explicit method, singular value decomposition, shallow water equation

The four-dimensional variational data assimilation (4DVAR) has been a very successful technique and used in operational numerical weather prediction (NWP) of some weather forecast centers[1,2]. In this method the optimal estimate of initial condition of a forecast model is obtained by fitting the forecasts to observations within a time window. The attractive features of 4DVAR include: (1) the full-model is set as a strong dynamical constraint, and (2) it has the ability to assimilate the data at multiple time. However, the control variables (initial state) are expressed implicitly in the cost function. In order to compute gradient of the cost function with respect to the control variables, one has to integrate the adjoint model of the forecast model. But coding the adjoint for the 4DVAR and maintaining the adjoint, updated with the model upgrading, are extremely labor-intensive, especially when the forecast model is nonlinear and the model physics contain parameterized discontinuities[3,4]. Some researchers try to avoid integrating the adjoint model or reducing the expensive computation[5−7]. But the linear or adjoint model is still required in the methods mentioned above. So the three-dimensional variational data assimilation (3DVAR) becomes the common practice in many numerical weather forecast centers. The 3DVAR can be considered as a simplification of the 4DVAR, but it has lost the two advantages of the 4DVAR mentioned above. In general, the analysis results rely heavily on the information from the background field. However, as we know, in 3DVAR the background error covariance matrix is usually simplified and not flow-dependent, and so the background error covariance matrix cannot reflect the characteristics

Received October 26, 2006; accepted December 31, 2006
doi: 10.1007/s11430-007-0050-8
†Corresponding author (email: qiucj@lzu.edu.cn)
Supported by the 973 Program (Grant No. 2004CB418305) and the National Natural Science Foundation of China (Grant No. 40575049)

of the forecast error in detail[8−11]. In 4DVAR, the background covariance matrix is also used as a constraint for the analysis, and it is also simplified greatly and not flow-dependent. In recent years, the methods based on the extended Kalman filter have been used in many applications. The Ensemble Kalman Filter (EnKF) method is one of them[12,13]. By forecasting the statistical characteristics of the background errors with the Monte-Carlo method, the EnKF can provide flow-dependent error estimates of background errors. Some researchers have related 4DVAR to the EnKF. In their researches the background error matrix used in 4DVAR was replaced by the flow-dependent background errors matrix obtained using EnKF method[14,15]. However, like the traditional four-dimensional data assimilation, the adjoint model is still required because it remains the basic characteristics of the 4DVAR. Recently, Qiu and Chou[16] proposed a new method for four-dimensional data assimilation. They pointed out that the solution of the data assimilation should be restricted to the attractors of atmosphere dynamic equations in the phase space in order to reduce the degree of underdetermined problem. The basic idea is that a base that supports the attractor can be obtained from the forecast ensemble by performing a SVD analysis. The atmosphere state will be expressed by a truncated expression of the basis function. Based on this work, Shao and Qiu[17] designed an ensemble-based three-dimensional data assimilation scheme. The results showed that this method performed much better than the traditional three-dimensional variational data assimilation method. However, the analysis is only performed in a three-dimensional spectral space. Cao et al.[18] apply the technique of Proper Orthogonal Decomposition (POD) to the 4DVAR. This technique was shown to perform well, but the adjoint integration is still necessary in this method. If we apply the technique of singular value decomposition (SVD) to a four-dimensional forecast ensemble, the singular vectors not only express the spatial structure of the atmosphere state but also reflect the time evolution of the atmosphere state. After the model status is expressed by a truncated expansion of the basis vectors obtained by SVD, the control variables in the cost function appear explicitly, so that the adjoint model is no longer needed. Based on this idea an explicit four-dimensional variational data assimilation method is proposed in this paper. The method is expected to not only simplify the data assimilation procedure but also maintain the main advantages of the traditional four-dimensional variational data assimilation. Seven numerical experiments are performed with a two-dimensional shallow water equation model and simulated observations. Then a comparison is made between this method and the traditional four-dimensional variational data assimilation method.

1 Description of methodology

In principle, the four-dimensional variational data assimilation (4DVAR) analysis $\boldsymbol{x}_a$ is obtained through the minimization of a cost function J that measures the misfit between the model trajectory $H_k(\boldsymbol{x}_k)$ and the observations $\boldsymbol{y}_k$ at a series of times $\boldsymbol{t}_k, k = 1,...,K$.

The 4DVAR method can be defined as a process of minimizing the following cost functional:

$$J(\boldsymbol{x}_0) = (\boldsymbol{x}_0 - \boldsymbol{x}_b)^T \mathbf{B}^{-1}(\boldsymbol{x}_0 - \boldsymbol{x}_b) + \sum_{k=0}^{K}[\boldsymbol{y}_k - H_k(\boldsymbol{x}_k)]^T \mathbf{R}_k^{-1}[\boldsymbol{y}_k - H_k(\boldsymbol{x}_k)] \quad (1)$$

with the forecast model M imposed as strong constraints, defined by

$$\boldsymbol{x}_k = M_k(\boldsymbol{x}_0). \quad (2)$$

In (1) and (2) the superscript T stands for a transpose, b is a background value, the index k defines observational times, H_k is the observational operator that transforms the vector $\boldsymbol{x}$ from the model space to the vector $\boldsymbol{y}$ in the observational space. Matrices $\mathbf{B}$ and $\mathbf{R}$ are background and observational error covariances, respectively. The control variable is the initial conditions $\boldsymbol{x}_0$ (at the beginning time of the assimilation time window) of the model.

In the cost function the control variable $\boldsymbol{x}_0$ is connected with $\boldsymbol{x}_k$ through forward model and expressed implicitly, so it is difficult to compute the gradient of the cost function with respect to $\boldsymbol{x}_0$. For convenience we call the traditional four-dimensional variational data assimilation as implicit four-dimensional variational data assimilation (Implicit 4DVAR or I-4DVAR) and the new method proposed in this paper is called explicit four-dimensional variational data assimilation (Explicit 4DVAR or E-4DVAR).

Like the I-4DVAR, the E-4DVAR also needs to choose an assimilation time window. The four-dimensional sample ensemble is obtained from the forecast ensembles at multiple times produced by using the Monte Carlo method, which is similar to that in the ensemble Kalman filter (EnKF). Then the basis vectors

are generated by applying the singular value decomposition (SVD) technique to the matrix composed of the four-dimensional sample ensemble. The model states are then expressed by a truncated expansion of the leading SVD basis vectors. The SVD expansion coefficients can be determined by using a linear combination of the basis vectors to fit 4D innovation (observation minus background) data with least-squares fitting method. In this way, the control variables are transformed to the expansion coefficients and are expressed explicitly in the cost function. The details are described as follows.

Assuming there are K+1 observations y_k(k=0, K) at time $t = t_0, ..., t_k, ..., t_K$ during the assimilation time window $(0, T)$. For simplicity, we assume that the analysis time levels are the same as the observation time. Integrate the model from t_τ, a time before the starting time t_0, to t_k to produce the background field over the analysis time window. Generate M random perturbation fields and add each perturbation field to the initial background field at $t = t_\tau$ and integrate the model to produce a perturbed 4D field over the analysis time window. The mth difference field is then given by $\delta \boldsymbol{x}_m = \boldsymbol{x}_m - \boldsymbol{x}_b$ at time $t = t_0, ..., t_k, ..., t_K$, where $\boldsymbol{x}_b$ and $\boldsymbol{x}_m$ denote the background and the perturbed fields, respectively. After scaling the difference fields by using the stand covariance of the perturbation fields, M normalized forecast samples are obtained in the 4D space. Consider an ensemble of column vectors represented by matrix $\boldsymbol{A} = (\delta \boldsymbol{x}_1, \delta \boldsymbol{x}_2, ..., \delta \boldsymbol{x}_M)$, where the mth column vector $\delta \boldsymbol{x}_M$ represents the mth sampled data field in a discredited four-dimensional (4D) analysis space. The length of vector $\delta \boldsymbol{x}_M$ is $N \times M$, where $N = N_g \times N_v \times K$, N_g is the number of the model spatial grid points, N_v is the number of the model variables, and K is the number of selected time levels over each analysis time window. The SVD of $\boldsymbol{A}$ yields

$$\boldsymbol{A} = \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{V}^T, \tag{3}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix composed of the singular values of $\boldsymbol{A}$ with $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_r \geqslant 0$ and $\lambda_{r+1} = \lambda_{r+2} = ... = 0$, $r \leqslant \min(M, N)$ is the rank of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{V}$ are orthogonal matrix composed of the left and right singular vectors of $\boldsymbol{A}$, respectively[19]. The SVD in (3) gives $\boldsymbol{C} = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^T$ and $\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{B}\boldsymbol{\Lambda}^2\boldsymbol{B}^T$. Thus, the ith column vector of $\boldsymbol{V}$, denoted by $\boldsymbol{V}_i$, is the ith eigenvector of $\boldsymbol{C}$, while the jth column vector of $\boldsymbol{B}$, denoted by $\boldsymbol{b}_j$, is the jth column vector of $\boldsymbol{Q}$ and is called the singular vector of $\boldsymbol{A}$.

The truncated reconstruction of analysis variable $\boldsymbol{x}_a$ in four-dimensional space is given by

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \sum_{j=1}^{p} \boldsymbol{\alpha}_j \boldsymbol{b}_j, \tag{4}$$

where p ($\leqslant r$) is the truncation number. The solution for the forward model is approximately expressed by a truncated expansion of the singular vectors in a four-dimensional space. Substituting (4) into (1), the control variable becomes $\boldsymbol{\alpha}$ instead of $\boldsymbol{x}_0$, so the control variable is expressed explicitly in the cost function and the computation of the gradient is simplified greatly.

2 Numerical experiments

In this section, seven identical observing system simulation experiments are performed with a two-dimensional shallow-water equation model to test the proposed method. In addition, comparison is performed between the I-4DVAR (traditional four-dimensional variational data assimilation) method and the E-4DVAR (explicit four-dimensional variational data assimilation) method.

2.1 Set-up of experiments

The two-dimensional shallow-water equation model are formulated in the f-plane by

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - fv + g\frac{\partial h}{\partial x} = 0, \tag{5.1}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + fu + g\frac{\partial h}{\partial y} = 0, \tag{5.2}$$

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} + \frac{\partial (vh)}{\partial y} = \frac{\partial (uh_s)}{\partial x} + \frac{\partial (vh_s)}{\partial y}. \tag{5.3}$$

Here, $f = 10^{-4}\text{s}^{-1}$ is the Coriolis parameter, and h_s is the terrain height and is defined as

$$h_s = h_0 \sin(4\pi x / L_x) \sin(\pi y / L_y), \tag{6}$$

where h_0 = 200 m, $L_x = L_y$=13200 km are the length of two sides of the model domain, respectively, and the grid distance is $\Delta x = \Delta y$ =300 km. The model domain is square with 45×45 grid points and the periodic boundary conditions at x = 0 and L_x as well as y = 0 and L_y are stipulated. The spatial derivatives are discretized by the two-order central finite difference scheme. The local time derivatives are discretized by using the two-step

backward difference scheme of Matsuno[20] to ensure the computational stability and restrain the effect of computational damping (on short waves in particular). The time step Δt =360 s (equal to 6 minutes). The model state vector consists of the height h and the horizontal velocity components u and v at the grid points.

For the experiments, the "true" state is produced by integrating the "true" model (h_0=200 m) with the following initial conditions at the very beginning of the integration (48 hours before the starting time of the first data assimilation cycle):

$$h = 3000 + 240\sin(\pi y / L_y) + 120\cos(2\pi x / L_x)\sin(2\pi y / L_y), \tag{7.1}$$

$$u = -f^{-1} g \partial h / \partial y, \tag{7.2}$$

$$v = -f^{-1} g \partial h / \partial x. \tag{7.3}$$

The model-produced "true" fields at t = 0 (after 48 hour integration to the starting time of the first assimilation cycle) are plotted in Figure 1(a). In all the experiments, we assume that the simulated "observations" are only the height h and available at selected grid points (the details will be described later). If the observations are complete, the model-produced "true" fields correspond to the simulated observations. If the observations are incomplete, the simulated observations are generated by adding random errors to the above model-produced "true" fields. The statistical covariance of the random errors is 100 m^2. The imperfect initial field at t=0 in the first assimilation cycle, as shown in Figure 1(b), is the temporal average of every 3-hour outputs of previous 240 hours. This initial state is significantly different from the "true" state in Figure 1(a). In particular, the rms errors are 30.3 m/s, 1.56 m/s and 1.81 m/s for the h-, u- and v-fields, respectively, in this initial state.

2.2 Design of experiments

In each experiment, the above imperfect initial field (at t = 0) is used to initialized the model and the model is integrated from t = 0 to $t = T$ to produce the 4D background field. The length of the data assimilation cycle is set to T = 12 hour. For E-4DVAR method, by adding perturbations to the above imperfect initial state, the same model is integrated from t = 0 to $t = T$ to produce the perturbed 4D fields over the time window between $0 \leqslant t \leqslant \tau$ in the first assimilation cycle. By using the background field and perturbed 4D fields, the analysis is then performed, and the analyzed field is used to update the background state at $t = T$ (the ending time of the first cycle). After the first assimilation cycle, the model is integrated from t=T to t=2T for the next assimilation cycle, and so on so forth. In each experiment, the procedure goes through 10 cycles. The observations are available every 3 hours. The background fields are saved every 3 hours at the same time as the observations over each analysis time window. Each perturbed integration is initialized by adding a random field to the updated background state at the starting time of each cycle. Because the outputs of the perturbed integration at t = 0 do not reflect the model constraint, the samples are taken from 3rd hour to the ending time of an analysis time window. It means that the analysis and the observations are at the same time levels and the actual assimilation time window is 9 hours for E-4DVAR. Therefore, there are only observations at 4 time steps can be used during an analysis cycle. The ensemble size is M = 150 and the truncation number is p = 75 in all the experiments for E-4DVAR to ensure that the truncation error of the

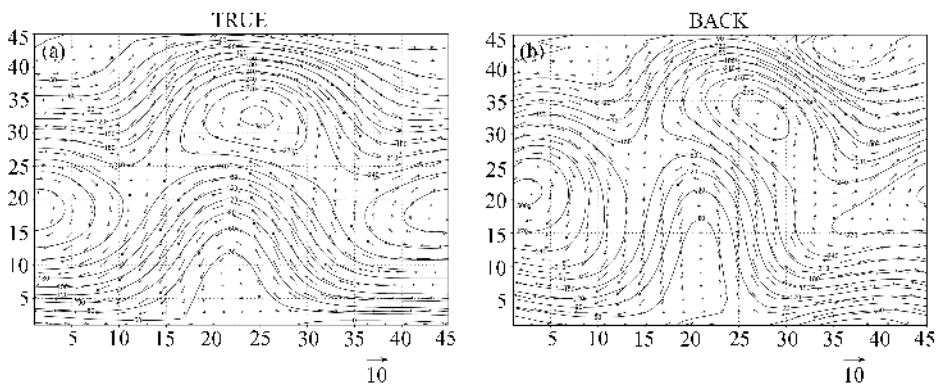

Figure 1 Model-produced "true" initial fields (a) and imperfect initial state (b) at the starting time (t = 0) of the first assimilation cycle. Contours are every 30 m for the height fields. Vectors are for the wind fields, and the vector scale (10 m/s) is labeled at the bottom of each panel.

sample fitting (quantified by the relative energy) is less than 5%. The perturbation fields are generated by using a quasi-random method proposed by Shao and Qiu[17] (see the appendix), which can guarantee appropriate co-relation to the perturbed field spatially. It is noted that, unlike EnKF, only one analyzed field is obtained in each analysis procedure in E-4DVAR and the initial condition should be perturbed at the starting time of the assimilation time window in each cycle.

For I-4DVAR, the assimilation time window is also set to $T=12$ hours and observations at 5 time levels can be used during an assimilation procedure. The same imperfect initial field (at $t=0$) is used to initialize the model in the first assimilation cycle.

If we consider the background constraint in 4DVAR, the analysis will heavily rely on the background error covariance which usually cannot be determined objectively. Because of this, all the experiments performed in this paper do not consider the background constraint. In addition, we assume that the observation errors are not co-related with each other. Under this situation the computation of the cost function becomes simple. In this paper, a memoryless quasi-Newton algorithm is used to find the minimum of the cost function in E-4DVAR and in I-4DVAR.

To evaluate the performance of the two algorithms, seven identical twin experiments are performed. The seven experiments are listed in Table 1. Here, 2025 observations imply that the height h observations are available at all the grid points within the model domain; 202 observations imply that only 202 observations (10 percent of all the grid points of the model domain) are available within the model domain. The locations of the observations are determined as the following fashion: 50 percent are concentrated randomly into the southwest quadrant of the domain; another 50 percent are distributed randomly within the rest area of the model domain. The distribution of 101 observations is similar to that of 202 observations except the number of the observations. One observation time implies that only the observations at the ending time of the assimilation time window are used; all observation times mean that all the observations at the observation times are used. If the model is imperfect, the maximum terrain height h_0 (see eq. (6)) in the forecast model is set to 300 m, instead of 200 m.

2.3 Experimental results

To evaluate the performance of the two algorithms, the relative error given by the rms deviation of the assimilated state from the true state divided by the rms deviation of the forecast state from the true state at the ending time of the first assimilation cycle and denoted as E, is considered separately for the three state fields, h, u and v. For h it is

$$E_n(h)=\frac{\sum_{i=1}^{S}(h_{n,i}^{a}-h_{n,i}^{t})^2}{\sum_{i=1}^{S}(h_{1,i}^{f}-h_{1,i}^{t})^2}.$$

Here, the index n defines the number of assimilation cycle, h_n^a and h_n^t are the analysis state and the "true" state at the ending time of nth assimilation cycle, respectively, h_1^f and h_1^t are the forecast and the "true" state of the forecast model at the ending time of the first assimilation cycle. The summation is made for all the grid points. For horizontal velocity components u and v the definition is similar. $E_n(V)=[E_n(u)+E_n(v)]/2$ is used to denote the total wind error.

The first and the second experiment are designed to demonstrate the advantages of assimilating multiple time observations in the four-dimensional variational data assimilation and to compare the performance of the two methods with respect to perfect model and complete observations. In both experiments, the height observations are available at all grid points of the model domain. The difference between the two experiments is that all the observations within the assimilation time window are assimilated in experiment 1 while only observations

Table 1 Experiments design

Experiment No.	Number of observations	Time level of observations	Observation errors	Model errors
1	2025	all	No	No
2	2025	1	No	No
3	202	all	No	No
4	101	all	No	No
5	202	all	Yes	No
6	202	all	No	Yes
7	202	all	Yes	Yes

at a particular time are used in experiment 2. Figure 2 shows the evolution of the relative error for height and wind through the assimilation cycles in each experiment. For I-4DVAR, when the height observations are available at all the grid points the height error decreases rapidly after first assimilation cycle and then keeps the small values through the assimilation cycles in both experiments. The wind field can be retrieved accurately when observations at 5 time steps are used, however, when only the observations at the ending time of the assimilation cycle are used in analysis procedure, the retrieved wind field is much worse than that of using all the observations, but it is still much better than that from E-4DVAR. For E-4DVAR, when only the height observations at the ending time of the assimilation cycle are assimilated, the height error is 0.3 after first assimilation cycle, which can be considered as the truncation error generated by using eq. (4). The height error decreases continually in the first three assimilation cycles and then is about the same in the later assimilation cycles. When the observations of 4 time levels are used, the height analyses are not so good as that when the observations at a single time level in the initial several cycles. The reason is that the larger truncation error is generated when the dimension of the variables is larger. However, after three cycles, the height analyses become better than that when using a single time observations, which is due to the obvious improvement of the wind field analyses through the assimilation. As shown in Figure 2(b), for E-4DVAR, the wind errors are much larger than the background errors in the initial three cycles when only the observations at a single time level are used. After four cycles, the analyses become better than the background. The results can be greatly improved if using four observation sets, even the analyses become better than that in I-4DVAR after several cycles. This implies that it is very useful to assimilate multiple time observations if we try to retrieve the variables which cannot be observed directly either for E-4DVAR or for I-4DVAR.

To evaluate the influence of the spacial density of observations on the two methods experiments 1, 3 and 4 are compared. The relative errors for experiments 3 and 4 are plotted in Figure 3. Although the observation density has an effect on both E-4DVAR and I-4DVAR, its influence on E-4DVAR is weaker than that on I-4DVAR. The height errors from E-4DVAR are not obvious difference between experiment 3 (202 observations) and experiment 4 (101 observations), and both of them are

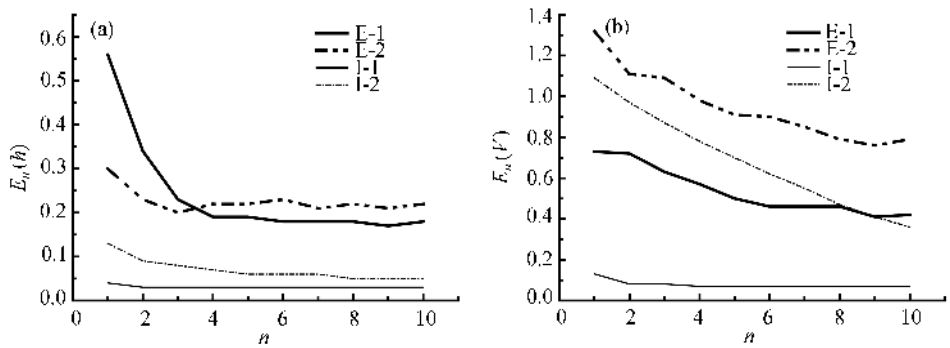

Figure 2 Relative error for height (a) and wind (b) plotted as functions of cycle number in experiments 1 and 2. E-1 and I-1 denote E-4DVAR and I-4DVAR methods for experiment 1, respectively, E-2 and I-2 denote E-4DVAR and I-4DVAR methods for experiment 2, respectively.

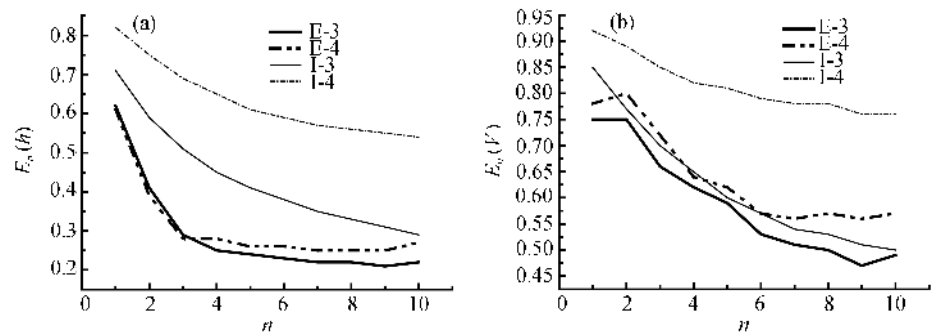

Figure 3 Relative error for height (a) and wind (b) plotted as functions of cycle number in experiments 3 and 4. E-3 and I-3 denote E-4DVAR and I-4DVAR methods for experiment 3, respectively, E-4 and I-4 denote E-4DVAR and I-4DVAR methods for experiment 4, respectively.

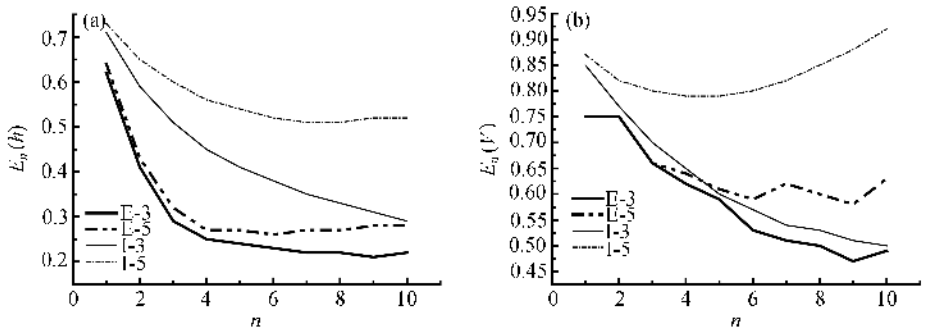

Figure 4 Relative error for height (a) and wind (b) plotted as functions of cycle number in experiments 3 and 5. E-3 and I-3 denote E-4DVAR and I-4DVAR methods for experiment 3, respectively, E-5 and I-5 denote E-4DVAR and I-4DVAR methods for experiment 5, respectively.

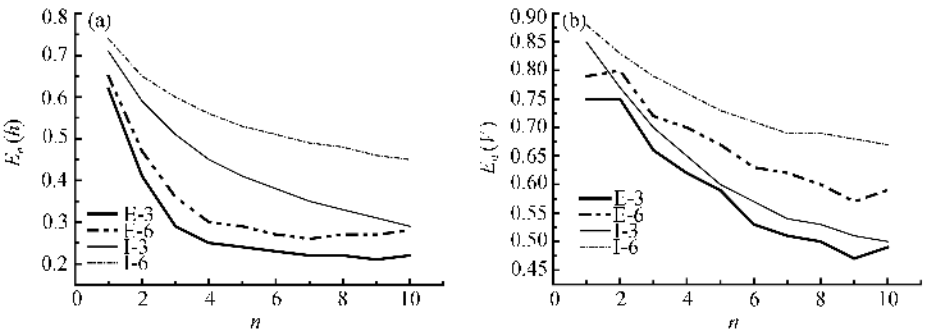

Figure 5 Relative error for height (a) and wind (b) plotted as functions of cycle number in experiments 3 and 6. E-3 and I-3 denote E-4DVAR and I-4DVAR methods for experiment 3, respectively, E-6 and I-6 denote E-4DVAR and I-4DVAR methods for experiment 6, respectively.

smaller than those from I-4DVAR. In particular, when the observations decrease from 202 (experiment 3) to 101 (experiment 4), the analyses become much worse in I-4DVAR. Similar situation can be seen in wind analyses. When there are 202 observations, the analyses of E-4DVAR are a little better than that of I-4DVAR. When there are 101 observations, the wind errors will increase a little through the assimilation cycles for E-4DVAR, but in I-4DVAR, the wind errors increase greatly.

To assess the effect of observation error on the analysis, two experiments (experiment 3 and experiment 5) are compared (Figure 4). The analysis from I-4dvar is much more sensitive to the observation errors than that from E-4DVAR, especially for the wind field. A possible explanation is that in I-4DVAR the analysis is performed in all the grid points and without background constraint, so the analysis heavily relies on the observations. However, in E-4DVAR, the analysis is performed in the truncated spectral space. Although a truncation error is generated in the analysis procedure, some observation noise can also be filtered.

In addition, a comparison is performed between experiments 3 and 6 to examine the possible effect of the imperfect model on the two methods. As shown in Figure 5, for E-4DVAR, when the model is imperfect, the height errors increase a bit, but the wind errors increase greatly, In particular, the larger the number of the assimilation cycles, the more obvious influence on analyses can be found through the 10 assimilation cycles. Although the model error has an effect on the analyses in E-4DVAR experiments, with the same experiment setup the imperfect model assumption has a greater influence on the analyses in I-4DVAR experiments. This implies that the E-4DVAR method can behave better than the I-4DVAR method with the imperfect model assumption.

When the errors in both the model and observations exist (experiment 7), the E-4DVAR method also behaves better than the I-4DVAR method. To compare analysis accuracy further of the two methods for all seven experiments, averaged relative errors over the last five cycles (from cycle 6 to cycle 10) are listed in Table 2. As shown by the first two columns in Table 2, only when the observations are very dense (for experiment 1 and experiment 2), which is difficult to realize, can the I-4DVAR perform better than E-4DVAR. Otherwise, the I-4DVAR always does not perform well compared with the E-4DVAR method.

Table 2 Averaged relative rms error for the analysis over the last five cycles (from cycle 6 to cycle 10) in seven experiments

Experiment No.		1	2	3	4	5	6	7
$\bar{E}_{6-10}(h)$	I-4DVAR	0.030	0.054	0.332	0.562	0.516	0.478	0.632
	E-4DVAR	0.178	0.218	0.220	0.256	0.272	0.207	0.320
$\bar{E}_{6-10}(V)$	I-4DVAR	0.070	0.482	0.530	0.774	0.854	0.688	0.960
	E-4DVAR	0.442	0.818	0.500	0.566	0.604	0.602	0.702

3 Summary and conclusions

In this paper, a new explicit four-dimensional variational data assimilation (E-4DVAR) method is proposed. In the method, the control variables in the cost function appear explicitly so the analysis is very straightforward and does not require the use of an adjoint integration. The method is robust even when the shallow-water equation model is imperfect and the observations are incomplete. The potential merits of this method and its comparison with the traditional four-dimensional variational data assimilation technique (I-4DVAR) are demonstrated by seven experiments. The main conclusions are summarized as follows:

(1) When the model is perfect and the observations are complete with a dense observation distribution, the E-4DVAR method does not perform as well as the I-4DVAR method, but when the observations are sparse, the E-4DVAR method performs much better than I-4DVAR method. In addition, the E-4DVAR method is less sensitive to the model errors and observation errors, so it is a very promising method and deserves further investigation in the future.

(2) Like the EnKF method, the quality of analysis relies on the number of the ensemble size used. The computation is expensive due to the perturbed integration which is required to produce the forecast sample ensemble during each assimilation cycle. But the parallel computation is easily applied in this method, and so the computation will not prevent it from applying to application in the long run. However, unlike EnKF, in E-4DVAR, the initial condition needs to be perturbed in each assimilation cycle. Therefore the quality of the results relies on the perturbation method. How to generate a reasonable perturbed field is a topic requiring further investigation.

(3) The truncation number has an effect on the analysis in E-4DVAR method, and it is associated with the observation variable, observation errors, ensemble size as well as the degree of freedom used. The choice of the optimal truncation number should be also handled carefully.

(4) If the background term is included in the cost function, it is expected that the quality of the analysis will be greatly improved, especially when the observations are sparse or contain a significant error. This can be easily implemented in the E-4DVAR, because only a few coefficients need to be determined. However, in the I-4DVAR method, the background covariance error usually is considered as homogeneous and needs to be inverted during the analysis procedure. In this paper, the potential of the assimilation methods may be underestimated, especially for the I-4DVAR method, since the background term is not included in the cost function.

Appendix: The method for generating perturbed field

The method for generating two-dimensional perturbed fields is expected to ensure that the generated perturbed fields approximately obey the Gaussian probability with variance σ and mean zero, and to have the ability to adjust the spatial correlation of the fields according to different requirements. The details are as follows:

(1) Given the number of samples is N, the stand deviation of the random field is σ, the admissible gross random values might be modeled over a predefined interval $[-3\sigma, +3\sigma]$. Divide the whole interval into m subintervals. Then estimate how many random values might be generated in each subinterval according to the Gaussian probability theory and generate the corresponding values. In this way, the generated perturbations approximately obey the Gaussian probability with variance σ and mean zero as long as the ensemble size is large enough.

(2) Assume the center point of the spatial domain is the starting point. The first value u_1 is generated randomly, and then the value of the neighbor grid point is generated randomly from the subinterval $[u_1 - \delta u_{1,2}, u_1 + \delta u_{1,2}]$, where $\delta u_{1,2} = r_{1,2}\sigma / L$, $r_{1,2}$ is the distance between two points, L is the characteristic length. The larger the characteristic length, the higher is the spatial

correlation of the random field. When $L = 0$, this method is equal to the Monte-Carlo method. In this paper, L is set to 1.5 to 3 grid distance. Similarly, the admissible interval of the kth grid point is given by $[u_1 - \delta u_{1,k}, u_1 + \delta u_{1,k}] \bigcap [u_2 - \delta u_{2,k}, u_2 + \delta u_{2,k}] \ldots \bigcap [u_{k-1} - \delta u_{k-1,k}, u_{k-1} + \delta u_{k-1,k}]$. After all random values at all grid points are generated, a two-dimensional perturbed field is obtained. In this way, each grid point related with its neighbor points and the correlation can be adjusted by changing the characteristic length.

(3) Repeat steps (1) and (2) to generate another perturbed field until all the perturbed fields are obtained. It is noted that the random values used in the previous procedure should not be used to generate later perturbed fields.

1 Lewis L, Deber J. The use of adjoint equations to solve a variational adjustment problem with advective constraint. Tellus, 1985, 37A: 309－322

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

3 Xu Q. Generalized adjoint for physical processes with parameterized discontinuities. Part I: basic issues and heuristic examples. J Atmos Sci, 1996, 53(8): 1123－1142

4 Mu M, Wang J. A method for adjoint variational data assimilation with physical "on-off" processes. J Atmos Sci, 2003, 60: 2010－2018

5 Kalnay E, Park S K, Pu Z X, et al. Application of the quasi-inverse method to data assimilation. Mon Wea Rev, 2000, 128: 864－875

6 Wang B, Zhao Y. A new data assimilation approach. Acta Meteorol Sin (in Chinese), 2005, 63: 694－700

7 Courtier P, Thepaut J N, Hollingsworth A. A strategy for operational implementation of 4D-Var using an incremental approach. Quart J R Meteor Soc, 1994, 120: 1367－1388

8 Courtier P, Andersson E, Heckley W, et al. The ECMWF implementation of three-dimensional Variational assimilation (3D-Var). I: Formulation. Quart J R Meteor Soc, 1998, 124: 1783－1807

9 Parrish D, Derber J. The National Meteorological Center's spectral statistical analysis system. Mon Wea Rev, 1992, 120: 1747－1763

10 Daley R, Barker E. NAVDAS: Formulation and diagnostics. Mon Wea Rev, 2001, 129: 869－883

11 Barker D M, Huang W, Guo Y R, et al. A three-dimensional variational data assimilation system for MM5: Implementation and initial results. Mon Wea Rev, 2004, 132: 897－914

12 Evensen G. Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res, 1994, 99(C5): 10143－10162

13 Burgers G, van Leeuwen P J, Evensen G. On the analysis scheme in the ensemble Kalman filter. Mon Wea Rev, 1998, 126: 1719－1724

14 Hamill T M, Snyder C. A hybrid ensemble Kalman Filter-3D Variational analysis scheme. Mon Wea Rev, 2000, 128: 2905－2919

15 Lorenc A C. Modelling of error covariances by 4D-Var Variational data assimilation. Quart J R Meteor Soc, 2003, 129: 3167－3182

16 Qiu C, Chou J. Four-dimensional data assimilation method based on SVD: Theoretical aspect. Theor Appl Climat, 2006, 83: 51－57

17 Shao A M, Qiu C. A numerical experiments study on a reduced-dimensional assemble assimilation method. Chin J Atmos Sci (in Chinese). 2006 (in press)

18 Cao Y H, Zhu J, Navon I M, et al. A reduced order approach to four-dimensional variational data assimilation using proper orthogonal decomposition. Int J Numer Meth Fluids, 2006 (in press)

19 Golub G H, Van Loan C F. Matrix Computations. Maryland: The Johns Hopkins Univ Press, 1983, 476

20 Matsuno T. Numerical integration of the primitive equations by a simulated backward difference method. J Meteor Soc, Japan, 1966, 44: 76－84