A Hybrid-Space Approach for Ensemble-Based 4DVar

Aimei Shao, Shuang Xi, Chongjian Qiu

Gansu Key Laboratory for Arid Climate Change and Reducing Disaster,

College of Atmospheric Sciences, Lanzhou University, Lanzhou, China

Qin Xu

NOAA/National Severe Storms Laboratory, Norman, Oklahoma, USA

(Submitted to *JGR*, January 14, 2008, revised 5/2/08, 7/16/08, 11/6/08)

Corresponding author address:

Dr. Qin Xu,

National Severe Storms Laboratory,

120 David L. Boren Blvd., Norman, Oklahoma 73072, USA

E-mail: Qin.Xu@noaa.gov

ABSTRACT

A new scheme is developed to improve the ensemble-based 4D variational data assimilation (En4DVar). In this new scheme, leading singular vectors are extracted from 4D ensemble perturbations in a hybrid space and then used to construct the analysis increment to fit the 4D innovation (observation minus background) data. The hybrid space combines the 4D observation space with only a gridded 3D subspace at the ending time of each assimilation cycle, so its dimension can be much smaller than the dimension of the fully gridded 4D space used in the original En4DVar. This improves the computational efficiency. With this hybrid-space approach, the background covariance matrix can be and only needs to be constructed in the 3D subspace while the analysis increment can fit the 4D innovation data in the observation space directly and also provide the necessary initial condition in the gridded 3D subspace directly for the model integration into the next assimilation cycle. In addition to the use of hybrid-space, the background error covariance model previously used in the original En4DVar is refined in consistency with the conventional ensemble Kalman filter and the analysis time window is shifted backward to avoid its caused forecast delay and shortened to avoid repeated uses of observations. The potential merits of the new scheme are demonstrated by assimilation experiments performed with an imperfect shallow-water equation model and simulated observations.

Key words: Data assimilation, Ensemble, Hybrid 4D space.

1. Introduction

In an ensemble-based filter, such as the ensemble Kalman filter (EnKF) (Evensen 1994, 2003; Burgers et al. 1998; Houtekamer and Mitchell 2001) and ensemble square root filter (EnSRF) – a variant of the EnKF (Anderson 2001; Bishop et al. 2001; Whitaker and Hamill 2002; Tippett et al. 2003), the probability density function of the model state is represented by an ensemble of state vectors, so the mean and covariance are directly estimated from and updated through the ensemble. This makes the filter easier to code than the three-dimensional variational assimilation (3DVar) (Lorenc 1981; Parrish and Derber 1992; Cohn et al. 1998; Daley and Barker 2001), and much easier to code and implement than the four-dimensional variational assimilation (4DVar) (Lewis and Derber 1985; LeDimet and Talagrand 1986; Bennett 1992; Courtier 1997), especially if the model physics (Talagrand 1997) contain parameterized discontinuities (Xu 1996, 1997; Xu and Qiu 1997). Computationally, however, an ensemble-based filter is still very expensive for operational applications due to large ensemble sizes required by the Monte Carlo method that the filter is based on. Thus, how to reduce the computation cost is a major problem for ensemble-based filters especially when they are designed for operational applications.

Another very challenging problem encountered by the EnKF concerns how to deal with model errors. The presence of unknown model errors, especially model biases, can cause the filter to diverge. To prevent the EnKF from filter divergence as a result of unknown model errors as well as other reasons, a variety of treatments, such as the covariance inflation, additive error and state augmentation method, have been proposed and examined. The covariance inflation and additive error treatments are empirical but effective if the unknown model errors are not large

(Anderson 2001; Hamill and Whitaker 2005). The state augmentation method appears to be promising as it can improve the filter performance substantially if the unknown true model errors can be adequately captured by the augmented state (Mitchell et al. 2002; Reichle et al. 2002; Baek et al. 2006; Zupanski and Zupanski 2006; Gillijns and Moor 2007). Nevertheless, the original model error problem remains largely unsolved. It is thus desirable to explore other possible approaches to alleviate this problem effectively from different perspectives.

Lorenc (2003) reviewed EnKF in comparison with 4DVar, and suggested that a hybrid method may be attractive for mesoscale NWP systems. Previously, a hybrid EnKF-3DVar scheme was proposed by Hamill and Snyder (2000), and a more sophisticated hybrid approach that combines EnKF and 3DVar was elaborated by Zupanski (2005). Inspired by these previous studies, a hybrid method, called ensemble-based 4DVar or En4DVar for short, was developed by Qiu et al. (2007) based on the singular value decomposition (SVD) approach proposed by Qiu and Chou (2006) for climate data assimilation. The SVD is also often called the proper orthogonal decomposition (POD). The use of SVD in Kalman and ensemble Kalman filters for atmospheric data assimilations started a decade ago with Todling and Cohn (1996, 1998), Heemink (2001), and more recently Uzunoglu et al. (2007). These previous studies showed the effectiveness of SVD in reducing the rank of the background covariance and improving the filter computational efficiency. The SVD can be also used to reduce the control-variable dimension and improve the efficiency of 4DVar (Cao et al. 2007). In the En4DVar of Qiu et al. (2007), the SVD was used to extract the leading singular vectors from an ensemble of 4D perturbation fields produced by the model, and then a linear combination of the extracted singular vectors is used to

fit 4D innovation (observation minus background) data to produce an incremental analysis in each assimilation cycle. The involved least-squares fitting is similar to that in 4DVar, but the leading singular vectors are used in place of the representer solutions (Bennett 1992), so the fitting is computationally much less expensive than 4DVar. As shown by the assimilation experiments performed with a shallow-water equation model in Qiu et al. (2007), the En4DVar can be also effective in reducing problems caused by model biases.

In the original En4DVar (Qiu et al. 2007), the analysis increment is a linear combination of leading singular vectors extracted from an ensemble of gridded 4D perturbations sampled from perturbed integrations of the prediction model, and this linear combination is projected into the observation space to fit the 4D innovation data in the costfunction. Note that the observation space dimension is often much smaller than the dimension of the gridded model state vector, and the gridded analysis increment is needed only at the ending time in each assimilation cycle to initialize the model integration into the next assimilation cycle. It is therefore not really necessary and can be wasteful to use fully-gridded 4D perturbations to construct the ensemble perturbation matrix and extract leading singular vectors. This implies that the dimension of the 4D perturbation fields sampled in the original En4DVar can be reduced substantially. Specifically, we can project the sampled perturbation fields into the 4D observation space and retain gridded perturbations only in a 3D subspace at the ending time of each assimilation cycle. The retained 3D perturbations and projected 4D perturbations can be combined into hybrid 4D perturbations to construct the ensemble perturbation matrix and extract leading singular vectors in a hybrid space. Since the hybrid-space dimension can be much smaller than the gridded 4D

space dimension in the original En4DVar, the computational efficiency can be improved. In addition, as will be demonstrated by the assimilation experiments in this paper, further improvements can be made to the En4DVar by refining the background error covariance model [see (8) in section 2.2] and by shifting the analysis time window backward to avoid its caused forecast delay (see the explanation at the end of section 2.1) and shortening the window to avoid repeated uses of observations (see the explanation at the end of section 2.2).

The hybrid-space approach considered in this study is similar to that proposed in Qiu and Chou (2006) but the ensemble is generated by a large number (about 150) of perturbed integrations and the background term is included in the costfunction, while the ensemble considered in Qiu and Chou (2006) was a time series sampled from a single integration of the model over a long time period and the background constraint was not considered explicitly. In this regards, the current hybrid-space approach is an extension of the original En4DVar (Qiu et al. 2007) for mesoscale and synoptic-scale data assimilation, while the hybrid-space approach in Qiu and Chou (2006) was proposed mainly for climate data assimilation and their proposed approach has not been tested numerically. The detailed method is described in the next section. Assimilation experiments are designed in section 3. The experiment results are presented in section 4 to demonstrate the potential merits of the new scheme in comparison with the original En4DVar. Conclusions follow in section 5.

2. Description of the method

2.1. Original En4DVar

In the original En4DVar, the ensemble perturbation matrix, denoted by $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_M)$, is composed of M column vectors, where M is the ensemble size. Each column vector in **A** is a 4D perturbation field (with respect to the ensemble mean field) produced by the model on the analysis grid over the analysis time window for each assimilation cycle. Each variable (height or velocity) field in **A** is normalized by its spatially averaged ensemble spread (RMS amplitude) computed at each observation time level in the analysis time window. The dimension of each column vector is $N_v \times N_x \times N_t$, where N_v is the number of the model variables, N_x is the number of model spatial grid points, and N_t is the number of observation time levels in each assimilation cycle. The singular vectors decomposition (SVD) of **A** yields

$$\mathbf{A} = \mathbf{B}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}, \tag{1}$$

where $\mathbf{\Lambda}$ is a diagonal matrix composed of the singular values of **A** with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ and $\lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_M = 0$, $r = \mathrm{rank}(\mathbf{A})$ is the rank of **A**, and **B** and **V** are orthogonal matrices composed of the left and right singular vectors of **A**, respectively (see theorem 2.3-1 of Golub and Van Loan 1983).

The background mean state, denoted by vector $\mathbf{u}_b$, is a gridded 4D field at the N_t observation time levels. The analysis state, denoted by vector $\mathbf{u}_a$, is also a gridded 4D field at the N_t observation time levels. The analysis increment, defined by $\delta\mathbf{u}_a = \mathbf{u}_a - \mathbf{u}_b$, is expressed by a linear combination of the leading singular vectors in **B**, that is,

$$\delta\mathbf{u}_a = \sum_{k=1}^{p} \mathbf{b}_k \beta_k = \mathbf{B}\boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{ß} = (\beta_1, \beta_2, \ldots \beta_M)^T$ with $\beta_{p+1} = \beta_{p+2} = \ldots = \beta_r = 0$ is a vector coefficient for the linear combination and p is the truncation number. For practical applications, we consider $M <$ or $<<$ $N_v \times N_x \times N_t$ and thus $p < r = \text{rank}(\mathbf{A}) \leq \min(M, N_v \times N_x \times N_t) = M$.

The truncated expansion in (2) is used to fit the observations at the N_t time levels by minimizing the following costfunction:

$$J = J_b + J_o = \delta\mathbf{u}_a^T \mathbf{P}^{-1} \delta\mathbf{u}_a + (\mathbf{H}\delta\mathbf{u}_a - \delta\mathbf{y})^T \mathbf{O}^{-1} (\boldsymbol{H}\delta\mathbf{u}_a - \delta\mathbf{y}), \quad (3)$$

where J_b and J_o denote the background and observation terms, that is, the first and second terms, respectively, on the righthand side, $\delta\mathbf{y} = \mathbf{y} - \boldsymbol{H}\mathbf{u}_b$ is the normalized 4D observation innovation, $\mathbf{y}$ is the normalized 4D observation vector, $\boldsymbol{H}$ denotes the normalized 4D observation operator, and $\mathbf{H}$ is the tangent-linearization of $\boldsymbol{H}$ at $\mathbf{u}_b$. These normalizations are done in the same way as $\mathbf{A}$ is normalized by scaling each variable (height or velocity) field with its spatially averaged RMS amplitude at each observation time level. For the experiments presented in Qiu et al. (2007) and in this paper, $\boldsymbol{H}$ is linear and thus $\mathbf{H} = \boldsymbol{H}$. $\mathbf{P}$ is the normalized background error covariance matrix, and $\mathbf{O}$ is the normalized observation error covariance matrix. In the original En4DVar, the normalized background error covariance was modeled by $\mathbf{P} = \mathbf{B}\mathbf{B}^T$, which assumed that the background error structures could be represented by the singular vectors of $\mathbf{A}$. As the background error variances were properly tuned (estimated and inflated) based on the true forecast RMS errors, the tuned background error variances should be close to the true error variances. Because of this, the original En4DVar appeared to be not very sensitive to the above

covariance model and performed better than the EnKF and EnSRF for the imperfect-model case.

Substituting (2) with $\mathbf{P} = \mathbf{B}\mathbf{B}^{\mathrm{T}}$ into (3) gives

$$J = \boldsymbol{ß}^{\mathrm{T}}\boldsymbol{ß} + (\boldsymbol{H}\mathbf{B}\boldsymbol{ß} - \delta\mathbf{y})^{\mathrm{T}}\mathbf{O}^{-1}(\boldsymbol{H}\mathbf{B}\boldsymbol{ß} - \delta\mathbf{y}), \quad (4)$$

where the vector coefficient **ß** is the control variable for the minimization of *J*. Once **ß** is determined by minimizing the costfunction in (4) and then substituted back into (2), the analysis increment becomes available for the model integration into the next assimilation cycle. Each assimilation cycle is performed and completed in three steps, and the detailed steps are described below (see sections 2 and 3.2 of Qiu et al. 2007).

I. Integrate the model by using the initial condition provided by the analysis at the ending time of the previous cycle to produce the 4D background field over the analysis time window of the current assimilation cycle. Generate *M* random perturbation fields that are not only homogeneous and isotropic but also spatially correlated for each variable (with a de-correlation length of $3d$ = 900 km), so they are spatially smooth fields with short-wave noise largely filtered. Add each of these perturbation fields to the same initial condition that produces the 4D background field, and then integrate the model to produce *M* perturbed 4D fields over the analysis time window. Obtain the perturbation ensemble matrix **A** by subtracting the 4D background field from each of the 4D perturbed field, and normalize each variable field in **A** (as described at the beginning as this subsection). In addition, a proper diagnostic should be done to detect and correct significant biases (see section 4.4).

II. Perform the SVD to extract the singular vectors from the normalized **A** and select the first *p*

singular vectors (with $M > p \geq M/2$). Use the selected first p basis vectors for the truncated expressions of the analysis increment field in (2). Use the limited memory quasi-Newton method (Liu and Nocedal, 1989) to find the optimal **ß** that minimizes the costfunction in (4). Because the truncated system dimension is relatively small ($p \leq 100$) in all the experiments, the descending algorithm converges rapidly (in less than 10 iterations).

III. Substitute the obtained **ß** into (2) to obtain the analysis increment field and then add it with the background field to obtain the analysis field.

In the original En4DVar, the analysis time window is centered at the ending time of the assimilation cycle, and the algorithm flow chart is plotted in Fig. 1a. For the j-th cycle ended at $t = jT$, the analysis time window is set over the time interval of $jT - \tau/2 \leq t \leq jT + \tau/2$, where τ is the time length of the analysis time window, and T is the time length of each assimilation cycle. With this setting, the analysis has to wait for observations collected from $t = jT$ to $jT + \tau/2$ beyond the ending time of the assimilation cycle, so the forecast is delayed by $\tau/2$ from the initial time of the forecast integration (that is, the ending time of the assimilation cycle). For $\tau = T = 12$ hours used in Qiu et al. (2007), the forecast delay is $\tau/2 = 6$ hours. This is clearly undesirable and disadvantageous for operational applications. To avoid such a delay, we need to shorten shift backward the analysis time window in the next subsection.

2.2. New scheme

As explained in the introduction, the original En4DVar can be improved if the ensemble perturbation matrix **A** is constructed by ensemble perturbations in a hybrid 4D space that

contains only a gridded 3D subspace in addition to the 4D observation space. In this case, the m-th column vector of matrix $\mathbf{A}$ is given by

$$\mathbf{a}_m = (\Delta\mathbf{u}_m{}^{\mathrm{T}}, \Delta\mathbf{d}_m{}^{\mathrm{T}})^{\mathrm{T}}. \tag{5}$$

Here, $\Delta\mathbf{u}_m = \mathbf{u}_m - \mathbf{u}_{\mathrm{b}}$ is the m-th perturbation field in the gridded 3D subspace at the ending time of the current assimilation cycle and the starting time of the next assimilation cycle. $\Delta\mathbf{d}_m{}^{\mathrm{T}} = (\Delta\mathbf{d}_{m,1}{}^{\mathrm{T}}, \Delta\mathbf{d}_{m,2}{}^{\mathrm{T}}, \ldots \Delta\mathbf{d}_{m,N_{\mathrm{t}}}{}^{\mathrm{T}})$ is the m-th perturbation field in the 4D observation space. $\Delta\mathbf{d}_{m,\mathrm{n}} = \boldsymbol{H}\mathbf{u}_{\mathrm{m}}(t_{\mathrm{n}}) - \boldsymbol{H}\mathbf{u}_{\mathrm{b}}(t_{\mathrm{n}})$ is the m-th perturbation field in the 3D observation subspace at the n-th observation time level ($t = t_{\mathrm{n}}$), while $\mathbf{u}_{\mathrm{m}}(t_{\mathrm{n}})$ and $\mathbf{u}_{\mathrm{b}}(t_{\mathrm{n}})$ denote the m-th perturbed field and the ensemble mean field, respectively, on the model grid at the n-th observation time level. The dimension of each column vector is $N_{\mathrm{v}}\times N_{\mathrm{x}} + N_{\mathrm{o}}\times N_{\mathrm{t}}$, where N_{o} is the number of observations (assumed to be the same) at each observation time level. Since $N_{\mathrm{o}} << N_{\mathrm{v}}\times N_{\mathrm{x}}$ and $N_{\mathrm{t}} > 1$, the column dimension ($= N_{\mathrm{v}}\times N_{\mathrm{x}} + N_{\mathrm{o}}\times N_{\mathrm{t}}$) of the ensemble perturbation matrix $\mathbf{A}$ is now much smaller than that ($= N_{\mathrm{v}}\times N_{\mathrm{x}}\times N_{\mathrm{t}}$) in the original En4DVar.

The SVD of matrix $\mathbf{A}$ is performed in the same way as in the original En4DVar but the space dimension is reduced and so is the computational cost due to the use of the hybrid space. From (1), it is easy to see that $\mathbf{B} = \mathbf{A}\mathbf{V}\boldsymbol{\Lambda}^{-1}$ or, equivalently, $\mathbf{b}_{\mathrm{k}} = \mathbf{A}\mathbf{v}_{\mathrm{k}}/\lambda_{\mathrm{k}}$ for the k-th column vector of $\mathbf{B}$, so each column vector of $\mathbf{B}$ is a linear combination of ($\mathbf{a}_1$, $\mathbf{a}_2$, … $\mathbf{a}_M$). Thus, similar to the m-th column vector of $\mathbf{A}$ in (5), the k-th column vector of $\mathbf{B}$ can be written into the following partitioned form:

$$\mathbf{b}_k = (\mathbf{b}^u{}_k{}^T, \mathbf{b}^d{}_k{}^T)^T, \quad (6)$$

where $\mathbf{b}^u{}_k$ and $\mathbf{b}^d{}_k$ correspond to $\Delta\mathbf{u}_m$ and $\Delta\mathbf{d}_m$ in (5), respectively.

With the partition in (6), the expression of the analysis increment vector in (2) is now reduced to the following form in the gridded 3D subspace:

$$\delta\mathbf{u}_a = \sum_{k=1}^{p} \mathbf{b}^u{}_k \beta_k = \mathbf{B}_u \boldsymbol{\beta}, \quad (7)$$

where $\mathbf{B}_u = (\mathbf{b}^u{}_1, \mathbf{b}^u{}_2, \ldots \mathbf{b}^u{}_M)$. Based on the partition introduced in (5), the background error covariance matrix is constructed directly in this paper by the partitioned ensemble perturbation sub-matrix $\mathbf{A}_u = (\Delta\mathbf{u}_1, \Delta\mathbf{u}_2, \ldots \Delta\mathbf{u}_M)$ essentially in the same way as in the conventional EnKF, which gives $\mathbf{P} = \mathbf{B}_u\boldsymbol{\Lambda}_p{}^2\mathbf{B}_u{}^T/(M - 1) \approx \mathbf{B}_u\boldsymbol{\Lambda}^2\mathbf{B}_u{}^T/(M - 1) = \mathbf{A}_u\mathbf{A}_u{}^T/(M - 1)$, where $\boldsymbol{\Lambda}_p = \mathrm{diag}(\lambda_1, \lambda_2, \ldots \lambda_p)$ denotes the truncated $\boldsymbol{\Lambda}$. Substituting this result with (7) and into (3) gives

$$J = J_b + J_o = (M - 1)\boldsymbol{\beta}^T\boldsymbol{\Lambda}_p{}^{-2}\boldsymbol{\beta} + (\mathbf{B}_d\boldsymbol{\beta} - \delta\mathbf{y})^T\mathbf{O}^{-1}(\mathbf{B}_d\boldsymbol{\beta} - \delta\mathbf{y}), \quad (8)$$

where $\mathbf{B}_d = (\mathbf{b}^d{}_1, \mathbf{b}^d{}_2, \ldots \mathbf{b}^d{}_M)$ and thus $\mathbf{B}_d\boldsymbol{\beta} = \sum_{k=1}^{p} \mathbf{b}^d{}_k \beta_k$. The vector coefficient $\boldsymbol{\beta}$ is the control variable for the minimization of J. The similarities and differences between the costfunction formulation in (8) and the conventional perfect-model 4DVar costfunction formulation [see (18) of Courtier 1997] are examined below.

When (8) is derived from (3), the sub-matrix $\mathbf{A}_u = \mathbf{A}_u(t_{N_t})$ is constructed by the gridded 3D ensemble perturbations at the ending time $t = t_{N_t}$ of each assimilation cycle and the background error covariance is modeled by $\mathbf{P} = \mathbf{P}(t_{N_t}) \approx \mathbf{A}_u(t_{N_t})\mathbf{A}_u(t_{N_t})^T/(M - 1)$, so the background term J_b in

(8) corresponds to $\delta\mathbf{u}_a(t_{N_t})^T\mathbf{P}(t_{N_t})\delta\mathbf{u}_a(t_{N_t})$. If $\mathbf{A}_u$ is constructed at the beginning time $t = t_1$ of each assimilation cycle, then the resulting costfunction will have the same form as in (8) but the background term will correspond to $\delta\mathbf{u}_a(t_1)^T\mathbf{P}(t_1)\delta\mathbf{u}_a(t_1)$, as the background term in the conventional perfect-model 4DVar. This implies

$$J_b = \delta\mathbf{u}_a(t_{N_t})^T\mathbf{P}(t_{N_t})\delta\mathbf{u}_a(t_{N_t}) \approx \delta\mathbf{u}_a(t_1)^T\mathbf{P}(t_1)\delta\mathbf{u}_a(t_1) \tag{9}$$

for the background term in (8). The approximate equivalence in (9) can be verified by substituting $\mathbf{P} = \boldsymbol{M}(t_{N_t}, t_1)\mathbf{P}_1\boldsymbol{M}(t_{N_t}, t_1)^T$ and $\delta\mathbf{u}_a(t_{N_t}) \approx \boldsymbol{M}(t_{N_t}, t_1)\delta\mathbf{u}_a(t_1)$ into $\delta\mathbf{u}_a(t_{N_t})^T\mathbf{P}(t_{N_t})\delta\mathbf{u}_a(t_{N_t})$. Here, $\boldsymbol{M}(t_{N_t}, t_1)$ is the matrix operator that maps $\mathbf{A}_u(t_1)$ to $\mathbf{A}_u(t_{N_t})$ and its inverse exists in the ensemble space to map $\mathbf{A}_u(t_{N_t})$ back to $\mathbf{A}_u(t_1)$, while $\delta\mathbf{u}_a(t_{N_t}) \approx \boldsymbol{M}(t_{N_t}, t_1)\delta\mathbf{u}_a(t_1)$ involves the following approximation. From (1) and (7), it is easy to see that $\delta\mathbf{u}_a(t_n) = \mathbf{A}_u(t_n)\boldsymbol{\alpha}(t_n)$, where $\boldsymbol{\alpha}(t_n) = \mathbf{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{ß}$ depends on the time level $t = t_n$ at which the sub-matrix $\mathbf{A}_u(t_n)$ is constructed for the SVD of $\mathbf{A} = (\mathbf{A}_u, \mathbf{A}_d)$ in (1) and $\mathbf{A}_d = (\Delta\mathbf{d}_1, \Delta\mathbf{d}_2, \ldots \Delta\mathbf{d}_M)$ according to (5). Note that $\mathbf{A}_d$ is invariant to t_n and is dissimilar to $\mathbf{A}_u(t_n)$ for any t_n, but $\mathbf{A}_u(t_{N_t})$ is similar to $\mathbf{A}_u(t_1)$ since $\mathbf{A}_u(t_{N_t}) = \boldsymbol{M}(t_{N_t}, t_1)\mathbf{A}_u(t_1)$. This implies that $\boldsymbol{\alpha}(t_{N_t}) \approx \boldsymbol{\alpha}(t_1)$, which leads to

$$\delta\mathbf{u}_a(t_{N_t}) = \boldsymbol{M}(t_{N_t}, t_1)\mathbf{A}_u(t_1)\boldsymbol{\alpha}(t_{N_t}) \approx \boldsymbol{M}(t_{N_t}, t_1)\mathbf{A}_u(t_1)\boldsymbol{\alpha}(t_1) = \boldsymbol{M}(t_{N_t}, t_1)\delta\mathbf{u}_a(t_1). \tag{10}$$

This approximation is used in the derivation of (9). The approximate equivalence in (9) provides a formal link between the background term in (8) and the background term in the conventional perfect-model 4DVar.

If the sub-matrix $\mathbf{A}_u$ is constructed by the gridded 3D ensemble perturbations at both t_1 and t_{N_t},

then $\boldsymbol{\alpha} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{ß}$ will be invariant to t_{N_t} and t_1, so $\delta\mathbf{u}_a(t_{N_t}) = \mathbf{A}_u(t_{N_t})\boldsymbol{\alpha} = \boldsymbol{M}(t_{N_t}, t_1)\delta\mathbf{u}_a(t_1)$ can be derived precisely. This implies that constructing $\mathbf{A}_u$ at both t_1 and t_{N_t} can be redundant for the background term (unless the model error is considered explicitly in the ensemble integration), so it will not necessarily improve the analysis accuracy (for the current perfect-model En4DVar) but will increase the computational cost. By formulating the background term and associated analysis increment at the ending time t_{N_t} only, the costfunction formulated in (8) will not only facilitate the model integration into the next cycle but also improve the computational efficiency. This is a major difference between the background term in (8) and the background term formulated at t_1 in the conventional perfect-model 4DVar. If the background term is formulated at t_1 and the minimization is achieved by adjusting the initial state at t_1, then the adjusted state $\delta\mathbf{u}_a(t_1)$ must be integrated to t_{N_t} to provide an initial background state for the assimilation in the next cycle. According to (10), this integrated state, that is, $\boldsymbol{M}(t_{N_t}, t_1)\delta\mathbf{u}_a(t_1)$ is close to $\delta\mathbf{u}_a(t_{N_t})$ but subject to additional errors in the perfect-model integration [represented by $\boldsymbol{M}(t_{N_t}, t_1)$] after $\delta\mathbf{u}_a(t_1)$ is optimally adjusted at t_1. By formulating the background term at t_{N_t}, $\delta\mathbf{u}_a(t_{N_t})$ can be optimally adjusted at t_{N_t} and then used directly to initialize the model integration into the next cycle, so the analysis is not subject to the aforementioned additional errors and hence can have an improved accuracy.

The observation term J_o in (8) is derived from the observation term $J_o = (\boldsymbol{H}\delta\mathbf{u}_a - \delta\mathbf{y})^T\mathbf{O}^{-1}(\boldsymbol{H}\delta\mathbf{u}_a - \delta\mathbf{y})$ in (3), while the latter with $\boldsymbol{H}$ linearized to $\mathbf{H}$ has the same form as the observation term in the conventional perfect-model 4DVar. Noted again from (1) that $\delta\mathbf{u}_a(t_{N_t}) = \mathbf{A}_u(t_{N_t})\boldsymbol{\alpha}$ and $\mathbf{B}_d\boldsymbol{ß} = \mathbf{A}_d\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{ß}$ is obtained with $\mathbf{A}_u$ constructed at $t = t_{N_t}$ for the SVD

of $\mathbf{A} = (\mathbf{A}_u, \mathbf{A}_d)$ in (1). Substituting $\mathbf{A}_d = (\Delta\mathbf{d}_1, \Delta\mathbf{d}_2, \ldots \Delta\mathbf{d}_M)$ with $\mathbf{d}_m = (\Delta\mathbf{d}_{m,1}{}^T, \Delta\mathbf{d}_{m,2}{}^T, \ldots \Delta\mathbf{d}_{m,N_t}{}^T)^T$ and $\Delta\mathbf{d}_{m,n} = \boldsymbol{H}_n\Delta\mathbf{u}_m(t_n) = \boldsymbol{H}_n\boldsymbol{M}(t_n, t_1)\Delta\mathbf{u}_m(t_1)$ into $\mathbf{B}_d\boldsymbol{ß} = \mathbf{A}_d\boldsymbol{\alpha}$ gives $\mathbf{B}_d\boldsymbol{ß} = \boldsymbol{HM}\delta\mathbf{u}_a(t_1)$, where $\boldsymbol{HM} = [\boldsymbol{H}_1, \boldsymbol{H}_1\boldsymbol{M}(t_1, t_1)^T, \boldsymbol{H}_2\boldsymbol{M}(t_2, t_1)^T, \ldots \boldsymbol{H}_{N_t}\boldsymbol{M}(t_{N_t}, t_1)^T]^T$. Substituting $\mathbf{B}_d\boldsymbol{ß} = \boldsymbol{HM}\delta\mathbf{u}_a(t_1)$ into the observation term J_o in (8) gives

$$J_o = (\mathbf{B}_d\boldsymbol{ß} - \delta\mathbf{y})^T\mathbf{O}^{-1}(\mathbf{B}_d\boldsymbol{ß} - \delta\mathbf{y}) = [\boldsymbol{HM}\delta\mathbf{u}_a(t_1) - \delta\mathbf{y}]^T\mathbf{O}^{-1}[\boldsymbol{HM}\delta\mathbf{u}_a(t_1) - \delta\mathbf{y}]. \quad (11)$$

When $\boldsymbol{H}$ is linear or tangent-linearized to $\mathbf{H}$, the righthand side of (11) recovers the observation term in (3). When $\boldsymbol{H}$ and $\boldsymbol{M}$ are both tangent-linearized, the righthand side of (11) becomes formally the same as the observation term in the conventional perfect-model 4DVar. In general, the observation increment $\boldsymbol{HM}\delta\mathbf{u}_a(t_1)$ is a nonlinear function of $\delta\mathbf{u}_a(t_1)$ but is linearized in the conventional 4DVar. As the entire tem $\boldsymbol{HM}\delta\mathbf{u}_a(t_1)$ is expressed in (8) by a truncated linear combination of the singular vectors, that is, $\mathbf{B}_d\boldsymbol{ß}$ in the space spanned by $\mathbf{B}_d = (\mathbf{b}^d{}_1, \mathbf{b}^d{}_2, \ldots \mathbf{b}^d{}_M)$, the nonlinearity is fully retained. Since $\boldsymbol{HM}$ is used and embedded in the computation of the ensemble perturbations in the observation space – a subspace of the hybrid-space, there is no need to linearize $\boldsymbol{HM}$ and code its adjoint. This simplifies the minimization algorithm. As in the original En4DVar, the vector coefficient $\boldsymbol{ß}$ is the only control variable for the minimization, but here it is given more explicitly in conjunction with the much reduced matrix $\mathbf{B}_d$ in the new costfunction in (8). Once $\boldsymbol{ß}$ is determined by minimizing the costfunction in (8) and substituted back into (7), the analysis increment becomes available on the model grid at the ending time level of the current assimilation cycle and thus can be used for the model integration into the next cycle.

By using the new formulations (7)-(8) in place of the original (2)-(3), each assimilation cycle can be performed with the new scheme through the same three steps as described in the previous subsection for the original En4DVar. The original analysis time window, however, needs to be shifted by $\tau/2 = 6$ hours to avoid its caused forecast delay (as explained at the end of the previous subsection). For the j-th assimilation cycle ended at $t = jT$, the shifted time window is over $jT - \tau \le t \le jT$. For the new test experiments in this paper, the analysis time window will be shortened to ensure that $\tau = (N_t - 1)\Delta\tau < T = 12$ hours, and the algorithm flow chart is plotted in Fig. 1b. The shortening will avoid the observations at the ending time level of the current assimilation cycle being used again by the analysis in the next cycle. If $\tau = T$ (= 12 hours) is used as in Qiu et al. (2007), then the analysis time window of the current assimilation cycle will end at the same time level as the analysis time window starts in the next cycle, so the observations at this time level (covered jointly by both the current and next analysis time windows) will be used twice. Repeated uses of observations are undesirable and should be avoided, as they reduce not only the computational efficiency but also the optimality of the analysis. In a realistic system, observations are distributed almost continuously in time, so using the observations in the interval of $jT - \tau \le t \le jT$ does not cause the above problem.

3. Design of assimilation experiments

To examine the performance of the new scheme in comparison with the original En4DVar, it is convenient to design the assimilation experiments by using the same shallow-water equation model with the same parameter settings as in Qiu et al. (2007). In particular, the model domain is

a square with periodic boundary conditions at $x, y = 0$ and D, where $D = 44d$ is the length of one side of the model domain and $d = \Delta x = \Delta y = 300$ km is the grid spacing. The terrain height is given by $h_s = h_0 \sin(4\pi x/D)\sin(\pi y/D)$ with the maximum height set to $h_0 = 250$ m for the "true-state" run but to $h_0 = 0$ for the imperfect model runs. The initial condition and ensemble perturbations are generated in the same way as described in section 3 of Qiu et al. (2007).

As in Qiu et al. (2007), the observation errors are assumed to be uncorrelated between different variables and different points in space and time. Simulated observations are generated by adding random errors to the model-produced "true" fields with the observation error standard deviations set to 12 m for h and 1.2 m s^{-1} for u and v. These simulated observations are available every $\Delta\tau = 3$ hours on coarse grid points spaced every $3d = 900$ km in the x- and y-direction. In Qiu et al. (2007), simulated observations were generated in three types: (i) both height and velocity observations (type-1), (ii) height h only observations (type-2), and (iii) velocity (u, v) only observations (type-3). Since the observation points were fixed, the number of observations was the smallest in type-2, which was only 1/3 of that in type-1 and 1/2 of that in type-2. The original En4DVar was previously tested with all three types of observations for both the perfect-model and imperfect-model cases. The results showed that the analysis RMS errors increased by nearly 70% from the perfect-model case to the imperfect-model case. With each type of observations, the original En4DVar performed significantly better than the conventional EnKF and EnSRF for the imperfect-model case. In view of the previous results, we only need to test and compare the new scheme with the original En4DVar for the imperfect-model case with type-1 observations in this paper.

As explained at the ends of sections 2.1 and 2.2, the analysis time window is centered at the ending time of the assimilation cycle in the original En4DVar and this time window needs to be shifted backward to avoid its caused forecast delay and shortened to avoid repeated uses of observations. The modified time window is over the range of $jT - \tau \leq t \leq jT$ for the j-th assimilation cycle with τ reduced to 6 hours < T (and thus $N_t = 3$). Note that this time window excludes the first observation time level (3 hours after the model initialization in each cycle). Expanding the time window to include the first observation time level should improve the analysis. This, however, is not always the case with the current En4DVar, because the ensemble integration is too short to generate significant flow-dependent covariance structures at the first observation time level (3 hours after the initialization). The situation can be improved (for example, the analysis RMS errors obtained by including the first observation time level can be reduced by 0.4% in height and by 10% in velocity for the control experiment TEN in Tables 1 and 2) if the ensemble integration is initialized earlier by 3 hours. This, however, will require the analysis time level (for the gridded 3D subspace in the hybrid 4D space) shift backward by 3 hours and hence will delay the forecast by 3 hours. To solve the problem effectively, we need to update not only the ensemble mean but also the ensemble perturbations in the En4DVar. This requires continued research beyond this study.

Based on the above considerations, two groups of experiments are designed to compare the new scheme with the original En4DVar for the imperfect-model case with type-1 observations. The first group, called group 1, contains one pair of reference experiments and one pair of test experiments plus an additional test experiment. The paired reference experiments, called REO

and REN, are designed for the original and new schemes, respectively, in the same way as the control experiment 3.1 with type-1 observations in Table 3 of Qiu et al. (2007) except that the assimilation is cycled for 25 days (through 50 cycles) instead of the original 5 days (through only 10 cycles). Note that REN is performed with the new scheme without modifying the original analysis time window. That is, the analysis time window is still over $jT - \tau/2 \le t \le jT + \tau/2$ for the j-th assimilation cycle with $\tau = (N_t - 1)\Delta\tau = T = 12$ hours (and thus $N_t = 5$) in REO and REN. These paired reference experiments are used to test and compare the original and new schemes over a sufficiently long period of assimilation without modifying the original analysis time window.

The paired test experiments in group 1, called TEO and TEN, are designed to examine the impact of the modified analysis time window. This pair is the same as the above pair of REO and REN but the analysis time window is modified as described earlier and the truncation number is increased from $p = 75$ to 100 (for $M = 150$). The additional (third) test experiment in group 1, called TEOa, is the same as TEO except that the normalized background error covariance matrix is modeled by $\mathbf{P} = \mathbf{A}\mathbf{A}^T/(M-1) = \mathbf{B}\mathbf{\Lambda}^2\mathbf{B}^T/(M-1)$ instead of $\mathbf{P} = \mathbf{B}\mathbf{B}^T$ [used in (4) for the original En4DVar] without inflating the estimated background error variances (by 2.5 times for the imperfect-model case, as mentioned earlier for the original En4DVar). This covariance model is similar to that used in (8) for the new scheme in consistency with the covariance model in the conventional EnKF. This additional experiment TEOa is designed to examine the impact of the refined background error covariance model (without using the hybrid space) on the performance of the En4DVar.

The second group, called group 2, contains one pair of control experiments, called the control pair, and seven pairs of test experiments, called the test pairs for short. The control pair duplicates the above TEOa and TEN. The first five test pairs are designed, similarly to the test experiments 3.1-3.4 in Table 3 of Qiu et al. (2007), to examine the sensitivities of the new scheme to the ensemble size, observation errors and observation density in comparison with the original scheme (but the background error covariance model is refined as explained above). These five test pair are the same as the control pair (TEOa and TEN) except for the following changes: The ensemble size and truncation number are reduced from $M = 150$ and $p = 100$ (in the control pair) to $M = 100$ and $p = 80$, respectively, in test pair 1 and further reduced to $M = 50$ and $p = 40$, respectively, in test pair 2. The observation error standard deviations are increased by 50% (from 12 m to 18 m for h and from 1.2 m s^{-1} to 1.8 m s^{-1} for u and v) in test pair 3. The observation spacing is increased from $3d$ (= 900 km) to $5d$ = (1500 km) in test pair 4, and further increased to $7d$ = (2100 km) in the x- and y-directions in test pair 5.

The last two test pairs (6 and 7) in group 2 are newly designed in this paper to examine the sensitivity of the new scheme to non-uniformly distributed observations in comparison with the original scheme. In these two test pairs, the observation points are non-uniformly redistributed (from those in the control pair) with a 50% of the points randomly selected from the grid points in the southwest quadrant of the model domain while the other 50% of the observation points are randomly selected from the remaining three quadrants. The number of observation points is 225 in test pair 6 (which is the same as that in the control pair) but is reduced to 112 in test pair 7. These two test pairs are designed to mimic the non-uniform distribution of operational

observations over the global continents and oceans.

4. Results and discussions

4.1. Analysis accuracy

To evaluate the analysis accuracy of the new scheme versus the original En4DVar, spatially averaged RMS errors are computed from the height and velocity difference (analysis minus "true") fields as functions of assimilation time. The two RMS errors are defined by

$$\sigma_h \equiv \langle (h - h^t)^2 \rangle^{1/2} \text{ and } \sigma_{\mathbf{v}} \equiv \langle |\mathbf{v} - \mathbf{v}^t|^2 \rangle^{1/2} \quad (12)$$

for the height h and velocity $\mathbf{v} = (u, v)$ fields, respectively, where $\langle\ \rangle$ denotes the spatial average, and $(\ )^t$ denotes the "true" value of $(\ )$.

The RMS errors from the reference pair (REO and REN) and test pair (TEO and TEN) are plotted as functions of assimilation time (for the 50 cycles over the entire 25-day assimilation period) in Fig. 2. As shown, the analysis errors produced by the new scheme are notably smaller than that produced by the original En4DVar especially for the test pair (TEO and TEN). When the original analysis time window is modified (as explained at the end of section 2.2 and described in section 3) and the truncation number is increased from p = 75 to 100, the new scheme and original En4DVar both produce improved analyses, and the new scheme outperforms the original En4DVar significantly. Also, as we can see from all the four experiments in Fig. 2, the RMS errors decrease rapidly during the first six assimilation cycles (up to day 3) and then become stabilized in the subsequent cycles and keep nearly stationary during

the last 20 cycles (from day 15 to 25). This suggests that the analysis errors averaged over the last 20 cycles can provide the required statistics for evaluating the analysis accuracy produced by the new scheme versus the original En4DVar.

The analysis RMS errors are averaged over the last 20 cycles (from day 15 to 25) of the total 50 cycles and listed in Table 1 for all the five experiments in group 1. As shown by the results listed for the reference pair, the height RMS error produced by the new scheme in REN is slightly larger (by 8.90/8.71 - 1 = 2.2%) than that produced by the original En4DVar in REO, but the velocity RMS error is reduced significantly (by 1 - 0.55/0.69 = 20.3%). For the test pair, the height and velocity RMS errors are both reduced significantly (1 - 6.75/7.61 = 11.3% in height and 1 - 0.54/0.64 = 15.6% in velocity) by the new scheme in TEN compared with the original scheme in TEO.

In the test pair, the analysis time window is modified with N_t reduced from 5 to 3, so the amount of observations used by the analyses is reduced (by 40%). Regardless of this reduced use of observations, the two schemes both produce more accurate analyses in the test pair than in the reference pair, as mentioned above and shown by the first two rows in Table 1. This improvement appears to be contradictory to the reduced use of observations but it can be explained by considering the following two factors: (i) The truncation number is increased from p = 75 to 100 in the test pair. (ii) The last two time levels of observations (that are used in the reference pair but not in the test pair) are beyond the ending time in each assimilation cycle, so their associated background states and ensemble perturbations have relatively large bias errors with significant deviations away from the true state as they are produced by integrating the

imperfect model (in which the maximum terrain height is reduced from the true value of $h_0 = 250$ m to $h_0 = 0$) to the last two time levels (3 and 6 hours after the ending time of each assimilation cycle). Thus, shortening the analysis time window with the last two time levels removed can not only avoid their caused forecast delay (as explained in section 2) but also avoid to use seriously biased background states and ensemble perturbations on these last two time levels.

As shown by the difference between TEO and TEOa in Table 1, the refined background error covariance model alone (without using the hybrid space) can improve the original En4DVar and the improvement is more significant than that produced by the use of the hybrid space (in TEN in comparison with TEOa). In particular, the RMS errors are reduced by 1 - 6.94/7.61 = 8.8% in height and 1 - 0.59/0.64 = 7.8% in velocity as the original background error covariance model used in TEO is refined in TEOa, and these error reductions are more significant than the reductions (1 - 6.75/6.94 = 2.7% in height and 1 - 0.54/0.59 = 8.5% in velocity) produced by use of the hybrid space in TEN in place of the original state space in TEOa.

Note that the truncation number was set to $p = 75$ in the control experiment with the original En4DVar in Table 3 of Qiu et al. (2007) and this truncation retained 95.6% (or more) of the total variance of the ensemble perturbations obtained in the first assimilation cycle (or subsequent cycles). With the modified analysis time window and other changes made in the above test pair, the original truncation can no longer retain the desired percentage (≈ 95%) of the total ensemble variance and thus needs to be adjusted accordingly for the new scheme. To see this, the retained variance percentages, measured by $(\sum_{k=1}^{p} \lambda_k^2)/(\sum_{k=1}^{r} \lambda_k^2)$ as in Qiu et al. (2007), are plotted as functions of the truncation number p in Fig. 3 for the ensemble perturbations obtained in the first

cycle by the two schemes in the test pair. Here, λ_k is the k-the singular value of **A** computed by the SVD in (1), r is the rank of **A** and $\sum_{k=1}^{r}\lambda_k^2$ is the total ensemble variance. As shown in Fig. 3, with the original truncation number (p = 75), the retained variance percentages are only 87.5% and 88.0% in the first cycle of TEOa and TEN, respectively. When p is increased to 100, the percentages increase to 93.3% and 94.1%, respectively, in the first cycle. In the subsequent cycles, the retained variance percentages become even higher and reach 96.4% and 97.5%, respectively, in the 30th cycle (not shown). To retain the desired percentage (≈ 95%) of the total ensemble variance in TEOa and TEN, p = 100 is used (in place of p = 75) in the test pair.

4.2. Sensitivity

The averaged analysis RMS errors from the seven pairs of experiments in group 2 are listed in Table 2. These experiments are designed to examine the sensitivities of the new scheme to the ensemble size, observation errors and observation distributions in comparison with the original En4DVar (with the refined background error covariance model). As shown by the results of test pair 1, in response to the reductions of the ensemble size (from M = 150 to 100) and truncation number (from p = 100 to 80), the averaged RMS errors are increased by 22.2% (= 8.48/6.94 - 1) in height and 23.7% (= 0.73/0.59 - 1) in velocity for the original En4DVar, and increased by 23.7% (= 8.35/6.75 - 1) in height and 27.8% (= 0.69/0.54 - 1) in velocity for the new scheme. Thus, compared with the original scheme, the new scheme is slightly more sensitive to the ensemble size reduction (to M =100) and truncation number reduction (to p = 80). This also implies that the error reductions made by the new scheme (with respect to the original scheme)

become slightly less significant in test pair 1 than in the control pair. In particular, the error reductions made by the new scheme in test pair 1 are 1.5% (= 1 - 8.35/8.48) in height and 5.5% (= 1 - 0.69/0.73) in velocity, while the error reductions made by the new scheme in the control pair are 2.7% (= 1 - 6.75/6.94) in height and 8.5% (= 1 - 0.54/0.59) in velocity. If the ensemble size is reduced more dramatically to $M = 50$ with the truncation number reduced to $p = 40$, then the averaged RMS errors will increase by 103% (= 14.10/6.94 - 1) in height and 108% (= 1.23/0.59 - 1) in velocity for the original En4DVar, and increase by 90% (= 12.83/6.75 - 1) in height and 113% (= 1.15/0.54 - 1) in velocity for the new scheme, as shown by the results of test pair 2 in Table 2. In this case, both the new and original schemes perform poorly, but the new scheme still outperforms the original En4DVar.

As shown by the results of test pair 3 in Table 2, when the observation errors are increased by 50% from the values used in the control pair, the analysis RMS errors are increased by 7.8% (= 7.48/6.94 - 1) in height and 10.2% (= 0.65/0.59 - 1) in velocity for the original En4DVar, and increased by 6.7% (= 7.20/6.75 - 1) in height and 11.1% (= 0.60/0.54 - 1) in velocity for the new scheme. Thus, with respect to such a degradation of observation accuracy, the new scheme is slightly less sensitive in the height analysis but slightly more sensitive in the velocity analysis than the original scheme. In this case, the new scheme still outperforms the original En4DVar, and the error reductions made by the new scheme are 3.7% (= 1 - 7.20/7.48) in height and 7.7% (= 1 - 0.60/0.65) in velocity.

In test pair 4, the observation spacing is increased to $5d$ (= 1500 km) from $3d$ in the control pair. In response to this observation density reduction, the analysis RMS errors are increased by

22.2% (= 8.48/6.94 - 1) in height and 23.7% (= 0.73/0.59 - 1) in velocity for the original En4DVar, and increased by 22.8% (= 8.29/6.75 - 1) in height and 33.3% (= 0.72/0.54 - 1) in velocity for the new scheme. The new scheme is more sensitive than the original En4DVar to such a moderate reduction of observation density. In this case, the new scheme slightly outperforms the original En4DVar, and the error reductions made by the new scheme are 2.2% (= 1 - 8.29/8.48) in the height analysis and 1.4% (= 1 - 0.72/0.73) in the velocity analysis, as shown by the results of test pair 4 in Table 2. When the observation spacing is increased more dramatically to 7*d* (= 2100 km) in test pair 5, the RMS errors are increased by 68% (= 11.68/6.94 - 1) in height and 83% (= 1.08/0.59 - 1) in velocity for the original En4DVar, and increased by 73% (= 11.69/6.75 - 1) in height and 113% (= 1.15/0.54 - 1) in velocity for the new scheme. In this case, the new scheme slightly underperforms the original En4DVar.

When the observation points (uniformly distributed in the control pair) are redistributed non-uniformly in test pair 6, the RMS errors are increased by 20% (= 8.35/6.94 - 1) in height and 22% (= 0.72/0.59 - 1) in velocity for the original En4DVar, and increased by 25% (= 8.44/6.75 - 1) in height and 31.5% (= 0.71/0.54 - 1) in velocity for the new scheme. The new scheme is thus more sensitive than the original En4DVar to the observation non-uniformity. In this case, the new scheme slightly underperforms the original scheme (by 1.1% = 1 - 8.44/8.35) in the height analysis but outperforms the original scheme (by 1.4% = 1 - 0.71/0.72) in the velocity analysis, as shown by the results of test pair 6 in Table 2. When the non-uniform observation points are thinned from 225 (in test pair 6) to 112 in test pair 7, the RMS errors are increased by 84% (= 12.77/6.94 - 1) in height and 95% (= 1.15/0.59 - 1) in velocity for the original En4DVar, and

increased by 99% (= 13.45/6.75 - 1) in height and 126% (= 1.22/0.54 - 1) in velocity for the new scheme. In this case, the new scheme clearly underperforms the original En4DVar (with the refined background error covariance)

4.3. Computational efficiency

As explained in the introduction and shown in section 2.2, the column dimension of the ensemble perturbation matrix **A** in the new scheme is substantially smaller than that in the original En4DVar, so the new scheme is computational more efficient than the original En4DVar. This is indeed seen from all the pairs of experiments in groups 1 and 2. In particular, as shown by the results from the reference pair (in group 1) listed in the first row of Table 3, the CPU time required by the new scheme in REN is only 26% of that required by the original En4DVar in REO. As shown by the second row in comparison with the first row of Table 3, when the analysis time window is modified (with N_t reduced from 5 to 3) and p is increased from 75 to 100 in the control pair, the CPU time percentage is reduced from 100% to 59% for the original En4DVar in TEOa but is increased slightly from 26% to 28% for the new scheme in TEN. Note that the column dimension of the ensemble perturbation matrix **A** is $N_v \times N_x \times N_t$ in the original En4DVar (see section 2.1). When N_t is reduced from 5 to 3 with the modified analysis time window, the column dimension is reduced by 40%, so the CPU time is reduced substantially although the row dimension of **A** is increased by 25% as p is increased from 75 to 100. With the new scheme, the column dimension of **A** is reduced to $N_v \times N_x + N_o \times N_t$. Since $N_o \approx N_v \times N_x/9 << N_v \times N_x$ in TEN, the column dimension is reduced merely by about 14% (= 2/14) when N_t is

reduced from 5 to 3 with the modified analysis time window. The row dimension of **A**, however, is increased by 25% as p is increased from 75 to 100, and this explains the slight increase of the CPU time percentage from 26% in REN to 28% in TEN, as shown in Table 3.

The CPU time percentage for the original En4DVar in TEO is not listed in Table 3 because it is the same as that in TEOa. As shown by the third row in comparison with the first row of Table 3, when the ensemble size and associated truncation number are reduced (to $M = 100$ and from $p = 80$) in test pair 1 of group 2, the CPU times required by the two schemes are both reduced, and the CPU time required by the new scheme is 43% of that required by the original En4DVar. When the ensemble size and associated truncation number are further reduced (to $M = 50$ and $p = 40$) in test pair 2 of group 2 (see row 4 of Table 3), the CPU times required by the two schemes are both further reduced, and the CPU time required by the new scheme is still 43% of that required by the original En4DVar. The CPU times required by the two schemes in test pair 3 of group 2 are not listed because they are the same as in the control pair.

When the observation spacing is increased (from 3d to 5d) in test pair 4 of group 2, the CPU times required by the two schemes are both reduced and, again, the CPU time required by the new scheme is still 43% of that required by the original En4DVar, as shown by row 5 in comparison with row 1 in Table 3. When the observation spacing is further increased (to 7d) in test pair 5 of group 2 (see the last row in Table 3), the CPU times required by the two schemes are both further reduced, and the CPU time required by the new scheme is about 40% of that required by the original En4DVar. The CPU times required by the two schemes in test pair 6 of group 2 are not listed because they are the same as in the control pair. Note that the number of

observation points in test pair 7 is the same as that (= 112) in test pair 5, so the CPU times required by the two schemes in test pair 7 are the same as in test pair 5 and thus not listed.

4.4. Diagnoses of innovation statistics and optimality consistencies

In the conventional 4DVar formulation, the background forecasts and observations are assumed to be free of bias, and this assumption has been adopted in both the original and new En4DVar schemes. The forecast model used in this paper, however, is not only imperfect but also biased (since it neglects the "true" terrain) and this can cause the conventional EnKF fail to converge (as shown in Fig. 3 of Qiu et al. 2007). The original En4DVar and new scheme can converge, but the forecasts and analyses may still more or less biased due to the model bias ignored in the En4DVar. In real applications, the true state is unknown, so the biases are hard to estimate, but the difference between the observation and forecast biases can be diagnosed in the observation space according to

$$<\delta\mathbf{y}> = \mathbf{b}_o - \mathbf{b}_b, \tag{13}$$

where <( )> denotes the expectation of ( ), $\delta\mathbf{y} = \mathbf{y} - \boldsymbol{H}\mathbf{u}_b$ is the innovation, $\mathbf{b}_o$ and $\mathbf{b}_b$ denote the observation bias and forecast bias, respectively, in the observation space (that is, the biases of $\mathbf{y}$ and $\boldsymbol{H}\mathbf{u}_b$, respectively, with respect to the same true state in the observation space). Similarly, the expectation of $\delta\mathbf{y} - \boldsymbol{H}\delta\mathbf{u}_a$ is the difference between the observation bias and analysis bias in the observation space, that is,

$$<\delta\mathbf{y} - \boldsymbol{H}\delta\mathbf{u}_a> = \mathbf{b}_o - \mathbf{b}_a, \quad (14)$$

where $\boldsymbol{H}\delta\mathbf{u}_a = \boldsymbol{H}\mathbf{u}_a - \boldsymbol{H}\mathbf{u}_b$ is the analysis increment in the observation space and $\mathbf{b}_a$ denotes the bias of $\boldsymbol{H}\mathbf{u}_a$. Note that $\boldsymbol{H}\delta\mathbf{u}_a = \mathbf{B}_d\boldsymbol{ß}$ in the new En4DVar scheme. With the ergodicity assumption, the expectations in (13) and (14) can be computed by averaging over a large number of assimilation cycles. If the observations are bias-free or can be made bias-free through adequate quality controls (as assumed in this paper), then the forecast and analysis biases can be diagnosed from (13) and (14), respectively.

As shown in Fig. 2, the RMS errors decrease rapidly in the first 10 cycles (the first 5 days) and then fluctuate around a constant level. Similar features are also seen for the minimized costfunction values in Fig. 4, so the ergodicity assumption may be valid for the subsequent 40 cycles. By taking averages over the last 40 cycles, $\mathbf{b}_b$ and $\mathbf{b}_a$ are diagnosed from the left-hand sides (LHSs) of (13) and (14), respectively, for the two control experiments TEOa and TEN. The diagnosed forecast and analysis biases are found to be mostly within ±0.5 m for h and within ±0.05 m s^{-1} for u and v. These diagnosed biases are small but very noisy in the observation space as they are obtained by averaging over merely 40 cycles. Examples of the diagnosed height forecast biases in TEOa and TEN are plotted by thick dashed curves in Fig. 5a and 5b, respectively. Note that $N_o = 225\times3$ is the number of observations at each observation time level and $N_t = 3$ is the number of observation time levels used in TEOa and TEN, so the vector dimension of $\mathbf{b}_b$ and $\mathbf{b}_a$ is $N_o\times N_t = 225\times3\times3= 2025$. The dimension for each component vector is $225\times3 = 675$, so each curve in Fig. 5 plots only the last 1/3 of the h-component vector, that is, the

height bias field over the 225 observation points at the last (third) observation time level. Similar noisy features are seen in the (u, v)-component fields of the diagnosed forecast and analysis biases for all the three observation time levels. These noisy structures appear to represent the uncertainties caused by the insufficient sample size (only 40 cycles) rather than the true bias variations in the observation space.

Since the model-simulated "true" state is known in these experiments, the forecast and analysis biases can be also directly computed in the observation space by subtracting the "true" state from the forecast and analysis, respectively, and then averaging over the last 40 cycles. The computed height forecast biases are plotted by thin solid curves in Fig. 5 for TEOa and TEN. The RMS differences between the computed and diagnosed biases are around 0.3 m for h and 0.03 m s^{-1} for u and v in TEOa and TEN. These RMS differences are about 60% of the variability ranges (±0.5 m for h and ±0.05 m s^{-1} for u and v) of the diagnosed biases, so the diagnosed biases are significantly different from the directly computed biases. Note from (13) and (14) that the differences the above computed and diagnosed biases are simply the observation errors averaged over the last 40 cycles. As revealed by the differences between the thick dashed and thin solid curves in Fig. 5, the averaged observation errors are very noisy and not close to the true zero observation bias. This suggests again that the sample size (40 cycles) is insufficient for the bias estimation. Further averaging over all the three observation time levels (with the sample size increased from 40 to 120) can reduce the averaged observation errors (revealed by the RMS differences) by a factor of about $\sqrt{3}$, but their spatial fluctuations remain to be significant. To filter the spurious fluctuations, we need to assume that the errors are statistically homogeneous in

the 4D observation space in addition to the ergodicity. This assumption is accurate for the simulated observations but very crude for the forecast and analysis errors. Under this assumption, the diagnosed forecast and analysis bias fields can be further averaged in the observation space, although their averaged values turn to be extremely small and the observation-space averaging appears to overly suppress the biases. As listed in Table 4, the averaged forecast and analysis biases in TEN are mostly smaller than those in TEOa. The true forecast and analysis biases and their spatial variations are difficult to diagnose, but their magnitude should not exceed the above estimated ranges (±0.5 m for h and ±0.05 m s^{-1} for u and v). The biases are thus not very significant (relative to the forecast error standard deviations (listed in Table 5).

For an optimal analysis, the expectation of the value of the costfunction at the minimum is proportional to the number of observations used in the analysis. This well-known criterion and its extension for partitioned observations have been used effectively for diagnosis and adaptive tuning of error parameters in variational data assimilation (Desroziers et al. 2001). For the costfunction defined in (3) or (8) (without multiplying ½), this criterion requires that

$$<J_{\min}> = N_o \times N_t, \tag{15}$$

where $N_o \times N_t$ is the number of observations used by each En4DVar analysis. With the ergodicity assumption, the expectation of $J_{\min}$ can be estimated by averaging $J_{\min}$ over a large number of assimilation cycles, so the criterion in (15) can be conveniently used to check the averaged optimality of the analyses without additional computational cost. In Fig. 4, $J_{\min}$ is plotted as functions of the assimilation cycle number for TEOa and TEN. As explained earlier, since $J_{\min}$

decreases rapidly in the first 10 cycles and then fluctuates around a constant level in each experiment, the ergodicity assumption can be valid for the subsequent 40 cycles. The averaged J over the 40 cycles gives $<J_{min}>$ = 2004.2 for TEOa and $<J_{min}>$ = 2062.4 for TEN. These estimated values are quite close to the optimality value of $N_o \times N_t$ = 2025. Thus, the optimality criterion in (15) is largely satisfied, although the forecast and analysis are not truly unbiased as estimated above.

The criterion in (15) is only a necessary condition for the analysis optimality. In addition to this criterion, several additional consistency criteria have been proposed by Desroziers et al. (2005) for diagnosis of observation, background and analysis error statistics. In our notations, these consistency criteria can be expressed below:

$$<\delta\mathbf{y}'\delta\mathbf{y}'^{\mathrm{T}}> = \mathbf{O} + \mathbf{P}_o, \tag{16}$$

$$<\boldsymbol{H}\delta\mathbf{u}_a'\delta\mathbf{y}'^{\mathrm{T}}> = \mathbf{P}_h, \tag{17}$$

$$<(\delta\mathbf{y}' - \boldsymbol{H}\delta\mathbf{u}_a')\delta\mathbf{y}'^{\mathrm{T}}> = \mathbf{O}, \tag{18}$$

$$<\boldsymbol{H}\delta\mathbf{u}_a'(\delta\mathbf{y}' - \boldsymbol{H}\delta\mathbf{u}_a')^{\mathrm{T}}> = \mathbf{Q}_h, \tag{19}$$

where $\delta\mathbf{y}' = \delta\mathbf{y} + \mathbf{b}_b$, $\boldsymbol{H}\delta\mathbf{u}_a' = \boldsymbol{H}\delta\mathbf{u}_a - \mathbf{b}_a + \mathbf{b}_b$, (13) and (14) are used with $\mathbf{b}_o = 0$, $\mathbf{P}_h$ and $\mathbf{Q}_h$ denote the background and analysis error covariance matrices, respectively, in the observation space. As explained above, since $\mathbf{b}_a$ and $\mathbf{b}_b$ are small and hard to accurately estimate from (13) and (14), we will simply set $\mathbf{b}_a = \mathbf{b}_b = 0$ in (16)-(19) for the subsequent computations in this section. Note that $\mathbf{P}_h = \mathbf{B}_d\boldsymbol{\Lambda}_p^2\mathbf{B}_d^{\mathrm{T}}/(M - 1)$ in the new En4DVar scheme, which is consistent with the background error covariance model $\mathbf{P} = \mathbf{B}_u\boldsymbol{\Lambda}_p^2\mathbf{B}_u^{\mathrm{T}}/(M - 1)$ used in (8). The original

consistency criteria derived in (1)-(4) of Desroziers et al. (2005) require $\boldsymbol{H}$ to be linear or tangent-linearized to $\mathbf{H}$. This tangent-linearization is avoided when the analysis is performed in the hybrid space as explained in section 2.2.

The En4DVar-estimated $\mathbf{P}_h$ is flow-dependent and non-stationary, so the expectations on the LHSs of (16) and (17) should be also non-stationary. In order to use the ergodicity assumption, the variations of $\mathbf{P}_h$ must be treated as random fluctuations and the number of assimilation cycles sampled for the statistical averaging should be constrained by the time window during which the flow regime does not change substantially. When the expectation on the LHS of (16) or (17) is computed by averaging over the last 40 cycles, $\mathbf{P}_h$ should be also averaged over the last 40 cycles. In this case, the rank of the computed matrices on the LHS is limited by the number (= 40) of cycles used for the averaging and thus is much smaller than the full rank of $N_o \times N_t = 2025$. The rank of $\mathbf{P}_h = \mathbf{B}_d \boldsymbol{\Lambda}_p^2 \mathbf{B}_d^T/(M - 1)$ computed in each cycle is limited by the SVD truncation number ($p = 100$ for TEOa and TEN). As these matrices are highly rank-deficient, their off-diagonal terms contain more spurious correlations than useful information. Furthermore, the current En4DVar does not update the ensemble in the analysis step, so it provides no estimate of $\mathbf{Q}_h$ to facilitate the use of the consistency criterion in (19) in real data assimilation. Due to the above limitations, we will only evaluate the diagonal parts of $\mathbf{O}$ and $\mathbf{P}_h$ diagnosed from the LHSs of (16)-(18), and compare them with the prescribed $\mathbf{O}$ and En4DVar-estimated $\mathbf{P}_h$ on the right-hand sides (RHSs). Note also that (16) is the sum of (17) and (18), so the three criteria in (16)-(18) are not independent and we will use (17) and (18) only. Since the diagonal terms of $\mathbf{O}$ and $\mathbf{P}_h$ are the observation and background error variances, respectively, in the observation space, it is

convenient to denote their associated standard deviations by vectors $\mathbf{s}_o$ and $\mathbf{s}_b$, respectively.

From the LHS of (18), $\mathbf{s}_o$ can be diagnosed by averaging the diagonal part of the matrix term inside the expectation over the last 40 cycles and taking square root. The last 1/3 of the height component of the diagnosed $\mathbf{s}_o$ in TEOa is plotted in Fig. 6 to show how the diagnosed height observation error standard deviation varies as a function of the observation point at the last observation time level. As explained in section 3, the simulated observations are produced by adding random errors (generated independently in each experiment) to the simulated "true" fields with the observation error standard deviations set to s_{oh} = 12 m for h and s_{ov} = 1.2 m s^{-1} for u and v, so the "true" $\mathbf{O}$ is diagonal and the "true" $\mathbf{s}_o$ is known. As shown in Fig. 6, the diagnosed s_{oh} is too noisy to match the "true" constant value of s_{oh} = 12 m. Similar noisy structures are seen for the diagnosed s_{oh} at the other two observation time levels and for the diagnosed s_{ov} at all the three observation time levels in both TEOa and TEN. Thus, as in the above bias estimation, we need to further average the diagnosed $\mathbf{s}_o$ in the observation space. The averaged values are s_{oh} = 12.14 m for h and s_{ov} = 1.224 and 1.221 ms^{-1} for u and v, respectively, in TEOa. The averaged values are s_{oh} = 12.16 m for h and s_{ov} = 1.223 and 1.222 ms^{-1} for u and v, respectively, in TEN. Clearly, the averaged values are very close to the "true" observation standard deviations (s_{oh} = 12 m for h and s_{ov} = 1.2 m s^{-1} for u and v).

From the LHS of (17), $\mathbf{s}_b$ can be diagnosed by averaging the diagonal part of the matrix term inside the expectation over the last 40 cycles and then taking square root. From the RHS of (17), $\mathbf{s}_b$ can be estimated by averaging the diagonal part of $\mathbf{P}_h = \mathbf{B}_d\mathbf{\Lambda}_p^2\mathbf{B}_d^T/(M - 1)$ in TEN or its equivalent in TEOa over the last 40 cycles and taking square root. The consistency criterion in

(17) requires the RHS estimated $\mathbf{s}_b$ match the LHS diagnosed $\mathbf{s}_b$. The required match, however, cannot be verified locally due to the spurious noisy structures caused by under-sampling in both the LHS diagnosed and RHS estimated forecast error variances. Here, again, we have to further average the diagnosed and estimated forecast error variances in the observation space. The averaged values are listed in Table 5, where the consistency accuracy is computed in percentage for each case by the absolute value of the difference between the LHS diagnosed and RHS estimated values divided by the LHS diagnosed value. As shown in Table 5, the consistency criterion in (17) is closely satisfied by the averaged values and is more closely satisfied in TEN than in TEOa.

5. Conclusions

In this paper, the ensemble-based four-dimensional variational data assimilation (En4DVar) method of Qiu et al. (2007) is revisited, and a new scheme is developed to improve the computational efficiency and analysis accuracy of the En4DVar. In this new scheme, leading singular vectors are extracted from hybrid 4D ensemble perturbations to fit observation innovation data and construct the analysis increment simultaneously in a hybrid space. This hybrid space combines the 4D observation space with only a gridded 3D subspace at the ending time of the assimilation cycle, so its dimension can be much smaller than that of the gridded 4D space in the original En4DVar [see Eqs. (5)-(6)]. This improves the computational efficiency.

The use of hybrid space also has two apparent advantages: (i) The background covariance matrix can be and only needs to be constructed in the 3D subspace [see the first term in Eq. (8)].

(ii) The analysis increment can fit the 4D innovation data in the observation space directly [see the second term in Eq. (8)] and also provide the necessary initial condition in the gridded 3D subspace directly [see Eq. (7)] for the model integration into the next assimilation cycle. In addition to the use of hybrid-space, the background error covariance model used in the original En4DVar is refined in consistency with the covariance model in the conventional EnKF [see Eq. (8) in section 3] and the analysis time window is shifted backward to avoid its caused forecast delay and shortened to avoid repeated uses of observations (as explained at the end of section 2).

The new scheme is compared with the original En4DVar in ten pairs of assimilation experiments performed in two groups with simulated observations generated by the same shallow-water equation model for the imperfect-model case as in Qiu et al. (2007). The schemes' performances are evaluated by their analysis RMS errors averaged over the last 20 cycles (10 days). The results show that the new scheme can significantly (or only slightly) outperform the original En4DVar if the original analysis time window is (or is not) modified. The modified analysis time window can improve the analysis accuracy (see Table 1) and computational efficiency (see Table 3) for both schemes. The refined background error covariance model can further improve the analysis accuracy (as shown by TEOa versus TEO in Table 1). The use of hybrid space can improve not only the computational efficient (see Table 3) but also the analysis accuracy unless the observations become too sparse and/or too non-uniform (see Table 2). The new scheme is slightly more sensitive than the original En4DVar to the ensemble size and truncation number and to the observation spacing and non-uniformity, and is about the same moderately sensitive to the observation errors as the original En4DVar (see Table 2).

As in Qiu et al. (2007), the forecast model used in this study neglects the "true" terrain, so the associated model bias can cause the conventional EnKF fail to converge (see Fig. 3 of Qiu et al. 2007). The En4DVar has no problem to converge but the forecast is still subject to bias and so is the analysis. By applying the ergodicity assumption to the last 40 cycles (after the assimilation is stabilized), the forecast and analysis biases are diagnosed in the observation space and found to be not very significant (see Fig. 5) relative to the diagnosed forecast error standard deviations (listed in Table 5). The diagnosed bias fields, however, are very noisy due to limited samples (merely 40 cycles) used in the averaging. The under-sampling problem is evidenced by spurious noisy structures in the observation bias (see Fig. 5) and error standard deviation (in Fig. 6) computed from the last 40 cycles. The optimality consistency criteria proposed by Desroziers and Ivanov (2001) and Desroziers et al. (2005) are used to examine the analysis optimality and to diagnose the observation and forecast error variances against the prescribed observation error variance and En4DVar-estimated background error variance, respectively. To filter the spurious noisy structures, the diagnosed error variances are further averaged in the observation space. The results show that (i) the optimality criterion [see (15)] can be largely satisfied by both the new scheme and original En4DVar; and (ii) the consistency criterion [see (17)] can be more closely satisfied by the averaged forecast error variances in the new scheme than the original En4DVar (see Table 5).

For a given prediction model system, the gridded space dimension is fixed, so the dimension reduction due to the use of the hybrid space depends on the difference between the observation space dimension and model gridded space dimension. A similar situation is seen when a standard

3DVar (or 4DVar) is formulated in observation space versus model space. For conventional observations, the observations space dimension is much smaller than the model space dimension, so the new scheme can be more efficient than the original En4DVar. For remotely sensed observations (such as those from satellites and radars), the observation space can be very large and locally dense. There is no advantage to use the new scheme or any other schemes to directly assimilate such a huge amount observations due to the high degrees of information redundancy in these densely distributed observations (Eyre 1990; Huang and Purser 1996; Xu 2007). Redundant observations should be compressed into fewer super-observations before they are assimilated, and this will reduce the observation space dimension substantially to allow the hybrid-space approach to regain its computational advantages over the original En4DVar.

In this paper, the hybrid-space approach is tested only with idealized observations simulated by the same two-dimensional shallow-water equation model as in Qiu et al. (2007). The potential merits and limitations of the new En4DVar scheme need to be further verified by testing with more realistic models. When the analysis time window reduces to a single time level of observations in each assimilation cycle, the new scheme reduces to an En3DVar. This En3DVar was tested recently with simulated temperature observations by the Pennsylvania State University-NCAR Mesoscale Model version 5 (MM5) for a real weather event and the results showed improved analyses due to the use of En3DVar instead of the original MM5 3DVar (Shao and Qiu, 2007). Similar experiments will be performed with MM5 to test the hybrid-space En4DVar to address issues (such as the need of covariance localization) in real data applications.

In a standard ensemble-based data assimilation method, such as the EnKF or EnSRF, the

entire ensemble is updated by the analysis in each assimilation cycle and then is integrated forward to generate the forecast ensemble for the analysis in the next cycle. In this way, the method is able to learn about the forecast error covariance from the data. In the En4DVar developed in Qiu et al. (2007) and improved in this paper, the ensemble perturbations are not updated by the analysis and thus not relayed into the next assimilation cycle. Not updating the ensemble perturbations makes the En4DVar less venerable to model errors (especially systematic biases) and computationally more efficient than the standard ensemble-based data assimilation methods, but for a price of not being able to learn about the forecast error covariance from the data. This learning ability can be very important and useful for future data assimilation as the prediction models become increasingly accurate. It is thus desirable and eventually necessary to incorporate such a leaning ability into the En4DVar without excessively comprising its computational efficient. This problem is under our investigation.

Acknowledgments. The authors are thankful to the anonymous reviewers for their comments and suggestions that greatly improved the presentation of the results. This work was supported by the Grants 40505022 and 40575049 from National Science Foundation of China to Lanzhou University and by the ONR Grant N000140410312 to University of Oklahoma.

REFERENCES

Anderson, J. L. (2001) An ensemble adjustment Kalman filter for data assimilation, *Mon. Wea. Rev.,* **129**, 2884–2903.

Baek, S.-J., B. R. Hunt, E. Kalnay, E. Ott, and I. Szunyogh (2006) Local ensemble Kalman filtering in the presence of model bias, *Tellus* A, **58**, 293–306.

Bennett, A. F. (1992) *Inverse Method in Physical Oceanography*, 346 pp., Cambridge University Press, New York.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar (2001) Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects, *Mon. Wea. Rev.*, **129**, 420–436.

Burgers, G., P. J. van Leeuwen, and G. Evensen (1998) Analysis scheme in the ensemble Kalman filter, *Mon. Wea. Rev.*, **126**, 1719–1724.

Cao, Y. H., J, Zhu, I. M. Navon, and Z. Luo (2007) A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition, *International Journal for Numerical Methods in Fluids*, **54**, 353–353.

Cohn, S. E., and R. Todling (1996) Approximate data assimilation schemes for stable and unstable dynamics, *J. Met. Soc. Japan*, **74**, 63–75.

Cohn, S. E., A. da Silva, J. Guo, M. Sienkiewicz, and D. Lamich (1998) Assessing the effects of data selection with the DAO Physical-Space Statistical Analysis System, *Mon. Wea. Rev*., 126, 2913–2926.

Courtier, P. (1997) Dual formulation of four dimensional variational assimilation, *Quart. J. Roy. Meteor. Soc.*, **123**, 2449–2461.

Daley, R., and E. Barker (2001) NAVDAS: Formulation and Diagnostics, *Mon. Wea. Rev*., **129**, 869–883.

Desroziers, G., and S. Ivanov (2001) Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Quart. J. Roy. Meteor. Soc.*, **127**, 1433-1452.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli (2005) Diagnosis of observation, background and analysis-error statistic in observation space. *Quart. J. Roy. Meteor. Soc.*, **131**, 3385-3396.

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

Evensen, G (2003) The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dynamics*, **53**, 343–367.

Eyre, J. R. (1990) The information content of data from satellite sounding systems. A simulation study, *Quart. J. Roy. Meteor. Soc.*, **116**, 401–434.

Gillijns, S., and B. De. Moor (2007) Model error estimation in ensemble data assimilation. Nonlin, Processes Geophys., **14**, 59–71.

Golub, G. H., and C. F. Van Loan (1983) *Matrix Computations*, 476 pp., Johns Hopkins Univ. Press, Baltimore, MD.

Hamill, T. M., and C. Snyder (2000) A hybrid ensemble Kalman filter-3D variational analysis scheme, *Mon. Wea. Rev*., **128**, 2905–2919.

Hamill, T. M., and J. S. Whitaker (2005) Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches, *Mon. Wea. Rev*., **133**, 3132–3147.

Heemink A. W., M, Verlaan, and A. J. Segers (2001) Variance reduced ensemble Kalman filtering, *Mon. Wea. Rev*., **129**, 1718–1728.

Houtekamer, P. L., and H. L. Mitchell (2001) A sequential ensemble Kalman filter for atmospheric data assimilation, *Mon. Wea. Rev.*, **129**, 123–137.

Huang, H.-L., and Purser, R. J. (1996) Objective measures of the information density of satellite data, *Meteorology and Atmospheric Physics*, **60,** 105–117.

LeDimet, F. X., and O. Talagrand (1986) Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, *Tellus*, **38**A, 97–111.

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

Lorenc, A. (1981) A global three-dimensional multivariate statistical analysis system, *Mon. Wea. Rev.*, **109**, 701–721.

Lorenc, A. (2003) The potential of the ensemble Kalman fllter for NWP – a comparison with 4D-Var, *Q. J. R. Meteorol. Soc.*, **129,** 3183–3203.

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

Mitchell H. L., P. L. Houtekamer, and G. Pellerin (2002) Ensemble size, balance and model-error representation in an ensemble Kalman filter, *Mon. Wea. Rev.*, **130**, 2791–2808.

Parrish, D., and J. Derber (1992) The National Meteorological Center's spectral statistical analysis system, *Mon. Wea. Rev.*, **120**, 1747–1763.

Qiu, C., and J. Chou (2006) Four-dimensional data assimilation method based on SVD:

Theoretical aspect, *Theoretical and Applied Climatology*, 83, 51–57.

Qiu, C., A. Shao, Q. Xu, and L. Wei (2007) Fitting Model Fields to Observations by Using Singular Value Decomposition – An Ensemble-Based 4DVar Approach, *J. Geophys. Res.*, **112**, D11105, doi:10.1029/2006JD007994.

Reichle, R.H., D.B. McLaughlin, and D. Entekhabi (2002) Hydrologic data assimilation with the ensemble Kalman filter, *Mon. Wea. Rev.*, **130**, 103–114.

Shao, A., and C, Qiu (2007) A numerical experiments study on an ensemble-based reduced-dimensional data assimilation method, *Chinese Journal of Atmospheric Sciences*, **31**, 117–126.

Talagrand, O. (1997) Assimilation of observations, an introduction, *J. Met. Soc. Japan* (Special Issue), **75**, 1B, 81–99.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker (2003) Ensemble square root filters, *Mon. Wea. Rev.*, **131**, 1485–1490.

Todling R., S. E. Cohn, and N. S. Sivakumaran (1998) Suboptimal schemes for retrospective data assimilation based on the fixed-lag Kalman smoother, *Mon. Wea. Rev.*, **126**, 2274–2286.

Uzunoglu B., S. J. Fletcher, M. Zupanski, and I. M. Navon (2007) Adaptive ensemble reduction and inflation, *Q. J. R. Meteorol. Soc.*, **133**, 1281–1294.

Whitaker, J. S., and T. M. Hamill (2002) Ensemble data assimilation without perturbed observations, *Mon. Wea. Rev.* **130**, 1913–1924.

Xu, L., T. Rosmond, and R. Daley (2005) Development of NAVDAS-AR: formulation and initial test of the linear problem, *Tellus*, **57A**, 546–559.

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

Xu, Q. (1997) Generalized adjoint for physical processes with parameterized discontinuities – Part IV: Problems in time discretization, *J. Atmos. Sci.*, **54**, 2722–2728.

Xu, Q. (2007) Measuring information content from observations for data assimilation: Relative entropy versus Shannon entropy difference, *Tellus*, **59A**, 198–209.

Xu, Q., and C. Qiu (1997) Adjoint matching condition for parameterized discontinuities – A derivation using Lagrangian-form costfunction, *Advances in Atmospheric Sciences.* **14**, 49–52.

Zupanski, M. (2005) Maximum likelihood Ensemble filter: theoretical aspects, *Mon. Wea. Rev.*, **133**, 1710–1726.

Zupanski, D., and M. Zupanski (2006) Model error estimation employing an ensemble data assimilation approach, *Mon. Wea. Rev*., **134**, 1337–1354.

Table 1. The RMS errors averaged over the last 20 cycles (from day 15 to 25) for the analyses produced by the original En4DVar and new scheme from the experiments in group 1 that includes the reference pair (REO and REN), test pair (TEO and TEN) and additional test experiment (TEOa). Here, σ_h and $\sigma_{\mathbf{v}}$ denote the averaged RMS errors for the height h and velocity $\mathbf{v} = (u, v)$ fields, respectively.

Experiment:	*Original En4DVar*		*New Scheme*	
	σ_h (m)	$\sigma_{\mathbf{v}}$ (ms^{-1})	σ_h (m)	$\sigma_{\mathbf{v}}$ (ms^{-1})
Reference pair (REO & REN): τ = 12 h (N_t = 5), M = 150, p = 75	8.71	0.69	8.90	0.55
Test pair (TEO & TEN): τ = 6 h (N_t = 3), M = 150, p = 100	7.61	0.64	6.75	0.54
TEOa (as TEO but with refined **P**)	6.94	0.59	N/A	N/A

Table 2. As in Table 1 but for the averaged analysis RMS errors from the experiments in group 2 that includes the control pair (TEOa and TEN) and seven test pairs. The control pair duplicates TEOa and TEN in group 1, so their RMS errors duplicate those listed for TEOa and TEN in Table 1.

Experiment	*Original En4DVar*		*New Scheme*	
	σ_h (m)	σ_v (ms^{-1})	σ_h (m)	σ_v (ms^{-1})
Control pair (TEOa & TEN): $M = 150$, $p = 100$, Obs spacing = $3d$	6.94	0.59	6.75	0.54
Test pair 1: $M = 100$, $p = 80$	8.48	0.73	8.35	0.69
Test pair 2: $M = 50$, $p = 40$	14.10	1.23	12.83	1.15
Test pair 3: Obs error ×1.5	7.48	0.65	7.20	0.60
Test pair 4: Obs spacing = $5d$	8.48	0.73	8.29	0.72
Test pair 5: Obs spacing = $7d$	11.68	1.08	11.69	1.15
Test pair 6: non-uniform obs, 225 pts	8.35	0.72	8.44	0.71
Test pair 7: non-uniform obs, 112 pts	12.77	1.15	13.45	1.22

Table 3. Percentages of CPU times required by the original En4DVar and new scheme from different pairs of experiments relative to that required by the original En4DVar in REO from the reference pair in group 1.

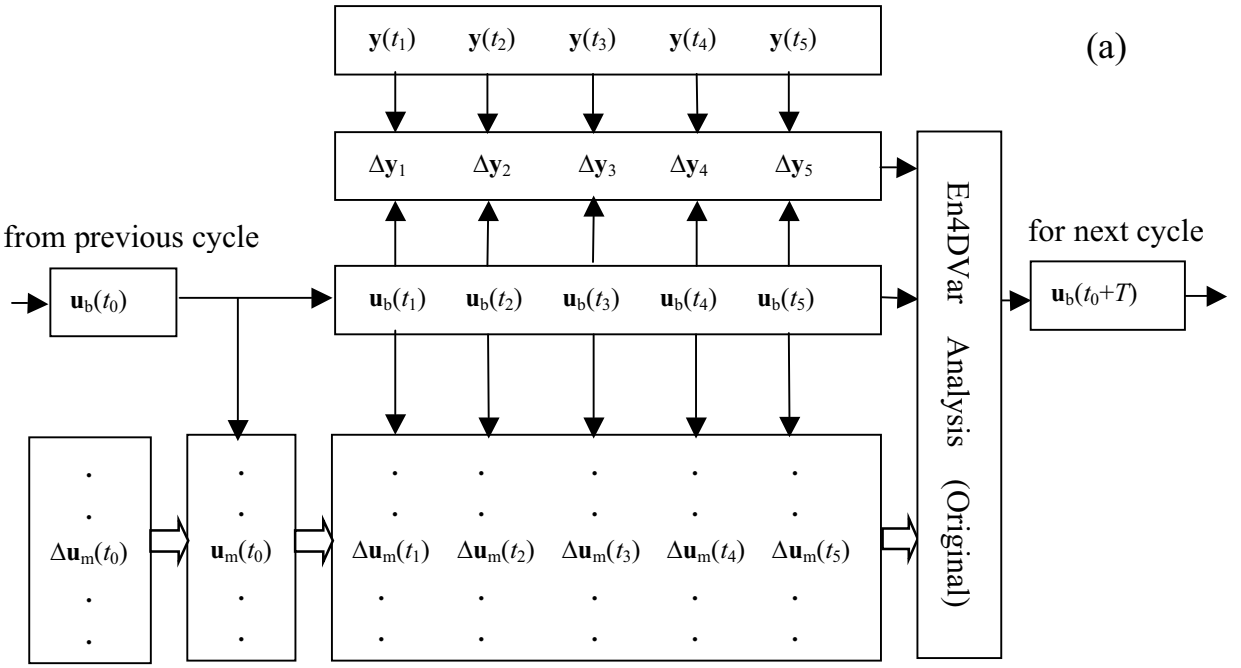
Experiment:	*Original scheme* *CPU in %*	*New Scheme* *CPU in %*
Reference pair (REO & REN) in group 1: $N_t = 5$, $M = 150$, $p = 75$, Obs spacing = $3d$	100	26
Control pair (TEOa & TEN) in group 2: $N_t = 3$, $M = 150$, $p = 100$, Obs spacing = $3d$	59	28
Test pair 1 in group 2: $M = 100$, $p = 80$	35	15
Test pair 2 in group 2: $M = 50$, $p = 40$	8.3	3.6
Test pair 4 in group 2: Obs spacing = $5d$ (225 pts)	58	25
Test pair 5 in group 2: Obs spacing = $7d$ (112 pts)	57	23

Table 4. Observation-space averaged values of the forecast and analysis biases diagnosed from the LHSs of (13) and (14), respectively, for TEOa and TEN.

	forecast bias			analysis bias		
	h (m)	u (ms^{-1})	v (ms^{-1})	h (m)	u (ms^{-1})	v (ms^{-1})
TEOa	-0.0139	-0.0002	0.0009	-0.0099	0.0010	0.0007
TEN	-0.0093	0.0004	0.0003	-0.0070	0.0007	0.0008

Table 5. Observation-space averaged values of the LHS diagnosed and RHS estimated forecast error variances from the two sides of Eq. (17) for TEOa and TEN. The consistency accuracy is computed in percentage for each data pair by the absolute value of the difference between the LHS diagnosed and RHS estimated values divided by the LHS diagnosed value.

	forecast error variance in h (m)			forecast error variance in $\mathbf{v}$ (ms^{-1})		
	LHS diagnosed	RHS estimated	Consistency accuracy	LHS diagnosed	RHS estimated	Consistency accuracy
TEOa	7.15	6.25	14%	0.95	0.94	1.1%
TEN	6.89	6.27	10%	1.19	1.18	0.1%



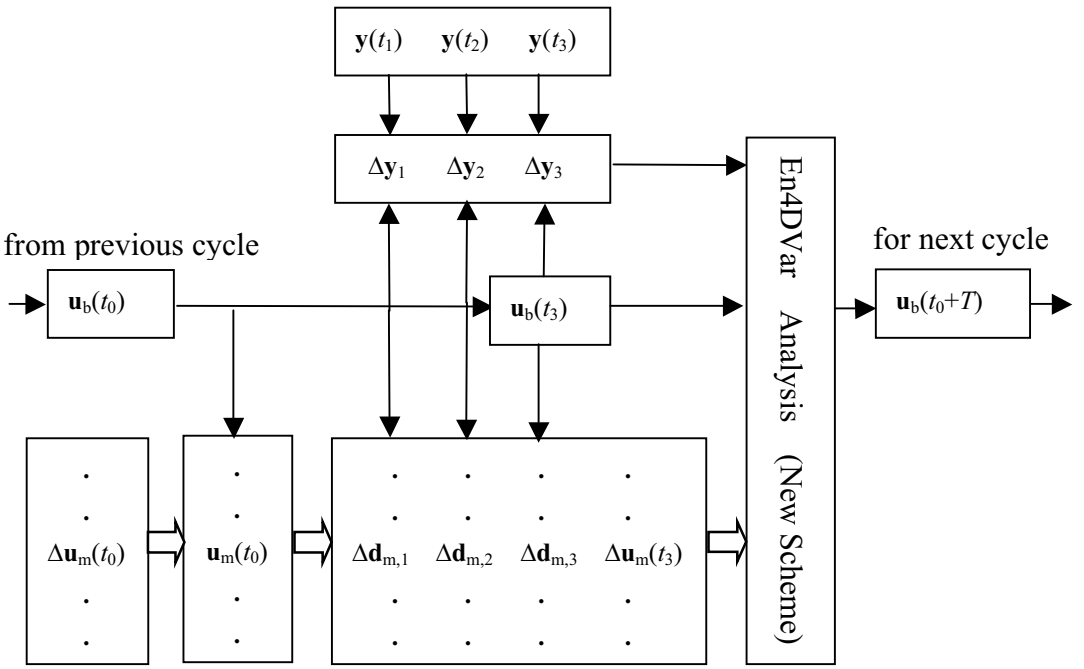
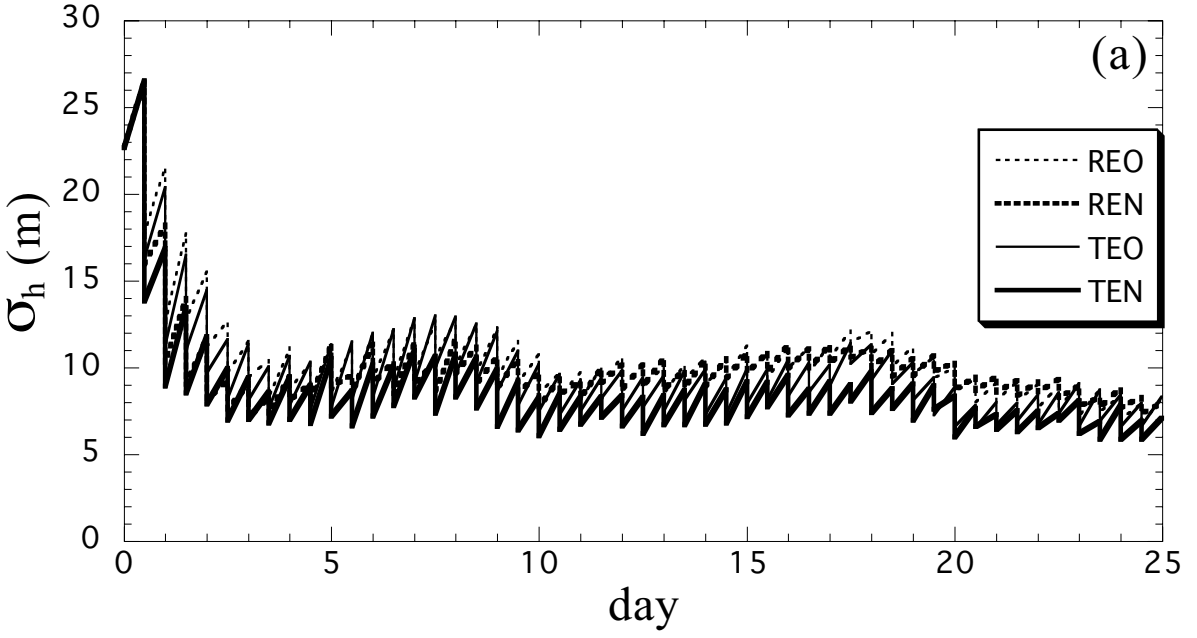


Fig. 1. Algorithm flow charts (a) for the original En4DVar used in Qiu et al. (2007) (as well as used by REO in this paper), and (b) for the new scheme (as well as the original scheme) used in

all the remaining experiments (except for REO) in this paper. Here, $\mathbf{u}_b(t_0)$ is the initial 3D background state vector (for the current cycle beginning at $t = t_0$) and $\mathbf{u}_b(t_n)$ is the 3D background state vector at the n-th observation time level $t = t_n$ obtained by integrating the model from $\mathbf{u}_b(t_0)$ at $t = t_0$; $\Delta\mathbf{u}_m(t_0)$ is the m-th member of the random perturbations generated initially at $t = t_0$ and $\mathbf{u}_m(t_0) = \mathbf{u}_b(t_0) + \Delta\mathbf{u}_m(t_0)$; $\Delta\mathbf{u}_m(t_n) = \mathbf{u}_m(t_n) - \mathbf{u}_b(t_n)$ and $\mathbf{u}_m(t_n)$ is the m-th ensemble state vector at $t = t_n$ obtained by integrating the model from $\mathbf{u}_m(t_0)$ at $t = t_0$; $\Delta\mathbf{y}_n$ is the 3D observation innovation vector at $t = t_n$, and $(\Delta\mathbf{y}_1^T, \Delta\mathbf{y}_2^T, \ldots \Delta\mathbf{y}_{N_t}^T)^T = \Delta\mathbf{y}$ is the 4D observation innovation vector [see Eqs. (3) and (8)]; $\Delta\mathbf{d}_{m,n}$ is the m-th member of the ensemble perturbations in the 3D observation subspace at $t = t_n$, and $(\Delta\mathbf{d}_{m,1}^T, \Delta\mathbf{d}_{m,2}^T, \ldots \Delta\mathbf{d}_{m,N_t}^T)^T = \Delta\mathbf{d}_m$ is the m-th member of the ensemble perturbations in the 4D observation space [see Eq. (5)].



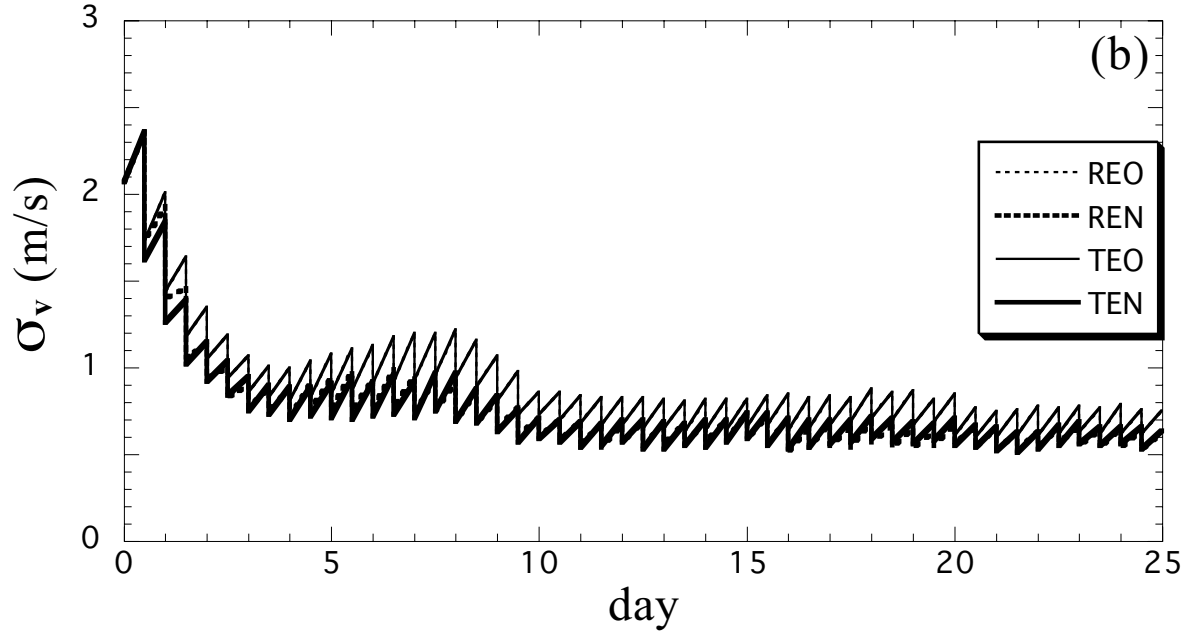


Fig. 2. Spatially averaged (a) height RMS errors and (b) velocity RMS errors plotted as functions of assimilation time for the analyses obtained by the original scheme (thin dotted curves) in REO and new scheme (thick dotted curves) in REN in the reference pair and by the original scheme

(thin solid curves) in TEO and new scheme (thick solid curves) in TEN in the test pair (all for the imperfect-model case. The drop of the error curve at each analysis time (every 12 hours) corresponds the error reduction made by the analysis. In panel (b), the thin dotted and solid curves obtained by the original scheme (in REO and TEO) are largely overlapped, and the thick dotted and solid curves obtained by the new scheme (in REN and TEN) are also largely overlapped and nearly indistinguishable.

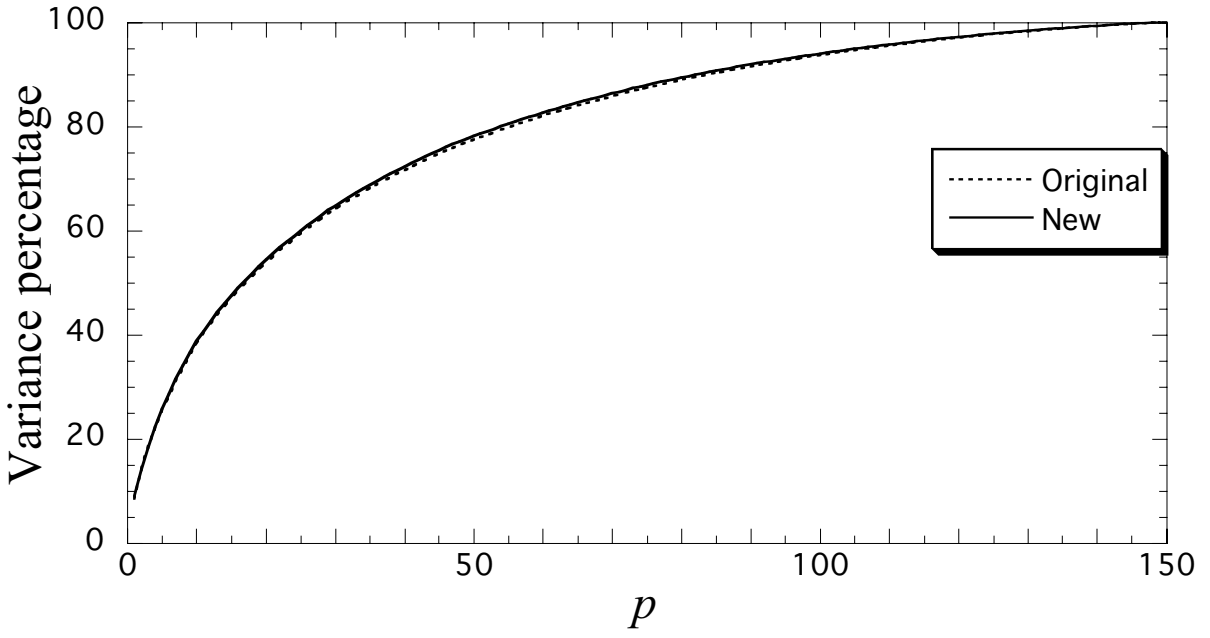


Fig. 3. Retained variance percentage, measured by $(\sum_{k=1}^{p}\lambda_k^2)/(\sum_{k=1}^{r}\lambda_k^2)$, plotted as functions of the truncation number p for the ensemble perturbations obtained in the first assimilation cycle by the original En4DVar (dotted curve) and new scheme (solid curve) in the test pair (TEO and TEN). Here, λ_k is the k-the singular value of **A** computed by the SVD in (1) and r = rank(**A**).

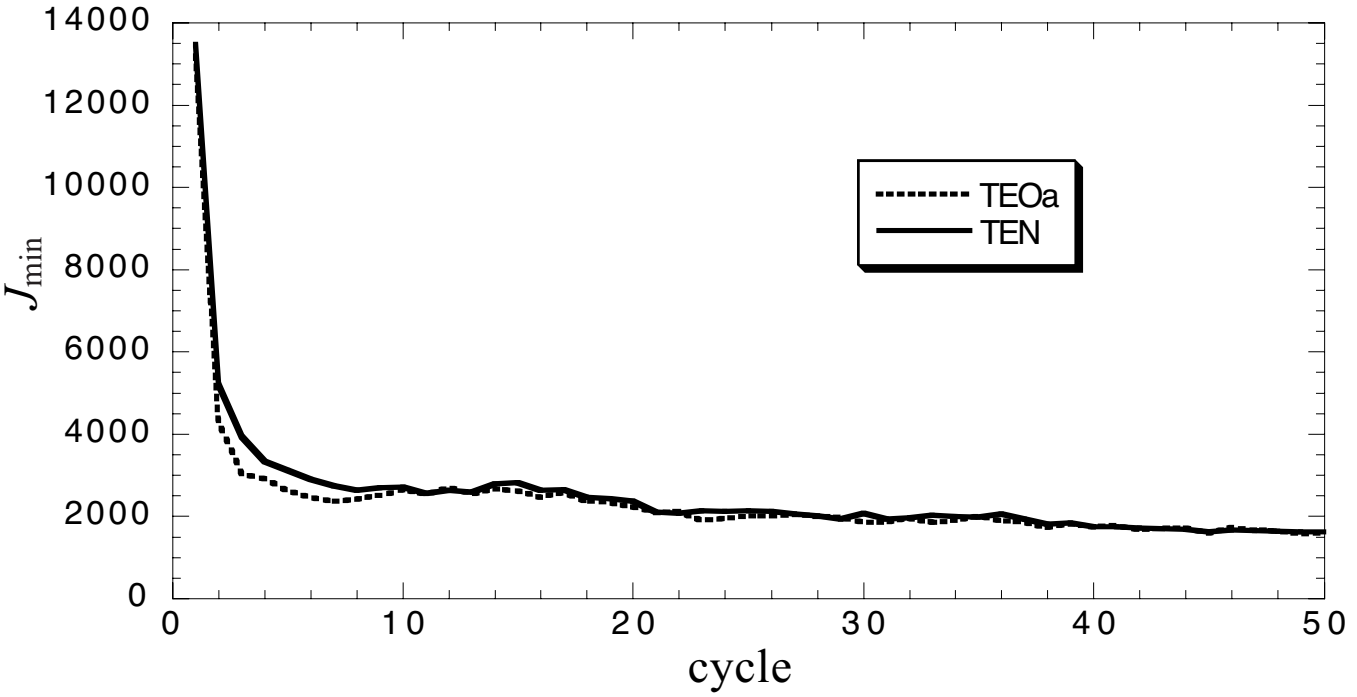


Fig. 4. J_{min} plotted as functions of the assimilation cycle number for the two control experiments: TEOa with the original scheme but new background error covariance model (dotted curve) and TEN with the new scheme (solid curve).

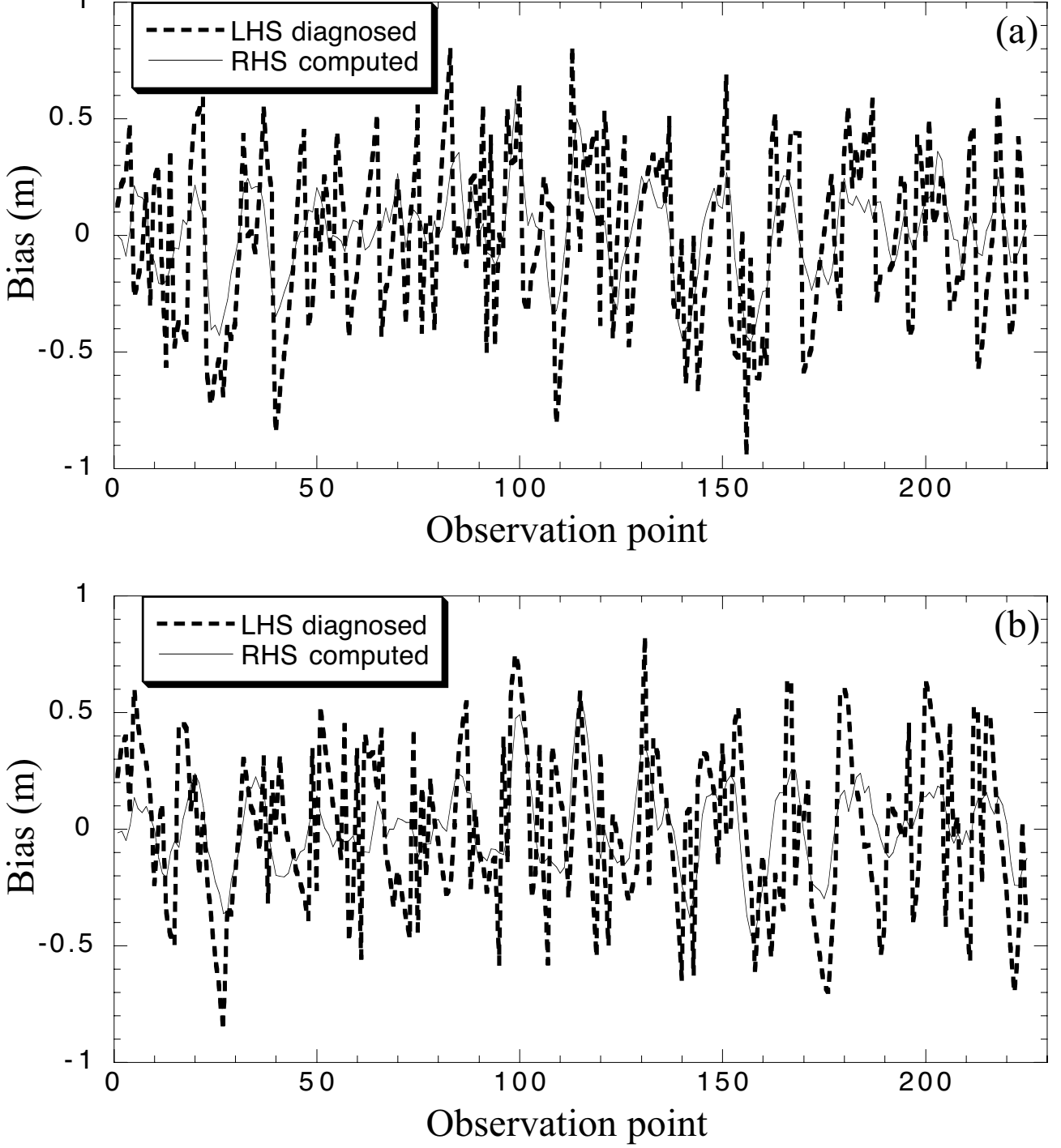


Fig. 5. Height forecast biases diagnosed from LHS of (13) and plotted by thick dashed curves as functions of the observation point at the last (third) observation time level for TEOa in (a) and TEN in (b). The thin solid curves are the directly computed height forecast biases by subtracting the "true" state from each forecast and then averaging over the last 40 cycles. The difference between the thick dashed and thin solid curves in each panel shows spurious observation bias caused by insufficient samples (from merely 40 cycles) in the averaging.

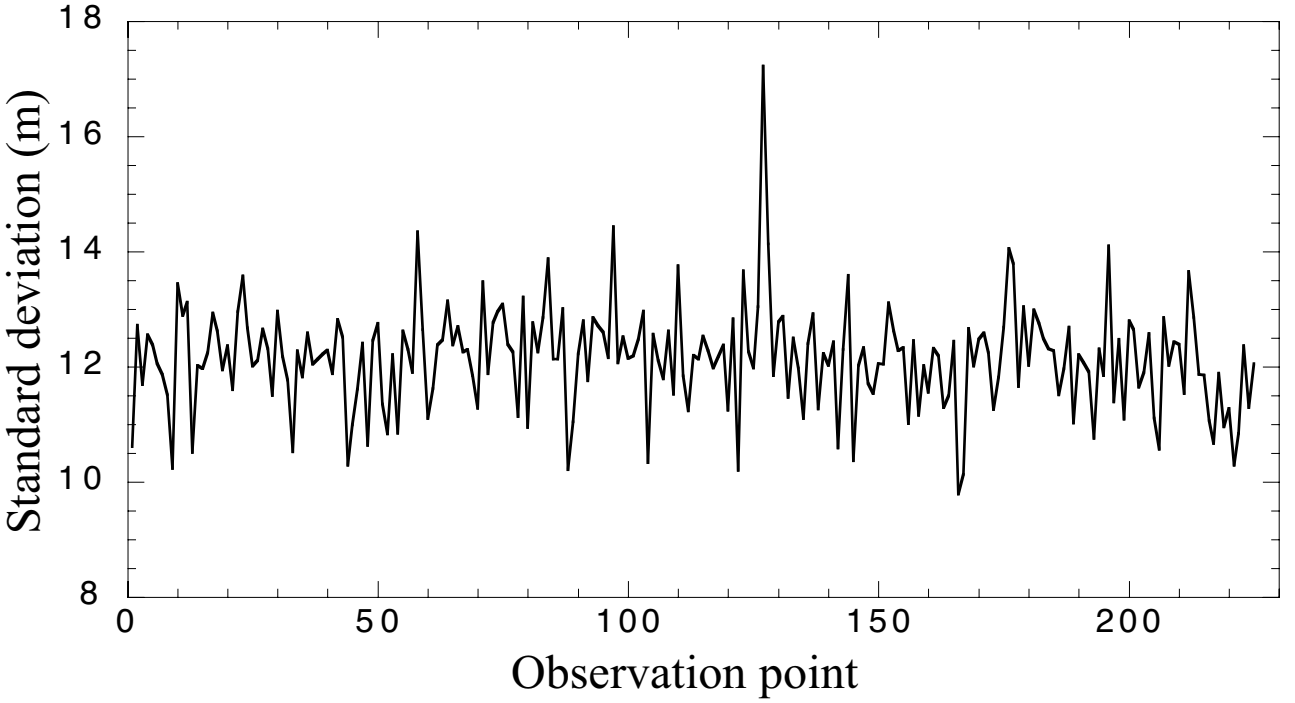


Fig. 6. Height observation error standard deviation diagnosed from LHS of (18) for TEOa at the last (third) observation time level.