


# Considering the impact of observation error correlation in ensemble square-root Kalman filter

Shaodong Zang<sup>1</sup>, Jichao Wang<sup>1</sup>\*

<sup>1</sup> China University of Petroleum, College of Science  
(Qingdao - shandong - China)

\*Corresponding author: wangjc@upc.edu.cn

## ABSTRACT

Data assimilation has been developed into an effective technology that can utilize a large number of multi-source unconventional data. It cannot only provide the initial field for the ocean numerical prediction model, but also construct the ocean reanalysis datasets and provide the design basis for the ocean observation plan. In data assimilation, the estimation of the observation error is of paramount importance, because the quality of the analysis depends on it. In general, the observation error covariance matrix is diagonal or assumed to be diagonal, which means that the observation errors are independent from one another. However, there are indeed correlations in the observation errors. A diagnostic method has been developed, which can estimate a correlated and more accurate observation error covariance matrix. The proposed method combines an ensemble square-root Kalman filter with the diagnostic method, providing an estimation of the observation error covariance matrix. In order to test the performance of the method, the numerical experiments are performed with the Lorenz 96 model and a Shallow water model. The more accurate observation error covariance matrix can be obtained to use in ensemble square-root Kalman filter by using the new method. We could find using the estimated correlated observation error in the data assimilation improves the analysis.

**Descriptors:** Correlated observation errors, Data assimilation, Ensemble square-root Kalman filter.

## INTRODUCTION

Data assimilation is the process of combining observations with a prior forecast state of the model, known as background, to produce an accurate estimate of the current state, known as analysis. It is important to effectively incorporate the observations into the numerical model to improve the accuracy of ocean prediction, when data assimilation is widely used in the field of ocean science. In recent years, more and more researchers concentrate on ensemble filters, as well as the ensemble variational method. In fact, most ocean numerical prediction models are high-dimensional nonlinear systems and then Kalman filter cannot do anything about it. But the ensemble filters are able to help us to deal with non-linear systems very well and the formulas of ensemble filters are much more computationally efficient than the Kalman filter. The

ensemble Kalman filter (EnKF) was originally introduced by Evensen (1994). EnKF is a good method but it comes at a cost: the filter divergence and numerical noise could be introduced in data assimilation. Then the numerical noise could affect the performance of the EnKF (Evensen, 2004). Whitaker and Hamill (2002) presented a new formula of ensemble filter called ensemble square-root Kalman filter in which the perturbations of measurements can be avoided. Based on this ensemble filter, a simpler and more straightforward variant of the square-root analysis scheme is presented by Evensen (2004). The number of calculations and storage has been greatly reduced in data assimilation due to the introduction of the idea of square-root. The ensemble square-root Kalman filter will be used for study in this paper and introduced in the subsequent section.

In order to provide an optimal estimate of the true state, the error covariances associated with the observations and background must be well understood and correctly specified (Houtekamer and Mitchell, 2005). But the correct

Submitted on: 13/April/2019

Approved on: 2/October/2019

<http://dx.doi.org/10.1590/S1679-87592019026106717>

Editor: Edmo Campos



error covariances of the observations and background are sometimes difficult to obtain in the real ocean system. Incorrect estimates of the observation errors would lead to a non-optimal analysis. Then these results can cause inaccurate estimate of analysis in data assimilation. Therefore, it is of vital importance to use the correct estimate of the error covariance matrix in data assimilation.

In general, much attention has been paid to the estimation of the background error covariance matrix. Significant progress has been made in this area and the background error covariance matrix often could be regarded as a flow-dependent matrix by taking covariance statistics of the differences between each ensemble member and the mean (Bannister, 2008). For atmospheric CO<sub>2</sub> data assimilation, however, the background errors cannot be obtained by ensemble-based techniques or other methods, then Chatterjee et al. (2013) proposed an approach in which the differences between two modeled CO<sub>2</sub> concentration fields, based on different but plausible CO<sub>2</sub> flux distributions and atmospheric transport models, are used as a proxy for the statistics of the background errors. The study of the Montmerle and Berre (2010) focused on diagnosing variations of background-error covariances between precipitating and non-precipitating areas. Hence, the development of the background error covariance matrix is a little better in recent years. In contrast, the development of the observation error covariance matrix is relatively slow. With the desire and need to make better use of the observations, especially for high-resolution forecasting, the understanding and accurate representation of observation error need to be addressed. The weight of the observation in data assimilation is determined by the observation error covariance matrix. Correlations in observation errors have a very different origin from those in the background (Fowler et al., 2018). In general, observation errors could be attributed to four different aspects: observation operator or forward model error, representativity error, pre-processing and instrument error. However, the instrumental error is often believed as uncorrelated error, because it can be eliminated by calibrating the instrument. The other sources of observation errors can lead to correlations (Hodyss and Satterfield, 2017; Janjić et al., 2017). Therefore, in order to improve the quantity and impact of observations used in data assimilation, it is necessary to consider the full and potentially correlated observation error statistics (Waller et al., 2016).

The observation error covariance matrix is not easy to directly provide and calculate in the real ocean system.

In general, it can be derived from the statistical average. When there is a lot of observation data, the storage and calculations (i.e. inverse matrix) of the matrix will become more complicated. Therefore, the observation error covariance matrix is assumed to be diagonal and invariant over time to save the computational time and simplify calculations, which means that the observation errors are independent from one another and without correlations. But in some real ocean numerical models and recent researches, this simple treatment is unreasonable and unrealistic in data assimilation. For example, observational data collected by the satellite radiation and radar have been proved to exist correlated and dependent observation errors (Bormann and Bauer, 2010a; Bormann et al., 2010b; Campbell et al., 2017; Ruggiero et al., 2016). When the correlated observation error covariance matrix is introduced, the analysis is better than the effect of assimilation with uncorrelated observation error covariance matrix (Stewart et al., 2013). Therefore it is not appropriate to treat the observation error covariance matrix as a diagonal matrix.

Despite the challenges existing in estimating correlated error, in order to solve the shortcoming of diagonal observation error covariance matrix and obtain the correct observation error covariance matrix, the study of the correlated observation error has been developed in recent years and various methods of constructing correlated observation error matrix have been presented to improve the performance of data assimilation. Desroziers et al. (2005) presented a popular and practical diagnostic method of estimating the error covariance matrix in the variational method filed (here and after denoted as the DBCP diagnostic method). The concrete application of the DBCP diagnostic method has been described and explained by Waller et al. (2016).

Based on combination of observation-minus-background (abbreviated as O-B) and observation-minus-analysis (abbreviated as O-A), the DBCP diagnostic method can be used to obtain a good estimation of the observation error covariance matrix (Desroziers et al., 2005). Then the observation error covariance matrix can be updated in the specified assimilated time, a more accurate analysis state can be obtained to provide the initial field for the ocean numerical prediction model. In recent years, the DBCP diagnostic method has been utilized in various variational method and ensemble filters scheme to provide estimates of the observation error covariance matrix and get a more accurate analysis in data assimilation. Some researchers introduced the DBCP diagnostic method into the local

ensemble transform Kalman filter (Li et al., 2009) and ensemble transform Kalman filter (Waller et al., 2014), while Cordoba et al. (2017) used the DBCP diagnostic method in atmospheric motion vector to provide an accurate analysis in the operational Met Office high-resolution data assimilation system. Their experimental results show that the better effects can be produced when using the DBCP diagnostic method in data assimilation. Due to the advantage of ensemble square-root Kalman filter in data assimilation, in the present paper the DBCP diagnostic method is introduced into the ensemble square-root Kalman filter to verify the performance of ensemble square-root Kalman filter that using correlated observation errors.

This paper is organized as follows. In section 2, the ensemble square-root Kalman filter (here and after denoted as the EnSRKF) is introduced in detail, as well as the DBCP diagnostic method. Then the experimental design is described in section 3 and some numerical results are demonstrated in section 4 to reveal the great effect of the combination of EnSRKF with the DBCP diagnostic method. Finally, conclusion and perspectives are given in section 5.

## METHODS

### THE ENSEMBLE SQUARE-ROOT KALMAN FILTER

A review article by Evensen (2003) covers many of the EnKF subsequent developments. With the ensemble filter applied to the data assimilation, the sample mean and the error covariance matrix calculated from the ensemble statistics result in a reduction in computational complexity, which greatly enhances the use of assimilation methods. Now numerous different approaches of data assimilation based on ensemble ideas have been developed to solve the real ocean system.

In this paper, the EnSRKF described in Evensen (2004) will be used for research and experiment. Now a brief overview of the EnSRKF and the notations used throughout this study are introduced. Suppose that observations are available at time  $t_n$ , and let them be assembled in the  $p$ -element vector  $y(t_n)$  (for  $p$  observations). The forecast state is represented by a  $m$ -element state vector  $x_k^f(t_n)$ , where the superscript “f” stands for forecast. Suppose that an ensemble of  $N$  such model states exists ( $1 \leq k \leq N$ ) and let these state vectors comprise the columns of the forecast ensemble matrix  $A^f(t_n)$  (a  $m \times N$  matrix) as follows:

$$A^f(t_n) = (x_1^f(t_n), x_2^f(t_n), \dots, x_N^f(t_n)). \quad (1)$$

For simplification and convenience of symbols, naturally, the time labels  $t_n$  can be omitted. The mean of ensemble members contained in  $A^f$  is defined as

$$\bar{x}^f = \frac{1}{N} \sum_{k=1}^N x_k^f. \quad (2)$$

Then subtracting the mean of ensemble members from the ensemble members can get the forecast perturbation matrix  $A^{f'}$  denoted as

$$A^{f'} = (x_1^f - \bar{x}^f, \dots, x_N^f - \bar{x}^f). \quad (3)$$

This symbol allows us to write the forecast error covariance matrix as

$$P^f = \frac{1}{N-1} A^{f'} A^{f'T}. \quad (4)$$

In the same way, the analysis error covariance matrix can be written as

$$P^a = \frac{1}{N-1} A^a A^{a'T}, \quad (5)$$

where the superscript “a” stands for analysis and  $A^a$  denotes the analysis perturbation matrix. In the Kalman filter, the analysis state is expressed as

$$x^a = x^f + P^f H^T (H P^f H^T + R)^{-1} (y - H x^f), \quad (6)$$

where  $H$  represents  $p \times m$  linear observation operator matrix (it provides a mapping from model space to observation space) and  $R$  is the  $p \times p$  observation error covariance matrix (the uncertainty of the observational data). Here  $m$  denotes the dimensions of the model state vectors. For the method of ensemble data assimilation, the single analysis value can be substituted by the analysis state ensemble matrix to provide the analysis ensemble matrix by using the Eq. (6),

$$A^a = A^f + P^f H^T (H P^f H^T + R)^{-1} (Y - H A^f), \quad (7)$$

where  $Y$  is the  $p \times N$  matrix of identical columns comprising the observation vector  $y$ . Therefore, by using Eq. (4) and Eq. (7), the mean of analysis ensemble members can be expressed as

$$\bar{x}^a = \bar{x}^f + A^f S^T C^{-1} (y - H \bar{x}^f), \quad (8)$$

where  $S = HA^f$  and  $C = SS^T + (N-1)R$ . In order to produce the square root  $A^{a^f}$  of the analysis error covariance matrix in Eq. (5) (the key step of the EnSRKF), that is, the analysis perturbation matrix, let the analysis error covariance matrix in Eq. (5) and the analysis error covariance matrix of the ordinary Kalman filter be equal. We can obtain the equation as follows,

$$\frac{1}{N-1}A^a A^{a^T} = (I - KH)P^f, \quad (9)$$

where K is the Kalman gain. Then we substitute Eq. (4) into Eq. (9) with the C and S that we defined before,

$$A^a A^{a^T} = A^f [I - S^T C^{-1} S] A^{f^T}. \quad (10)$$

Note that the matrix C is considered to be invertible and positively defined. The analysis ensemble perturbation matrix can be produced by taking the square root of  $I - S^T C^{-1} S$  in Eq. (10), which is  $A^a = A^f [I - S^T C^{-1} S]^{1/2}$ . Now the focus is to quickly find the square root of  $I - S^T C^{-1} S$ . The eigen-decomposition may be a great choice to tackle this problem. For more details of this method, please see Evensen (2004).

By introducing the eigen-decomposition into EnSRKF, the forecast ensemble perturbation matrix can be used for representing the analysis ensemble perturbation matrix. Meanwhile, the square-root filter is a deterministic filter that not requires the addition of perturbations to the observation ensemble and thus it does not introduce numerical noise in the observation. For the ensemble Kalman filter, however, it is necessary to add perturbations to the observation ensemble. This treatment will introduce numerical errors and affect the performance of the ensemble Kalman filter, especially when the number of observations for a single analysis is limited. Therefore, the EnSRKF is a good scheme of data assimilation.

It is generally considered that the observation error covariance matrix is a diagonal matrix, but here we don't recommend doing this and it has been explained by Bormann et al. (2010b). Here we present a method that combines EnSRKF and the DBCP diagnostic method. The following part would describe this diagnostic method.

#### THE DBCP DIAGNOSTIC METHOD

The DBCP diagnostic method by using the combination of O-B and O-A in variational methods to provide an estimate of the observation error covariance matrix

and update the observation error covariance matrix at the current time, where O-B is  $d^b = y - H(x^f)$  and O-A is  $d^a = y - H(x^a)$ . Assume that the errors of observation and forecast are always uncorrelated and independent in data assimilation, the observation error covariance matrix can be represented by the expectation of the product of O-B and O-A,

$$E[d^a d^{b^T}] \approx R. \quad (11)$$

Moreover, the DBCP diagnostic method can potentially provide information on imperfectly known observation and background error statistics (Desroziers et al., 2005). Another advantage is that it is nearly cost-free and can be applied to any analysis scheme in data assimilation.

Although the DBCP diagnostic method does not fully and explicitly account for the errors, it has been successfully used in a complex model to get an approximate observation error covariance matrix. The DBCP diagnostic method is first applied to 4D-Var data assimilation and has been great developed in variational methods. Then with the development of ensemble data assimilation methods, the DBCP diagnostic method has been applied to ensemble Kalman filter gradually. Waller et al. (2014) have applied this method to the ensemble transform Kalman filter and achieved great assimilation effect compared with the ordinary diagonal ensemble transform Kalman filter. Here, we will combine the DBCP diagnostic method with the EnSRKF to verify whether the DBCP diagnostic method still has a good effect on the EnSRKF.

#### THE ENSRKF WITH THE DBCP DIAGNOSTIC METHOD

Now the algorithmic formulas of combining the EnSRKF with DBCP diagnostic method are presented in the subsequent part. In this article, the EnSRKFR is deemed to the combination of EnSRKF and the DBCP diagnostic method. The EnSRKFR is used to estimate a correlated observation error covariance matrix and to update the currently known observation error covariance matrix. In this way, a more accurate error covariance matrix with correlated errors for the current observations can be produced in each assimilation step.

The EnSRKFR is mainly divided into two phases: start-up stage and the observation error covariance matrix update stage. Start-up stage: At this stage, before it reaches a preset number of assimilation steps  $N^s$ , we still use the pre-specified observation error covariance

matrix  $R_0$  (often it is a diagonal matrix) with each element unchanged and observation errors are not correlated. Next, observation error covariance matrix update stage: once the assimilation steps  $N^s$  have been executed, sufficient observation information has been obtained, so the update stage begins. Here the correlated observation error covariance matrix can be constructed at each assimilation step by using the statistical mean of O-B and O-A produced in the previous assimilation process. The EnSRKFR method in detail is given below. The observation operator  $H$  is chosen to be linear, but the method could be extended to a non-linear observation operator (Evensen, 2003). Begin with the initial ensemble members  $\{x_k^a(0)\}$  ( $1 \leq k \leq N$ ). Then we assume that the initial analysis error covariance matrix is  $P_0^a$  and the initial observation error covariance matrix is  $R_0$ , it is possible that this error could just consist of the instrument error. The specific implementation is demonstrated below.

The forecast stage: the updated ensemble obtained in the analysis (at time  $t_{n-1}$ ) is propagated by the model (under the perfect model assumption) for the next time step  $t_n$  to produce the forecast ensemble members,

$$x_i^f(t_n) = M(x_i^a(t_{n-1})) \text{ for } i = 1, \dots, N, \quad (12)$$

where  $M$  may represent a non-linear model. Again, for the sake of simplifying the notation, the time symbol  $t_n$  is omitted here. The forecast is the mean of the forecast ensemble members,

$$\bar{x}^f = \frac{1}{N} \sum_{i=1}^N x_i^f. \quad (13)$$

The forecast perturbation matrix is

$$A^f = (x_1^f - \bar{x}^f, \dots, x_N^f - \bar{x}^f). \quad (14)$$

Once the observation  $y$  of current time  $t_n$  is captured, the O-B in current time is obtained from the equation  $d^b = y - H(x^f)$ .

The analysis stage: some symbols of the previous sections are considered again. First we use the equation  $\bar{x}^a = \bar{x}^f + A^f S^T C^{-1} (y - H\bar{x}^f)$  to produce the estimate of analysis, where  $C = SS^T + (N-1)R$  and  $S = HA^f$ . Then the twice eigen-decompositions are used to provide the analysis ensemble perturbation matrix. The specific method is as follows: first the matrix  $C$  is carried out eigen-decomposition  $[Z, \Lambda] = \text{eig}(C)$ , where  $Z$  and

$\Lambda$  respectively represent the corresponding matrix of eigenvectors and eigenvalues. Then we perform the second eigen-decomposition and it can be given by the following formula  $[V, \Sigma^T \Sigma] = \text{eig}(X^T X)$ , where the matrix  $X^T X$  is constructed by  $X = \Lambda^{-1/2} Z^T S$ . Therefore, the analysis perturbation matrix can be obtained by using Eq. (10) as follows,

$$A^a = A^f V [I - \Sigma^T \Sigma]^{1/2} V^T. \quad (15)$$

Once the analysis ensemble perturbation matrix is found, it can be added to the  $\bar{X}^a$  to give the full analysis ensemble. That is  $A^a = \bar{X}^a + A^{a'}$ , where the  $\bar{X}^a$  is the  $m \times N$  matrix of identical analysis states  $\bar{x}^a$  in each column. Finally, the analysis is used for calculating the O-A in current time, i.e.  $d^a = y - H(x^a)$ .

Covariance diagnostic stage: if the current assimilation step  $n$  is more than the pre-specified number of steps  $N^s$  (i.e.  $n > N^s$ ), the observation error covariance matrix  $R$  is updated by using the following equation,

$$R = \frac{1}{(N^s - 1)} \sum_{k=n-N^s+1}^{k=n} d^a d^{b^T} \quad (16)$$

Note that since the observation error covariance matrix  $R$  generated by the DBCP diagnostic method may be not symmetric matrix, therefore the matrix also needs to be symmetrized by the following formula,

$$R = \frac{1}{2} (R + R^T), \text{ otherwise } R = R_0. \quad (17)$$

It can be seen from the above steps, the computational procedures of EnSRKFR are basically the same as those of EnSRKF. Only the background innovation (O-B) and the analysis innovation (O-A) are needed to obtain the estimate of the observation error covariance matrix. At every assimilation step, we can see from the Eq. (16) of the correlated observation error covariance matrix that updating the observation error matrix only utilizes the latest information and discards the previous information. At the same time, when calculating and analyzing the ensemble perturbation matrix, the eigenvalue decomposition is used, which reduces the amount of calculation and improves the computational efficiency.

This method produces a slowly time-varying estimate of the observation error covariance matrix, and we should take into account the most recent information relating to the observations (Waller et al., 2014). How to find an optimal pre-specified number of assimilation steps  $N^s$  is not

easy to carry out. In general, in order to obtain sufficient information from the preceding data, the pre-specified number would be larger than the number of entries to be estimated. However, if we use a larger pre-specified number, the estimate of the observation error covariance matrix will be an average over a large period of time. Therefore, the  $N^s$  must be a compromise between the large number of samples required to get a good approximation to the matrix and the limited number of samples that allows the time-varying nature of the observation error covariance matrix to be captured (Waller et al., 2014).

Due to the limited number of samples obtained in practice, the variation of the variables cannot be fully represented, the observation error covariance matrix may appear to be under-ranked. Hence a regularization method can be adopted to eliminate this phenomenon. Significant progress and development have been made in dealing with this problem. Weston et al. (2014) have successfully used a method called reconditioning techniques to obtain full rank approximation of this matrix. Regularization method is a complementary to the DBCP diagnostic method and promotes the development of the DBCP diagnostic method. The regularization method for Waller et al. (2014) is adopted in the present paper.

## CONFIGURATION OF NUMERICAL EXPERIMENTS

### THE MODELS

The numerical experiments are performed with the Lorenz 96 model (Lorenz and Emanuel, 1998) and a shallow water model (Krysta et al., 2011). These models are simple, but they exhibit strong nonlinear behavior. Because the dynamics of the two models are different, the comparison of the results from both models provides insight to which extent the combination of EnSRKF with the DBCP diagnostic method.

#### a. The Lorenz 96 model

The Lorenz 96 model has been widely used to exam performance of different ensemble filters. It has  $m$  variables  $\{x_i\}$ ,  $i = 1, \dots, m$ . The dynamic system is represented by the following ordinary differential equations:

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F \quad i = 1, \dots, m. \quad (18)$$

Note that the domain on which is defined the 40 variables is circle-like, so that  $x_{-1} = x_{39}$ ,  $x_0 = x_{40}$ ,  $x_1 = x_{41}$

The constant forcing term  $F$  is configured as eight and it can cause chaotic behavior. In this paper, this model is solved by the fourth-order Runge-kutta scheme with a time step  $\Delta t$  of 0.01 units (the time unit equal to five days).

#### b. The Shallow water model

A 2D model using the shallow water equations is used to assess the DBCP diagnostic method in the case of a multivariate model. For simplicity, frictional effects and the Coriolis force are ignored and so are the nonlinear terms. The equations consist of the horizontal and vertical velocities  $(u, v)$  and the water height  $h$ . Under these assumptions, the momentum equations can be formulated as:

$$\begin{aligned} d_t u &= -g * d_x h \\ d_t v &= -g * d_y h \\ d_t h &= d_x(hu) - d_y(hv) \end{aligned} \quad (19)$$

where  $g$  denotes the gravity acceleration. The model domain is chosen as the square domain  $[0, L] \times [0, L]$  with length  $L = 2200km$  and the  $22km$  resolution in both directions. The equations are solved by the Lax-Wendroff finite difference method with a time step  $\Delta t$  of 0.01 units (the time unit equal to 1000 minutes).

## EXPERIMENTAL DESIGN

In these experiments, some indispensable initial conditions are determined by using the similar methods. First the true state is produced by evolving the perfect (without model errors) model equations forward from known initial state value. Then the observation is obtained by adding the observation error to the true state. As for ensemble members, they are also evolved by using the perfect model but beginning from perturbed initial state. Note that the observation error covariance matrix is considered to be isotropic and homogeneous. Hence, the observation error covariance matrix obtained from EnSRKFR method needs have a cyclic structure. To regularize the estimated observation error covariance matrix, the method used in Waller et al. (2014) is adopted after each the DBCP diagnostic method. Now, the experiments will be specifically described.

#### a. The observations

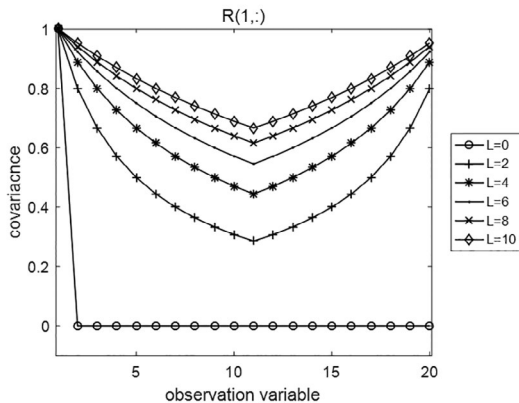


In general, the observations are generated by adding Gaussian distributed random numbers to the true states. The  $R^t$  is used to denote the true observation error covariance matrix. There are different methods to structure an observation error covariance matrix as exact observation error covariance matrix. Such as the SOAR function which has been used to calculate the correlation matrix can be seen in Bormann et al. (2003) and Waller et al. (2016). In the present method, a simple choice of the non-diagonal  $R^t$  is,

$$R^t(i,j) = \frac{\nu}{1 + \frac{d(i,j)}{2L}}, \quad (20)$$

where the observation error variance is fixed at  $\nu = 1.0$ ,  $d(i,j)$  denotes the grid distance between the  $i$ th and  $j$ th grid points with considering the cyclic boundary.  $L$  is the correlation strength parameter, so that a larger value of  $L$  corresponds to stronger observation error correlations in observations. If  $L = 0$ ,  $R^t$  is diagonal. The  $R^t$  varies with the different  $L$  values is shown in Figure 1. The vertical axis represents the covariance between first variable and other variables, the horizontal axis represents the variable. Then for conducting comparative experiments, the diagonal matrix  $diag(R^t)$  is chosen to be the initial observation error covariance matrix  $R_0$  in the EnSRKFR. Hence observations containing correlated observation errors could be obtained by adding correlated observation errors to the true states. The specific formula is as follows,

$$y = Hx^t + C \cdot N(0, 1), \quad (21)$$



**Figure 1.** The first row of the  $R$  matrix with different  $L$ . The vertical axis represents the covariance between first variable and other variables, the horizontal axis represents the variable.

where  $R = CC^T$  is the Cholesky decomposition of  $R^t$  and  $N(0,1)$  is a standard normal distribution.

### b. Experiments with the Lorenz 96 model

In order to perform the EnSRKFR method in Lorenz 96 model, we assume that the initial condition is from a vector that all elements are eight but the 0.2 is added to the 20<sup>th</sup> element. Similarly,  $H$  is the  $p \times m$  observation operator matrix, then the 20 equally spaced direct observations can be obtained by adding the observation errors to the true states at every assimilation step (using the method mentioned in section 3.2.1). The assimilation time step  $\Delta t$  of 0.01 units, or 1.2 hours  $\Delta t = 0.01$ , and the final time is  $T = 30$ , or 150 days. Then the observations can be obtained every 10 and 20 time steps, that is, every 0.1 and 0.2 time units, respectively. Next in order to generate the initial ensemble,  $N = 40$  pseudo-random samples from the normal distribution  $N(0, \sigma_b^2 I)$ , where the  $\sigma_b^2$  (we set the  $\sigma_b^2 = 2$ ) is the background error variance, are added to the true initial states. A large number of ensemble members is used to minimize the risk of ensemble collapse and to help obtain an accurate forecast error covariance matrix. For the purpose of this initial study, we wish to avoid using techniques of covariance inflation and localization so as not to contaminate the estimate of  $R$ . Note that, the pre-specified number of steps  $N^s$  is considered to be 85 in this experiment (We have tested that when  $N^s$  is greater than the number of variables, the assimilation effect is also good).

### c. Experiments with the Shallow water model

The shallow water model is a multivariate model, so an additional degree of complexity is introduced. At each grid point, the water height ( $h$ ), the horizontal ( $u$ ) and the vertical velocities ( $v$ ) are defined. But only the water height ( $h$ ) is observed in the experiments, and the 200 observational data are generated by adding the observation errors to the true states at every assimilation step. The experiment is initialized by integrating the initial state  $u=v=0$  and the initial water height ( $h$ ) is defined as follows:

$$\begin{aligned} h &= \sqrt{x^2 + \gamma^2} \\ h &= \sin(h)/h \\ h &= \max(h, 0) \end{aligned} \quad (22)$$

Here the assimilation time step  $\Delta t$  of 0.01 units, or ten minutes  $\Delta t = 0.01$ , and the final time is  $T = 15$ , or

15000 minutes. Next in order to generate the initial ensemble,  $N = 200$  pseudo-random samples from the normal distribution  $N(0, \sigma_b^2 I)$ , where the  $\sigma_b^2$  (we set the  $\sigma_b^2 = 1$ ) is the background error variance, are added to the true initial states. Here, the larger ensemble is also used to avoid using techniques of covariance inflation and localization so as not to contaminate the estimate of  $R$ . Note that, the pre-specified number of steps  $N^s$  is considered to be 55 in the experiments. We next present experimental results of applying EnSRKFR to these models.

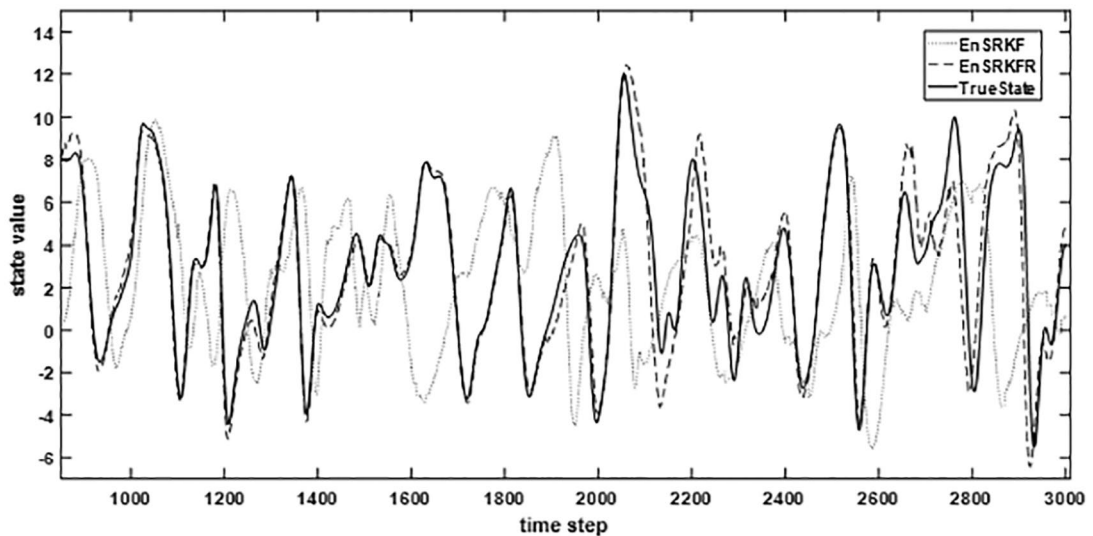
## RESULTS

In order to verify the performance of the EnSRKFR method compared with the EnSRKF method, we combine the mentioned models with the EnSRKFR method (using the  $R_0 = \text{diag}(R^t)$ ). As a comparison, the standard EnSRKF method (using the  $R = \text{diag}(R^t)$ ) is also applied to the models. The two methods are similar in the operation before the start-up stage of the DBCP diagnostic method. So only the analysis produced after the beginning of the start-up stage need to be compared in the EnSRKFR and the standard EnSRKF.

In the Lorenz 96 model, with the chosen frequencies being observations available every 10 and 20 time steps, then the twenty-fourth and twenty-fifth variables of the Lorenz 96 model are used as examples for analysis. In different observation

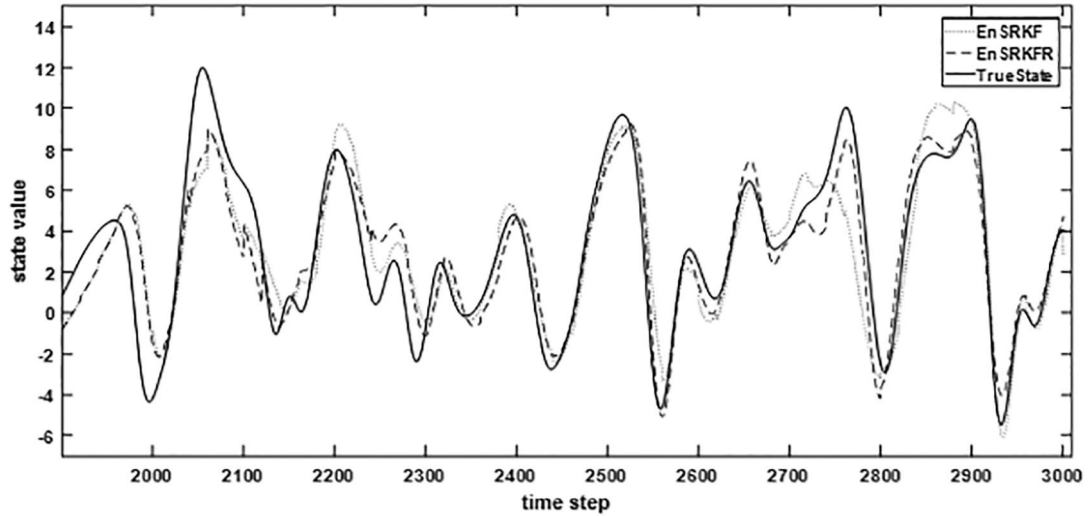
frequencies, the Figure 2 and Figure 3 show the analysis states of using EnSRKF method (using the  $R = \text{diag}(R^t)$ ) as an observation error matrix, the analysis states of using EnSRKFR method (using the  $R_0 = \text{diag}(R^t)$ ) and the true states, they are all generated after the start-up stage of the DBCP diagnostic method. It is clear that when the DBCP diagnostic method is added to EnSRKF, the analysis states can better fit the trajectory of true states compared with the analysis states without the DBCP diagnostic method, which is due to the introduction of correlated observation errors in the EnSRKFR. By using the two diagonal observation error matrix as the initial error matrix in these experiments, the results show that even if the initial diagonal observation error matrix is inaccurate, a better assimilation performance is shown in EnSRKFR compared with EnSRKF. When the chosen frequency of observations varies 20 time steps from 10 time steps, the analysis of EnSRKF deviates from the true state, but the analysis of EnSRKFR still shows excellent assimilation effect. All similar features can also be verified in other variables.

Then, we can quantify the difference between the analyses provided by the two different methods (EnSRKFR and EnSRKF) and the true state for different variables. The root-mean-square error (RMSE) is used to measure the deviation between the analysis state and the true state. Firstly, let us take the twenty-first to twenty-fifth variables as examples to show the changes of the RMSE for a single variable in EnSRKF and EnSRKFR, where observations are available



**Figure 2.** The analysis values in EnSRKF and EnSRKFR compared with the true states, where observations are available every 10 time steps and the twenty-fourth variable is used.





**Figure 3.** The analysis values in EnSRKF and EnSRKFR compared with the true states, where observations are available every 20 time steps and the twenty-fourth variable is used.

from 10 and 20 time steps in experiments, respectively. The formula is given by equation (23)

$$RMSE = \sqrt{\frac{\sum_{i=1}^T d_i^2}{T}}, \quad (23)$$

where the  $T$  represents the all assimilation time steps and  $d_i$  is the deviation of the analysis value from the true value at each time step for a single variable. Then, in order to get the RMSE of all variables at a certain assimilation time, the RMSE of the ensemble mean (E1) is defined as

$$E1 \sqrt{\frac{1}{m} \sum_{i=1}^m \left( \frac{1}{N} \sum_{j=1}^N X_i^j - X_i^{true} \right)^2}, \quad (24)$$

where  $N$  is the number of ensemble members,  $m$  is the number of state variables,  $X_i^j$  is the  $j$ th ensemble member for the  $i$ th variable and  $X_i^{true}$  is the “true” state. Here we only calculate the E1 at the final assimilation time. All the numerical results are shown in Table 1.

From Table 1, when the chosen frequencies of observations are identical, the RMSE of EnSRKFR has significantly decreased in each variable compared with that of EnSRKF. This result shows that the EnSRKFR works better than the EnSRKF in the test and the introduction of the correlated observation error covariance matrix has a positive influence on the assimilated effect. Meanwhile, the E1

of EnSRKFR shown in Table 1 are also decreasing, when compared with that of EnSRKF. Therefore, the use of correlated observation errors can better estimate the states of the simple ocean and atmospheric system.

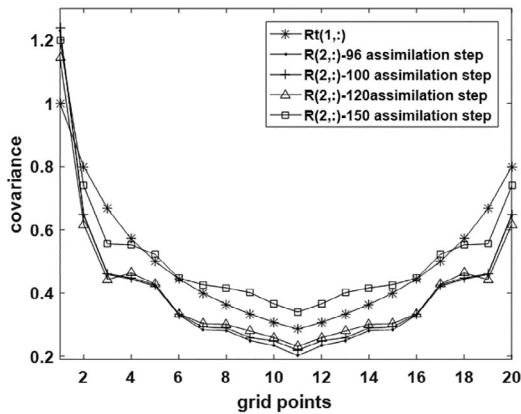
Then, in order to compare the difference between the observation error covariance matrix obtained from EnSRKFR and true observation error covariance matrix,  $R$  provided by the EnSRKFR in different assimilation time steps are shown in Figure 4. Obviously, as assimilation proceeds,  $R$  becomes closer and closer to the true observation error covariance matrix  $R^t$ . This result shows that DBCP diagnostic method is effective in estimating the observation error covariance matrix.

From the above experimental results, the EnSRKFR used in Lorenz 96 model demonstrates a better assimilation effect compared with the EnSRKF and the application of DBCP diagnostic method in EnSRKF is successful. Next, the EnSRKFR is used in a complex multivariate shallow water model.

In shallow water model, with the chosen frequencies being observations available every five and ten time steps, the water height ( $h$ ) analyses of final assimilation are shown in Figure 5 and Figure 6. From the figures, in different observation frequencies, the difference between analysis of EnSRKFR and true state is smaller than the analysis of EnSRKF. However, when we are concerned about the comparisons between the horizontal velocities ( $u$ ) of EnSRKFR and true states, it can be seen that the assimilation effect is not ideal compared with EnSRKF. The horizontal velocities

**Table 1.** The RMSE and E1 of EnSRKF compared with the RMSE of EnSRKFR in Lorenz 96 model.

Variables	Obs Freq. (time steps)	RMSE (EnSRKF)	RMSE (EnSRKFR)	Obs Freq. (time steps)	E1 (EnSRKFR)	E1 (EnSRKF)
$x_{21}$	10	1.02	0.63			
$x_{22}$	10	0.95	0.81			
$x_{23}$	10	1.17	0.95	10	2.61	1.74
$x_{24}$	10	1.71	1.38			
$x_{25}$	10	2.01	1.54			
$x_{21}$	20	1.12	0.98			
$x_{22}$	20	1.28	0.92			
$x_{23}$	20	1.54	1.16	20	3.0	2.1
$x_{24}$	20	1.91	1.60			
$x_{25}$	20	2.49	1.63			



**Figure 4.** Rows of the true and estimated covariance matrices, where the chosen frequency of observations is 20 time steps.

(u) from EnSRKF are closer to the true states in some variables, this may be due to the lack of horizontal and vertical velocity observations for the multiple shallow water model. The specific comparisons can be observed in Figure 7.

Then, for evaluating the assimilation performance in shallow water model, the RMSE is also used in different observation frequencies. Table 2 shows the RMSE of some variables in different observation frequencies. Compared with the EnSRKF, the RMSE in EnSRKFR are reduced for most variables. But when the observation frequency is five time steps, the RMSE of some variables increase a little bit. Therefore, the experimental results show that the application of the EnSRKFR method to the multivariate model may not be as effective as the application to the unary model. But in general, the effect of the EnSRKFR method is good.

## CONCLUSIONS

For a data assimilation process, in order to obtain an optimal estimation of the true state for ocean numerical models, the observation error covariance and the background error covariance must have more accurate estimation. The observation error covariance matrix is treated as a diagonal matrix in previous, but in the recent study, the observation error has been shown to be correlated. Miyoshi et al. (2013) introduced and demonstrated the beneficial effect of correlated observation error covariance matrix used in data assimilation, which means the observation error correlation is worthy of further study. Therefore, the method of obtaining correlated observation error covariance matrix has been greatly developed in recent years.

In this paper, we introduce a diagnostic method for constructing the observation error covariance matrix. This diagnostic method is combined with EnSRKF and a correlated observation error covariance matrix can be obtained by the combination of O-A and O-B at each assimilation step. The DBCP diagnostic method has been proved to give an approximate estimate of the true observation error covariance matrix. We can update the observation error covariance matrix at each assimilation step and then use it in the next assimilation. Hence a new observation error covariance matrix can be obtained at each assimilation step, which changes with the assimilation time and approximates the exact observation error covariance matrix.

In a simple data assimilation framework which is based on the EnSRKF, we use the Lorenz 96 model and neglect model error. Meanwhile, an isotropic and homogeneous observation error covariance matrix is considered

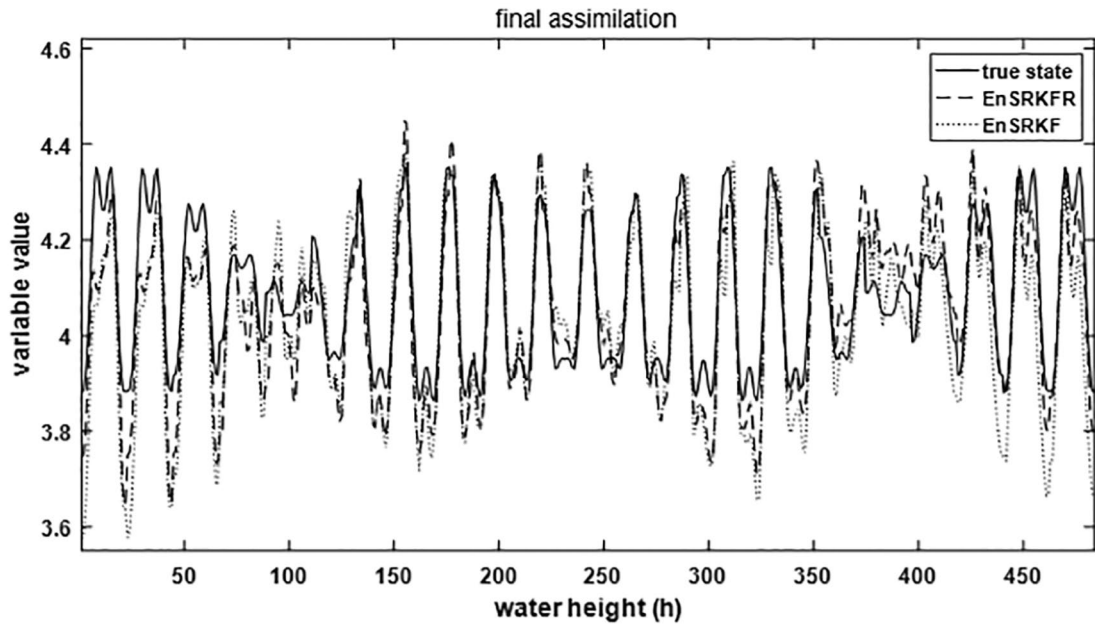


Figure 5. The analysis values in EnSRKF and EnSRKFR compared with the true states, where observations are available every 10 time steps.

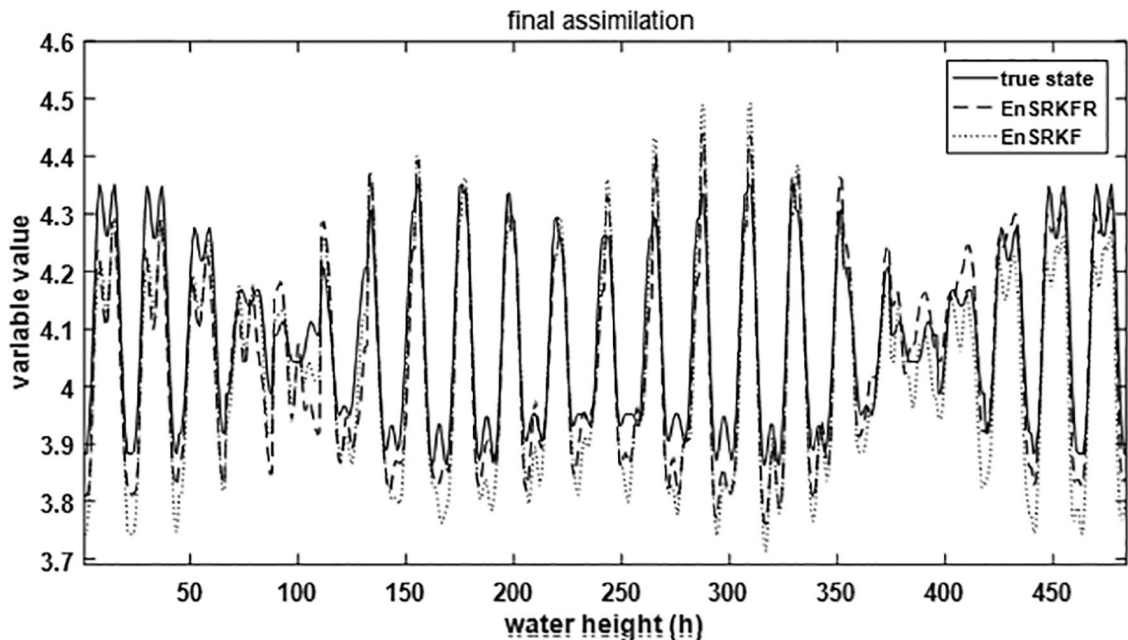


Figure 6. The analysis values in EnSRKF and EnSRKFR compared with the true states, where observations are available every 5 time steps.

to be used in the experiments. Here in order to verify the effect of the proposed method, so the model of assimilation is simplified for convenience. The results from these experiments demonstrate that the analysis after joining the DBCP diagnostic method is better than that without

the DBCP diagnostic method. Meanwhile, the RMSE of each variable is further significantly reduced. These all mean that EnSRKFR outperforms EnSRKF in these experiments. Therefore, under our basic assumption, the application of the DBCP diagnostic method to EnSRKF is

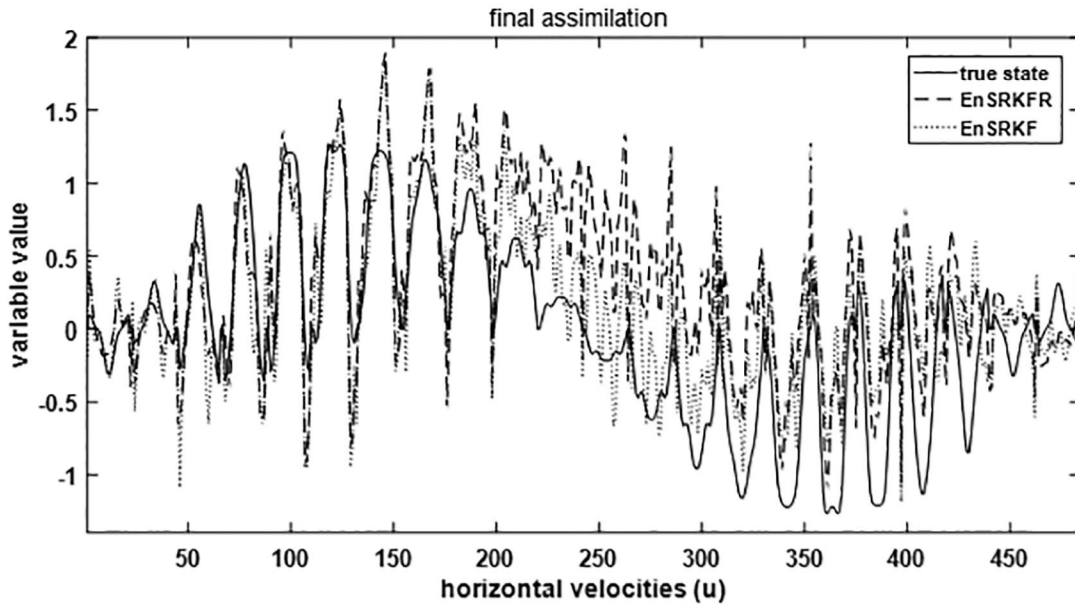


Figure 7. The analysis values in EnSRKF and EnSRKFR compared with the true states, where observations are available every ten time steps.

Table 2. Comparisons of RMSE between EnSRKFR and EnSRKF in shallow water model.

Variables	Obs Freq. (time steps)	RMSE (EnSRKF)	RMSE (EnSRKFR)	Obs Freq. (time steps)	RMSE (EnSRKF)	RMSE (EnSRKFR)
$h_{23}$	5	0.14	0.12	10	0.20	0.16
$h_{254}$	5	0.11	0.09	10	0.18	0.12
$h_{484}$	5	1.18	0.14	10	0.18	0.13
$u_{88}$	5	0.24	0.20	10	0.90	0.33
$u_{113}$	5	0.40	0.43	10	0.85	0.73
$u_{364}$	5	0.58	0.58	10	0.80	0.60
$v_{20}$	5	0.21	0.24	10	0.62	0.14
$v_{213}$	5	0.44	0.38	10	0.65	0.43
$v_{482}$	5	0.29	0.15	10	0.60	0.37

feasible and shows good results, that is, the inclusion of an approximate correlation structure in the observation error covariance matrix is generally better than the assumption of uncorrelated error. When the EnSRKF with correlated observation errors is applied to the ocean numerical model, the numerical model will show better numerical simulation results, because more accurate analyses are obtained in the process of data assimilation.

At the same time, in the one-dimensional Lorenz 96 model, for the EnSRKF method the  $R = \text{diag}(R^t)$  is chosen to be the observation error covariance matrix and for the EnSRKFR method an error covariance matrix  $\text{diag}(R^t)$  is treated as the initial error matrix. From the results of the experiments, however, the EnSRKFR

method still shows good properties, which shows that even if the initial observation error matrix used in EnSRKFR is not very accurate, it still will show better results than EnSRKF. Owing to the previous experimental results, we can clearly understand that the assimilation effect of EnSRKFR is indeed better than EnSRKF. The results demonstrate the importance of introducing a correlated observation error covariance matrix in data assimilation.

When the EnSRKFR method is used in the multi-variate shallow water model the analyses of water height are closer to true states and the RMSE is smaller than RMSE of EnSRKF method. However, from comparisons of horizontal velocities between EnSRKF method and EnSRKFR method, the EnSRKFR method is not as good

as the EnSRKF method in the multivariate model. These phenomena may be due to the correlation between the various variables in the multivariate model. Although the EnSRKFR method does not work well for some variables in the multivariate model, the error is generally acceptable. Due to the complexity of the multivariate model, the further work is required to improve the EnSRKFR method to better apply to the multivariate model.

## ACKNOWLEDGEMENTS

The authors thank the National Key Research and Development Project of China under contract No. 2016YFC1401800, the National Natural Science Foundation of China under contract No. 41406007 and the Fundamental Research Funds for the Central Universities under contract No. 19CX05003A-5 for support.

## REFERENCES

- BANNISTER, R. N. 2008. A review of forecast error covariance statistics in atmospheric variational data assimilation. I: characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, 134, 1951-1970. DOI: 10.1002/qj.339
- BORMANN, N. & BAUER, P. 2010a. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: methods and application to ATOVS data. *Quarterly Journal of the Royal Meteorological Society*, 136, 1036-1050. DOI: 10.1002/qj.616
- BORMANN, N., COLLARD, A. & BAUER, P. 2010b. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: application to AIRS and IASI data. *Quarterly Journal of the Royal Meteorological Society*, 136, 1051-1063. DOI: 10.1002/qj.615
- BORMANN, N., SAARINEN, S., KELLY, G. & THÉPAUT, J. N. 2003. The spatial structure of observation errors in atmospheric motion vectors from geostationary satellite data. *Monthly Weather Review*, 131, 706-718. DOI: 10.1175/1520-0493(2003)131<0706:tssooe>2.0.co;2
- CAMPBELL, W. F., SATTERFIELD, E. A., RUSTON, B. & BAKER, N. L. 2017. Accounting for correlated observation error in a dual-formulation 4D variational data assimilation system. *Monthly Weather Review*, 145, 1019-1032. DOI: 10.1175/mwr-d-16-0240.1
- CHATTERJEE, A., ENGELEN, R. J., KAWA, S. R., SWEENEY, C. & MICHALAK, A. M. 2013. Background error covariance estimation for atmospheric CO<sub>2</sub> data assimilation. *Journal of Geophysical Research: Atmospheres*, 118, 10,140-10,154. DOI: 10.1002/jgrd.50654
- CORDOBA, M., DANCE, S. L., KELLY, G. A., NICHOLS, N. K. & WALLER, J. A. 2017. Diagnosing atmospheric motion vector observation errors for an operational high-resolution data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 143, 333-341. DOI: 10.1002/qj.2925
- DESROZIERS, G., BERRE, L., CHAPNIK, B. & POLI, P. 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, 131, 3385-3396.
- EVENSEN, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99, 10143-10162. DOI: 10.1029/94JC00572
- EVENSEN, G. 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53, 343-367. DOI: 10.1007/s10236-003-0036-9
- EVENSEN, G. 2004. Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, 54, 539-560. DOI: 10.1007/s10236-004-0099-2
- FOWLER, A. M., DANCE, S. L. & WALLER, J. A. 2018. On the interaction of observation and prior error correlations in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144, 48-62. DOI: 10.1002/qj.3183
- HODYSS, D. & SATTERFIELD, E. 2017. The treatment, estimation, and issues with representation error modelling. In: PARK, S. K. & XU, L. (eds.) *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. III)*. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-43415-5\_8
- HOUTEKAMER, P. L. & MITCHELL, H. L. 2005. Ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, 131, 3269-3289. DOI: 10.1256/qj.05.135
- JANJÍČ, T., BORMANN, N., BOCQUET, M., CARTON, J. A., COHN, S. E., DANCE, S. L., LOSA, S. N., NICHOLS, N. K., POTTHAST, R., WALLER, J. A. & WESTON, P. 2017. On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144, 1257-1278. DOI: 10.1002/qj.3130
- KRYSTA, J. S., COSME, E. & VERRON, J. 2011. A consistent hybrid variational-smoothing data assimilation method: Application to a simple shallow-water model of the turbulent midlatitude Ocean. *Monthly Weather Review*, 139, 3333-3347
- LI, H., KALNAY, E. & MIYOSHI, T. 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 135, 523-533. DOI: 10.1002/qj.371
- LORENZ, E. N. & EMANUEL, K. A. 1998. Optimal sites for supplementary weather observations: simulation with a small model. *Journal of the Atmospheric Sciences*, 55, 399-414. DOI: 10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2
- MIYOSHI, T., KALNAY, E. & LI, H. 2013. Estimating and including observation-error correlations in data assimilation. *Inverse Problems in Science and Engineering*, 21, 387-398. DOI: 10.1080/17415977.2012.712527
- MONTMERLE, T. & BERRE, L. 2010. Diagnosis and formulation of heterogeneous background-error covariances at the mesoscale. *Quarterly Journal of the Royal Meteorological Society*, 136, 1408-1420. DOI: 10.1002/qj.655
- RUGGIERO, G. A., COSME, E., BRANKART, J.-M., LE SOMMER, J. & UBELMANN, C. 2016. An efficient way to account for observation error correlations in the assimilation of data from the future SWOT high-resolution altimeter mission. *Journal of Atmospheric and Oceanic Technology*, 33, 2755-2768. DOI: 10.1175/jtech-d-16-0048.1

- STEWART, L. M., DANCE, S. L. & NICHOLS, N. K. 2013. Data assimilation with correlated observation errors: experiments with a 1-D shallow water model. *Tellus A: Dynamic Meteorology and Oceanography*, 65, 19546. DOI: 10.3402/tellusa.v65i0.19546
- WALLER, J. A., DANCE, S. L., LAWLESS, A. S. & NICHOLS, N. K. 2014. Estimating correlated observation error statistics using an ensemble transform Kalman filter. *Tellus A: Dynamic Meteorology and Oceanography*, 66, 23294. DOI: 10.3402/tellusa.v66.23294
- WALLER, J. A., DANCE, S. L. & NICHOLS, N. K. 2016. Theoretical insight into diagnosing observation error correlations using observation-minus-background and observation-minus-analysis statistics. *Quarterly Journal of the Royal Meteorological Society*, 142, 418-431. DOI: 10.1002/qj.2661
- WESTON, P. P., BELL, W. & EYRE, J. R. 2014. Accounting for correlated error in the assimilation of high-resolution sounder data. *Quarterly Journal of the Royal Meteorological Society*, 140, 2420-2429. DOI: 10.1002/qj.2306
- WHITAKER, J. S. & HAMILL, T. M. 2002. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130, 1913-1924. DOI: 10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2