An Estimate of the Inflation Factor and Analysis Sensitivity in the Ensemble Kalman Filter

Guocan Wu[1,2] and Xiaogu Zheng[3]

1 College of Global Change and Earth System Science, Beijing Normal University, Beijing, China

2 Joint Center for Global Change Studies, Beijing, China

3 Key Laboratory of Regional Climate-Environment Research for East Asia, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

Abstract

The Ensemble Kalman Filter is a widely used ensemble-based assimilation method, which estimates the forecast error covariance matrix using a Monte Carlo approach that involves an ensemble of short-term forecasts. While the accuracy of the forecast error covariance matrix is crucial for achieving accurate forecasts, the estimate given by the EnKF needs to be improved using inflation techniques. Otherwise, the sampling covariance matrix of perturbed forecast states will underestimate the true forecast error covariance matrix because of the limited ensemble size and large model errors, which may eventually result in the divergence of the filter.

In this study, the forecast error covariance inflation factor is estimated using a generalized cross-validation technique. The improved EnKF assimilation scheme is tested on the atmosphere-like Lorenz-96 model with spatially correlated observations, and is shown to reduce the analysis error and increase its sensitivity to the observations.

Key words: data assimilation; ensemble Kalman filter; forecast error inflation; analysis sensitivity; cross validation

1. Introduction

For state variables in geophysical research fields, a common assumption is that systems have "true" underlying states. Data assimilation is a powerful mechanism for estimating the true trajectory based on the effective combination of a dynamic forecast system (such as a numerical model) and observations (Miller et al. 1994). Data assimilation provides an analysis state that is usually a better estimate of the state variable because it considers all of the information provided by the model forecasts and observations. In fact, the analysis state can generally be treated as the weighted average of the model forecasts and observations, while the weights are approximately proportional to the inverse of the corresponding covariance matrices (Talagrand 1997). Therefore, the performance of a data assimilation method relies significantly on whether the error covariance matrices are estimated accurately. If this is the case, the assimilation can be accomplished with the rapid development of supercomputers (Reichle 2008), although finding the appropriate analysis state is a much difficult problem when the models are nonlinear.

The ensemble Kalman filter (EnKF) is a practical ensemble-based assimilation scheme that estimates the forecast error covariance matrix using a Monte Carlo method with the short-term ensemble forecast states (Burgers et al. 1998; Evensen 1994). Because of the limited ensemble size and large model errors, the sampling covariance matrix of the ensemble forecast states usually underestimates the true forecast error covariance matrix. This finding indicates that the filter is over reliant on

1   the model forecasts and excludes the observations. It can eventually result in the

2   divergence of the filter (Anderson and Anderson 1999; Constantinescu et al. 2007;

3   Wu et al. 2014).

4   The covariance inflation technique is used to mitigate filter divergence by

5   inflating the empirical covariance in EnKF, and it can increase the weight of the

6   observations in the analysis state (Xu et al. 2013). In reality, this method will perturb

7   the subspace spanned by the ensemble vectors and better capture the sub-growing

8   directions that may   not have been captured by the original ensemble (Yang et al.

9   2015). Therefore, using the inflation technique to enhance the estimate accuracy of

10   the forecast error covariance matrix is increasingly important.

11   A widely used inflation technique involves multiplying the forecast error matrix

12   by an inflation factor, which must be chosen appropriately. In early studies,

13   researchers usually tuned the inflation factor by repeated assimilation experiments

14   and selected the estimated inflation factor according to their experience and prior

15   knowledge (Anderson and Anderson 1999). However, such methods are very

16   empirical and subjective. It is not appropriate to use the same inflation factor during

17   all the assimilation procedure. Too small or too large an inflation factor will cause the

18   analysis state to over rely on the model forecasts or observations, and can seriously

19   undermine the accuracy and stability of the filter.

20   In later studies, the inflation factor is estimated online based on the innovation

21   statistic (observation-minus-forecast; (Dee 1995; Dee and Silva 1999)) with different

22   conditions. Moment estimation can facilitate the calculation by solving an equation of

the innovation statistic and its realization (Li et al. 2009; Miyoshi 2011; Wang and Bishop 2003). Maximum likelihood approach can obtain a better estimate of the inflation factor than moment approach, although it must calculate a high dimensional matrix determinant (Liang et al. 2012; Zheng 2009). Bayesian approach assumes a prior distribution for the inflation factor but is limited by spatially independent observational errors (Anderson 2007, 2009). This study seeks to address the estimation of the inflation factor from the perspective of cross validation (CV).

The concept of CV was first introduced for linear regressions (Allen 1974) and spline smoothing (Wahba and Wold 1975), and it represents a common approach that can be applied to estimate tuning parameters in generalized additive models, nonparametric regressions and kernel smoothing (Eubank 1999; Gentle et al. 2004; Green and Silverman. 1994; Wand and Jones 1995). Usually, the data are divided into subsets some of which are used for modelling and analysis while others for verification and validation. The most widely used technique removes only one data point and uses the remainder to estimate the value at this point to test the estimation accuracy, which is also called the leave-one-out cross validation (Gu and Wahba 1991).

The basic motivation behind CV is to minimize the prediction error at the sampling points. The generalised cross validation (GCV) is a modified form of ordinary CV, that has been found to possess several favourable properties and is more popular for selecting tuning parameters (Craven and Wahba 1979). For instance, Gu and Wahba (1991) applied the Newton's method to optimize the GCV score with

multiple smoothing parameters in a smoothing spline model. Wahba (1995) briefly reviewed the properties of the GCV and conducted an experiment to choose smoothing parameters in the context of variational data assimilation schemes with numerical weather prediction models. Zheng and Basher (1995) also applied the GCV in a thin-plate smoothing spline model of spatial climate data to deal with South Pacific rainfalls.

Actually, the GCV criterion is based on a predictive mean-square error criterion that attempts to obtain a best estimate (Wahba et al. 1995). It has a rotation-invariant property that is relative to the orthogonal transformation of the observations and is a consistent estimate of the relative loss (Gu 2002). For the inverse problems in such fields as meteorological data assimilation, GCV method can choose parameters systematically by minimizing a given objective function that will improve the assimilation results. It can particularly select parameters that reflect not only measurement accuracies from different sources but also model capability (Krakauer et al. 2004).

This study proposes a new method for choosing the inflation factor using GCV method. The suitability of this choice is assessed using a statistic known as the analysis sensitivity, which apportions uncertainty in the output to different sources of uncertainty in the input (Saltelli et al. 2004; Saltelli et al. 2008). In the context of statistical data assimilation, this quantity describes the sensitivity of the analysis to the observations, which is complementary to the sensitivity of the analysis to model forecasts (Cardinali et al. 2004; Liu et al. 2009).

This study focuses on a methodology that can be potentially applied to geophysical applications of data assimilation in the near future. This paper consists of four sections. The conventional EnKF scheme is summarized and the improved EnKF with GCV inflation scheme is proposed in Section 2; the verification and validation processes are conducted on an idealized model in Section 3; the discussions are presented in Section 4 and conclusions are given in Section 5.

2. Methodology

2.1. EnKF algorithm

For consistency, a nonlinear discrete-time dynamical forecast model and linear observation system can be expressed as follows (Ide et al. 1997):

$$\mathbf{x}_i^t = M_{i-1}\left(\mathbf{x}_{i-1}^a\right) + \boldsymbol{\eta}_i, \tag{1}$$

$$\mathbf{y}_i^o = \mathbf{H}_i \mathbf{x}_i^t + \boldsymbol{\varepsilon}_i, \tag{2}$$

where i represents the time index; $\mathbf{x}_i^t = \left\{x_{i,1}^t, x_{i,2}^t, ..., x_{i,n}^t\right\}^T$ represents the n-dimensional true state vector at the i-th time step; $\mathbf{x}_{i-1}^a = \left\{x_{i-1,1}^a, x_{i-1,2}^a, ..., x_{i-1,n}^a\right\}^T$ represents the n-dimensional analysis state vector, which is an estimate of $\mathbf{x}_{i-1}^t$; M_{i-1} represents a nonlinear dynamical forecast operator such as a numerical weather prediction model; $\mathbf{y}_i^o = \left\{y_{i,1}^o, y_{i,2}^o, ..., y_{i,p_i}^o\right\}^T$ represents a p_i-dimensional observation vector; $\mathbf{H}_i$ represents the observation operator matrix; and $\boldsymbol{\eta}_i$ and $\boldsymbol{\varepsilon}_i$ represent the forecast and observation error vectors, which are assumed to be time-uncorrelated,

statistically independent of each other and have mean zero and covariance matrices $\mathbf{P}_i$ and $\mathbf{R}_i$, respectively. The EnKF assimilation result is a series of analysis states $\mathbf{x}_i^a$ that is an accurate estimate of the corresponding true states $\mathbf{x}_i^t$ based on the information provided by M_i and $\mathbf{y}_i^o$.

Suppose the perturbed analysis state at a previous time step $\mathbf{x}_{i-1}^{a(j)}$ has been estimated ($1 \le j \le m$ and m is the ensemble size), the detailed EnKF assimilation procedure is summarized as the following forecast step and analysis step (Burgers et al. 1998; Evensen 1994).

Step 1. Forecast step.

The perturbed forecast states are generated by running dynamical model forward:

$$\mathbf{x}_i^{f(j)} = M_{i-1}\left(\mathbf{x}_{i-1}^{a(j)}\right). \tag{3}$$

The forecast state $\mathbf{x}_i^f$ is defined as the ensemble mean of $\mathbf{x}_i^{f(j)}$, and the forecast error covariance matrix is initially estimated as the sampling covariance matrix of perturbed forecast states:

$$\mathbf{P}_i = \frac{1}{m-1}\sum_{j=1}^{m}\left(\mathbf{x}_i^{f(j)} - \mathbf{x}_i^f\right)\left(\mathbf{x}_i^{f(j)} - \mathbf{x}_i^f\right)^T. \tag{4}$$

Step 2. Analysis step.

The analysis state is estimated by minimizing the following cost function:

$$J(\mathbf{x}) = \left(\mathbf{x} - \mathbf{x}_i^f\right)^T \mathbf{P}_i^{-1}\left(\mathbf{x} - \mathbf{x}_i^f\right) + \left(\mathbf{y}_i^o - \mathbf{H}_i\mathbf{x}\right)^T \mathbf{R}_i^{-1}\left(\mathbf{y}_i^o - \mathbf{H}_i\mathbf{x}\right), \tag{5}$$

which has the analytic form

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{P}_i\mathbf{H}_i^T\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^T + \mathbf{R}_i\right)^{-1}\mathbf{d}_i, \tag{6}$$

where

1    $$\mathbf{d}_i = \mathbf{y}_i^o - \mathbf{H}_i \mathbf{x}_i^f \qquad (7)$$

2   is the innovation statistic (observation-minus-forecast residual in observation space).

3   To complete the ensemble forecast, the perturbed analysis states are calculated using

4   perturbed observations (Burgers et al. 1998):

5    $$\mathbf{x}_i^{a(j)} = \mathbf{x}_i^{f(j)} + \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} \left(\mathbf{H}_i \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1} \left(\mathbf{d}_i + \boldsymbol{\varepsilon}_i^{'(j)} \right), \qquad (8)$$

6   where $\boldsymbol{\varepsilon}_i^{'(j)}$ is a normally distributed random variable with mean zero and covariance

7   matrix $\mathbf{R}_i$. Here, $\left(\mathbf{H}_i \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1}$ can be easily calculated using the

8   Sherman-Morrison-Woodbury formula (Golub and Loan 1996; Liang et al. 2012;

9   Tippett et al. 2003). Finally, set $i = i + 1$, return to Step 1 for the model forecast at

10   the next time step and repeat until the model reaches the last time step N.

11

12   *2.2. Influence matrix and forecast error inflation*

13   The forecast error inflation procedure should be added to any ensemble-based

14   assimilation scheme to prevent the filter from diverging (Anderson and Anderson

15   1999; Constantinescu et al. 2007). Multiplicative inflation is one of the commonly

16   used inflation techniques, and it adjusts the initially estimated forecast error

17   covariance matrix $\mathbf{P}_i$ to $\lambda_i \mathbf{P}_i$ after estimating the inflation factors λ_i properly.

18   In this study, a new procedure for estimating multiplicative inflation factors λ_i

19   is proposed based on the following GCV function (Craven and Wahba 1979)

20    $$GCV_i(\lambda) = \frac{\frac{1}{p_i} \mathbf{d}_i^{\mathrm{T}} \mathbf{R}_i^{-1/2} \left(\mathbf{I}_{p_i} - \mathbf{A}_i(\lambda) \right)^2 \mathbf{R}_i^{-1/2} \mathbf{d}_i}{\left[\frac{1}{p_i} \mathrm{Tr} \left(\mathbf{I}_{p_i} - \mathbf{A}_i(\lambda) \right) \right]^2}, \qquad (9)$$

21   where $\mathbf{I}_{p_i}$ is the identity matrix with dimension $p_i \times p_i$; $\mathbf{R}_i^{-1/2}$ is the square root

matter of $\mathbf{R}_i$; and

$$\mathbf{A}_i(\lambda) = \mathbf{I}_{p_i} - \mathbf{R}_i^{1/2} \left(\mathbf{H}_i \lambda \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1} \mathbf{R}_i^{1/2} \qquad (10)$$

is the influence matrix (see Appendix for details).

The inflation factor λ_i is estimated by minimizing the GCV (Eq. (9)) as an objective function, and it is implemented between Steps 1 and 2 in Section 2.1. Then, the perturbed analysis states are modified to

$$\mathbf{x}_i^{\mathrm{a}(j)} = \mathbf{x}_i^{\mathrm{f}(j)} + \lambda_i \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} \left(\mathbf{H}_i \lambda_i \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1} \left(\mathbf{d}_i + \boldsymbol{\varepsilon}_i^{\prime(j)} \right). \qquad (11)$$

The flowchart of the EnKF equipped with the proposed forecast error inflation based on the GCV method is shown in Figure 1.

2.3. Analysis sensitivity

In the EnKF, the analysis state (Eq. (6)) is a weighted average of the observation and forecast. That is:

$$\mathbf{x}_i^{\mathrm{a}} = \mathbf{K}_i \mathbf{y}_i^{\mathrm{o}} + \left(\mathbf{I}_n - \mathbf{K}_i \mathbf{H}_i \right) \mathbf{x}_i^{\mathrm{f}} \qquad (12)$$

where $\mathbf{K}_i = \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} \left(\mathbf{H}_i \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1}$ is the Kalman gain matrix and $\mathbf{I}_n$ is the identity matrix with dimension $n \times n$. Then, the normalized analysis vector can be expressed as follows:

$$\tilde{\mathbf{y}}_i^{\mathrm{a}} = \mathbf{R}_i^{-1/2} \mathbf{H}_i \mathbf{K}_i \mathbf{R}_i^{1/2} \tilde{\mathbf{y}}_i^{\mathrm{o}} + \mathbf{R}_i^{-1/2} \left(\mathbf{I}_{p_i} - \mathbf{H}_i \mathbf{K}_i \right) \mathbf{R}_i^{1/2} \tilde{\mathbf{y}}_i^{\mathrm{f}} \qquad (13)$$

where $\tilde{\mathbf{y}}_i^{\mathrm{f}} = \mathbf{R}_i^{-1/2} \mathbf{H}_i \mathbf{x}_i^{\mathrm{f}}$ is the normalized projection of the forecast on the observation space. The sensitivities of the analysis to the observation and forecast are defined by Eq. (14) and (15), respectively:

$$\mathbf{S}_i^{\mathrm{o}} = \frac{\partial \tilde{\mathbf{y}}_i^{\mathrm{a}}}{\partial \tilde{\mathbf{y}}_i^{\mathrm{o}}} = \mathbf{R}_i^{1/2} \mathbf{K}_i^{\mathrm{T}} \mathbf{H}_i^{\mathrm{T}} \mathbf{R}_i^{-1/2}, \qquad (14)$$

1 $$\mathbf{S}_i^{\mathrm{f}} = \frac{\partial \tilde{\mathbf{y}}_i^{\mathrm{a}}}{\partial \tilde{\mathbf{y}}_i^{\mathrm{f}}} = \mathbf{R}_i^{1/2} \left(\mathbf{I}_{p_i} - \mathbf{K}_i^{\mathrm{T}} \mathbf{H}_i^{\mathrm{T}} \right) \mathbf{R}_i^{-1/2}, \qquad (15)$$

2 which satisfy $\mathbf{S}_i^{\mathrm{o}} + \mathbf{S}_i^{\mathrm{f}} = \mathbf{I}_{p_i}$.

3 The elements of the matrix $\mathbf{S}_i^{\mathrm{o}}$ reflect the sensitivity of the normalized analysis

4 state to the normalized observations; its diagonal elements are the analysis

5 self-sensitivities and the off-diagonal elements are the cross-sensitivities. On the

6 other hand, the elements of the matrix $\mathbf{S}_i^{\mathrm{f}}$ reflect the sensitivity of the normalized

7 analysis state to the normalized forecast state. The two quantities are complementary,

8 and the GCV function can be interpreted as minimizing the normalized forecast

9 sensitivity because the inflation scheme will increase the observation weight

10 appropriately.

11 In fact, the sensitivity matrix $\mathbf{S}_i^{\mathrm{o}}$ is equal to the influence matrix $\mathbf{A}_i$ (see

12 Appendix B for detailed proof), whose trace can be used to measure the "equivalent

13 number of parameters" or "degrees of freedom for the signal" (Gu 2002; Pena and

14 Yohai 1991). Similarly, the sensitivity matrix $\mathbf{S}_i^{\mathrm{o}}$ can be interpreted as a

15 measurement of the amount of information extracted from the observations (Ellison

16 et al. 2009). Trace diagnostics can be used to analyse the sensitivities to observations

17 or forecast vectors (Cardinali et al. 2004). The Global Average Influence (GAI) at the

18 i-th time step is defined as the globally averaged observation influence:

19 $$GAI = \frac{\mathrm{Tr}(\mathbf{S}_i^{\mathrm{o}})}{p_i}, \qquad (16)$$

20 where p_i is the total number of observations at the i-th time step.

21 In the conventional EnKF, the forecast error covariance matrix $\mathbf{P}_i$ is initially

estimated using a Monte Carlo method with short-term ensemble forecast states. However, because of the limited ensemble size and large model errors, the sampling covariance matrix of perturbed forecast states usually underestimate the true forecast error covariance matrix. This will cause the analysis to over rely on the forecast state and exclude useful information from the observations. This is captured by the fact that the GAI values are rather small for the conventional EnKF scheme. Adjusting the inflation of the forecast error covariance matrix alleviates this problem to some extent, as will be shown in the following simulations.

2.4 Forecast ensemble spread and analysis RMSE

The spread of the forecast ensemble at the i-th step is defined as follows:

$$\text{Spread} = \sqrt{\frac{1}{n(m-1)} \sum_{j=1}^{m} \left\| \mathbf{x}_{i,j}^{\text{f}} - \mathbf{x}_i^{\text{f}} \right\|^2} \ . \tag{17}$$

Roughly speaking, the forecast ensemble spread is usually underestimated for the conventional EnKF, which also dramatically decreases until the observations ultimately have an irrelevant impact on the analysis states. The inflation technique can effectively compensate for the underestimation of the forecast ensemble spread, and thereby can improve the assimilation results.

In the following experiments, the "true" state $\mathbf{x}_i^{\text{t}}$ is non-dimensional and can be obtained by a numerical solution of partial differential equations. In this case, the distance of the analysis state to the true state can be defined as the analysis root-mean-square error (RMSE), which is used to evaluate the accuracy of the assimilation results. The RMSE at the i-th time step is defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\mathrm{x}_{i,k}^{\mathrm{a}} - \mathrm{x}_{i,k}^{\mathrm{t}}\right)^{2}}\,. \tag{18}$$

where $\mathrm{x}_{i,k}^{\mathrm{a}}$ and $\mathrm{x}_{i,k}^{\mathrm{t}}$ are the k-th components of the analysis state and true state at

the i-th time step. In principle, a smaller RMSE indicates a better performance of the

assimilation scheme.

3. Numerical Experiments

The proposed data assimilation scheme was tested using the Lorenz-96 model

(Lorenz 1996) with model errors and a linear observation system as a test bed. The

performances of the assimilation schemes described in Section 2 were evaluated via

the following experiments.

3.1. Dynamical forecast model and observation systems

The Lorenz-96 model (Lorenz 1996) is a quadratic nonlinear dynamical system

that has properties relevant to realistic forecast problems and is governed by the

equation:

$$\frac{dX_{k}}{dt} = (X_{k+1} - X_{k-2})X_{k-1} - X_{k} + F\,, \tag{19}$$

where $k = 1,2,\cdots,40$. The cyclic boundary conditions $X_{-1} = X_{K-1}$, $X_{0} = X_{K}$, and

$X_{K+1} = X_{1}$ were applied to ensure that Eq. (19) is well defined for all values of k.

The Lorenz-96 model is "atmosphere-like" because the three terms on the right-hand

1   side of Eq. (19) are analogous to a nonlinear advection-like term, a damping term,

2   and an external forcing term, respectively. The model can be considered

3   representative of an atmospheric quantity (e.g., zonal wind speed) distributed on a

4   latitude circle. Therefore, the Lorenz-96 model has been widely used as a test bed to

5   evaluate the performance of assimilation schemes in many studies (Wu et al. 2013).

6   The true state is derived by a fourth-order Runge-Kutta time integration scheme

7   (Butcher 2003). The time step for generating the numerical solution was set at 0.05

8   non-dimensional units, which is roughly equivalent to 6 hours in real time assuming

9   that the characteristic time-scale of the dissipation in the atmosphere is 5 days

10   (Lorenz 1996). The forcing term was set as $F = 8$, so that the leading Lyapunov

11   exponent implies an error-doubling time of approximately 8 time steps and the fractal

12   dimension of the attractor was 27.1 (Lorenz and Emanuel 1998). The initial value was

13   chosen to be $X_k = F$ when $k \neq 20$ and $X_{20} = 1.001F$.

14   In this study, the synthetic observations were assumed to be generated by

15   adding random noises that were multivariate-normally distributed with mean zero and

16   covariance matrix $\mathbf{R}_i$ to the true states. The frequency was every 4 time steps, which

17   can be used to mimic daily observations in practical problems, such as satellite data.

18   The observation errors were assumed to be spatially correlated, which is common in

19   applications involving remote sensing and radiance data. The variance of the

20   observation at each grid point was set to $\sigma_o^2 = 1$, and the covariance of the

21   observations between the j-th and k-th grid points was as follows:

22
$$\mathbf{R}_i(j,k) = \sigma_o^2 \times 0.5^{\min\{|j-k|,40-|j-k|\}}. \tag{20}$$

3.2. Assimilation scheme comparison

Because model errors are inevitable in practical dynamical forecast models, it is reasonable to add model errors to the Lorenz-96 model in the assimilation process. The Lorenz-96 model is a forced dissipative model with a parameter F that controls the strength of the forcing. Modifying the forcing strength F changes the model forecast states considerably. For values of F that are larger than 3, the system is chaotic (Lorenz and Emanuel 1998). To simulate model errors, the forcing term for the forecast was set to 7, while using $F=8$ to generate the "true" state. The initially selected ensemble size was 30.

The Lorenz-96 model was run for 2000 time steps, which is equivalent to approximately 500 days in realistic problems. The synthetic observations were assimilated at every grid point and every 4 time steps using the conventional EnKF, the constant inflated EnKF and the improved EnKF schemes for comparisons. The time series of estimated inflation factors are shown in Figure 2. It can be seen that, the estimated inflation factors vary between 1 and 6 in most instances, although the values smaller than 1 are estimated in several assimilation time steps. The median of the estimated inflation factors was 1.88, which was used as the inflation factor in the constant inflated EnKF scheme. Since the median is a robust and highly efficient statistic of the central tendency, this can ensure a relative fair comparison between the constant inflated EnKF and the improved EnKF schemes.

The forecast ensemble spread of the conventional EnKF, constant inflated

EnKF and improved EnKF are plotted in Figure 3. For the conventional EnKF, because the forecast states usually shrink together, the forecast ensemble spread was quite small and had a mean value of 0.36. The mean spread value of the improved EnKF was 3.32, which was larger than that of the constant inflated EnKF (3.25). These findings illustrate that the underestimation of forecast ensemble spread can be effectively compensated for by the two EnKF schemes with forecast error inflation and that the improved EnKF is more effective than the constant inflated EnKF.

To evaluate the analysis sensitivity, the GAI statistics (Eq. (16)) were calculated, and the results are plotted in Figure 4. The GAI value increases from 10% for the conventional EnKF to 30% for the improved EnKF, indicating that the latter relies more on the observations. This finding is important because the observations can play a significant role in combining the results with the model forecasts to generate the analysis state. In addition to small fluctuations, the mean GAI value of the constant inflated EnKF was 27.80%, which was smaller than that of the improved EnKF.

To evaluate the analysis estimate accuracy, the analysis RMSE (Eq. (18)) and the corresponding values of the GCV functions (Eq. (9)) were calculated and plotted in Figures 5 and 6, respectively. The results illustrate that the analysis RMSE and the values of the GCV functions decrease sharply for the two EnKF with forecast error inflation schemes. However, the GCV function and the RMSE values of the improved EnKF were about 15% smaller than those of the constant inflated EnKF, indicating that the online estimate method performs better than the simple multiplicative inflation techniques with a constant value. The correlation coefficient of the analysis

RMSE and the value of the GCV function at the assimilation time step were approximately 0.76, which indicates that the GCV function is a good criterion to estimate the inflation factor.

The ensemble analysis state members of the conventional EnKF, constant inflated EnKF and improved EnKF are shown in Figure 7, and the results indicate the uncertainty of the analysis state to some extent. The true trajectory obtained by the numerical solution is also plotted. It illustrates that a larger difference occurred between the true trajectory and the ensemble analysis state members for the conventional EnKF than for the improved EnKF and constant inflated EnKF. In addition, the analysis state was more consistent with the true trajectory for the improved EnKF than that for the constant inflated EnKF. Therefore, the GCV inflation can lead to a more accurate analysis state than the simple constant inflation.

The time-mean values of the forecast ensemble spread, the GAI statistics, the GCV functions and the analysis RMSE over 2000 time steps are listed in Table 1. These results illustrate that the forecast error inflation technique using the GCV function performs better than the constant inflated EnKF, which can indeed increase the analysis sensitivity to the observations and reduce the analysis RMSE.

3.3 Influence of ensemble size and observation number

Intuitively, for any ensemble-based assimilation scheme, a large ensemble size will lead to small analysis errors; however, the computational costs are high for practical problems. The ensemble size in the practical land surface assimilation

1   problem is usually several tens of members (Kirchgessner et al. 2014). The

2   preferences of the proposed inflation method and the constant inflation method with

3   respect to different ensemble sizes (10, 30 and 50) were evaluated, and the results are

4   listed in Table 1. It shows that for each scheme, using a 10-member ensemble

5   produced a threefold increase in the analysis RMSE, while using a 50-member

6   ensemble reduced the analysis RMSE by 20% relative to the analysis RMSE obtained

7   using a 30-member ensemble. The forecast ensemble spread increased slightly from a

8   10-member ensemble to a 50-member ensemble. The GAI and GCV function values

9   changed sharply from a 10-member ensemble to a 30-member ensemble, and they

10   became relatively stable from a 30-member ensemble to a 50-member ensemble.

11   Ensembles less than 10 were unstable, and no significant changes occurred for

12   ensembles greater than 50. Considering the computational costs for practical

13   problems, a 30-member ensemble may be necessary for Lorenz-96 model to estimate

14   statistically robust results. In the realistic problem, a system in which the errors grow

15   in multiple directions will need more ensembles to produce statistically robust results.

16   To evaluate the preferences of the inflation method with respect to different

17   numbers of observations, synthetic observations were generated at every other grid

18   point and for every 4 time steps. Hence, a total of 20 observations were performed at

19   each observation step in this case. The assimilation results with ensemble sizes of 10,

20   30 and 50 are listed in Table 2, which shows that the GAI values were larger than

21   those with 40-observations in all assimilation schemes. This finding may be related to

22   the relatively small denominator of the GAI statistic (Eq. (16)) in the 20-observation

experiments. The forecast ensemble spread does not change much but the GCV function and the RMSE values increase greatly in the 20-observation experiments with respect to those in the 40-observation experiments, which illustrates that more observations will lead to less analysis error.

4. Discussions

4.1 Performance of the GCV inflation

Accurate estimates of the forecast error covariance matrix are crucial to the success of any data assimilation scheme. In the conventional EnKF assimilation scheme, the forecast error covariance matrix is estimated as the sampling covariance matrix of the ensemble forecast states. However, limited ensemble size and large model errors often cause the matrix to be underestimated, which produces an analysis state that over relies on the forecast and excludes observations. This can eventually cause the filter to diverge. Therefore, the forecast error inflation with proper inflation factors is increasingly important.

The use of multiplicative covariance inflation techniques can mitigate this problem to some extent. Several methods have been proposed in the literature, and each has different assumptions. For instance, the moment approach can be easily conducted based on the moment estimation of the innovation statistic. The maximum likelihood approach can obtain a more accurate inflation factor than the moment

1   approach, but requires computing high dimensional matrix determinants. The

2   Bayesian approach assumes a prior distribution for the inflation factor but is limited

3   to spatially independent observational errors. In this study, the inflation factor was

4   estimated based on cross-validation and the analysis sensitivity was detected. The

5   estimated inflation factor by minimizing the GCV function is not affected by the

6   observation unit and can optimize the analysis sensitivity to the observation.

7   In fact, the GCV method can evaluate and compare learning algorithms and

8   represents a widely used statistical method. It can be applied in inverse problems in

9   such fields as meteorological data assimilation (Wahba et al. 1995). Specifically,

10   GCV provides a well-characterized method, which can select a regularization

11   parameter by minimizing the predictive data errors with rotation-invariant in a least

12   squares solution (MacCarthy et al. 2011). In data assimilation research fields,

13   observation data such as in-situ observation and remote sensing data are usually from

14   different sources. GCV is particularly useful for choosing relative parameters that

15   reflect not only measurement accuracies from different sources but also model

16   capability (Krakauer et al. 2004). Apparently, GCV method requires calculating the

17   trace of a large matrix, which may be commonly computationally prohibitive for

18   large inverse problems (MacCarthy et al. 2011).

19   In this study, the GCV concept was adopted for the inflation factor estimation

20   in the improved EnKF assimilation scheme and was validated with the Lorenz-96

21   model. The assimilation results showed that inflating the conventional EnKF using

22   the factor estimated by minimizing the GCV function can indeed reduce the analysis

1 RMSE. Therefore, the GCV function can accurately quantify the goodness of fit of

2 the error covariance matrix. The values of the GCV function obviously decreased in

3 the proposed approach compared the conventional EnKF and constant inflated EnKF

4 schemes. The analysis RMSE of the proposed approach was also much smaller than

5 those of the conventional EnKF and constant inflated EnKF schemes, which suggests

6 that the GCV criterion works well for estimating the inflation factor.

7 　　The analysis sensitivities in the proposed approach and in the conventional

8 EnKF scheme were also investigated in this study. The time-averaged GAI statistic

9 increases from about 10% in the conventional EnKF scheme to about 30% using the

10 proposed inflation method. This illustrates that the inflation mitigates the problem of

11 the analysis depending excessively on the forecast and excluding the observations.

12 The relationship of the analysis state to the forecast state and the observations are

13 more reasonable.

14

15 *4.2 Computational cost*

16 　　The highest computational cost when minimizing the GCV function is related

17 to calculating the influence matrix $\mathbf{A}_i(\lambda)$. Since the matrix multiplication is

18 commutative for the trace, the GCV function can be easily re-expressed as follows:

19
$$GCV_i(\lambda) = \frac{p_i \mathbf{d}_i^{\mathrm{T}} \left(\mathbf{H}_i \lambda \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1} \mathbf{R}_i \left(\mathbf{H}_i \lambda \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1} \mathbf{d}_i}{\left[\mathrm{Tr} \left(\left(\mathbf{H}_i \lambda \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1} \mathbf{R}_i \right) \right]^2}. \quad (21)$$

20 Because both the numerator and denominator of the GCV function are scalars, the

21 inverse matrix is needed only in $\left(\mathbf{H}_i \lambda \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1}$, which can be effectively

calculated using the Sherman-Morrison-Woodbury formula. Furthermore, the inverse matrix calculation and the multiplication process are also indispensable for the conventional EnKF (Eq. (6)). Essentially, no additional computational burden is associated with the improved EnKF for the inverse matrix. Therefore, the total computational costs of the improved EnKF are feasible.

For the Lorenz-96 experiments in this study, the conventional EnKF, constant inflated EnKF and proposed improved EnKF assimilation schemes were conducted using R language on a computer with Intel Core i5 CPU and 8 GB RAM. The running times with different observation numbers and ensemble sizes were listed in Tables 1 and 2. It shows that for each assimilation scheme, the computational cost increases as the ensemble size grows. For the fixed observation number and ensemble size, the conventional EnKF, which does not involve the forecast error inflation, has the least running time but at a cost of losing assimilation accuracy. The proposed EnKF scheme is about 15% smaller in analysis RMSE, but only about 5% longer in running time than the constant inflated EnKF scheme. For the operational meteorological/ocean models, the most computational cost is in the ensemble model integrations (Ravazzani et al. 2016). Therefore, the proposed EnKF scheme does not significantly increase computational cost.

4.3 Notes

It is worth noting that the inflation factor is assumed to be constant in space in this study, which may be not the case in realistic assimilation problems. Forcing all

components of the state vector to use the same inflation factor could systematically

overinflate the ensemble variances in sparsely observed areas, especially when the

observations are unevenly distributed. In the presence of sparse observations, the

state that is not observed can be improved only by the physical mechanism of the

forecast model, although this improvement is limited. Therefore, a multiplicative

inflation may not be sufficiently effective to enhance the assimilation accuracy. In

this case, the additive inflation and the localization technique can be applied to

further improve the assimilation quality in the presence of sparse observations

(Miyoshi and Kunii 2011; Yang et al. 2015).

5. Conclusions

In this study, the approach for using GCV as a metric to estimate the covariance

inflation factor was proposed. In the case studies conducted in Section 3, the

observations were relatively evenly distributed and the assimilation accuracy could

indeed be improved by the forecast error inflation technique. These findings provide

insights on the methodology and validation of the Lorenz-96 model and illustrate the

feasibility of our approach. In the near future, methods of modifying the adaptive

procedure to suit the system with unevenly distributed observations and applying to

more sophisticated dynamic and observation systems will be investigated.

1   **Appendix A**

2   From Eq. (2), the normalized observation equation can be defined as follows:

3    $$\tilde{\mathbf{y}}_i^o = \mathbf{R}_i^{-1/2}\mathbf{H}_i\mathbf{x}_i^t + \tilde{\boldsymbol{\varepsilon}}_i \,, \tag{A1}$$

4   where $\tilde{\mathbf{y}}_i^o = \mathbf{R}_i^{-1/2}\mathbf{y}_i^o$ is the normalized observation vector and $\tilde{\boldsymbol{\varepsilon}}_i \sim N(\mathbf{0},\mathbf{I})$; $\mathbf{I}_{p_i}$ is

5   the identity matrix with the dimensions $p_i \times p_i$. Similarly, the normalized analysis

6   vector is $\tilde{\mathbf{y}}_i^a = \mathbf{R}_i^{-1/2}\mathbf{H}_i\mathbf{x}_i^a$ and the influence matrix $\mathbf{A}_i$ relates the normalized

7   observation vector to the normalized analysis vector, thereby ignoring the normalized

8   forecast state in the observation space (Gu 2002):

9    $$\tilde{\mathbf{y}}_i^a - \mathbf{R}_i^{-1/2}\mathbf{H}_i\mathbf{x}_i^f = \mathbf{A}_i\left(\tilde{\mathbf{y}}_i^o - \mathbf{R}_i^{-1/2}\mathbf{H}_i\mathbf{x}_i^f\right). \tag{A2}$$

10   Because the analysis state $\mathbf{x}_i^a$ is given by Eq. (5), the influence matrix $\mathbf{A}_i$ can be

11   verified as follows:

12   $$\mathbf{A}_i = \mathbf{I}_{p_i} - \mathbf{R}_i^{1/2}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^T + \mathbf{R}_i\right)^{-1}\mathbf{R}_i^{1/2}. \tag{A3}$$

13   If the initial forecast error covariance matrix is inflated as described in Section 2.2,

14   then the influence matrix is treated as the following function of λ

15   $$\mathbf{A}_i(\lambda) = \mathbf{I}_{p_i} - \mathbf{R}_i^{1/2}\left(\mathbf{H}_i\lambda\mathbf{P}_i\mathbf{H}_i^T + \mathbf{R}_i\right)^{-1}\mathbf{R}_i^{1/2}, \tag{A4}$$

16   The principle of CV is to minimize the estimated error at the observation grid

17   point. Lacking an independent validation data set, a common alternative strategy is to

18   minimize the squared distance between the normalized observation value and the

19   analysis value while not using the observation on the same grid point, which is the

20   following objective function:

21   $$V_i(\lambda) = \frac{1}{p_i}\sum_{k=1}^{p_i}\left(\tilde{\mathbf{y}}_{i,k}^o - \left(\mathbf{R}_i^{-1/2}\mathbf{H}_i\mathbf{x}_i^{a[k]}\right)_k\right)^2, \tag{A5}$$

22   where $\mathbf{x}_i^{a[k]}$ is the minima of the following "delete-one" objective function:

1

$$\left(\mathbf{x}-\mathbf{x}_i^{\mathrm{f}}\right)^{\mathrm{T}}(\lambda\mathbf{P}_i)^{-1}\left(\mathbf{x}-\mathbf{x}_i^{\mathrm{f}}\right)+\left(\mathbf{y}_i^{\mathrm{o}}-\mathbf{H}_i\mathbf{x}\right)_{-k}^{\mathrm{T}}\mathbf{R}_{i,-k}^{-1/2}\left(\mathbf{y}_i^{\mathrm{o}}-\mathbf{H}_i\mathbf{x}\right)_{-k}. \tag{A6}$$

2 The subscript $-k$ indicates a vector (matrix) with its k-th element (k-th row and

3 column) deleted. Instead of minimizing Eq. (A6) p_i times, the objective function

4 (Eq. (A5)) has another more simple expression (Gu 2002):

5
$$V_i\left(\lambda\right)=\frac{1}{p_i}\sum_{k=1}^{p_i}\frac{\left(\tilde{\mathbf{y}}_{i,k}^{\mathrm{o}}-\left(\mathbf{R}_i^{-1/2}\mathbf{H}_i\mathbf{x}_i^{\mathrm{a}}\right)_k\right)^2}{\left(1-a_{k,k}\right)^2}, \tag{A7}$$

6 where $a_{k,k}$ is the element at the site pair (k, k) of the influence matrix $\mathbf{A}_i(\lambda)$. Then,

7 $a_{k,k}$ is substituted with the average $\dfrac{1}{p_i}\displaystyle\sum_{k=1}^{p_i}a_{k,k}=\dfrac{1}{p_i}\mathrm{Tr}(\mathbf{A}_i(\lambda))$ and the constant is

8 ignored to obtain the following GCV statistic (Gu 2002):

9
$$GCV_i(\lambda)=\frac{\dfrac{1}{p_i}\mathbf{d}_i^{\mathrm{T}}\mathbf{R}_i^{-1/2}\left(\mathbf{I}_{p_i}-\mathbf{A}_i(\lambda)\right)^2\mathbf{R}_i^{-1/2}\mathbf{d}_i}{\left[\dfrac{1}{p_i}\mathrm{Tr}\left(\mathbf{I}_{p_i}-\mathbf{A}_i(\lambda)\right)\right]^2}. \tag{A8}$$

10

11 **Appendix B**

12 The sensitivities of the analysis to the observation are defined as follows:

13
$$\mathbf{S}_i^{\mathrm{o}}=\frac{\partial\tilde{\mathbf{y}}_i^{\mathrm{a}}}{\partial\tilde{\mathbf{y}}_i^{\mathrm{o}}}=\mathbf{R}_i^{1/2}\mathbf{K}_i^{\mathrm{T}}\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-1/2}, \tag{B1}$$

14 Substitute the Kalman gain matrix $\mathbf{K}_i=\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}+\mathbf{R}_i\right)^{-1}$ into $\mathbf{S}_i^{\mathrm{o}}$, then:

15 $\quad\mathbf{S}_i^{\mathrm{o}}=\mathbf{R}_i^{1/2}\mathbf{K}_i^{\mathrm{T}}\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-1/2}$

16 $\quad\quad=\mathbf{R}_i^{1/2}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}+\mathbf{R}_i\right)^{-1}\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-1/2}$

17 $\quad\quad=\mathbf{R}_i^{1/2}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}+\mathbf{R}_i\right)^{-1}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}+\mathbf{R}_i-\mathbf{R}_i\right)\mathbf{R}_i^{-1/2}$

18 $\quad\quad=\mathbf{R}_i^{1/2}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}+\mathbf{R}_i\right)^{-1}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}+\mathbf{R}_i\right)\mathbf{R}_i^{-1/2}-\mathbf{R}_i^{1/2}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}+\mathbf{R}_i\right)^{-1}\mathbf{R}_i\mathbf{R}_i^{-1/2}$

1 $$= \mathbf{I}_{p_i} - \mathbf{R}_i^{1/2} \left(\mathbf{H}_i \lambda \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1} \mathbf{R}_i^{1/2}$$

2 $$= \mathbf{A}_i \qquad\qquad\qquad\qquad\qquad (\mathrm{B2})$$

3 Therefore, the sensitivity matrix $\mathbf{S}_i^o$ is equal to the influence matrix $\mathbf{A}_i$.

4

11

1 **References**

2 Table 1. Time-mean values of the forecast ensemble spread, GAI statistics, GCV

3 functions and analysis RMSE over 2000 time steps, as well as the running times

4 (second) for different assimilation schemes. The observation number is 40 and the

5 ensemble size is selected as 10, 30 and 50, respectively.

6

Scheme	Ensemble Size	Spread	GAI	GCV	RMSE	Running Time
Conventional EnKF	10	0.23	4.56%	36.38	4.50	70.73
	30	0.36	10.78%	31.14	4.01	215.92
	50	0.41	13.58%	25.21	3.52	346.69
Constant inflated EnKF	10	3.15	4.78%	35.91	4.38	77.41
	30	3.25	27.48%	5.56	1.41	238.25
	50	3.27	19.67%	5.03	1.14	384.63
Improved EnKF	10	3.26	5.24%	35.56	3.74	81.31
	30	3.32	29.21%	3.29	1.10	251.06
	50	3.45	35.63%	2.30	0.88	405.68

7

1   Table 2. Same as in Table 1 but for 20 observations.

2

Scheme	Ensemble Size	Spread	GAI	GCV	RMSE	Running Time
Conventional EnKF	10	0.41	10.77%	33.64	4.85	67.75
	30	0.59	20.92%	22.89	4.10	181.27
	50	0.68	26.41%	14.97	3.29	295.92
Constant inflated EnKF	10	3.03	11.73%	33.39	4.64	71.22
	30	3.18	30.07%	17.12	3.92	203.64
	50	3.27	39.51%	12.74	3.37	322.29
Improved EnKF	10	3.33	13.25%	32.17	4.39	74.84
	30	3.36	35.09%	14.99	3.46	213.81
	50	3.48	41.28%	5.19	2.86	339.41

3

Figure captions

Figure 1. Flowchart of the proposed assimilation scheme.

Figure 2. Time series of the estimated inflation factors by minimizing the GCV function. The median of the estimated inflation factors is 1.88.

Figure 3. Forecast ensemble spread of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88.
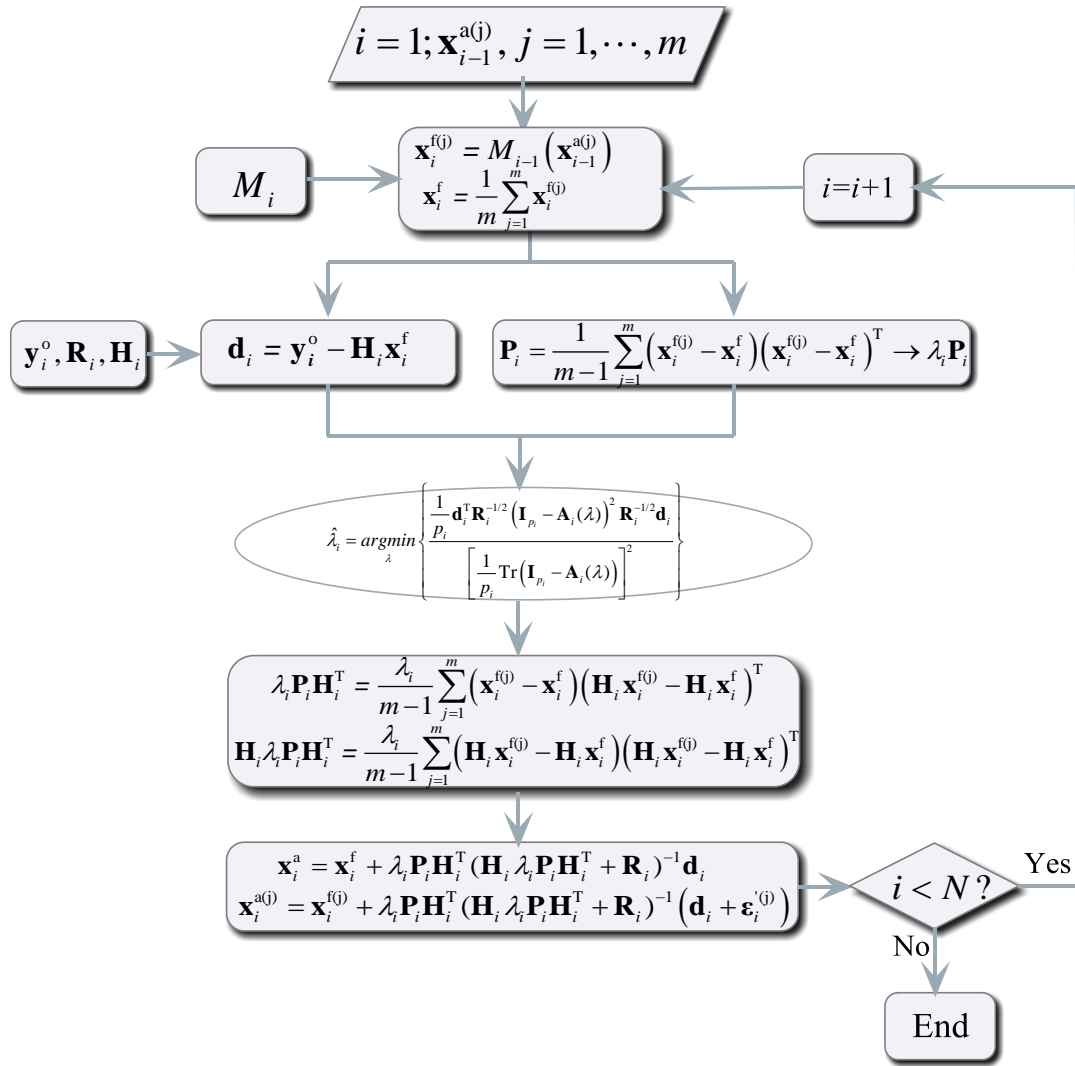
Figure 4. GAI statistics of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88.

Figure 5. Analysis RMSE of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88.

Figure 6. GCV function values of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88.

Figure 7. Ensemble analysis state members of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the
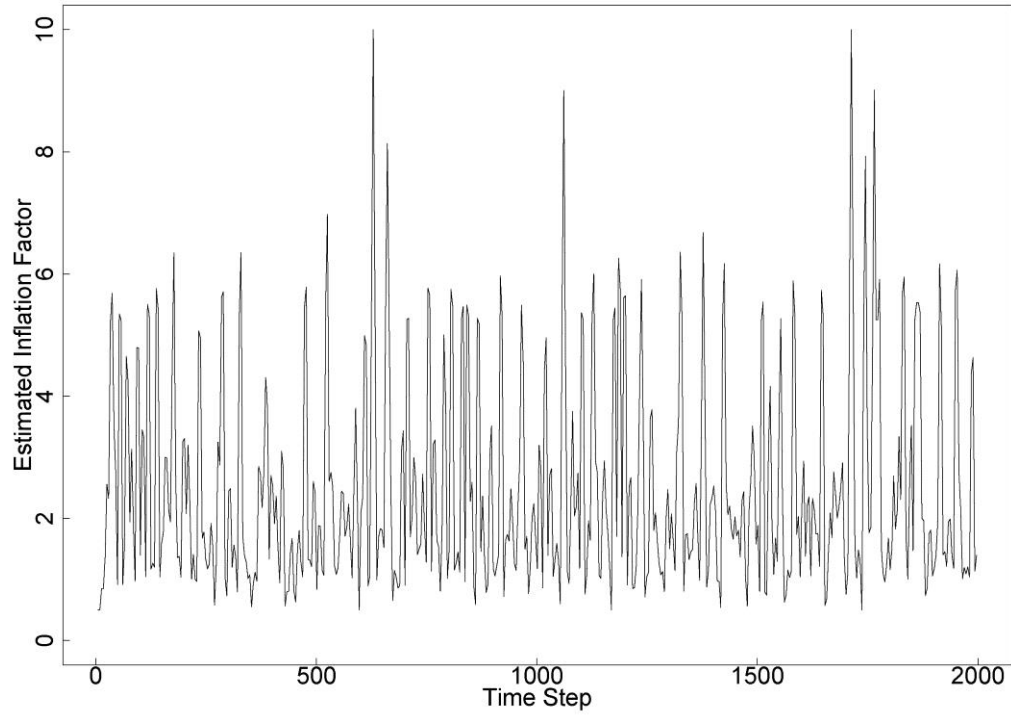
1   Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant

2   multiplicative inflation factor is set as 1.88. The green line refers to the true trajectory

3   obtained by the numerical solution.

4

Figure 1. Flowchart of the proposed assimilation scheme.

$$i = 1; \mathbf{x}_{i-1}^{a(j)}, \ j = 1, \cdots, m$$

$$\mathbf{x}_i^{f(j)} = M_{i-1}\left(\mathbf{x}_{i-1}^{a(j)}\right)$$
$$\mathbf{x}_i^{f} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_i^{f(j)}$$

M_i

$i = i + 1$

$\mathbf{y}_i^{o}, \mathbf{R}_i, \mathbf{H}_i$

$$\mathbf{d}_i = \mathbf{y}_i^{o} - \mathbf{H}_i\mathbf{x}_i^{f}$$

$$\mathbf{P}_i = \frac{1}{m-1}\sum_{j=1}^{m}\left(\mathbf{x}_i^{f(j)} - \mathbf{x}_i^{f}\right)\left(\mathbf{x}_i^{f(j)} - \mathbf{x}_i^{f}\right)^{T} \to \lambda_i \mathbf{P}_i$$

$$\hat{\lambda}_i = \underset{\lambda}{argmin}\left\{\frac{\dfrac{1}{p_i}\mathbf{d}_i^{T}\mathbf{R}_i^{-1/2}\left(\mathbf{I}_{p_i} - \mathbf{A}_i(\lambda)\right)^2 \mathbf{R}_i^{-1/2}\mathbf{d}_i}{\left[\dfrac{1}{p_i}\mathrm{Tr}\left(\mathbf{I}_{p_i} - \mathbf{A}_i(\lambda)\right)\right]^2}\right\}$$

$$\lambda_i \mathbf{P}_i \mathbf{H}_i^{T} = \frac{\lambda_i}{m-1}\sum_{j=1}^{m}\left(\mathbf{x}_i^{f(j)} - \mathbf{x}_i^{f}\right)\left(\mathbf{H}_i\mathbf{x}_i^{f(j)} - \mathbf{H}_i\mathbf{x}_i^{f}\right)^{T}$$
$$\mathbf{H}_i \lambda_i \mathbf{P}_i \mathbf{H}_i^{T} = \frac{\lambda_i}{m-1}\sum_{j=1}^{m}\left(\mathbf{H}_i\mathbf{x}_i^{f(j)} - \mathbf{H}_i\mathbf{x}_i^{f}\right)\left(\mathbf{H}_i\mathbf{x}_i^{f(j)} - \mathbf{H}_i\mathbf{x}_i^{f}\right)^{T}$$

$$\mathbf{x}_i^{a} = \mathbf{x}_i^{f} + \lambda_i \mathbf{P}_i \mathbf{H}_i^{T}(\mathbf{H}_i \lambda_i \mathbf{P}_i \mathbf{H}_i^{T} + \mathbf{R}_i)^{-1}\mathbf{d}_i$$
$$\mathbf{x}_i^{a(j)} = \mathbf{x}_i^{f(j)} + \lambda_i \mathbf{P}_i \mathbf{H}_i^{T}(\mathbf{H}_i \lambda_i \mathbf{P}_i \mathbf{H}_i^{T} + \mathbf{R}_i)^{-1}\left(\mathbf{d}_i + \boldsymbol{\varepsilon}_i^{'(j)}\right)$$

$i < N$?

Yes

No

End

1

2   Figure 1. Flowchart of the proposed assimilation scheme.
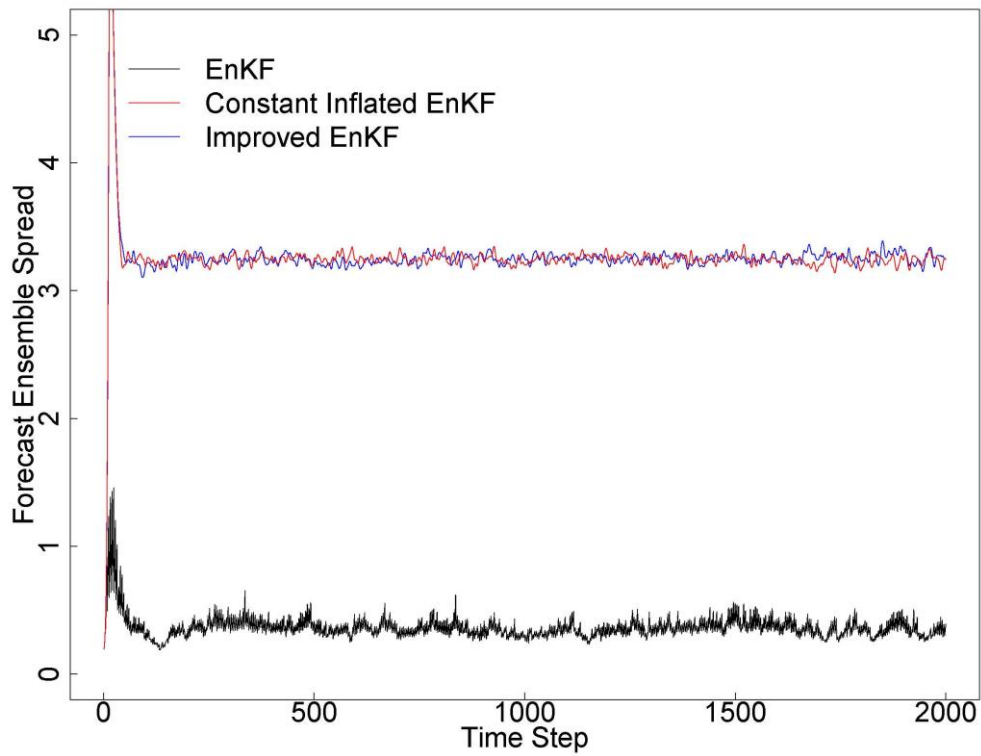
3

1

2   Figure 2. Time series of the estimated inflation factors by minimizing the GCV

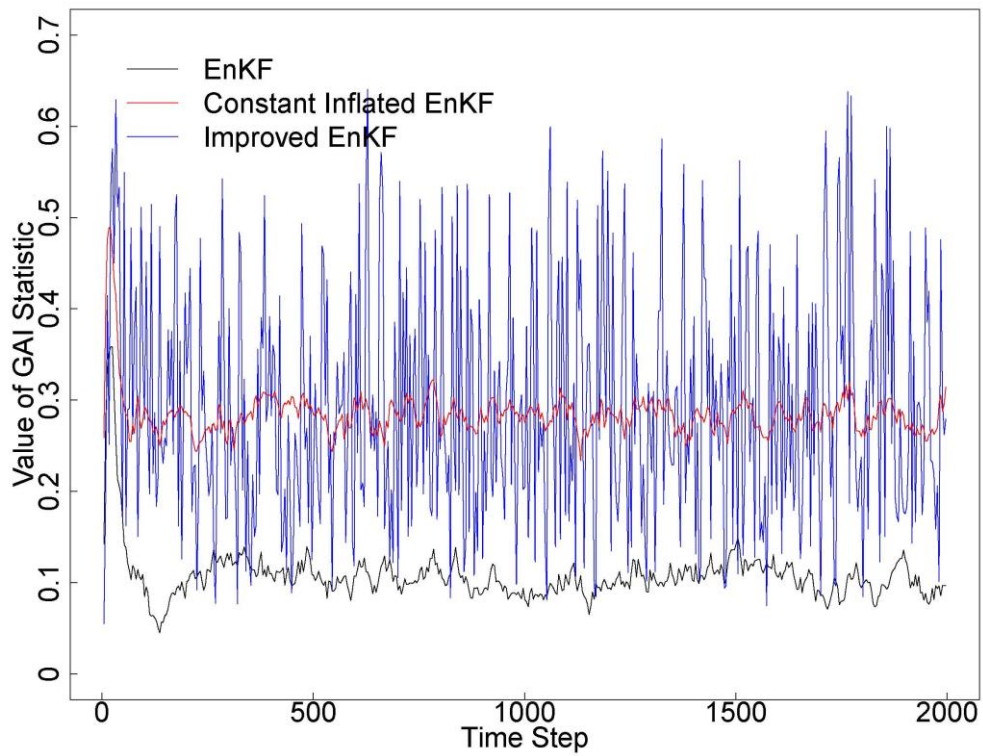3   function. The median of the estimated inflation factors is 1.88.

4

1

2 Figure 3. Forecast ensemble spread of the conventional EnKF (black line), the

3 constant inflated EnKF (red line) and the improved EnKF (blue line) for the

4 Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant
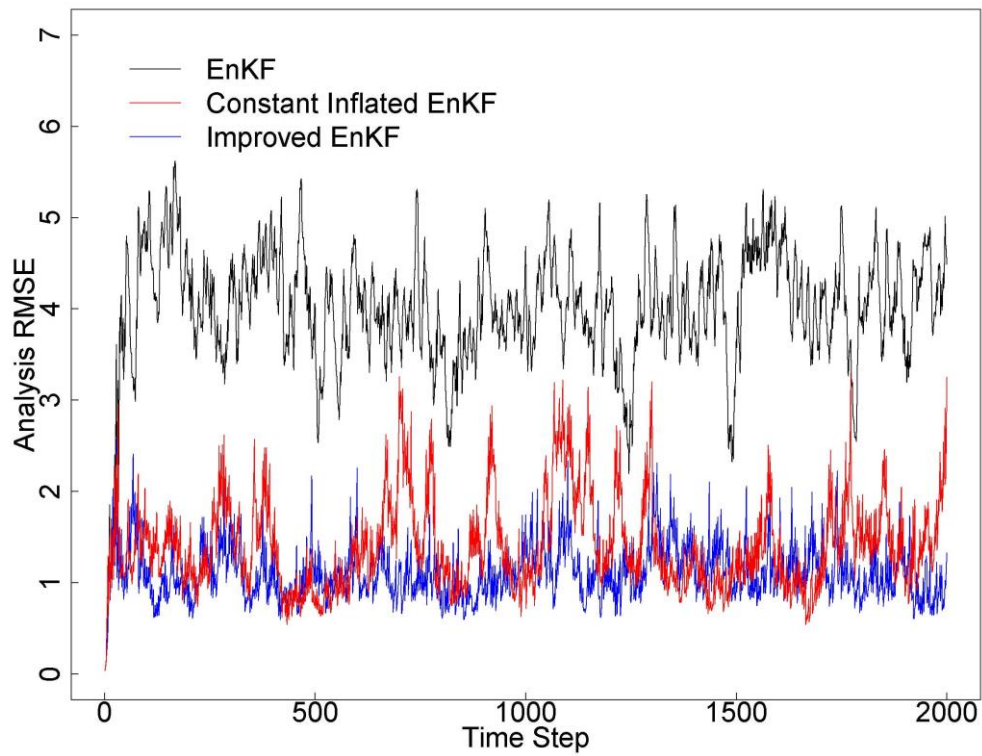
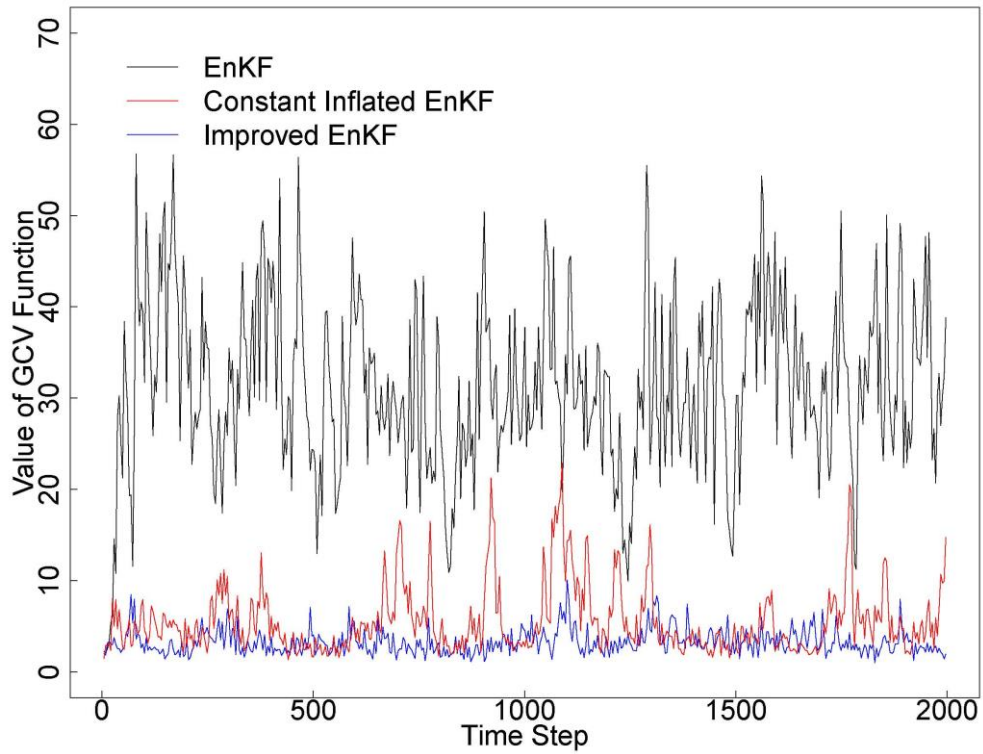5 multiplicative inflation factor is set as 1.88.

6

1

2 Figure 4. GAI statistics of the conventional EnKF (black line), the constant inflated

3 EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment

4 with 40-observation and 30-ensemble member. The constant multiplicative inflation

5 factor is set as 1.88.

6

34

1

Figure 5. Analysis RMSE of the conventional EnKF (black line), the constant inflated

EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment

with 40-observation and 30-ensemble member. The constant multiplicative inflation
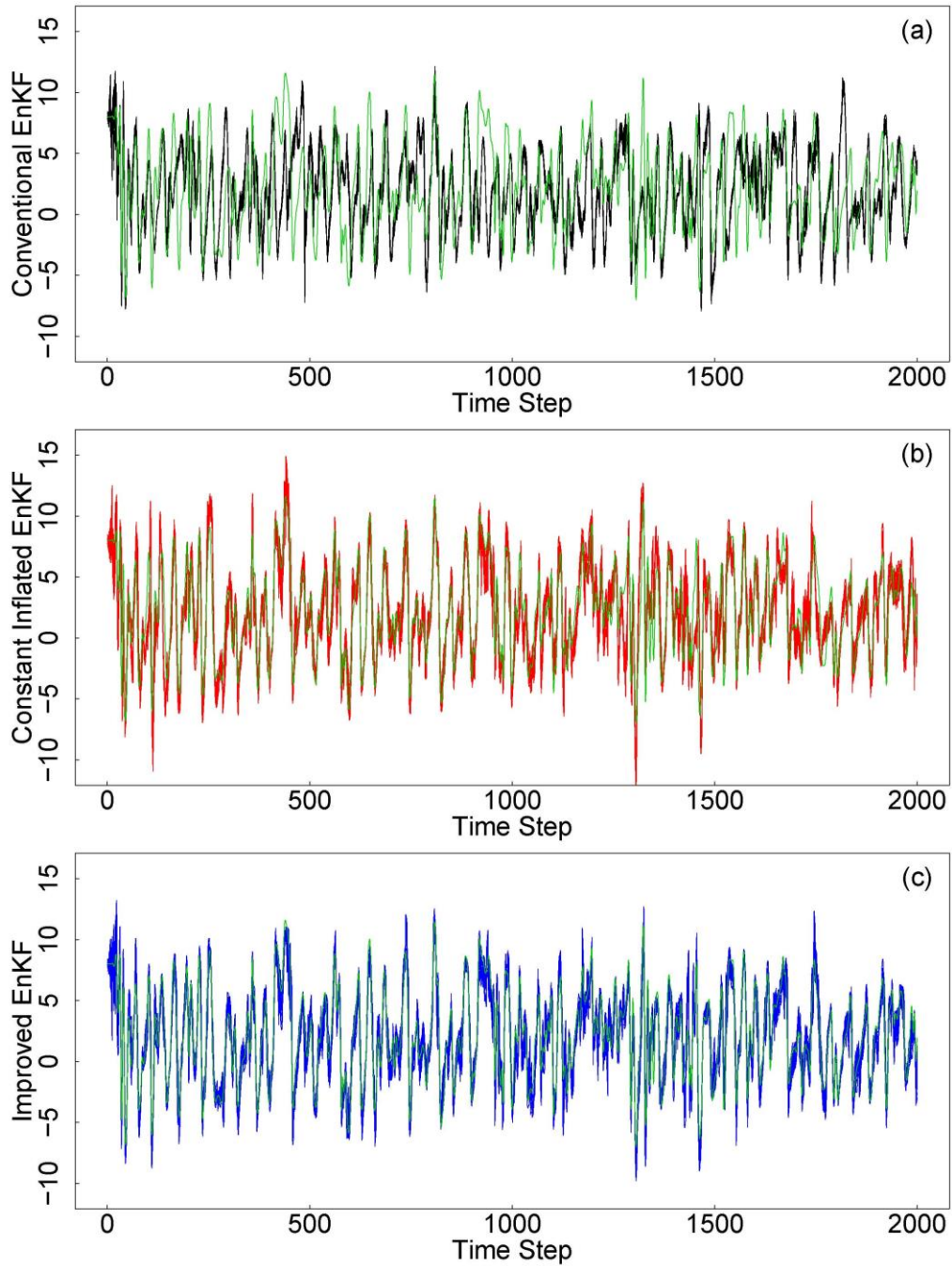
factor is set as 1.88.

6

1

2 Figure 6. GCV function values of the conventional EnKF (black line), the constant

3 inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96

4 experiment with 40-observation and 30-ensemble member. The constant

5 multiplicative inflation factor is set as 1.88.

6

Figure 7. Ensemble analysis state members of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88. The green line refers to the true trajectory obtained by the numerical solution.

References

Allen, D. M., 1974: The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16,** 125-127.

Anderson, J. L., 2007: An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus*, **59A,** 210-224.

Anderson, J. L., 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus*, **61A,** 72-83.

Anderson, J. L., and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear fltering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, **127,** 2741-2758.

Burgers, G., P. J. Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble kalman filter. *Monthly Weather Review*, **126,** 1719-1724.

Butcher, J. C., 2003: *Numerical methods for ordinary differential equations.* JohnWiley & Sons, 425 pp.

Cardinali, C., S. Pezzulli, and E. Andersson, 2004: Influence‐matrix diagnostic of a data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **130,** 2767-2786.

Constantinescu, E. M., A. Sandu, T. Chai, and G. R. Carmichael, 2007: Ensemble-based chemical data assimilation I: general approach. *Quarterly Journal of the Royal Meteorological Society*, **133,** 1229-1243.

Craven, P., and G. Wahba, 1979: Smoothing noisy data with spline functions. *Numerische Mathematik*, **31,** 377-403.

Dee, D. P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review*, **123,** 1128-1145.

Dee, D. P., and A. M. Silva, 1999: Maximum-likelihood estimation of forecast and observation error covariance parameters part I: methodology. *Monthly Weather Review*, **127,** 1822-1834.

Ellison, C. J., J. R. Mahoney, and J. P. Crutchfield, 2009: Prediction, Retrodiction, and the Amount of Information Stored in the Present. *Journal of Statistical Physics*, **136,** 1005-1034.

Eubank, R. L., 1999: *Nonparametric regression and spline smoothing.* Marcel Dekker, Inc., 338 pp.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99,** 10143-10162.

Gentle, J. E., W. Hardle, and Y. Mori, 2004: *Handbook of computational statistics: concepts and methods.* Springer, 1070 pp.

Golub, G. H., and C. F. V. Loan, 1996: *Matrix Computations.* The Johns Hopkins University Press: Baltimore.

Green, P. J., and B. W. Silverman., 1994: *Nonparametric Regression and Generalized Additive Models.* Vol. 182, Chapman and Hall,.

Gu, C., 2002: *Smoothing Spline ANOVA Models.* Springer-Verlag, 289 pp.

Gu, C., and G. Wahba, 1991: Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computation*, **12,** 383-398.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation for data assimilation operational sequential and variational. *Journal of the Meteorological Society of Japan*, **75,** 181-189.

Kirchgessner, P., L. Berger, and A. B. Gerstner, 2014: On the choice of an optimal localization radius in ensemble Kalman filter methods. *Monthly Weather Review*, **142,** 2165-2175.

Krakauer, N. Y., T. Schneider, J. T. Randerson, and S. C. Olsen, 2004: Using generalized cross-validation

to select parameters in inversions for regional carbon fluxes. *Geophysical Research Letters*, **31**.

Li, H., E. Kalnay, and T. Miyoshi, 2009: Simultaneous estimation of covariance inflatioin and observation errors within an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, **135,** 523-533.

Liang, X., X. Zheng, S. Zhang, G. Wu, Y. Dai, and Y. Li, 2012: Maximum Likelihood Estimation of Inflation Factors on Error Covariance Matrices for Ensemble Kalman Filter Assimilation. *Quarterly Journal of the Royal Meteorological Society*, **138,** 263-273.

Liu, J., E. Kalnay, T. Miyoshi, and C. Cardinali, 2009: Analysis sensitivity calculation in an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, **135,** 1842-1851.

Lorenz, E. N., 1996: Predictability a problem partly solved.

Lorenz, E. N., and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations simulation with a small model. *Journal of the Atmospheric Sciences*, **55,** 399-414.

MacCarthy, J. K., B. Borchers, and R. C. Aster, 2011: Efficient stochastic estimation of the model resolution matrix diagonal and generalized cross–validation for large geophysical inverse problems. *Journal of Geophysical Research*, **116**.

Miller, R. N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical systems. *Journal of the Atmospheric Sciences*, **51,** 1037-1056.

Miyoshi, T., 2011: The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Monthly Weather Review*, **139,** 1519-1534.

Miyoshi, T., and M. Kunii, 2011: The Local Ensemble Transform Kalman Filter with the Weather Research and Forecasting Model: Experiments with Real Observations. *Pure & Applied Geophysics*, **169,** 321-333.

Pena, D., and V. J. Yohai, 1991: The detection of influential subsets in linear regression using an influence matrix. *Journal of the Royal Statistical Society*, **57,** 145-156.

Ravazzani, G., A. Amengual, A. Ceppi, V. Homar, R. Romero, G. Lombardi, and M. Mancini, 2016: Potentialities of ensemble strategies for flood forecasting over the Milano urban area. *Journal of Hydrology*, **539,** 237-253.

Reichle, R. H., 2008: Data assimilation methods in the Earth sciences. *Advances in Water Resources*, **31,** 1411-1418.

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto, 2004: *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models.* JohnWiley & Sons, 219 pp.

Saltelli, A., and Coauthors, 2008: *Global Sensitivity Analysis: The Primer.* John Wiley & Sons, 292 pp.

Talagrand, O., 1997: Assimilation of Observations, an Introduction. *Journal of the Meteorological Society of Japan*, **75,** 191-209.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Notes and correspondence ensemble square root filter. *Monthly Weather Review*, **131,** 1485-1490.

Wahba, G., and S. Wold, 1975: A completely automatic french curve. *Communications in Statistics*, **4,** 1-17.

Wahba, G., D. R. Johnson, F. Gao, and J. Gong, 1995: Adaptive tuning of numerical weather prediction models randomized GCV in three- and four-dimensional data assimilation. *Monthly Weather Review*, **123,** 3358-3369.

Wand, M. P., and M. C. Jones, 1995: *Kernel Smoothing.* Chapman and Hall, 212 pp.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform kalman filter ensemble forecast schemes. *Journal of the Atmospheric Sciences*, **60,** 1140-1158.

1   Wu, G., X. Zheng, L. Wang, S. Zhang, X. Liang, and Y. Li, 2013: A New Structure for Error Covariance
2   Matrices and Their Adaptive Estimation in EnKF Assimilation. *Quarterly Journal of the Royal*
3   *Meteorological Society*, **139,** 795-804.

4   Wu, G., X. Yi, X. Zheng, L. Wang, X. Liang, S. Zhang, and X. Zhang, 2014: Improving the Ensemble
5   Transform Kalman Filter Using a Second-order Taylor Approximation of the Nonlinear Observation
6   Operator. *Nonlinear Processes in Geophysics*, **21,** 955-970.

7   Xu, T., J. J. Gómez-Hernández, H. Zhou, and L. Li, 2013: The power of transient piezometric head data
8   in inverse modeling: An application of the localized normal-score EnKF with covariance inflation in a
9   heterogenous bimodal hydraulic conductivity field. *Advances in Water Resources*, **54,** 100-118.

10 Yang, S.-C., E. Kalnay, and T. Enomoto, 2015: Ensemble singular vectors and their use as additive
11 inflation in EnKF. *Tellus A*, **67**.

12 Zheng, X., 2009: An adaptive estimation of forecast error statistic for Kalman filtering data
13 assimilation. *Advances in Atmospheric Sciences*, **26,** 154-160.

14 Zheng, X., and R. Basher, 1995: Thin-plate smoothing spline modeling of spatial climate data and its
15 application to mapping south Pacific rainfall. *Monthly Weather Review*, **123,** 3086-3102.

16
17