The author gratefully acknowledges the two anonymous reviewers for their insightful comments and constructive suggestions that lead to the significant improvement of the quality of this manuscript. The author is also grateful to the editors for their very kind help and comments. The author has checked the manuscript carefully and tried the best to address all the comments. Below boldface is used to indicate the comments from the reviewers and editor, and italics is for the point-by-point responses.

The main changes in the revised version are: add the assimilation results of constant inflated EnKF scheme to Tables 1 and 2; add the running times of different schemes to Tables 1 and 2 for comparison. The scientific discussions/explanations and the English grammar are also modified.

Comments from editor

I would like to suggest that you make the revisions suggested by the referees. The main revision that would be essential is the first question raised by referee 2 (but it also appeared in both the referee reports for the first version). In particular, please include a discussion comparing the computational costs of the proposed method with those of the constant inflation factor method, both in the context of the model studied here and in the context of operational models. Please also address the second and third points raised by referee 2.

Response: *Thanks for your constructive comments. In the revised version, the validation statistics and running times of different schemes were added to Tables 1 and 2 for comparison. Generally speaking, 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. Therefore, the proposed EnKF scheme does not significantly increase computational cost. (P21 L11-P22 L2)*

The constant multiplicative inflation factor is set as the median of the estimated inflation factor by minimizing the GCV function. For instance, it is 1.88 for the 40-observation and 30-ensemble assimilation experiment. 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. (P15 L11-15)

Anonymous Reviewer #1

Recommendation: accepted subject to technical corrections.

Additional questions:

1) On P3, what method do you use to optimize the inflation factor? Are derivatives required?

Response: *The optimization usually needs Newton's method, requiring the derivatives.*

2) Is there a reason why you did not compare the EnKF with constant vs adaptive covariance inflation in Tables 1,2? Based on the figures, the results look rather similar.

Response: *Thanks for your suggestion. In the revised version, the results of the EnKF with constant covariance inflation scheme are listed in Tables 1 and 2. It shows that the analysis RMSE values of the EnKF with adaptive inflation are about 15% smaller than those of the EnKF with constant inflation scheme.*

Technical corrections:

P3L7: Data assimilation provides

Response: *The words have been revised.*

P3L14-17: Can be attributed to technical aspect? Forecast models are most often nonlinear.

Response: *The sentence has been deleted.*

P4L18: Moment estimation...

Response: *The words have been revised.*

P4L20: "obtain a better estimate"... Better than what?

Response: *Better than moment approach. The sentence has been revised to*

“Maximum likelihood approach can obtain a better estimate of the inflation factor than moment approach…”

P5L18: Newton's method

Response: *The words have been revised.*

P6L5-8: Suggested rewording:

"Covariance inflation is a common method for enhancing the performance of ensemble filtering schemes. It involves multiplying the forecast error matrix by inflation factor, which is usually larger than 1. Its value must be chosen appropriately however. Too small or too large an inflation factor will cause the analysis state to over rely on the model forecasts or observations, and can seriously undermine the accuracy and stability of the filter."

Response: *Thanks for your suggestion. This text has been inserted into the introduction section (P4 L16-19)*

P6L8: Delete "Hence ... accurately"

Response: *The sentence has been deleted.*

P6L9-10: I still don't understand what is meant by "After this, the weights... reassigned"

Response: *The sentence has been deleted*

P6L10-15: The discussion of analysis sensitivity should be rewritten in a new paragraph. Something like this:

"This study proposes a new method for choosing the inflation factor using GCV. 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. In the context..."

Response: *Thanks for your suggestion. This text has been inserted into the* *introduction section (P6 L9-12)*

P11L16: Trace diagnostics can also be used...

Response: *The words have been revised.*

Again, thanks for your review and positive recommendation.

Anonymous Reviewer 2

Recommendation: reconsidered after major revisions

The author has come up with a new way of estimating inflation factor necessary to arrest the collapse of ensembles in Ensemble Kalman Factor (EnKF) Schemes. This method fares considerably better than a textbook EnKF scheme but exhibits little improvements when pitched against a simple constant multiplicative inflation schemes. The author is advised to show that this little benefit comes at no significant computational cost. If the author can demonstrate that, this paper may be considered for publication. The author is also advised to revise the language.

Response: *Thanks for your review. Following your comments, the computational cost of the proposed scheme is detected and discussed. The manuscript language has been polished in the revised version. Please see the responses to the following detailed points.*

1 General Comments

This paper presents a novel way of estimating model error covariance inflation factor to be used in ensemble-based filters. It is known that the spread of ensembles in ensemble-based filters tend to collapse unless explicitly taken care of. A conventional way to arrest this collapse is to inflate the model error covariance using an inflation factor. There are a number of ways to estimate this inflation factor. This paper borrows its ideas of estimating the inflation factor from the realm of generalized cross validation techniques. The author demonstrates using a Lorenz model that this method is considerably better than a basic Ensemble Kalman Filter. However, when pitched against a constant multiplicative inflation factor scheme, this method fares slightly better but involves more computational cost.

Response: *Thanks for your thorough review and constructive comments. The inflation factor is determined by the researchers' experience and prior knowledge in* *the constant inflated EnKF scheme. Therefore, it is very empirical and subjective. It also seems quite unreasonable to use the same inflation factor during all the assimilation procedure. The computational cost of different inflation schemes are detected and compared in the revised version. Generally speaking, the proposed scheme is about 15% smaller in analysis RMSE, but only about 5% longer in running time than the EnKF with constant inflation.*

The author should address the following questions:

1) How does the computational cost fare when compared to the constant multiplicative inflation factor method? Keeping the number of processors fixed, how much more time (in percentage) does this new method take in each analysis cycle when the ensemble sizes vary from 10 to 30 to 50? The author is also advised to discuss this aspect in the context of operational meteorological/ocean models.

Response: *Thanks for your constructive comments. The following text is added to the discussion section in the revised version (P21 L11-P22 L2)*

For the Lorenz-96 experiments in this study, the conventional EnKF, constant inflated EnKF and proposed 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.

Table 1. Time-mean values of the forecast ensemble spread, GAI statistics, GCV functions and analysis RMSE over 2000 time steps for different assimilation schemes. The ensemble size is selected as 10, 30 and 50, respectively.

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*

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

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*

2) It is not clear what is the value of the constant multiplicative inflation factor.

The author is advised to clearly mention it in the text and also in the figure captions.

Response: *The constant multiplicative inflation factor is set as the median of the estimated inflation factor by minimizing the GCV function. For instance, it is 1.88 for the 40-observation and 30-ensemble assimilation experiment. 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.*

This has been added in the text and in the figure captions in the revised version (P15 L11-15).

3) The author is advised to include statistics of constant multiplicative inflation factor scheme in Table 1. This will help in understanding how the new method fares with respect to the constant multiplicative inflation factor scheme for larger ensemble sizes. This is relevant because most of the weather models use ensemble sizes much larger than 30 which is being used here.

Response: *Thanks for your comments. The statistics of constant inflated EnKF scheme are added in Tables 1 and 2. The results show that, for the fixed observation number and ensemble size, the proposed EnKF scheme performs better than the constant inflated EnKF scheme, although at cost of about 5% increase of the computational time.*

The language in this paper needs revision. If the author can demonstrate that there is no significantly extra computational cost involved compared to the simple multiplicative inflation factor scheme, then this paper may be considered for publication.

Response: *Thanks for your comments and consideration. The manuscript language has been polished in the revised version.*

2 Specific Comment

1) P3 L8: What does the author mean by fully considers?

Response: *It means data assimilation can fuse all the information from dynamic model forecasts and observations. The word "fully" has been deleted in the revised version.*

2) P3 L13-17: The concept of global minima is introduced without any prior background thereby introducing discontinuity in the flow of the introduction.

Response: *Thanks for your comment. The term "global minima" has been changed to "appropriate analysis state" in the revised version.*

3) P5 L15-16: What are the favorable properties of GCV and how are these properties relevant in the present context?

Response: *In the machine learning literature, The GCV criterion has a rotation-invariant property that is relative to the orthogonal transformation of the observations and is a consistent estimate of the relative loss. In the present context, the estimated inflation factor by minimizing the GCV function is not affected by the observation unit and can optimize the analysis sensitivity to the observation.*

This has been added in the discussion section in the revised version (P19 L20-22).

4) P9 L12-15: "In previous studies, a number of methods ...". This has already been mentioned in the introduction and sounds repetitive.

Response: *The sentence has been deleted.*

5) P15 L14-15: What is the meaning of "greatly majority"? Also there are instances when λ <1 in Fig(2). The significance of this should be discussed in the paper.

Response: *Thanks for your comments. The text has been changed as following in the revised version (P15 L9-11): "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."*

6) P15 L15-18: Is the value of constant multiplicative inflation factor set at λ =1.88? If yes, what is the motivation of choosing the median? The sentences used are not very clear.

Response: *Yes, the constant multiplicative inflation factor is set as 1.88 for the 40-observation and 30-ensemble assimilation experiment. 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.*

3 Technical Comment

1) P6 L11-13: "Generally speaking, analysis sensitivity is used to apportion . . . ". It's not clear what the author wants to convey.

Response: *The misleading sentence has been deleted.*

2) P7 L12: Change "numeric" to "numerical".

Response: *The word has been changed.*

3) P9 L7-9: "The forecast error inflation scheme should be included ..." is grammatically incorrect.

Response: *The sentence is revised to "The forecast error inflation procedure should be added to …".*

4) P14 L17: "The frequency was set as..." is not grammatically correct.

Response: *The sentence is revised to "The frequency was every 4 time steps …".*

5) P14 L20: "The variance of the observation on each grid ..." may be changed to "The variance of the observation at each grid ...".

Response: *The word has been changed.*

Again, thanks for your thorough review and constructive comments.

The reference in this reply is listed below.

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.

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

Guocan Wu[1,2]

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

2 Joint Center for Global Change Studies, 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 a “true” underlying state. 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 the model forecasts and excludes the observations. It can eventually result in the divergence of the filter (Anderson and Anderson 1999; Constantinescu et al. 2007; Wu et al. 2014).

The covariance inflation technique is used to mitigate filter divergence by inflating the empirical covariance in EnKF, and it can increase the weight of the observations in the analysis state (Xu et al. 2013). In reality, this method will perturb the subspace spanned by the ensemble vectors and better capture the sub-growing directions that may be missed in the original ensemble (Yang et al. 2015). Therefore, using the inflation technique to enhance the estimate accuracy of the forecast error covariance matrix is increasingly important.

A widely used inflation technique involves multiplying the forecast error matrix by inflation factor, which must be chosen appropriately. In early studies, researchers usually tuned the inflation factor by repeated assimilation experiments and selected the estimated inflation factor according to their experience and prior knowledge (Anderson and Anderson 1999). However, such methods are very empirical and subjective. It also seems quite unreasonable to use the same inflation factor during all the assimilation procedure. Too small or too large an inflation factor will cause the analysis state to over rely on the model forecasts or observations, and can seriously undermine the accuracy and stability of the filter.

In later studies, the inflation factor can be estimated online based on the innovation statistic (observation-minus-forecast; (Dee 1995; Dee and Silva 1999)) with different 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 applied the Newton's method to optimize the GCV score with multiple smoothing parameters in a smoothing spline model (Gu and Wahba 1991). 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 also applied the GCV in a thin-plate smoothing spline model of spatial climate data to deal with South Pacific rainfalls (Zheng and Basher 1995). The GCV criterion 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).

This study proposes a new method for choosing the inflation factor using GCV. 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):

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

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

where i represents the time index; $\mathbf{x}_i^{\mathrm{t}} = \left\{\mathrm{x}_{i,1}^{\mathrm{t}}, \mathrm{x}_{i,2}^{\mathrm{t}}, \ldots, \mathrm{x}_{i,n}^{\mathrm{t}}\right\}^{\mathrm{T}}$ represents the n-dimensional true state vector at the i-th time step; $\mathbf{x}_{i-1}^{\mathrm{a}} = \left\{\mathrm{x}_{i-1,1}^{\mathrm{a}}, \mathrm{x}_{i-1,2}^{\mathrm{a}}, \ldots, \mathrm{x}_{i-1,n}^{\mathrm{a}}\right\}^{\mathrm{T}}$ represents the n-dimensional analysis state vector, which is an estimate of $\mathbf{x}_{i-1}^{\mathrm{t}}$; M_{i-1} represents a nonlinear dynamical forecast operator such as a numerical weather prediction model; $\mathbf{y}_i^{\mathrm{o}} = \left\{\mathrm{y}_{i,1}^{\mathrm{o}}, \mathrm{y}_{i,2}^{\mathrm{o}}, \ldots, \mathrm{y}_{i,p_i}^{\mathrm{o}}\right\}^{\mathrm{T}}$ 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 $\mathbf{x}_i^{\mathrm{a}}$ that is an accurate estimate of the corresponding true states $\mathbf{x}_i^{\mathrm{t}}$ based on the information provided by M_i and $\mathbf{y}_i^{\mathrm{o}}$.

Suppose the perturbed analysis state at a previous time step $\mathbf{x}_{i-1}^{\mathrm{a(j)}}$ 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:

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

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

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

Step 2. Analysis step.

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

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

which has the analytic form

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

where

$$\mathbf{d}_i = \mathbf{y}_i^{\mathrm{o}} - \mathbf{H}_i\mathbf{x}_i^{\mathrm{f}} \tag{7}$$

is the innovation statistic (observation-minus-forecast residual). To complete the ensemble forecast, the perturbed analysis states are calculated using perturbed observations (Burgers et al. 1998):

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

where $\boldsymbol{\varepsilon}_i^{'\mathrm{(j)}}$ is a normally distributed random variable with mean zero and covariance 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 Sherman-Morrison-Woodbury formula (Golub and Loan 1996; Liang et al. 2012; Tippett et al. 2003). Finally, set $i=i+1$, return to Step 1 for the model forecast at the next time step and repeat until the model reaches the last time step N.

2.2. Influence matrix and forecast error inflation

The forecast error inflation procedure should be added to any ensemble-based assimilation scheme to prevent the filter from diverging (Anderson and Anderson 1999; Constantinescu et al. 2007). Multiplicative inflation is one of the commonly used inflation techniques, and it adjusts the initially estimated forecast error covariance matrix $\mathbf{P}_i$ to $\lambda_i\mathbf{P}_i$ after estimating the inflation factors λ_i properly.

In this study, a new procedure for estimating multiplicative inflation factors λ_i is proposed based on the following GCV function (Craven and Wahba 1979)

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

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

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

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

$$\mathbf{x}_i^{\mathrm{a(j)}} = \mathbf{x}_i^{\mathrm{f(j)}} + \lambda_i \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} \left(\mathbf{H}_i \lambda_i \mathbf{P}_i \mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i \right)^{-1} \left(\mathbf{d}_i + \boldsymbol{\varepsilon}_i^{\mathrm{'(j)}} \right). \tag{11}$$

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:

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

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:

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

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:

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

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

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

The elements of the matrix $\mathbf{S}_i^{\mathrm{o}}$ reflect the sensitivity of the normalized analysis state to the normalized observations; its diagonal elements are the analysis self-sensitivities and the off-diagonal elements are the cross-sensitivities. On the other hand, the elements of the matrix $\mathbf{S}_i^{\mathrm{f}}$ reflect the sensitivity of the normalized analysis state to the normalized forecast state. The two quantities are complementary, and the GCV function can be interpreted as minimizing the normalized forecast sensitivity because the inflation scheme will increase the observation weight appropriately.

In fact, the sensitivity matrix $\mathbf{S}_i^{\mathrm{o}}$ is equal to the influence matrix $\mathbf{A}_i$ (see Appendix B for detailed proof), whose trace can be used to measure the “equivalent number of parameters” or “degrees of freedom for the signal” (Gu 2002; Pena and Yohai 1991). Similarly, the sensitivity matrix $\mathbf{S}_i^{\mathrm{o}}$ can be interpreted as a measurement of the amount of information extracted from the observations (Ellison et al. 2009). Trace diagnostics can be used to analyse the sensitivities to observations or forecast vectors (Cardinali et al. 2004). The Global Average Influence (GAI) at the *i*-th time step is defined as the globally averaged observation influence:

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

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

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:

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

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 $\mathbf{x}_i^{\mathrm{t}}$ 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:

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

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:

$$\frac{d\boldsymbol{X}_k}{dt} = (\boldsymbol{X}_{k+1} - \boldsymbol{X}_{k-2})\boldsymbol{X}_{k-1} - \boldsymbol{X}_k + F , \tag{19}$$

where $k = 1, 2, \cdots, 40$. The cyclic boundary conditions $\boldsymbol{X}_{-1} = \boldsymbol{X}_{K-1}$, $\boldsymbol{X}_0 = \boldsymbol{X}_K$, and $\boldsymbol{X}_{K+1} = \boldsymbol{X}_1$ were applied to ensure that Eq. (19) was well defined for all values of *k*. The Lorenz-96 model is “atmosphere-like” because the three terms on the right-hand side of Eq. (19) are analogous to a nonlinear advection-like term, a damping term, and an external forcing term. The model can be considered representative of an atmospheric quantity (e.g., zonal wind speed) distributed on a latitude circle. Therefore, the Lorenz-96 model has been widely used as a test bed to evaluate the performance of assimilation schemes in many studies (Wu et al. 2013).

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

(Butcher 2003). The time step for generating the numerical solution was set at 0.05 non-dimensional units, which is roughly equivalent to 6 hours in real time assuming that the characteristic time-scale of the dissipation in the atmosphere is 5 days (Lorenz 1996). The forcing term was set as F = 8, so that the leading Lyapunov exponent implies an error-doubling time of approximately 8 time steps and the fractal dimension of the attractor was 27.1 (Lorenz and Emanuel 1998). The initial value was chosen to be $\boldsymbol{X}_k = F$ when $k \neq 20$ and $\boldsymbol{X}_{20} = 1.001F$.

In this study, the synthetic observations were assumed to be generated by adding random noises that were multivariate-normally distributed with mean zero and covariance matrix $\mathbf{R}_i$ to the true states. The frequency was every 4 time steps, which can be used to mimic daily observations in practical problems, such as satellite data. The observation errors were assumed to be spatially correlated, which is common in applications involving remote sensing and radiance data. The variance of the observation at each grid point was set to $\sigma_{\mathrm{o}}^2 = 1$, and the covariance of the observations between the j-th and k-th grid points was as follows:

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

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 variability of the analysis RMSE was consistent with that of the GCV function for the EnKF with the forecast error inflation scheme. 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 problem is usually several tens of members (Kirchgessner et al. 2014). The preferences of the proposed inflation method and the constant inflation method with respect to different ensemble sizes (10, 30 and 50) were evaluated, and the results are listed in Table 1. It shows that for each scheme, using a 10-member ensemble produced a threefold increase in the analysis RMSE, while using a 50-member ensemble reduced the analysis RMSE by 20% relative to the analysis RMSE obtained using a 30-member ensemble. The forecast ensemble spread increased slightly from a 10-member ensemble to a 50-member ensemble. The GAI and GCV function values changed sharply from a 10-member ensemble to a 30-member ensemble, and they became relatively stable from a 30-member ensemble to a 50-member ensemble. Ensembles less than 10 were unstable, and no significant changes occurred for ensembles greater than 50. Considering the computational costs for practical problems, a 30-member ensemble may be necessary to estimate statistically robust results.

To evaluate the preferences of the inflation method with respect to different numbers of observations, synthetic observations were generated at every other grid point and for every 4 time steps. Hence, a total of 20 observations were performed at each observation step in this case. The assimilation results with ensemble sizes of 10, 30 and 50 are listed in Table 2, which shows that the GAI values were larger than those with 40-observations in all assimilation schemes. This finding may be related to 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 approach, but requires computing high dimensional matrix determinants. The Bayesian approach assumes a prior distribution for the inflation factor but is limited to spatially independent observational errors. In this study, the inflation factor was estimated based on cross-validation and the analysis sensitivity was detected. The estimated inflation factor by minimizing the GCV function is not affected by the observation unit and can optimize the analysis sensitivity to the observation.

In fact, the CV method can evaluate and compare learning algorithms and represents a widely used statistical method. In this study, the CV concept was adopted for the inflation factor estimation in the improved EnKF assimilation scheme and was validated with the Lorenz-96 model. The assimilation results showed that inflating the conventional EnKF using the factor estimated by minimizing the GCV function can indeed reduce the analysis RMSE. Therefore, the GCV function can accurately quantify the goodness of fit of the error covariance matrix. The values of the GCV function obviously decreased in the proposed approach compared the conventional EnKF and constant inflated EnKF schemes. The analysis RMSE of the proposed approach was also much smaller than those of the conventional EnKF and constant inflated EnKF schemes, which suggests that the GCV criterion works well for estimating the inflation factor.

The analysis sensitivities in the proposed approach and in the conventional EnKF scheme were also investigated in this study. The time-averaged GAI statistic increases from about 10% in the conventional EnKF scheme to about 30% using the proposed inflation method. This illustrates that the inflation mitigates the problem of the analysis depending excessively on the forecast and excluding the observations. The relationship of the analysis state to the forecast state and the observations are more reasonable.

4.2 Computational cost

The highest computational cost when minimizing the GCV function is related to calculating the influence matrix $\mathbf{A}_i(\lambda)$. Since the matrix multiplication is commutative for the trace, the GCV function can be easily re-expressed as follows:

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

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

Appendix A

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

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

where $\tilde{\mathbf{y}}_i^{\mathrm{o}} = \mathbf{R}_i^{-1/2}\mathbf{y}_i^{\mathrm{o}}$ is the normalized observation vector and $\tilde{\boldsymbol{\varepsilon}}_i \sim N(\mathbf{0},\mathbf{I})$; $\mathbf{I}_{p_i}$ is the identity matrix with the dimensions $p_i \times p_i$. Similarly, the normalized analysis vector is $\tilde{\mathbf{y}}_i^{\mathrm{a}} = \mathbf{R}_i^{-1/2}\mathbf{H}_i\mathbf{x}_i^{\mathrm{a}}$ and the influence matrix $\mathbf{A}_i$ relates the normalized observation vector to the normalized analysis vector, thereby ignoring the normalized forecast state in the observation space (Gu 2002):

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

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

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

If the initial forecast error covariance matrix is inflated as described in Section 2.2, then the influence matrix is treated as the following function of λ

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

The principle of CV is to minimize the estimated error at the observation grid point. Lacking an independent validation data set, a common alternative strategy is to minimize the squared distance between the normalized observation value and the analysis value while not using the observation on the same grid point, which is the following objective function:

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

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

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

The subscript *–k* indicates a vector (matrix) with its *k*-th element (*k*-th row and column) deleted. Instead of minimizing Eq. (A6) p_i times, the objective function (Eq. (A5)) has another more simple expression (Gu 2002):

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

where $a_{k,k}$ is the element at the site pair (*k*, *k*) of the influence matrix $\mathbf{A}_i(\lambda)$. Then, $a_{k,k}$ is substituted with the average $\frac{1}{p_i}\sum_{k=1}^{p_i}a_{k,k} = \frac{1}{p_i}\mathrm{Tr}(\mathbf{A}_i(\lambda))$ and the constant is ignored to obtain the following GCV statistic (Gu 2002):

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

Appendix B

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

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

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:

$$\begin{aligned}
\mathbf{S}_i^{\mathrm{o}} &= \mathbf{R}_i^{1/2}\mathbf{K}_i^{\mathrm{T}}\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-1/2} \\
&= \mathbf{R}_i^{1/2}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right)^{-1}\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-1/2} \\
&= \mathbf{R}_i^{1/2}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right)^{-1}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i - \mathbf{R}_i\right)\mathbf{R}_i^{-1/2} \\
&= \mathbf{R}_i^{1/2}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right)^{-1}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right)\mathbf{R}_i^{-1/2} - \mathbf{R}_i^{1/2}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right)^{-1}\mathbf{R}_i\mathbf{R}_i^{-1/2} \\
&= \mathbf{I}_{p_i} - \mathbf{R}_i^{1/2}\left(\mathbf{H}_i\lambda\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right)^{-1}\mathbf{R}_i^{1/2} \\
&= \mathbf{A}_i
\end{aligned} \tag{B2}$$

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

Acknowledgements. This work is supported by the National Natural Science Foundation of China (Grant No. 91647202), the National Basic Research Program of China (Grant No. 2015CB953703) and the National Natural Science Foundation of China (Grant No. 41405098).

References

Table 1. Time-mean values of the forecast ensemble spread, GAI statistics, GCV functions and analysis RMSE over 2000 time steps, as well as the running times (second) for different assimilation schemes. The observation number is 40 and the ensemble size is selected as 10, 30 and 50, respectively.

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

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

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

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

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.

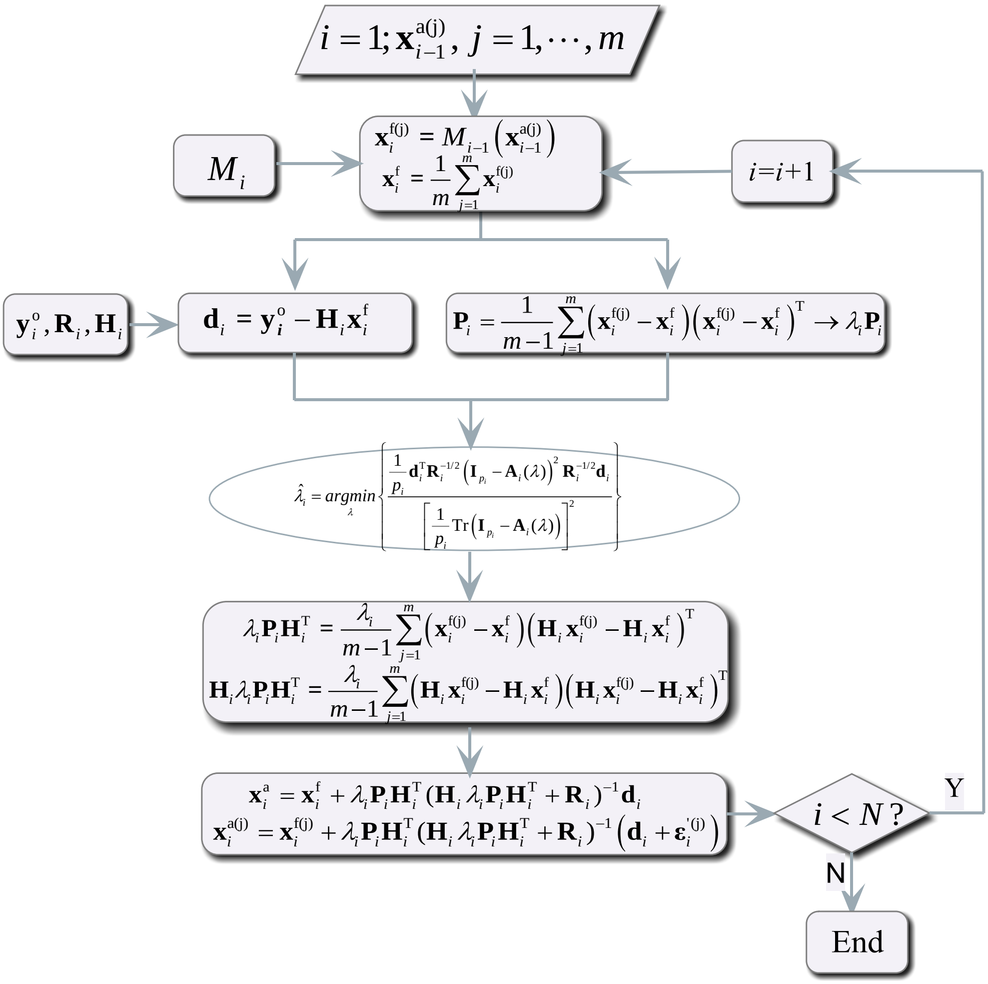

Figure 1. Flowchart of the proposed assimilation scheme.

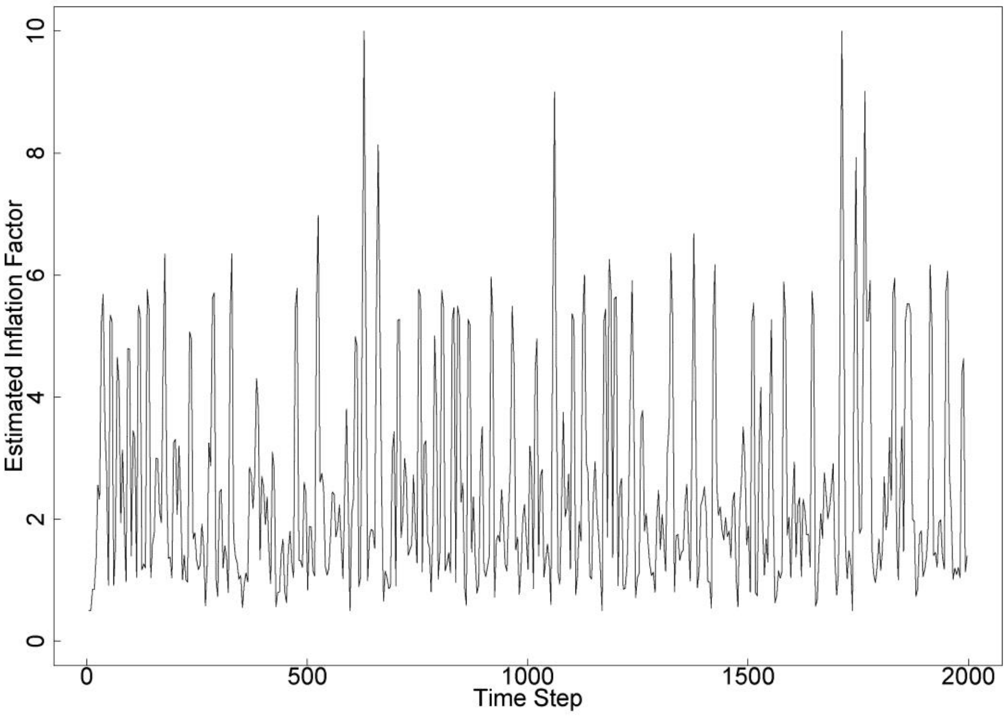

Figure 2. Time series of the estimated inflation factors by minimizing the GCV function. The median of the estimated inflation factors is 1.88.

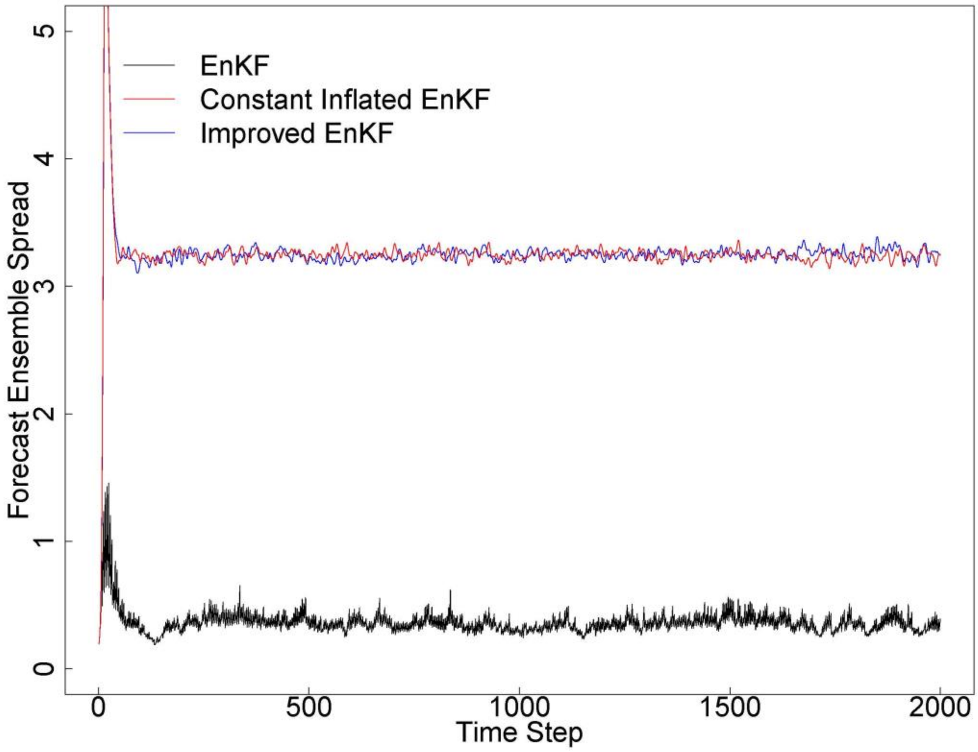

Figure 3. Forecast ensemble spread of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88.

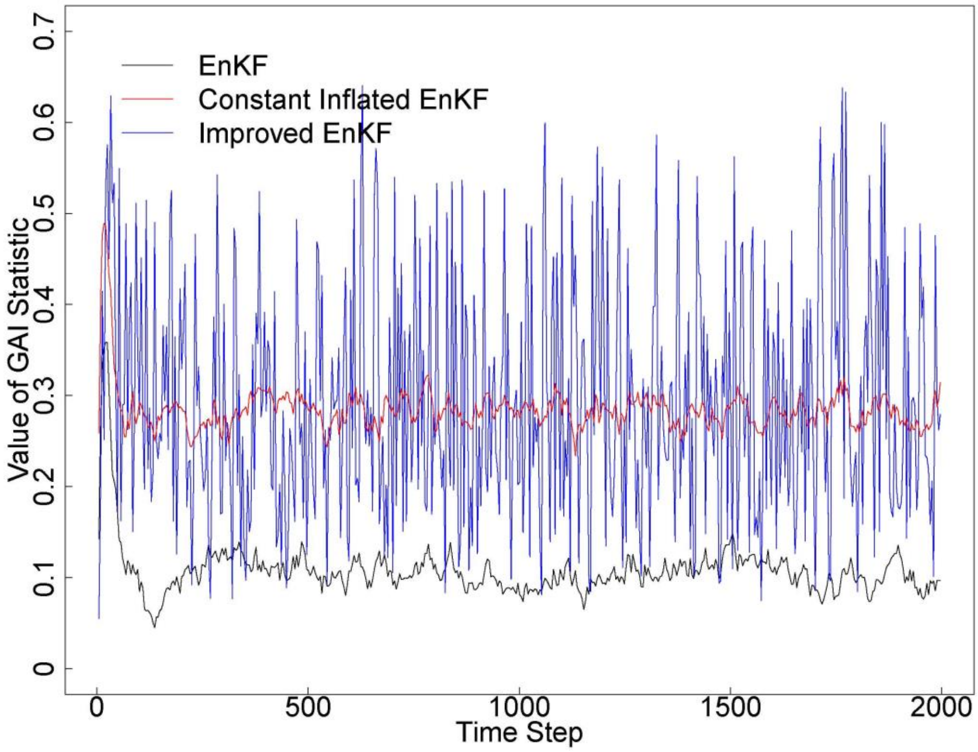

Figure 4. GAI statistics of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88.

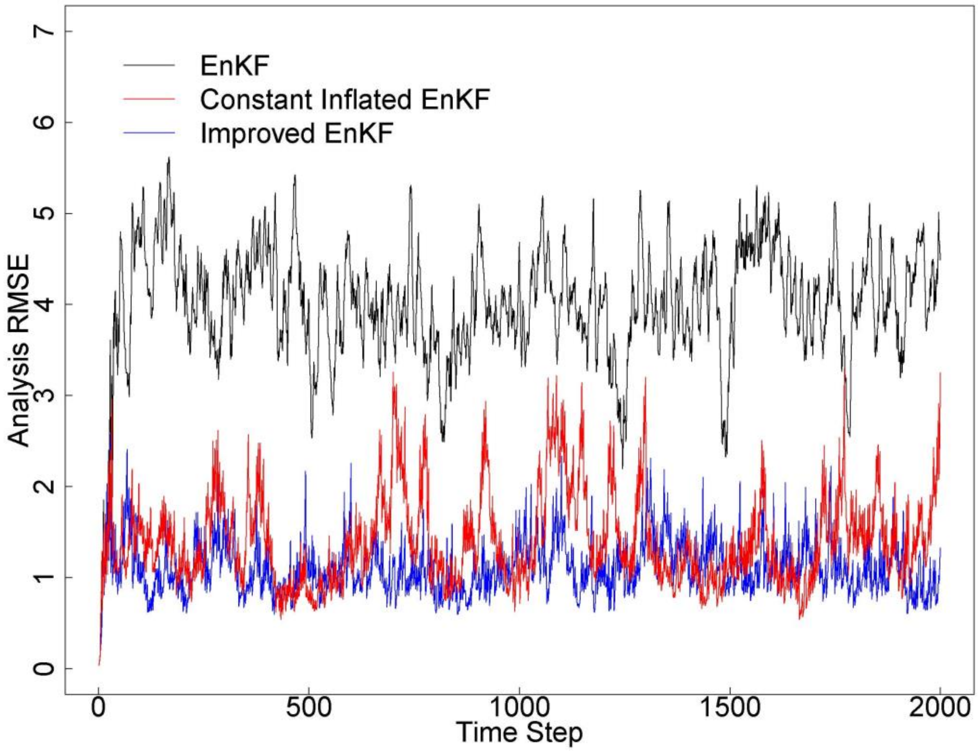

Figure 5. Analysis RMSE of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88.

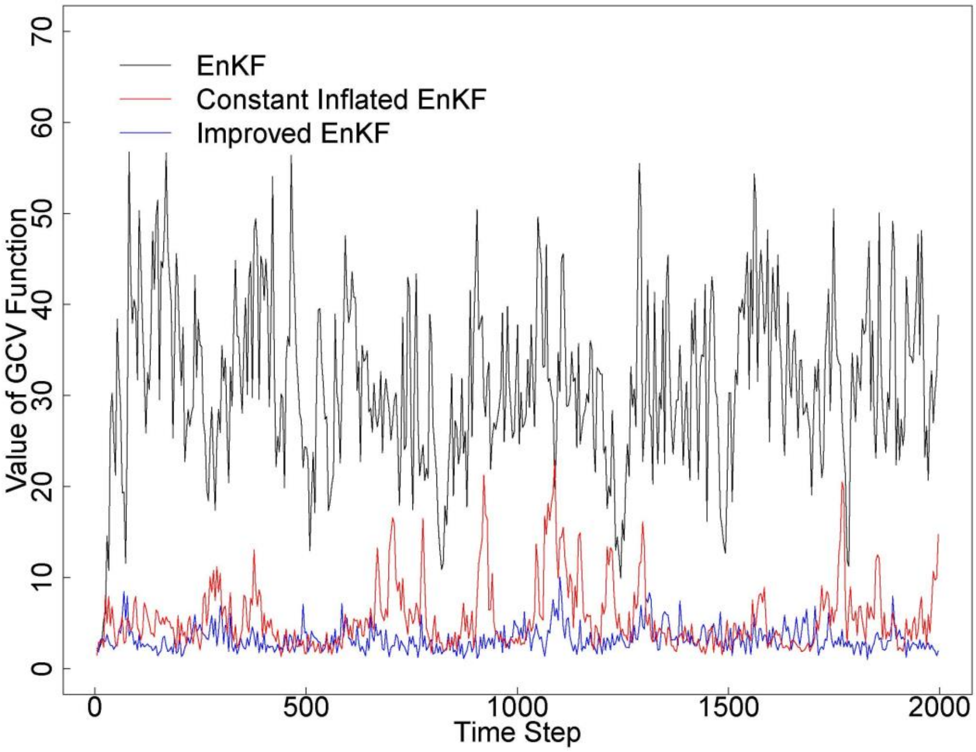

Figure 6. GCV function values of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88.

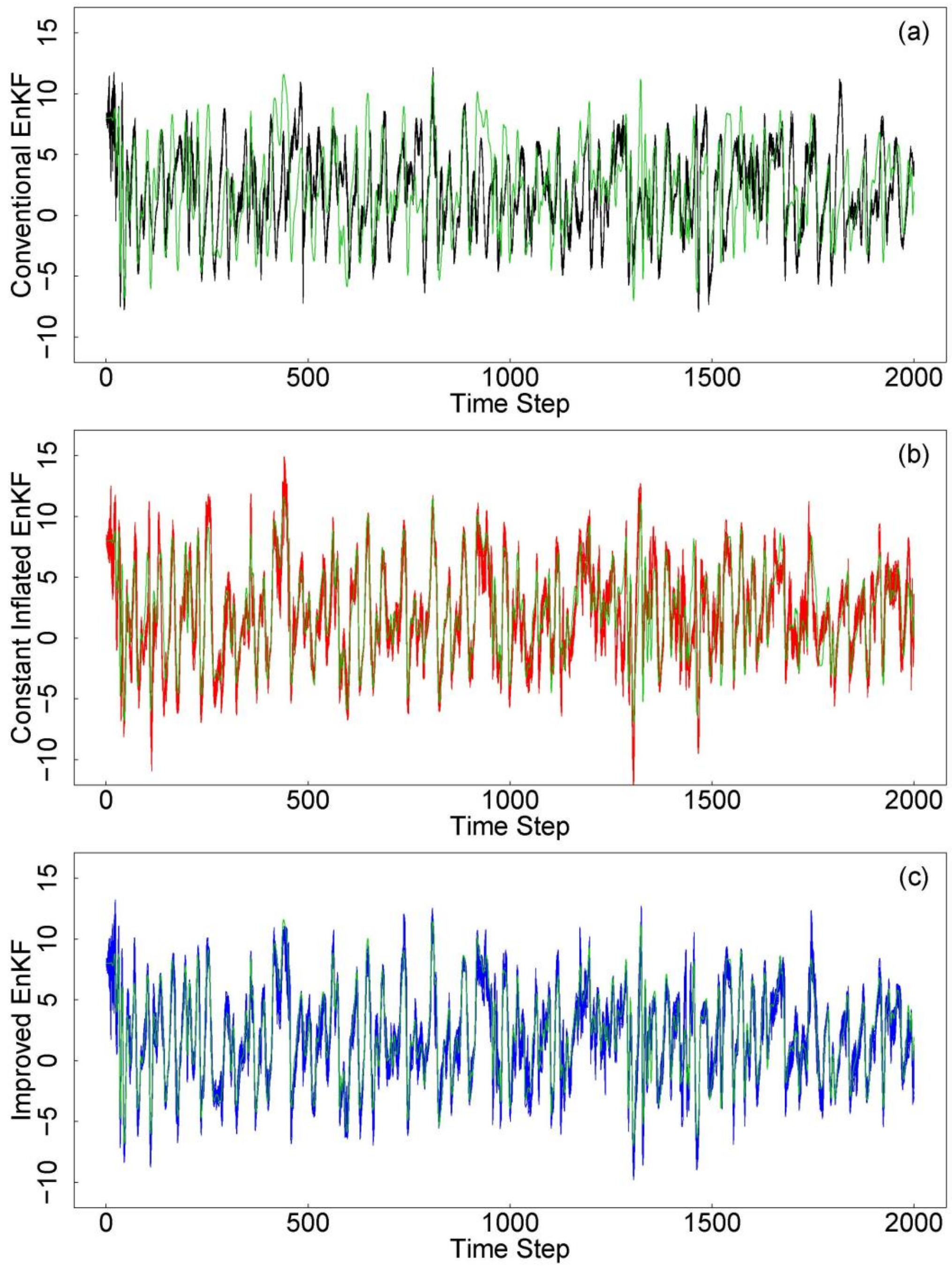

Figure 7. Ensemble analysis state members of the conventional EnKF (black line), the constant inflated EnKF (red line) and the improved EnKF (blue line) for the Lorenz-96 experiment with 40-observation and 30-ensemble member. The constant multiplicative inflation factor is set as 1.88. The green line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

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.

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., R. J. Donald, F. Gao, and J. Gong, 1995: Adaptive Tuning of Numerical Weath er 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.

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

Wu, G., X. Yi, X. Zheng, L. Wang, X. Liang, S. Zhang, and X. Zhang, 2014: Improving the Ensemble Transform Kalman Filter Using a Second-order Taylor Approximation of the Nonlinear Observation

Operator. *Nonlinear Processes in Geophysics*, **21,** 955-970.

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

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

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

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