Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access

An Estimate of Inflation Factor and Analysis Sensitivity in 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

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Abstract

The estimation accuracy of forecast error matrix is crucial to the assimilation result. Ensemble Kalman filter (EnKF) is a widely used ensemble based assimilation method, which initially estimate the forecast error matrix using a Monte Carlo method with the short-term ensemble forecast states. However, this estimate needs to be further improved using inflation technique.

In this study, the forecast error inflation factor is estimated based on cross validation and the analysis sensitivity is also investigated. The improved EnKF assimilation scheme is validated by assimilating spatially correlated observations to the atmosphere-like Lorenz-96 model. The experiment results show that, the analysis error is reduced and the analysis sensitivity to observations is improved.

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

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

1. Introduction

In the mathematical and physical research fields, data assimilation is a powerful mechanism to estimate the true trajectory of a state variable, based on the effective combination of the dynamic forecast system (numerical model) and the observations (Miller et al. 1994). It can provide an analysis state, which is generally treated as the weighted average of the model forecasts and observations, and is more close to the true state than either of them. The weights are approximately proportional to the inverse of the corresponding covariance matrices (Talagrand 1997). Therefore, the performance of a data assimilation method significantly relies on whether the error covariance matrices are estimated accurately. If this is the case, the analysis state can be technically easily obtained by minimizing a cost function with many existing optimization methods (Reichle 2008).

Ensemble Kalman filter (EnKF) is a very practical ensemble based assimilation scheme, which 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 error, the sampling covariance matrix of the ensemble forecast states is usually an underestimate of the true forecast error covariance matrix. It can lead that the filter over trusts the model forecasts and excludes the observations, and can eventually result in the divergence of the filter (Anderson; Anderson 1999; Constantinescu et al. 2007; Wu et al. 2014). Therefore, using the inflation technique to enhance the estimate accuracy of the

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

forecast error covariance matrix becomes gradually important.

In early studies in forecast error inflation, researchers usually tune the inflation factor by repeated assimilation experiments and select the estimated inflation factor according to their experiences and prior knowledge (Anderson; Anderson 1999). Hence such experimental determining is very empirical and subjective. In later studies, the inflation factor can be estimated on-line based on the innovation statistic (observation-minus-forecast (Dee 1995; Dee; Silva 1999)) with some different conditions. The moment estimation can facilitate the calculation by solving an equation of the innovation statistic and its realization (Li et al. 2009; Miyoshi 2011; Wang; Bishop 2003). The maximum likelihood approach can obtain a better inflation but has to calculate a high dimensional matrix determinant (Liang et al. 2012; Zheng 2009). The Bayesian approach assumes a prior distribution for the inflation factor but limited to the spatially independent observational errors (Anderson 2007, 2009). This study seeks to address the estimation of the inflation factor from the point view of Cross Validation (CV).

In fact, the idea of Cross Validation (CV) is firstly involved in linear regression (Allen 1974) and smoothing spline (Wahba; Wold. 1975). It is a common approach that can be applied to estimate tuning parameters in generalized additive models, nonparametric regression and kernel smoothing (Eubank 1999; Gentle et al. 2004; Green; Silverman. 1994; Wand; Jones 1995). In cross validation, sample is cut into several smaller data subsets, and some of them are used for modeling and analysis while others are used for verification and validation. The widely used technique is to

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

remove only one data point and use the rest to estimate the value at this point so as to test the estimation accuracy, which is also called Leave-Out-One Cross Validation (Gu; Wahba 1991).

The basic motivation behind the Cross Validation is minimizing the prediction error at the sampling points. The Generalized Cross Validation (GCV) is the modified form of Cross Validation, which is more popular for choosing these turning parameters (Craven; Wahba 1979). For instance, Gu and Wahba applied the Newton method to optimize the Generalized Cross Validation score with multiple smoothing parameters in a smoothing spline model (Gu; Wahba 1991). Wahba briefly reviewed the properties of Generalized Cross Validation and carried out an experiment to choose smoothing parameters in the context of variational data assimilation schemes with Numerical Weather Prediction models (Wahba et al. 1995). Zheng and Basher also used Generalized Cross Validation in thin-plate smoothing spline modeling of spatial climate data and applied to south Pacific rainfalls (Zheng; Basher 1995). The Generalized Cross Validation criterion also has been found to possess several favorable properties, such as consistency of the relative loss (Gu 2002).

Intuitively, if the forecast error matrix is inflated properly, the assimilation procedure can reassign the weights of the model forecasts and observations. Therefore the analysis sensitivity is also investigated in this study. Generally speaking, analysis sensitivity is how uncertainty in the output can be apportioned to different source of uncertainty in the input (Saltelli et al. 2004; Saltelli et al. 2008). The quantity can be introduced to the context of statistical data assimilation framework. It indicates that the sensitivity of analysis to observations, which is complementary to the sensitivity of analysis to model forecasts (Cardinali et al. 2004; Liu et al. 2009)..

This study focuses on the methodology part that can be potentially applied in the geophysical research fields in the near future. This paper consists of 4 sections. The conventional EnKF scheme is summarized and the improved EnKF with forecast error inflation scheme is proposed in Section 2. The verification and validation are conducted on an idealized model in Section 3. Discussion and conclusions are given in Section 4.

2. Methodology

2.1. EnKF Algorithm

For the sake of consistency, a nonlinear discrete-time dynamic forecast and linear observation system can be expressed as (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 stands for the time index; $\mathbf{x}_i^{\mathrm{t}} = \left\{ \mathrm{x}_{i,1}^{\mathrm{t}}, \mathrm{x}_{i,2}^{\mathrm{t}}, ..., \mathrm{x}_{i,n}^{\mathrm{t}} \right\}^{\mathrm{T}}$ is the n-dimensional true state vector at i-th time step; $\mathbf{x}_{i-1}^{\mathrm{a}} = \left\{ \mathrm{x}_{i-1,1}^{\mathrm{a}}, \mathrm{x}_{i-1,2}^{\mathrm{a}}, ..., \mathrm{x}_{i-1,n}^{\mathrm{a}} \right\}^{\mathrm{T}}$ is the n-dimensional analysis state vector which is an estimate of $\mathbf{x}_{i-1}^{\mathrm{t}}$, M_{i-1} is a nonlinear dynamic forecast operator such as a numeric weather prediction model; $\mathbf{y}_i^{\mathrm{o}} = \left\{ \mathrm{y}_{i,1}^{\mathrm{o}}, \mathrm{y}_{i,2}^{\mathrm{o}}, ..., \mathrm{y}_{i,p_i}^{\mathrm{o}} \right\}^{\mathrm{T}}$ is a p_i -dimensional observation vector; $\mathbf{H}_i$ is the

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

observation operator matrix, $\boldsymbol{\eta}_i$ and $\boldsymbol{\varepsilon}_i$ are 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 state $\mathbf{x}_i^{\mathrm{a}}$ that are sufficiently close to 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 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 dynamic model forecast 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 to be 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

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

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

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). In order to complete the ensemble forecast, the perturbed analysis state are calculated using perturbed observations (Burgers et al. 1998), that is

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

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 (Liang et al. 2012; Tippett et al. 2003). Finally, set $i = i + 1$ and return to Step 1 for the model forecast at next time step.

2.2. Influence matrix and forecast error inflation

It is recognized that the forecast error inflation scheme should be included in any ensemble based assimilation scheme, otherwise, the filter could diverge (Anderson; Anderson 1999; Constantinescu et al. 2007). The multiplicative inflation is one of the commonly used inflation techniques, which adjusts the initially estimated forecast error covariance matrix $\mathbf{P}_i$ to $\lambda_i\mathbf{P}_i$ and then estimates the inflation factors λ_i properly.

In previous studies, there are many methods for estimating the inflation factor, such as the maximum likelihood approach (Liang et al. 2012; Zheng 2009), moment approach (Li et al. 2009; Miyoshi 2011; Wang; Bishop 2003) and Bayesian approach

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access

(Anderson 2007, 2009). In this study, a new procedure for estimating the multiplicative inflation factors λ_i is proposed based on the following Generalized Cross Validation (GCV) function (Craven; 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 estimation procedure of inflation factors λ_i is implemented between the Step 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 forecast error inflation based on GCV method is shown in Figure 1.

2.3. Analysis sensitivity

In EnKF, the analysis state (Eq. (6)) can be treated as 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

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

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

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 as

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

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

respectively, 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 off-diagonal elements are 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 vector. The two quantities are complementary.

In fact, the sensitivity matrix $\mathbf{S}_i^{\mathrm{o}}$ is equal to the influence matrix $\mathbf{A}_i$ (see Appendix B for detail proof), whose trace can be used to measure the "equivalent number of parameters" or "degrees of freedom for signal". Similarly, it can be interpreted as a measurement of the amount of information extracted from the observations. The trace diagnostic can be used to analyze the sensitivities to observation or forecast vector (Cardinali et al. 2004). The Global Average Influence (GAI) at *i*-th time step is defined as the globally averaged observation influence, that is

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

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

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 the short-term ensemble forecast states. However, due to the limited ensemble size and large model error, the sampling covariance matrix of perturbed forecast states is usually an underestimation of the true forecast error covariance matrix. This will cause the assimilation systems over trust the forecast state, and then exclude the observational information. That is why the values of GAI are too small in conventional EnKF scheme. It will show that in simulations, this problem will be alleviated to some extent through the inflation adjustment on forecast error covariance matrix.

2.4 Analysis RMSE

In the following experiments, the "true" state $\mathbf{x}_i^{\mathrm{t}}$ is non-dimensional and can be obtained by 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 i-th time step is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\mathrm{x}_{i,k}^{\mathrm{a}} - \mathrm{x}_{i,k}^{\mathrm{t}}\right)^2}\,. \tag{17}$$

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

3. Experimentations

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Open Access
Discussions

The proposed data assimilation scheme is validated using the Lorenz-96 model (Lorenz 1996) with model error and a linear observation system as a test bed in this section. The performances of the assimilation schemes described in Section 2 are evaluated through the following experiments.

3.1. The dynamic forecast model and observation systems

The Lorenz-96 model (Lorenz 1996) is a quadratic nonlinear dynamical system, which has the properties relevant to realistic forecast problems and is governed by the equation

$$\frac{d\boldsymbol{X}_k}{dt} = (\boldsymbol{X}_{k+1} - \boldsymbol{X}_{k-2})\boldsymbol{X}_{k-1} - \boldsymbol{X}_k + F\,, \tag{18}$$

where $k = 1, 2, \cdots, 40$. The cyclic boundary conditions $\boldsymbol{X}_{-1} = \boldsymbol{X}_{K-1}$, $\boldsymbol{X}_0 = \boldsymbol{X}_K$, $\boldsymbol{X}_{K+1} = \boldsymbol{X}_1$ is applied to make Eq. (18) to be well-defined for all values of *k*. The Lorenz-96 model is “atmosphere-like”, since the three terms on the right-hand side of Eq. (18) can be analogized to a nonlinear advection-like term, a damping term, and an external forcing term respectively. It can be thought of as some atmospheric quantity (e.g. zonal wind speed) distributed on a latitude circle. Therefore the Lorenz-96 model is widely used as a test bed to evaluate the performances of assimilation schemes in many studies (Wu et al. 2013).

The time step is set as 0.05 non-dimensional unit when generate the numeric solution, 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).

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access

The true state is derived by a fourth-order Runge-Kutta time integration scheme (Butcher 2003). The forcing term is set as F = 8, so that the leading Lyapunov exponent implies an error-doubling time of about 8 time steps, and the fractal dimension of the attractor is 27.1 (Lorenz; Emanuel 1998). The initial value is chosen to be $\boldsymbol{X}_k = F$ when $k \neq 20$ and $\boldsymbol{X}_{20} = 1.001F$.

In this study, the synthetic observations are assumed to be generated at all of the 40 model grids but every 4 time steps by adding random noises which are multivariate-normally distributed with mean zero and covariance matrix $\mathbf{R}_i$ to the true states. The observation errors are assumed to be spatially correlated, which is the most cases in remote sensing observations and radiances data. The variance of the observation on each grid point is set $\sigma_o^2 = 1$ and the covariance of the observations between the j-th and k-th grid points is

$$\mathbf{R}_i(j,k) = \sigma_o^2 \times 0.5^{\min\{|j-k|, 40-|j-k|\}}. \tag{19}$$

The heat map of the observation error covariance matrix is shown in Figure 2.

3.2. Assimilation schemes comparison

Since model error is inevitable in practical dynamic forecast models, it is reasonable for us to add model error to the Lorenz-96 model in the assimilation process. Lorenz-96 model is a forced dissipative model with a parameter F that controls the strength of the forcing (Eq. (18)). The model forecast changes very much along with the change of F and is chaos with integer values of F larger than 3. Therefore the forcing term is set as 7 to simulate the range of model error in the assimilation schemes while retaining $F = 8$ when generating the "true" state. The ensemble size is selected as 30.

To evaluate the analysis sensitivity, the GAI statistics (Eq. (16)) are calculated and the results are plotted in Figure 3 over 2000 time steps, which is about equivalent to 500 days in realistic problems. It clearly shows that, the percentage of the analysis result relied on the observation is about 10% for the conventional EnKF, which is increased to about 30% for the EnKF with forecast error inflation.

To evaluate the assimilation result, the analysis RMSE (Eq. (17)) and the corresponding values of the GCV functions (Eq. (9)) are calculated and plotted in Figures 4 and 5, respectively. It illustrates that, the analysis RMSE, as well as the values of the GCV functions increase sharply if the forecast error inflation is adopted. The variety of the analysis RMSE is very consistent with that of the value of the GCV function for the EnKF with forecast error inflation scheme. The correlation coefficient of the analysis RMSE and the value of the GCV function at the assimilation time step is about 0.76, which indicting that, the GCV function seems to be a good criterion to estimate the inflation factor.

The time-mean values of 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 can indeed increase the analysis sensitivity to the observations and reduce the analysis RMSE.

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

4. Discussion and Conclusions

As we all know that accurately estimating the error covariance matrix is one of the most important steps in data assimilation, which has curial influence to the assimilation results. In conventional EnKF assimilation scheme, the forecast error covariance matrix is estimated as the sampling covariance matrix of the ensemble forecast states. But due to the limited ensemble size and large model error, this initial estimate is usually an underestimation, which can lead to that the filter over trusts the forecasts and excludes the observations, and eventually the divergence of the filter. Therefore the forecast error inflation with proper inflation factor is increasingly important.

The multiplicative inflation is an effective inflation technique and the inflation factor can be estimated under different assumptions. 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 with complicated calculations of high dimensional matrix determinant. The Bayesian approach assumes a prior distribution for the inflation factor but limited to the spatially independent observational errors. In this study, the inflation factor is estimated from the point of view of cross validation and the analysis sensitivity is detected.

The GCV function seems to be a good objective function that can well quantify the goodness of fit of the error covariance matrix. In fact, cross validation, which can

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access

evaluate and compare learning algorithms, is a widely used statistical method. The most common form of cross validation is leave-out-one cross validation. For this algorithm, all the data except for a single observation are used for training and the comparison is made on that single observation. Generalized Cross Validation estimate is a modified form of ordinary Cross Validation, which has the rotation-invariant property relative to an orthogonal transform of the observations and is a consistent estimate of the relative loss.

In this study, the idea of Cross Validation is adopted to the inflation factor estimation in the improved EnKF assimilation scheme and validated on the Lorenz-96 model. The values of the GCV function obviously decrease in the proposed approach comparing with that in the conventional EnKF scheme. The analysis RMSE in the proposed approach also is much smaller than that in the conventional EnKF scheme. This suggested that the estimate inflation factor method through minimizing the GCV function works very well.

The varieties of analysis sensitivity in the proposed approach and in the conventional EnKF scheme are also investigated in this study. The influence matrix is treated as the partial derivative of the normalized analysis state vector to the normalized observational vector, which is also used in the GCV function. Additionally, the time-mean Global Average Influence statistic is increased from about 10% in the conventional EnKF scheme to about 30% in the proposed improved EnKF assimilation scheme. This illustrated that the shortcoming of the assimilation result excessively depending on the forecast and excluding the observation can be

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

mitigated in some extent. The relations of analysis state to forecast state and observation are more reasonable.

It is notable that, the inflation factor is assumed to be constant in space in this study, which may be not the case in the realistic assimilation problems. Therefore persistently adjust all the state vectors using the same inflation factor could systematically overinflate the ensemble variances in sparsely observed areas, especially when the observations are unevenly distributed. In the case studies conducted in Section 3, the observations are relatively evenly distributed and the assimilation accuracy can be indeed improved by the forecast error inflation technique. It mainly sheds light on the methodology and validate on Lorenz-96 model to illustrate the feasibility in this study. In the near future, it will investigated that how to modify the adaptive procedure to suit the system with unevenly distributed observations and apply the proposed methodologies using more sophisticated dynamic and observation systems.

Appendix A

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

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

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 dimension $p_i \times p_i$. 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, ignoring the normalized forecast state in the

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

observation space (Gu 2002), i.e.,

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

Since the analysis state $\mathbf{x}_i^{\mathrm{a}}$ is given by Eq. (5), it can be easily checked that the influence matrix $\mathbf{A}_i$ is given by

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

If the initial forecast error covariance matrix is inflated as described in Section 2.2, 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 cross validation aims at minimizing the estimated error at the observation grid point. Lacking an independent validation data set, the alternative strategy commonly used 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, that is the following objective function

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

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

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

The subscript $-k$ means 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)

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

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

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Open Access
Discussions

substituting $a_{k,k}$ by 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 ignoring the constant to get the following generalized cross validation (GCV) statistic (Gu 2002)

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

Appendix B

The sensitivities of the analysis to the observation is defined as

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

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 Basic Research Program of China (Grant Nos. 2015CB953703 and 2010CB950703) and the National Natural Science Foundation of China (Grant No. 41405098).

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.

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.

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.

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.

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.

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.

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics Discussions

Table 1. The time-mean values of the GAI statistics, the GCV functions and the analysis RMSE over 2000 time steps.

EnKF schemes	Conventional EnKF	EnKF with forecast inflation
GAI	10.78%	29.21%
GCV	31.14	3.29
RMSE	4.01	1.10

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Figure captions

Figure.1. Flowchart of the proposed assimilation scheme.

Figure 2. The heat map of the observation error covariance matrix.

Figure 3. The GAI statistics of the conventional EnKF scheme and the improved EnKF with forecast error inflation scheme.

Figure 4. The analysis RMSE of the conventional EnKF scheme and the improved EnKF with forecast error inflation scheme.

Figure 5. The values of the GCV functions of the conventional EnKF scheme and the improved EnKF with forecast error inflation scheme.

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access

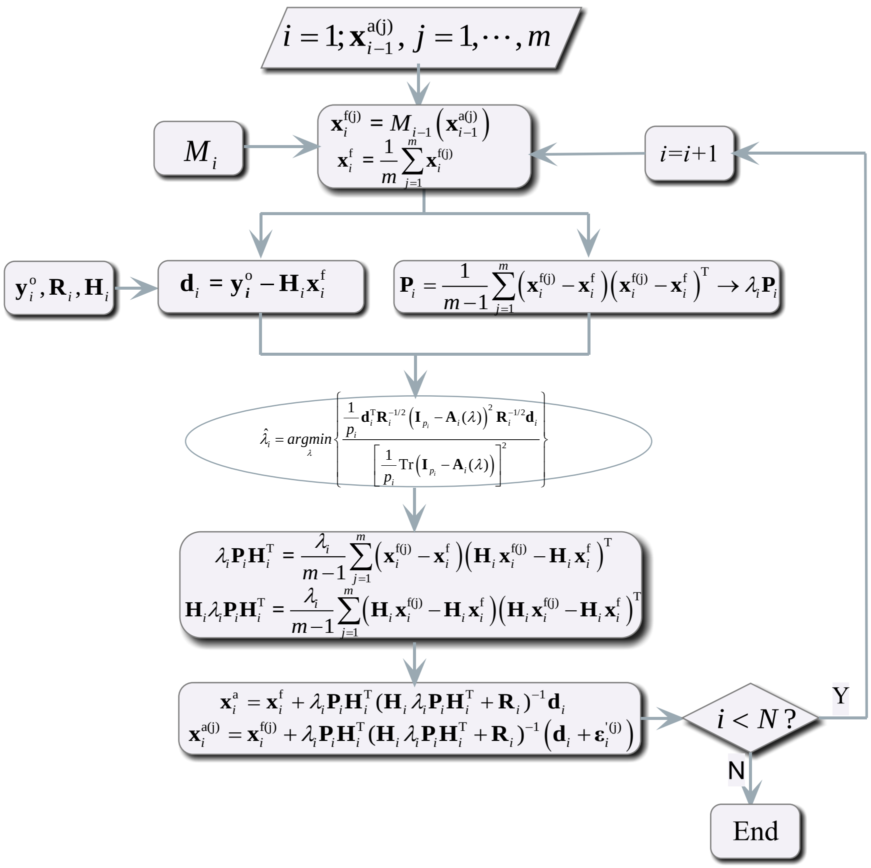


Figure.1. Flowchart of the proposed assimilation scheme.

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access

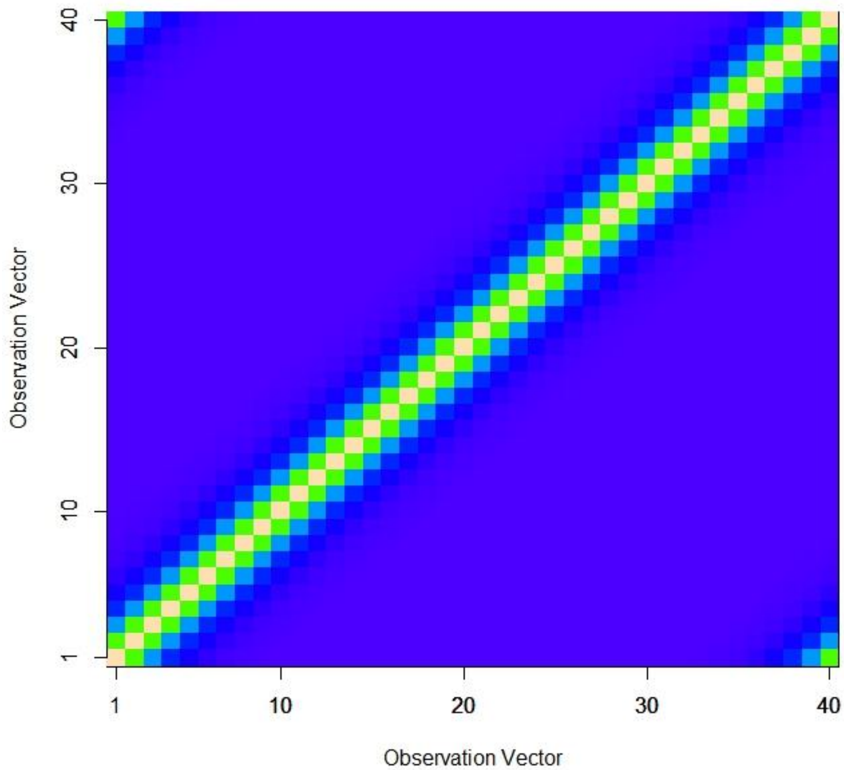
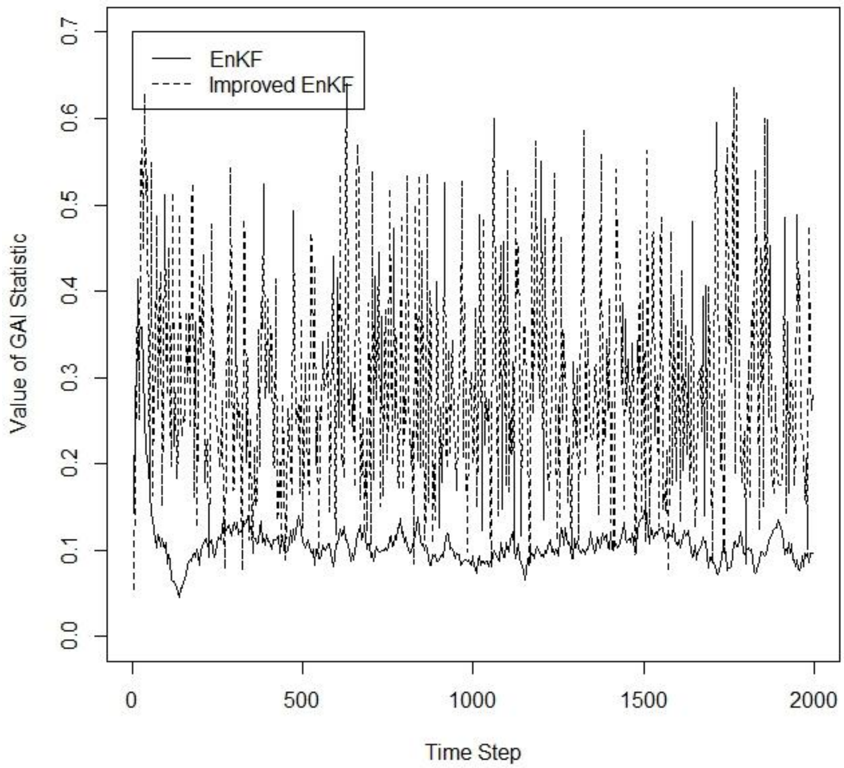


Figure 2. The heat map of the observation error covariance matrix.

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Open Access
Discussions



1

2 Figure 3. The GAI statistics of the conventional EnKF scheme and the improved

3 EnKF with forecast error inflation scheme.

4

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access

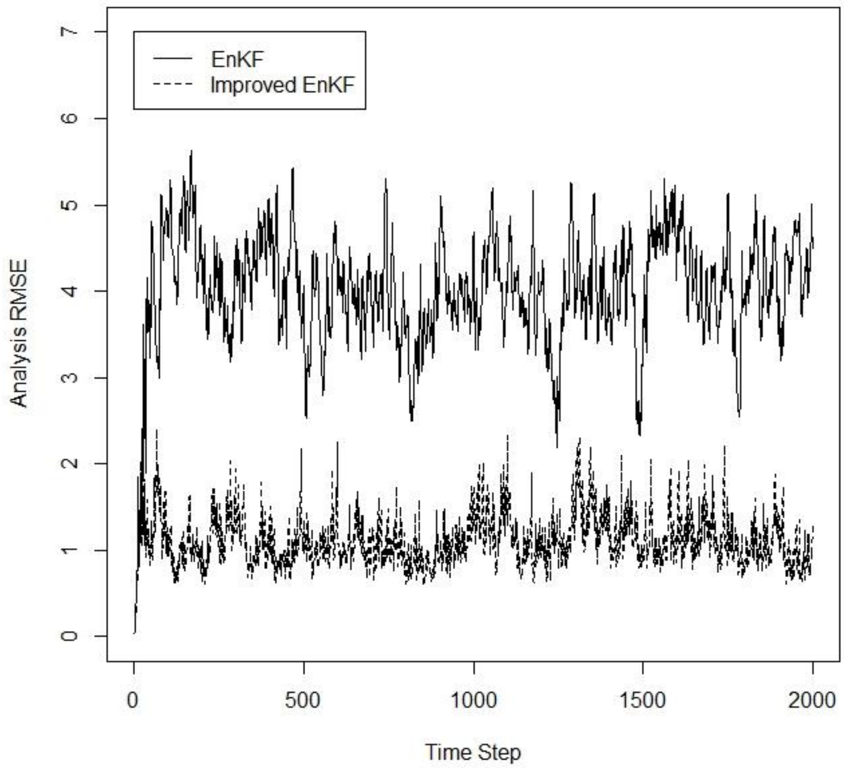


Figure 4. The analysis RMSE of the conventional EnKF scheme and the improved EnKF with forecast error inflation scheme.

Nonlin. Processes Geophys. Discuss., doi:10.5194/npg-2016-44, 2016
Manuscript under review for journal Nonlin. Processes Geophys.
Published: 4 October 2016

Nonlinear Processes in Geophysics
Discussions
Open Access

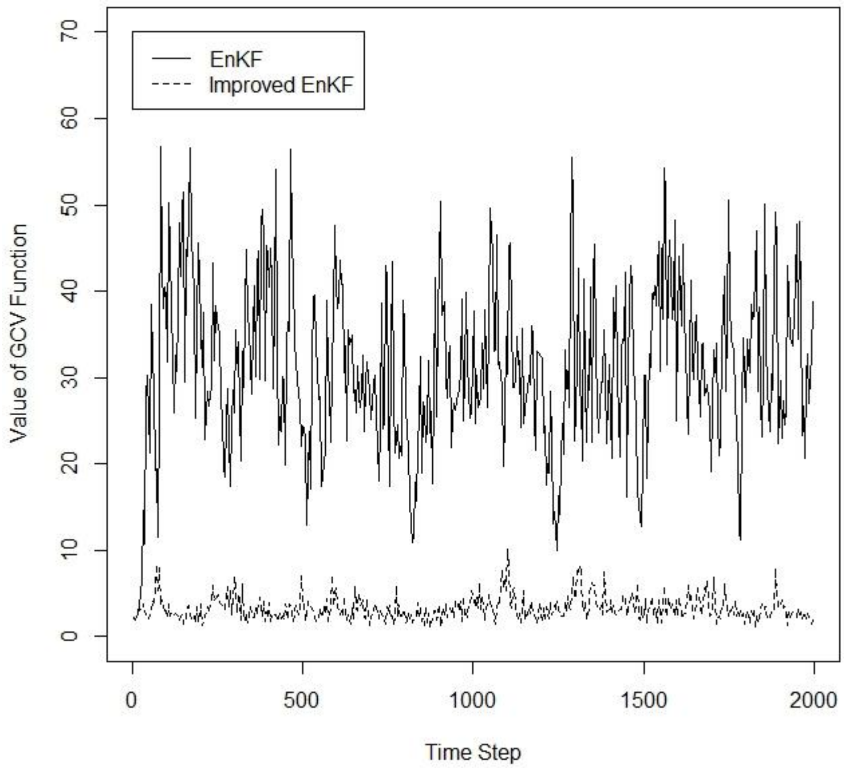


Figure 5. The values of the GCV functions of the conventional EnKF scheme and the improved EnKF with forecast error inflation scheme.