The author gratefully acknowledges the two anonymous reviewers for their insightful comments and thorough corrections that lead to the significant improvement of the quality of this manuscript. The author has checked the manuscript carefully and tried the best to address all the comments.

The main changes in the revised version are: add the spread statistic, comparison of schemes, influences of ensemble size and observation number and the discussion of computational cost and sparse observations. The scientific discussions/explanations and the English grammar are modified.

Below, the italic is used for quoting the comments from the reviewer and following with the point-by-point responses. Next is the list of all relevant changes made in the manuscript, and a marked-up manuscript version with the changes in blue text.

Reviewer 1

General comments:

This paper introduces a new technique for estimating the covariance inflation factor needed to help mitigate the problem of filter divergence often encountered when using EnKF data assimilation methods. Their approach is novel in the sense that computes these estimates by minimizing an objective function based on the concept of generalized cross validation, a technique commonly used in the machine learning literature. There is considerable overlap between the fields of data assimilation and machine learning and I appreciate that the author is trying to bridge the gap between these two fields. In my opinion, there is much that we can learn from one another.

However, the paper falls short of offering good comparisons for their new technique. Rather, it is more a proof-of-concept that this technique works better than the basic EnKF, which is known not to work well without inflation. There are a few additional questions that I would like to see answered:

Response: Thank you for your review and constructive comments.

1. How does it depend on the ensemble size and the number of observations?

Response: Thank you for your comment. These dependences are explored and discussed. The following texts have been added in the revised version as section 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 with respect to different ensemble sizes (10, 30 and 50) were evaluated, and the results are listed in Table 1, which shows that 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.

2. How does it compare with other the inflation schemes mentioned? What are the computational tradeoffs?

Response: The comparisons with the constant inflated EnKF are as follows, and have been added to section 3.2 in the revised version.

The constant was particularly selected as the median of the estimated inflation factor by minimizing the GCV function. In addition to small fluctuation, the mean GAI value of the constant inflated EnKF was 27.80%, which was smaller than that of the improved EnKF. The mean spread value of improved EnKF was 3.32, which is slightly 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.

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

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

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 (Golub; Loan 1996). 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 by minimizing the GCV function. Therefore, the total computational costs of the improved EnKF are feasible.

3. What does the time series of the inflation factor look like? Is it smooth?

Response: The time series of estimated inflation factors are shown in Figure 2, which vary between 1 and 6 with greatly majority. The median was 1.88, which was used in the following comparison of the improved EnKF and the simple multiplicative inflation techniques like setting a constant inflation factor.

This has been added to section 3.2 in the revised version.

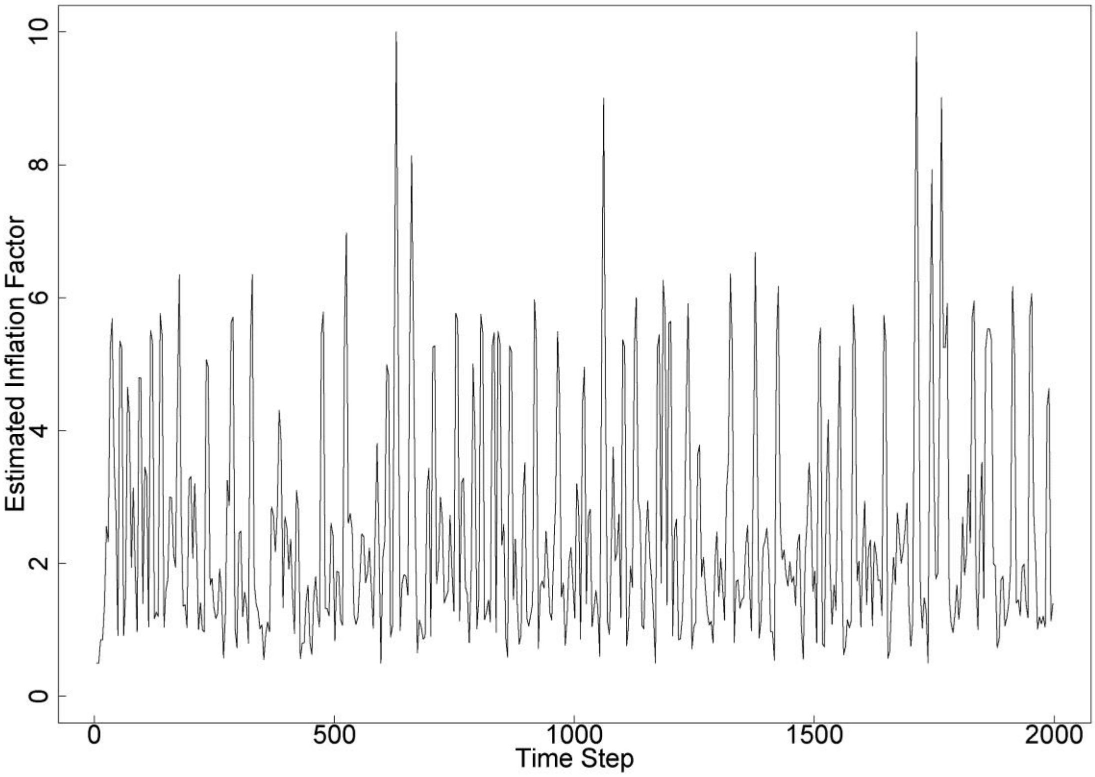

Figure 2. Time series of estimated inflation factors by minimizing GCV function.

4. Does it prevent the problem of ensemble divergence?

Response: The ensemble analysis state members of the conventional EnKF, improved EnKF and constant inflated 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, and 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 forecast error inflation can lead to a more accurate analysis state than the constant inflated EnKF.

This has been added to section 3.2 in the revised version.

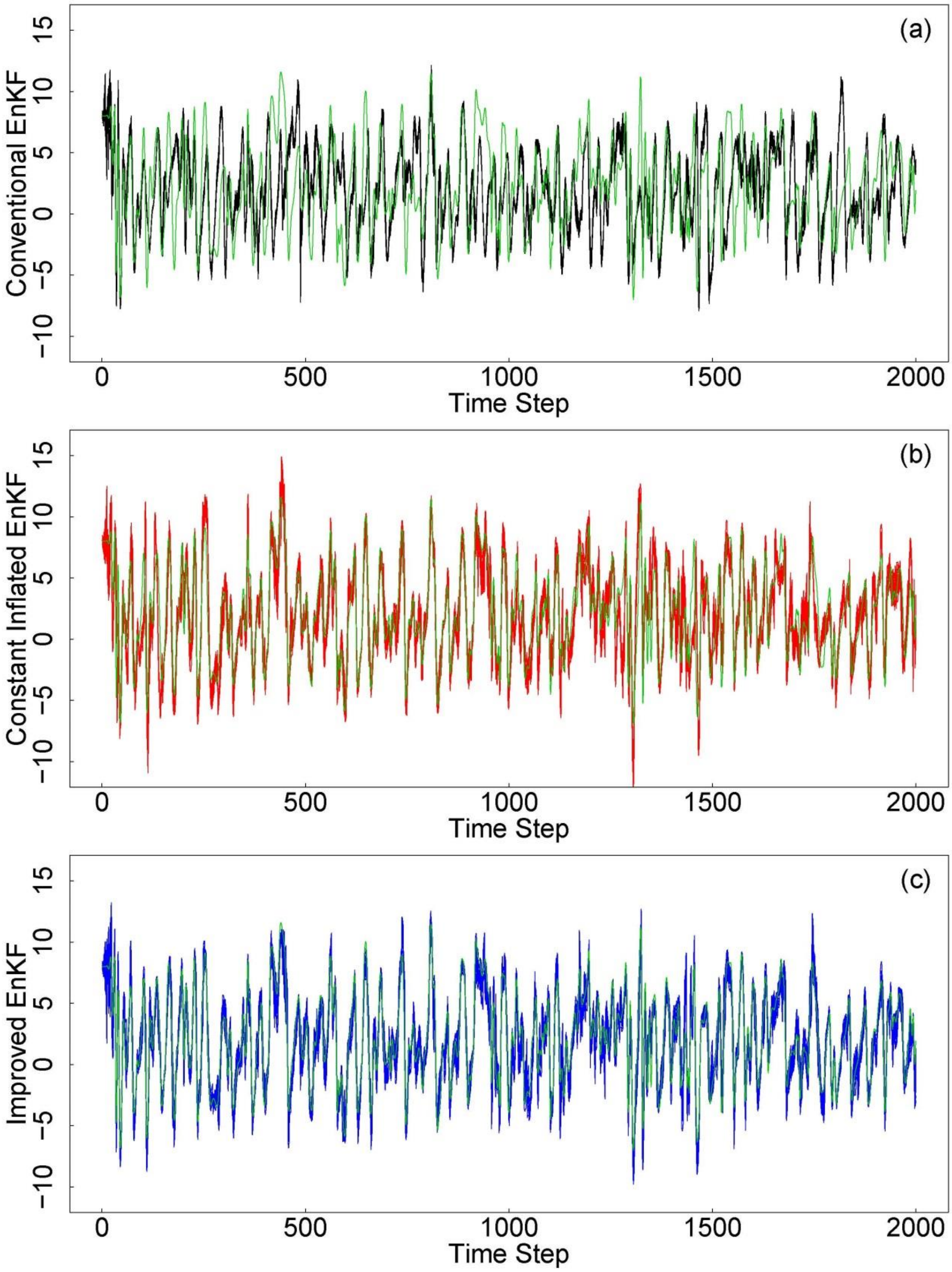

Figure 7. Ensemble analysis state members of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line). The green line refers to the true trajectory obtained by the numerical solution.

I understand that some of these questions fall under the future work category. But I do think it should be explicitly stated that this paper is intended just to show proof-of-concept, and that a more thorough comparison will be forthcoming in the near future.

Response: Thank you for your comment. Some comparisons of these questions have been added in the revised version. The more thorough comparison and application will be conducted in the near future studies.

The paper also needs considerable grammatical revision. In my following comments, I have tried to be as explicit as possible in offering suggestions.

Response: Thank you for your comment. The grammars have been checked carefully and the language has been polished in the revised version.

Specific comments:

1. P3 L3-13:

The author implicitly assumes the existence of a 'true' underlying state of the system. While this assumption is common, it is still an assumption. Also, be careful about saying: "are more close to the true state than either of them..." and "can be technically easily obtained by minimizing a cost function...". The former is not necessarily true, the latter depends on what you mean by 'easy'. While it is easy enough to run an optimization algorithm to minimize the cost function, you have no guarantees that the solution is unique when the models are nonlinear. Finding the most appropriate analysis state (i.e the global minimum) is a much more difficult problem.

Response: Thank you for your comments. We agree that the existence of a "true" state is a common assumption and finding the global minimum is a very difficult problem. This paragraph has been written as follows:

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 assimilations provide an analysis state that is usually a better estimate of the state variable because it fully 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 attributed to technical aspect and can be accomplished with the rapid development of supercomputers (Reichle 2008), although finding the global minimum is a much more difficult problem when the models are nonlinear.

2. P5 L1: Would leave-one-out cross validation be applicable to data assimilation, where the data is a time-series?

Response: Since EnKF is a sequential assimilation method, the observations at current step (40-dimensioal vector in the experiments of this manuscript) are assimilated to the forecast model for a given assimilation step. Therefore the cross validation could also be applicable.

3. P5 L4-16: Why is Generalized Cross Validation better? What makes it generalized? What are these "favorable properties" of "consistency of the relative loss"? I understand it has not been used much in data assimilation, but I think this should be more explicitly motivated, as it is the core method of this article.

Response: The 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; Wahba 1979). This criterion has been widely used in the linear regression and smoothing spline fields (Allen 1974; Gu; Wahba 1991; Wahba; Wold 1975; Wahba et al. 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). For the problem of estimating the inflation factor in this study, the objective function based on CV principle is (the detailed derivation is listed in Appendix A)

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

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

$$GCV_i(\lambda) = \frac{\frac{1}{p_i}\mathbf{d}_i^{\mathrm{T}}\mathbf{R}_i^{-1/2}\left(\mathbf{I}_{p_i} - \mathbf{A}_i(\lambda)\right)^2\mathbf{R}_i^{-1/2}\mathbf{d}_i}{\left[\frac{1}{p_i}\mathrm{Tr}\left(\mathbf{I}_{p_i} - \mathbf{A}_i(\lambda)\right)\right]^2} \tag{3}$$

It is easy to see that the GCV criterion is a weighted version. Originally proposed to reduce the computational burden, GCV is one of a number of criteria which all involve an adjustment to the average mean-squared-error over the training set (Craven; Wahba 1979).

4. P5 L17-19: I don’t understand these statements. What does it mean to be “inflated properly”? How does it “reassign the weights”? The segway to analysis sensitivity does not follow logically to me.

Response: In the covariance inflation scheme, the forecast error matrix is multiplied by an appropriate inflation factor. Usually, the inflation factor is larger than 1. Too small or too large an inflation factor will cause the analysis state to be over reliant on model forecasts or observations. Hence, the inflation factor should be estimated accurately. After this, the weights of the model forecasts and observations in the analysis state can be reassigned.

These explanations have been written in the revised version.

5. P9 L9: Please specifically state how you actually compute lambda_i. I assume you minimize the GCV as an objective function?

Response: Yes, this has been specifically stated in the revised version.

6. P10: If S^f=I-S^o=I-A_i, then can't the GCV_i function be interpreted as minimizing the normalized forecast sensitivity?

Response: The GCV function can be interpreted like this, because the aim of inflation scheme is increasing the observation weight appropriately. Since the sensitivities of analysis state to the model forecasts and observations are complementary, it will decrease the normalized forecast sensitivity.

This has been added to the revised version.

7. P14 L2: Any motivation for setting the ensemble size at 30?

Response: The ensemble size in the practical land surface assimilation problem is usually several tens of members (Kirchgessner et al. 2014). Too large ensemble size will significantly increase the computational cost. Therefore the ensemble size in this study in selected as 30 to ensure stable results.

The explanation has been added in the revised version. Also following your general comment 1, the inflation scheme with different ensemble size (10, 30 and 50) is investigated. Please see the response to the general comment 1.

8. P14 L15: Are there other examples in the literature to compare this correlation coefficient to? Should ideally it be as close to 1 as possible?

Response: There some comparisons between the analysis RMSE and the objective function, such as (Liang et al. 2012; Zheng 2009). This correlation coefficient is an indicator that can show whether the choice of the objective function is appropriate. In the ideal case, it is as close to 1 as possible.

9. P15 L22: Can you be more precise than this: "seems to be a good objective *function"?*

Response: The sentence has been changed to "The assimilation results showed that inflating the conventional EnKF using the factor estimated by minimizing the GCV function can indeed reduce the analysis RMSE".

Technical corrections:

1. Title: An estimate of ***the*** *inflation factor and analysis sensitivity in* ***the*** *ensemble Kalman filter*

Response: The correction has been followed.

2. P2 L7: Why does it "need" to be inflated? What happens otherwise?

Response: 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.

This has been added to the abstract of the revised version.

3. P2 L10: I would say the method is "tested" not "validated". Validation to me implies a more thorough comparison.

My suggestion for the 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…?

In this study, the forecast error covariance inflation factor is estimated using a generalized cross-validation technique. The improved EnKF assimilation scheme is tested with the atmosphere-like Lorenz-96 model with spatially correlated

observations, and is shown to reduce both the analysis error and its sensitivity to the observations.

Response: Thank you for your suggestion. The abstract of the revised version has been rewritten as follows:

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 with 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.

4. P3 L3-13: This paragraph needs some revision. See scientific comment 1 above.

Response: This paragraph has been written in the revised version as follows:

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 assimilations provide an analysis state that is usually a better estimate of the state variable because it fully 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 attributed to technical aspect and can be accomplished with the rapid development of supercomputers (Reichle 2008), although finding the global minimum is a much more difficult problem when the models are nonlinear.

5. P4 L1: What does it mean "gradually important"? I also think this needs a more motivation about what why inflation is used. Assume the reader has never used an EnKF before.

Response: It means researchers realize that, the covariance inflation is becoming more and more important. The following texts have been added in the revised version.

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.

6. P4 L1-15: Past tense seems more appropriate here: "tune" > tuned, "select" > selected.

Response: The words have been changed.

7. P4 L5: ***However, such methods are very empirical and subjective****.*

Response: The correction has been followed.

8. P4 L8: How does moment estimation "facilitate the calculation"?

Response: The moment estimation just obtains the estimated inflation factor by solving an equation of the innovation statistic and its realization. It does not need expensive calculation such as the determinant of high dimensional matrix in the maximum likelihood estimation. Therefore the moment estimation can facilitate the calculation.

9. P4 L10: "obtain a better ***estimate of the inflation factor****, but…"*

Response: The correction has been followed.

10. P4 L16: ***The idea of cross validation was first introduced in linear regression and spline smoothing****.*

Response: The correction has been followed.

11. P4 L20: ***In cross validation, the data is divided into subsets, some of which are used for modeling and analysis while others are used for verification and validation****.*

Response: Thank you for your suggestion. The correction has been followed.

12. P5 L21: "sources"

Response: The word has been changed.

*13. P5 L22: Replace "The quantity can be introduced…" with: "****In the context of statistical data assimilation, this quantity describes the sensitivity of the analysis to the observations, which is complementary…****"*

Response: Thank you for your suggestion. The sentence has been replaced.

14. P6 L3: ***This study focuses on methodology that can be potentially applied to geophysical applications of data assimilation in the near future****.*

Response: The correction has been followed.

15. P6 L14: ***dynamical forecast model***

Response: The correction has been followed.

16. P7 L4: series of analysis ***states***

Response: The word has been corrected.

17. P8 L16: "~~The~~ multiplicative inflation"

Response: "The" has been deleted.

18. P8 L18: ***by estimating the inflation factors \lambda_i***

Response: The correction has been followed.

19. P9 L16: In **the** *EnKF, "can be treated" >* ***is***

Response: The correction has been followed.

20. P9 L17: and forecast. ***That is,***

Response: The correction has been followed.

21. P10 L14: ***detailed*** *proof*

Response: The word has been corrected.

22. P10 L14-15: Why quotes? Should there be a reference?

Response: Yes, the references (Gu 2002; Pena; Yohai 1991) have been added.

23. P10 L15: "degrees of freedom for **the** *signal"*

Response: The correction has been followed.

24. P10 L16: Reference for its interpretation as "amount of information"? Is this heuristic or in an information theoretic sense?

Response: The phrase "amount of information" is in an information theoretic sense. The reference (Ellison et al. 2009) has been added.

*25. P11 L5: … **states usually underestimate** the true forecast…*

Response: The correction has been followed.

*26. P11 L6-10: **This will cause the analysis to over rely on the forecast state, excluding useful information from the observations. This is captured by the fact that for the conventional EnKF scheme the GAI values are rather small. Adjusting the inflation of the forecast error covariance matrix alleviates this problem to some extent, as will be shown in the following simulations**.*

Response: Thank you for your suggestion. The correction has been followed.

*27. P11 L22: I would say **Numerical Experiments***

Response: The correction has been followed.

28. P12 L2: validated > tested

Response: The word has been changed.

29. P12 L4: performance~~s~~

Response: The word has been changed.

30. P12 L12: Cyclic boundary conditions

Response: The correction has been followed.

31. P12 L13: ~~to be~~

Response: The words have been deleted.

32. P12 L15: are analogous to

Response: The correction has been followed.

33. P12 L18: performance~~s~~

Response: The word has been changed.

*34. P12 L20: **The time step for generating the numerical solution is set at 0.05 non-dimensional units, which is roughly**…*

Response: The correction has been followed.

35. P13 L1: I would maybe move this sentence up, before you discuss the time step.

Response: The sentence has been moved to the front of this paragraph.

*36. P13 L10: correlate, which is **common in applications involving remote sensing and radiance data**.*

Response: The correction has been followed.

*37. P13 L20-22: **Modifying the forcing strength F changes the model forecast considerably. For values of F larger than 3 the system is chaotic. To simulate model error, the forcing term for the forecast is set to 7, while using F=8 to generate the 'true' state**.*

Response: Thank you for your suggestion. The correction has been followed.

*38. P14 L5-7: **The increase in GAI from 10% for the conventional EnKF to 30% for the EnKF with forecast error inflation indicates that the latter relies more on the observations. This is important because…***

Response: Following your correction, the sentences have been rewritten as follows:

The increase in GAI from 10% for the conventional EnKF to 30% for the EnKF with forecast error inflation indicates that the latter relies more on the observations. This finding is important because the observations can play a more significant role in combining the results with the model forecasts to generate the analysis state.

*39. P14 L8: **To evaluate the resulting estimate, ...***

Response: The correction has been followed.

40. P14 L10-11. ... values of the GCV functions **decrease sharply** *... right?*

Response: Yes. The typo has been corrected.

41. P14 L12: I don’t understand this statement. Did you mean to say “The **variance** *of the analysis”?*

Response: The mean in the manuscript is “The variability of the analysis”.

42. P14 L15: ***... which indicates that the GCV function is a good criterion to estimate the inflation factor****.*

Response: The correction has been followed.

43. P15 L3-6: ***Accurate estimates of the forecast error covariance matrix are crucial to the success of any data assimilation scheme. In the conventional EnKF ...***

Response: The correction has been followed.

44. P15 L7-8: ***But limited ensemble size and large model error often cause it to be underestimated. This produces an analysis state that over relies on the forecast and excludes the observations, which can eventually cause the filter to diverge****.*

Response: The correction has been followed.

45. P15 L10: Begin new paragraph with this sentence. ***The use of multiplicative covariance inflation techniques can mitigate this problem to some extent. Several methods have been proposed in the literature, each with different assumptions. For instance, the moment****...*

Response: Thank you for your suggestion. The correction has been followed.

46. P15 L16: ... ***but requires computing high dimensional matrix determinants****.*

Response: The correction has been followed.

47. P15 L18: but **is limited to spatially independent** *...*

Response: The correction has been followed.

48. P15 L19: ***is estimated using generalized cross validation****.*

Response: The correction has been followed.

49. P16 L2-7: These sentences are perhaps better suited for the introduction, say on p5.

Response: Following your suggestion, these sentences have been moved to the introduction section.

50. P16 L11: ... compared with the conventional EnKF scheme.

Response: The correction has been followed.

51. P16 L13: ***This suggests that this method of minimizing the GCV works well for estimating the inflation factor****.*

Response: The sentence has been corrected.

52. P16 L15: What do you mean by "varieties"?

Response: The word has been deleted and the sentence is changed to "The analysis sensitivities in the proposed approach and in the conventional EnKF scheme…"

53. P16 L16: "The influence matrix…" this does not need to be restated.

Response: The sentence has been deleted.

54. P16 L19-22: ***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****.*

Response: Thank you for your suggestion. The correction has been followed.

55. P17 L1-2: What do you mean they are "more reasonable"?

Response: In the conventional EnKF the analysis over relies on the forecast state and excludes useful information form the observation (The GAI statistic is only about 10%). Adjusting the inflation of the forecast error covariance matrix in EnKF can increase the GAI to about 30%. Therefore more information form the observation is contained in the analysis.

56. P17 L3: ***It is also worth noting that the inflation****...*

Response: The correction has been followed.

57. P17 L4-5: ***Forcing all components of the state vector to use the same inflation factor could systematically overinflate the ensemble variances*** *...*

Response: The correction has been followed.

58. P17 L7: Start a new paragraph here. ***The examples shown here using the Lorenz-96 model illustrate the feasibility of this approach for using GCV as a metric to estimate the covariance inflation factor****.*

Response: Thank you for your suggestion. The correction has been followed.

Comments on Figures:

1. Figures 1 and 2 are not really that instructive for me.

Response: Figures 1 shows the detailed procedure of the assimilation scheme. The flowchart is elaborated exhaustively in section 2.1 and 2.2. Figure 2 has been deleted in the revised version.

2. Figures 3-5 would benefit from using different colors to distinguish between the traces.

Response: All figures of assimilation results have been plotted using different colors.

Again, thanks for your constructive comments and thorough corrections.

The references in this reply are listed as follows (Some of them are already in the original manuscript and some are newly added in the revised version).

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

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

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.*

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

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.*

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.*

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.*

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.*

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.*

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

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

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

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

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.*

Reviewer 2

1 General Comments

This paper presents a new technique in estimating model error covariance inflation factor which is widely used in ensemble-based filters. An inflated model error covariance is necessary to arrest the divergence of the filter. This paper relies on estimating the inflation factor from an objective function inspired from the domain of generalized cross validation (GCV) techniques widely used in the field of machine-learning. The author also shows that this method, in comparison to a basic Ensemble Kalman Filter, considerably improves the root-mean-squared error and enhances the influence of observations on the analysis when applied to the Lorenz 96 model.

However, it is well known that a basic Ensemble Kalman Filter (EnKF) falls short on many accounts and a mere improvement with respect to it does not give much credence to this new technique. Even introducing a simple constant multiplicative inflation factor to the basic EnKF considerably improves the analysis. The author should address the following questions:

Response: Thank you for your review and constructive comments.

1) How does this method fare when compared to simple multiplicative inflation techniques like setting a constant inflation factor in the basic ensemble Kalman filter? It will be more interesting to see this method pitted against other sophisticated inflation schemes.

Response: Thank you for your comment. The comparisons with the constant inflated EnKF are as follows, and have been added to section 3.2 in the revised version.

The improved EnKF was compared with the constant inflated EnKF in the revised version. The constant was particularly selected as the median of the estimated inflation factor by minimizing the GCV function. 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. The mean spread value of the improved EnKF was 3.32, which was slightly 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. 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 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.

2) How does the time-series of the inflation factor look like?

Response: The time series of estimated inflation factors are shown in Figure 2, which vary between 1 and 6 with greatly majority. The median was 1.88, which was used in the following comparison of the improved EnKF and the simple multiplicative inflation techniques like setting a constant inflation factor.

This has been added to section 3.2 in the revised version.

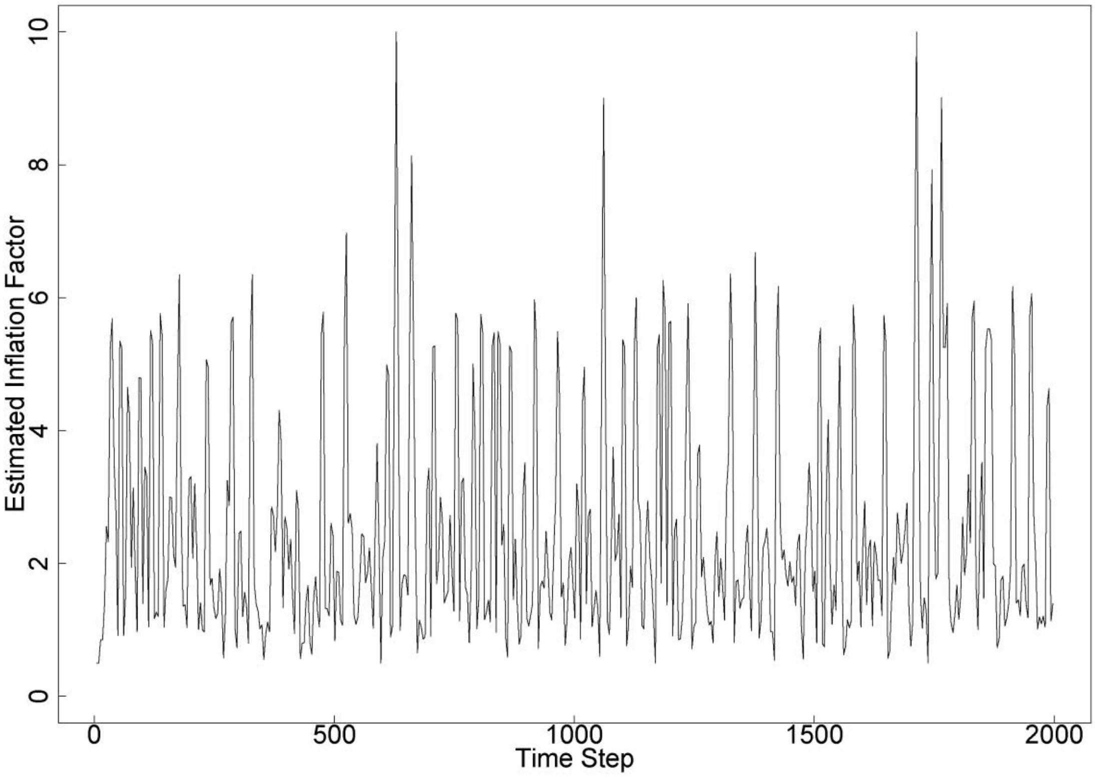

Figure 2. Time series of estimated inflation factors by minimizing GCV function.

3) What are the improvements in other statistical measures like spatial correlation, correlation coefficients, measure of ensemble spread, etc?

Response: The forecast ensemble spread of the conventional EnKF, improved EnKF and constant inflated 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 slightly 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.

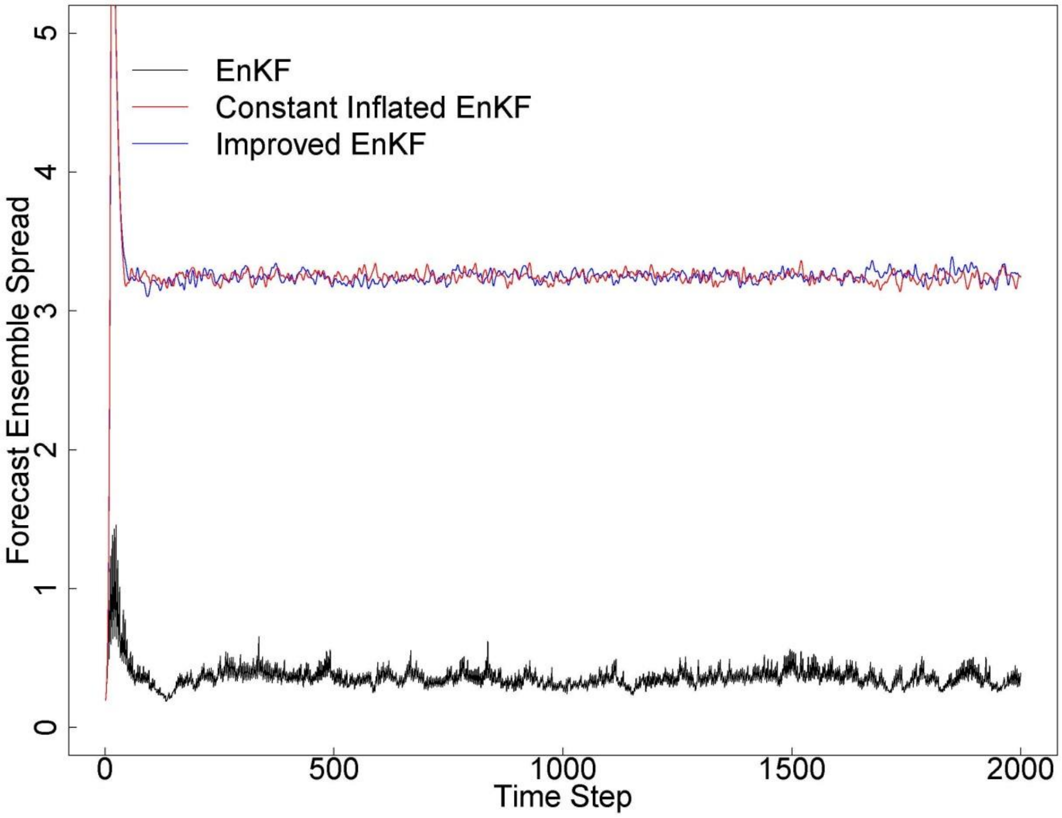

Figure 3. Forecast ensemble spread of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line).

4) What are the computational challenges in estimating the inflation factor?

Response: For the aspect of computational cost in minimizing the GCV function, t he highest computational cost when minimizing the GCV function is related to calculating the influence matrix $\mathbf{A}_i(\lambda)$. Because the matrix multiplication is commutative for the trace, the GCV function can be easily re-expressed as follows:

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

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 (Golub; Loan 1996). 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 by minimizing the GCV function. Therefore, the total computational costs of the improved EnKF are feasible.

5) How does this method fare in presence of sparse observations?

Response: Thank you for your comment. 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; Kunii 2011; Yang et al. 2015).

This discussion has been added to section 4 in the revised version.

It is clear that English is not the native language of the author. One of the suggestions is to seek help and revise the language of this paper. This paper, in its present version, is more of a curious exercise and needs to be considerably revised to make it *publishable. If these revisions are made, this paper may be accepted for publication.*

Response: Thank you for your comment. The grammars have been checked carefully and the language has been polished in the revised version.

2 Specific Comment

1) "The objective function needs to be minimized to estimate the inflation parameter" is not explicitly mentioned when it is introduced in the main text. It is however casually mentioned in the Discussion section.

Response: The inflation factor λ_i is estimated by minimizing the GCV (Eq. (9)) as an objective function. This has been explicitly mentioned in section 2.2 in the revised version.

2) What is the motivation of generating observations at every 4 time-steps?

Response: In realistic problems, the observation cannot be obtained every time. The time step for generating the numerical solution is 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). Therefore the frequency was set as every 4 time steps, which can be used to mimic daily observations in practical problems, such as satellite data.

This explanation has been added to section 3.1 in the revised version.

3) The simplest observation error covariance matrix is a diagonal R. There are many in-situ observation systems in which R is diagonal. What happens when R is diagonal? Also, what is the motivation behind choosing that particular expression of R? What is the harm in introducing a parameter in the expression of R which may be conveniently tuned to set R diagonal?

Response: In in-situ observation systems, the observation error covariance matrix is diagonal. But for remote sensing and radiances data, it is usually spatially correlated.

The correlation coefficient of two observation grids is inversely proportional to their distance. Therefore the particular expression of the observation error covariance matrix is chosen, which may potentially be applied to assimilate remote sensing observations and radiance data. The performances of these assimilation schemes are very similar in the case of diagonal *R*, which is a special form of the general case in the manuscript.

*4) P4 L1: What does the author mean by "**gradually important**"?*

Response: It means researchers realize that, the covariance inflation is becoming more and more important. The following texts have been added in the revised version.

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.

5) P5 L9-10: Kindly elaborate on what the "favorable properties" of GCV are?

Response: 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; Wahba 1979). 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).

*6) P6 L20-21: What does the author mean by "The EnKF assimilation result is . . . **sufficiently close** to the corresponding true state . . ."?*

Response: A common assumption is the existence of a "true" underlying state of the system for state variables in the geophysical research fields. Data assimilation is a powerful mechanism for estimating the true trajectory. The difference of the estimated vector to the truth (such as the root-mean-square-error used in the literature) is used to evaluate the accuracy of an assimilation scheme. The text “sufficiently close” means “accurate enough”. The sentence has been changed to “The EnKF assimilation result is . . . an accurate estimate of the corresponding true state …”

7) Please elaborate the flowchart in the main text as well.

Response: The flowchart is elaborated exhaustively in section 2.1 and 2.2 in the revised version. The inflation factor is estimated by minimizing the GCV function is specifically stated.

8) What is N in Fig 1?

Response: The scalar *N* is the total time step in the forecast procedure. This has been explained in section 3.1 in the revised version.

9) P13, L8-9: “The model forecast changes very much along with the change in F . . ..”. This is a very general statement. Please be a little more specific. Also cite references that show the model to be chaotic for F>3.

Response: The text has been changed to as follows:

Modifying the forcing strength *F* changes the model forecast considerably. For values of *F* that are larger than 3, the system is chaotic (Lorenz; Emanuel 1998).

3 Technical Comment

1) There are many notable absences of articles, usage of wrong prepositions and a few grammatical corrections to point out. One such instance is in P13 L8-9. “The model forecast . . . and is chaos with integer values of F larger than 3”. This should be “The model forecast . . . and is **chaotic for** *integer values of F>3”.*

Response: Thank you for your comment. The grammars have been checked carefully in the revised version.

2) P13 L22: "The variety of the analysis RMSE . . .". It is not clear what the author wants to convey.

Response: It has been changed to "The variability of the analysis RMSE is very consistent with that of the GCV function …".

Again, thanks for your constructive comments and helpful suggestions.

The references in this reply are listed as follows (Some of them are already in the original manuscript and some are newly added in the revised version).

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

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

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

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.*

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.*

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***.*

List of all relevant changes made in the manuscript (blue text in the following marked-up version)

P1: Change the title following the reviewer 1.

P2: Rewrite the abstract.

P3, L3-17: Add the common assumption and background of this field. Grammar modification.

P4, L5-9: Add a more motivation about what why inflation is used. Grammar modification.

P5: Grammar modification.

P6, L5-8: Add the explanation of the analysis weight reassignment. Grammar modification.

P7-9: Grammar modification.

P10 L3-4: Add explicitly mention of the inflation factor estimation. Grammar modification.

P11, L8-10: Add the explanation of GCV function. Grammar modification.

P12, L11-17: Add the spread statistic. Grammar modification.

P13-14: Grammar modification.

P15-17: Add the comparison of schemes. Grammar modification.

P17-18: Add the Influence of ensemble size and observation number.

P19: Grammar modification.

P20: Add the discussion of computational cost. Grammar modification.

P21: Add the discussion of sparse observations. Grammar modification.

P22-36: Grammar modification.

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 with 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 assimilations provide an analysis state that is usually a better estimate of the state variable because it fully 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 attributed to technical aspect and can be accomplished with the rapid development of supercomputers (Reichle 2008), although finding the global minimum is a much more 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, and can eventually result in the divergence of the filter (Anderson; 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.

In early studies on forecast error inflation, researchers usually tuned the inflation factor by repeated assimilation experiments and selected the estimated inflation factor according to their experience and prior knowledge (Anderson; Anderson 1999). However, such methods are very empirical and subjective. In later studies, the inflation factor can be estimated online based on the innovation statistic (observation-minus-forecast; (Dee 1995; Dee; Silva 1999)) with 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 estimate of the inflation factor, although it must calculate a high dimensional matrix determinant (Liang et al. 2012; Zheng 2009). The 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; 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; Silverman. 1994; Wand; Jones 1995). In CV, 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. This most commonly used form is also called the leave-one-out cross validation (Gu; 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; Wahba 1979). For instance, Gu and Wahba applied the Newton method to optimize the GCV score with multiple smoothing parameters in a smoothing spline model (Gu; Wahba 1991). Wahba 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 (Wahba et al. 1995). Zheng and Basher also used a GCV in a thin-plate smoothing spline model of spatial climate data and applied it to South Pacific rainfalls (Zheng; 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).

In the covariance inflation scheme, the forecast error matrix is multiplied by an appropriate inflation factor. Usually, the inflation factor is larger than 1. Too small or too large an inflation factor will cause the analysis state to be over reliant on model forecasts or observations. Hence, the inflation factor should be estimated accurately. After this, the weights of the model forecasts and observations in the analysis state can be reassigned. Therefore, the analysis sensitivity was also investigated in this study. Generally speaking, analysis sensitivity is used to apportion uncertainty in the output can 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 a forecast error inflation scheme is proposed in Section 2; the verification and validation processes are conducted on an idealized model in Section 3; and the discussion and conclusions are given in Section 4.

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 numeric 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 $\mathbf{y}_i^{\mathrm{o}}$.

Supposing 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 dynamical model forecast forward:

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

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:

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

Step 2. Analysis step.

The analysis state is estimated by minimizing the following cost function

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

which has the analytic form

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

where

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

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):

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

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 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 scheme should be included in any ensemble-based assimilation scheme or else to prevent the filter from diverging (Anderson; 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 $\lambda_i\mathbf{P}_i$ by estimating the inflation factors λ_i properly.

In previous studies, a number of methods were used to estimate 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 (Anderson 2007, 2009). In this study, a new procedure for estimating multiplicative inflation factors λ_i is proposed based on the following GCV function (Craven; Wahba 1979)

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

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 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, as follows:

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

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

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 vector. 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; 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 diagnostic 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, excluding useful information from the observations. This is captured by the fact that for the conventional EnKF scheme the GAI values are rather small. 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 in the conventional EnKF, which also dramatically decreases until the observations ultimately have an irrelevant impact on the analysis states. Applying the inflation technique, an underestimation of the forecast ensemble spread can be effectively compensated for, thereby improving 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:

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

where $\mathrm{x}_{i,k}^{\mathrm{a}}$ and $\mathrm{x}_{i,k}^{\mathrm{t}}$ are the *k*-th components of the analysis state and true state at the *i*-th time step. In principle, a smaller RMSE indicates a better performance of the assimilation scheme.

3. Numerical Experiments

The proposed data assimilation scheme was tested using the Lorenz-96 model (Lorenz 1996) with model errors and a linear observation system as a test bed. The performances of the assimilation schemes described in Section 2 were evaluated via the following experiments.

3.1. Dynamical forecast model and observation systems

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

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

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; 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 at all of the 40 model grids 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 set as 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 on each grid point was set to $\sigma_o^2 = 1$, and the covariance of the observations between the j-th and k-th grid points was as follows:

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

3.2. Assimilation scheme comparison

Because model errors are inevitable in practical dynamical forecast models, it is reasonable for us to add model error 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 (Eq. (19)). Modifying the forcing strength *F* changes the model forecast considerably. For values of *F* that are larger than 3, the system is chaotic (Lorenz; 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 every 4 time steps using the conventional EnKF and the improved EnKF with forecast error inflation. The time series of estimated inflation factors are shown in Figure 2, which vary between 1 and 6 with greatly majority. The median was 1.88, which was used in the following comparison of the improved EnKF and the simple multiplicative inflation techniques like setting a constant inflation factor.

The forecast ensemble spread of the conventional EnKF, improved EnKF and constant inflated 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 slightly 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 increase in GAI from 10% for the conventional EnKF to 30% for the EnKF with forecast error inflation indicates that the latter relies more on the observations. This finding is important because the observations can play a more 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 resulting estimate, 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 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, improved EnKF and constant inflated 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, and 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 forecast error inflation can lead to a more accurate analysis state than the constant inflated EnKF.

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 with respect to different ensemble sizes

(10, 30 and 50) were evaluated, and the results are listed in Table 1, which shows that 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. Discussion and Conclusions

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, a 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, which 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 using a GCV and the analysis sensitivity was detected.

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. 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 values of the GCV function obviously decreased in the proposed approach compared the conventional EnKF scheme. The analysis RMSE in the proposed approach was also much smaller than that in the conventional EnKF scheme, which suggests that this method of minimizing the GCV works well for estimating the inflation factor.

The highest computational cost when minimizing the GCV function is related to calculating the influence matrix $\mathbf{A}_i(\lambda)$. Because 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 (Golub; Loan 1996). 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 by minimizing the GCV function. Therefore, the total computational costs of the improved EnKF are feasible.

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.

It is also 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; Kunii 2011; Yang et al. 2015).

The examples shown here using the Lorenz-96 model illustrate the feasibility of this approach for using GCV as a metric to estimate the covariance inflation factor. 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 the proposed methodologies using 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:

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

where $\mathbf{x}_i^{\mathrm{a}[\mathrm{k}]}$ 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* 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):

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

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):

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

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

	Conventional EnKF			**EnKF with forecast inflation**		
Ensemble Size	10	30	50	10	30	50
Spread	0.23	0.36	0.41	3.26	3.32	3.45
GAI	4.56%	10.78%	13.58%	5.24%	29.21%	35.63%
GCV	36.38	31.14	25.21	35.56	3.29	2.30
RMSE	4.50	4.01	3.52	3.74	1.10	0.88

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

	Conventional EnKF			**EnKF with forecast inflation**		
Ensemble Size	10	30	50	10	30	50
Spread	0.41	0.59	0.68	3.33	3.36	3.48
GAI	10.77%	20.92%	26.41%	13.25%	35.09%	41.28%
GCV	33.64	22.89	14.97	32.17	14.99	5.19
RMSE	4.85	4.10	3.29	4.39	3.46	2.86

Figure captions

Figure 1. Flowchart of the proposed assimilation scheme.

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

Figure 3. Forecast ensemble spread of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line).

Figure 4. GAI statistics of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line).

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

Figure 6. GCV function values of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line).

Figure 7. Ensemble analysis state members of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line). The green line refers to the true trajectory obtained by the numerical solution.

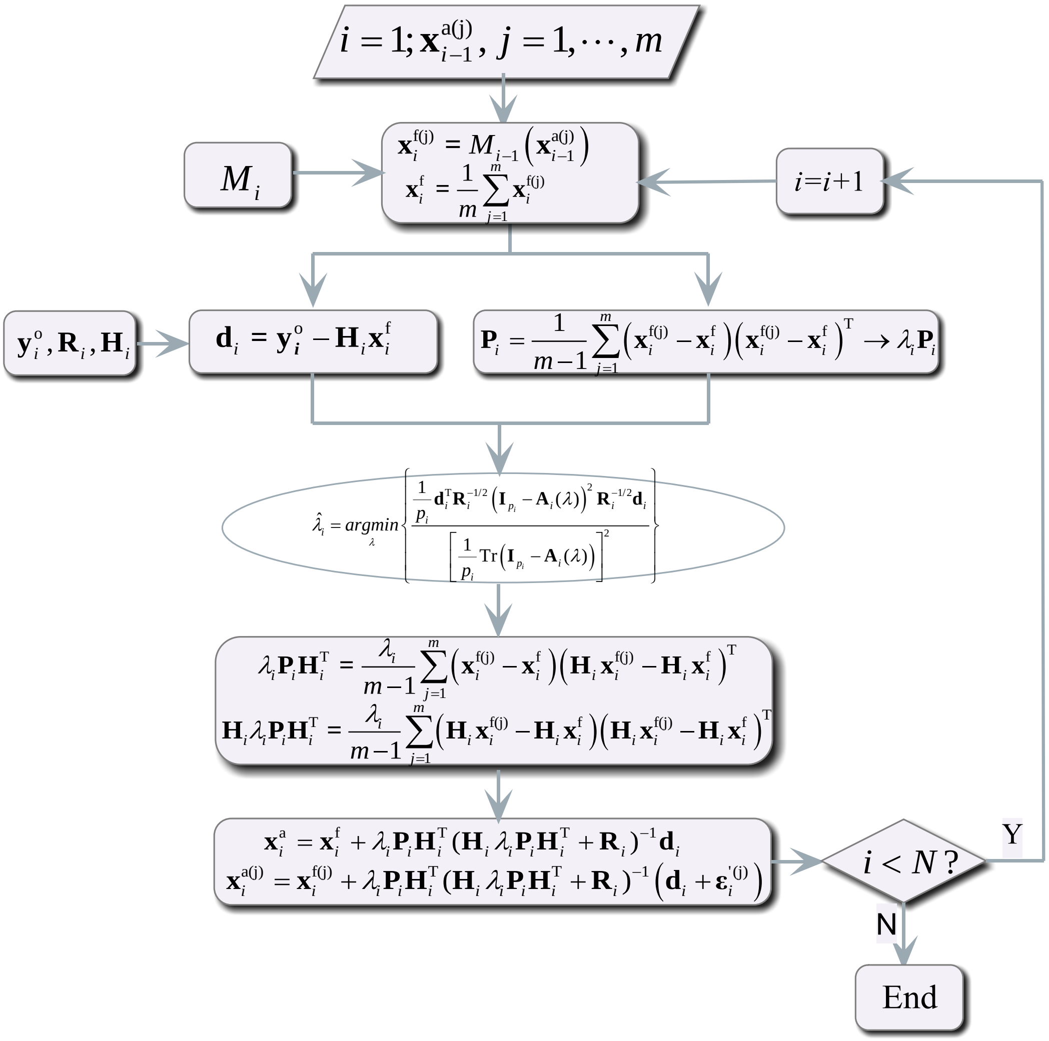
$i=1; \mathbf{x}_{i-1}^{\mathrm{a(j)}}, j=1,\cdots,m$
M_i
$\mathbf{x}_i^{\mathrm{f(j)}} = M_{i-1}\left(\mathbf{x}_{i-1}^{\mathrm{a(j)}}\right)$
$\mathbf{x}_i^{\mathrm{f}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_i^{\mathrm{f(j)}}$
$i=i+1$
$\mathbf{y}_i^{\mathrm{o}}, \mathbf{R}_i, \mathbf{H}_i$
$\mathbf{d}_i = \mathbf{y}_i^{\mathrm{o}} - \mathbf{H}_i\mathbf{x}_i^{\mathrm{f}}$
$\mathbf{P}_i = \frac{1}{m-1}\sum_{j=1}^{m}\left(\mathbf{x}_i^{\mathrm{f(j)}} - \mathbf{x}_i^{\mathrm{f}}\right)\left(\mathbf{x}_i^{\mathrm{f(j)}} - \mathbf{x}_i^{\mathrm{f}}\right)^{\mathrm{T}} \to \lambda_i\mathbf{P}_i$
$\hat{\lambda}_i = \underset{\lambda}{argmin}\left\{\frac{\frac{1}{p_i}\mathbf{d}_i^{\mathrm{T}}\mathbf{R}_i^{-1/2}\left(\mathbf{I}_{p_i} - \mathbf{A}_i(\lambda)\right)^2\mathbf{R}_i^{-1/2}\mathbf{d}_i}{\left[\frac{1}{p_i}\mathrm{Tr}\left(\mathbf{I}_{p_i} - \mathbf{A}_i(\lambda)\right)\right]^2}\right\}$
$\lambda_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} = \frac{\lambda_i}{m-1}\sum_{j=1}^{m}\left(\mathbf{x}_i^{\mathrm{f(j)}} - \mathbf{x}_i^{\mathrm{f}}\right)\left(\mathbf{H}_i\mathbf{x}_i^{\mathrm{f(j)}} - \mathbf{H}_i\mathbf{x}_i^{\mathrm{f}}\right)^{\mathrm{T}}$
$\mathbf{H}_i\lambda_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} = \frac{\lambda_i}{m-1}\sum_{j=1}^{m}\left(\mathbf{H}_i\mathbf{x}_i^{\mathrm{f(j)}} - \mathbf{H}_i\mathbf{x}_i^{\mathrm{f}}\right)\left(\mathbf{H}_i\mathbf{x}_i^{\mathrm{f(j)}} - \mathbf{H}_i\mathbf{x}_i^{\mathrm{f}}\right)^{\mathrm{T}}$
$\mathbf{x}_i^{\mathrm{a}} = \mathbf{x}_i^{\mathrm{f}} + \lambda_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}(\mathbf{H}_i\lambda_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i)^{-1}\mathbf{d}_i$
$\mathbf{x}_i^{\mathrm{a(j)}} = \mathbf{x}_i^{\mathrm{f(j)}} + \lambda_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}(\mathbf{H}_i\lambda_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i)^{-1}\left(\mathbf{d}_i + \boldsymbol{\varepsilon}_i^{'(j)}\right)$
$i < N$?
Y
N
End

Figure 1. Flowchart of the proposed assimilation scheme.

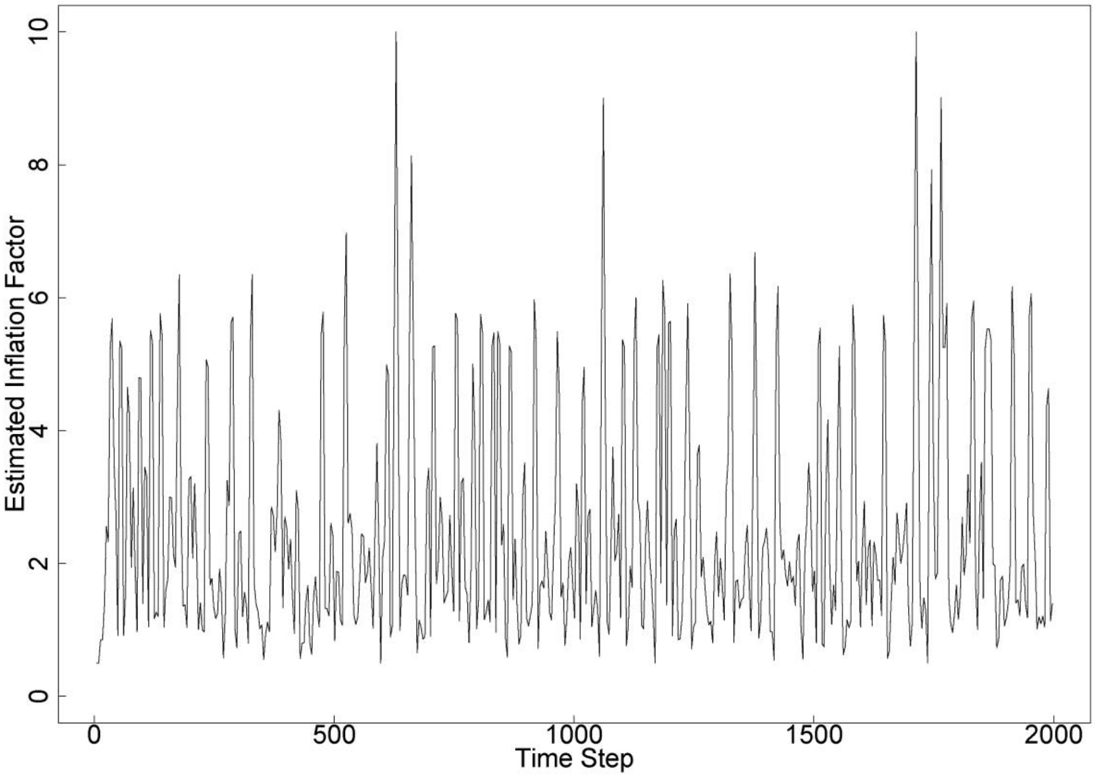

Figure 2. Time series of the estimated inflation factors by minimizing the GCV function.

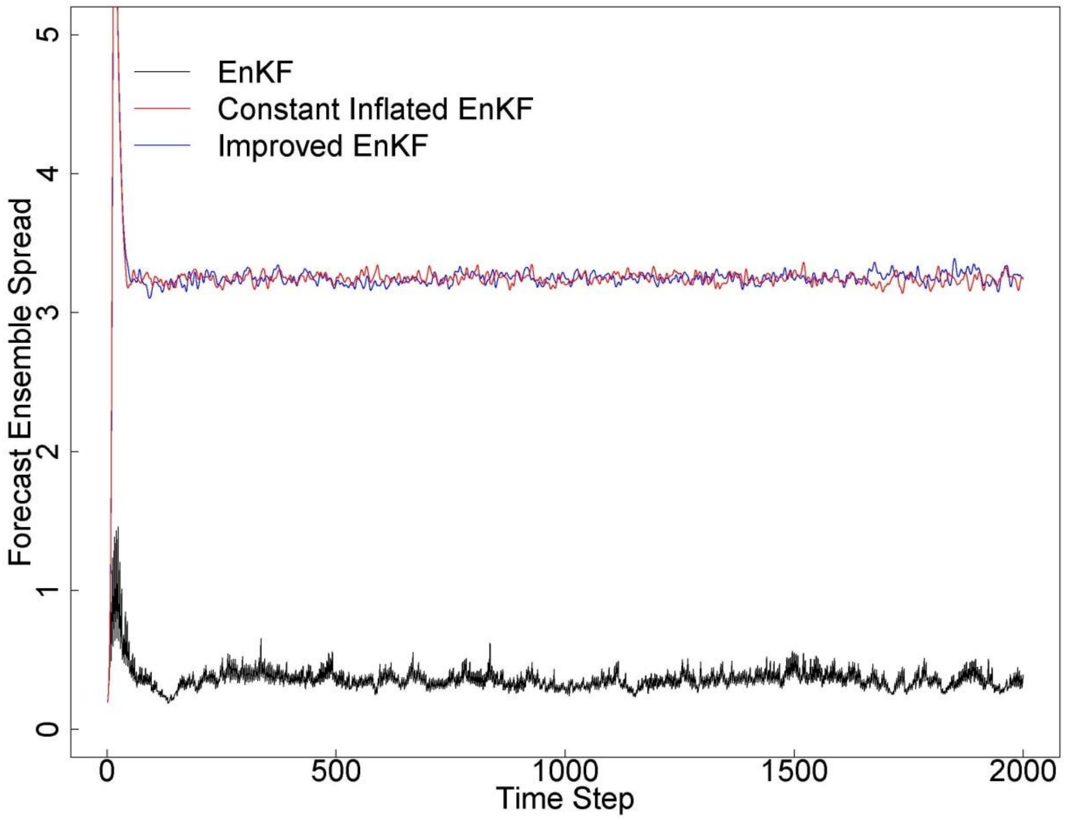

Figure 3. Forecast ensemble spread of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line).

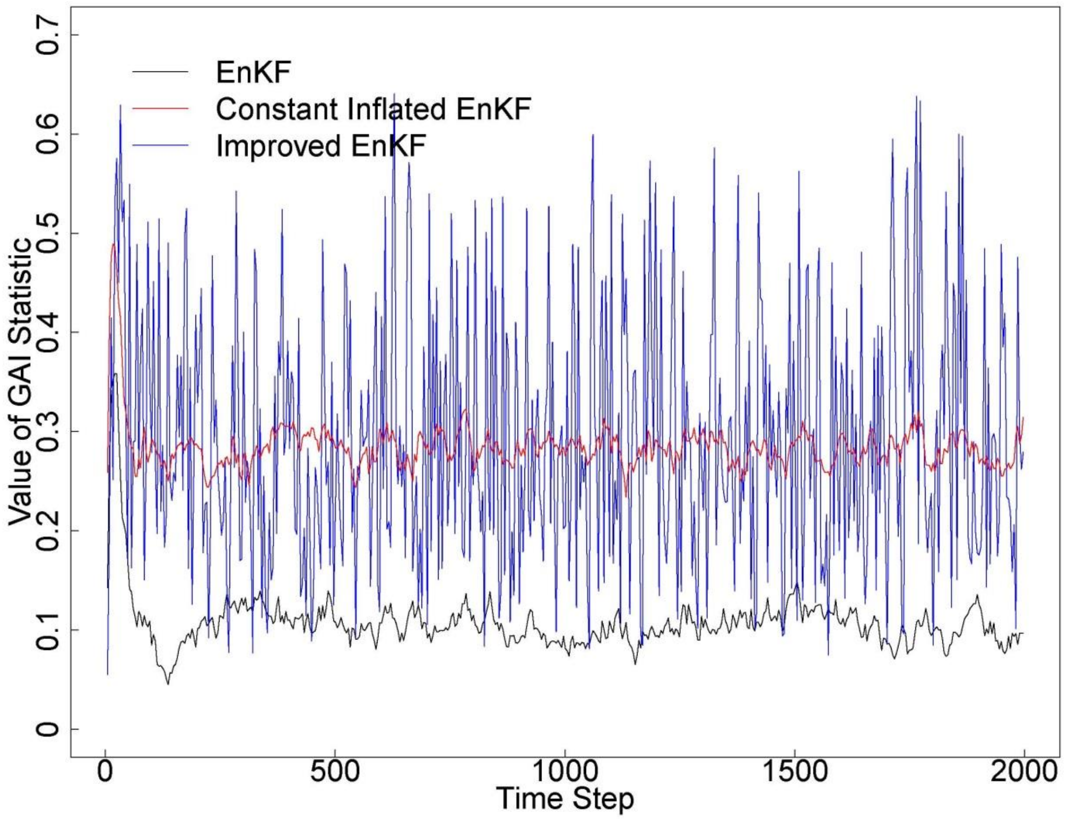

Figure 4. GAI statistics of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line).

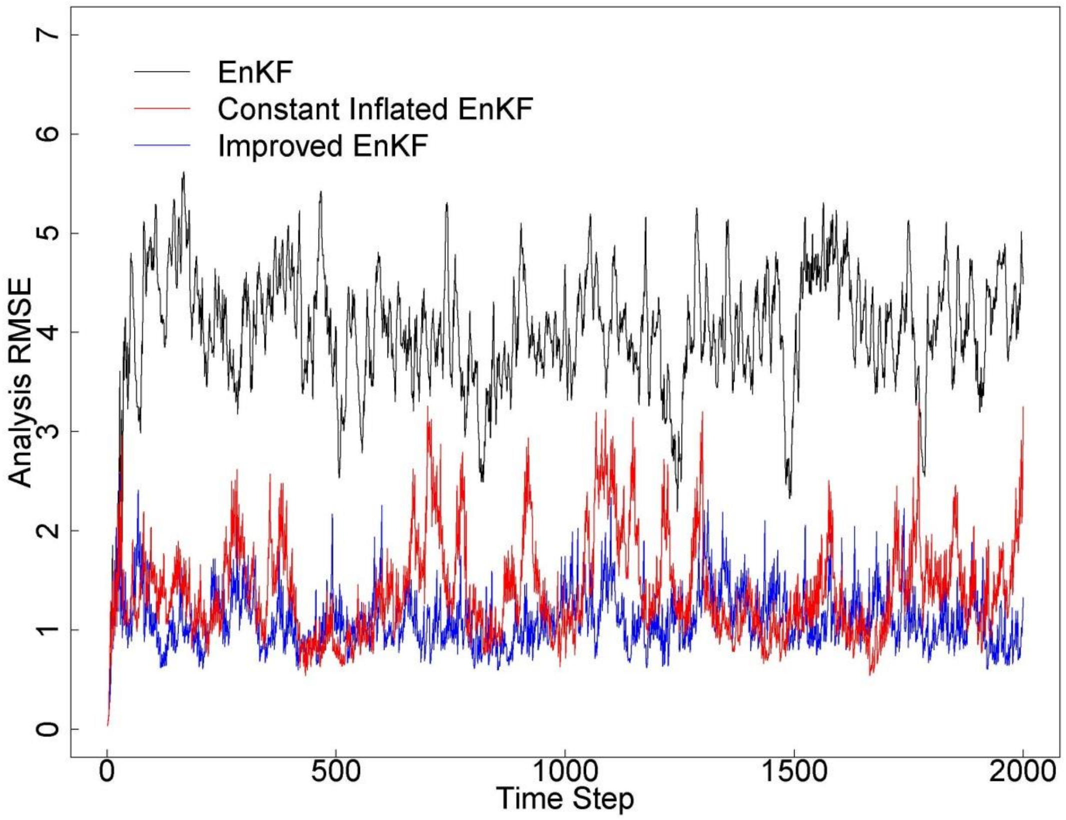

Figure 5. Analysis RMSE of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line).

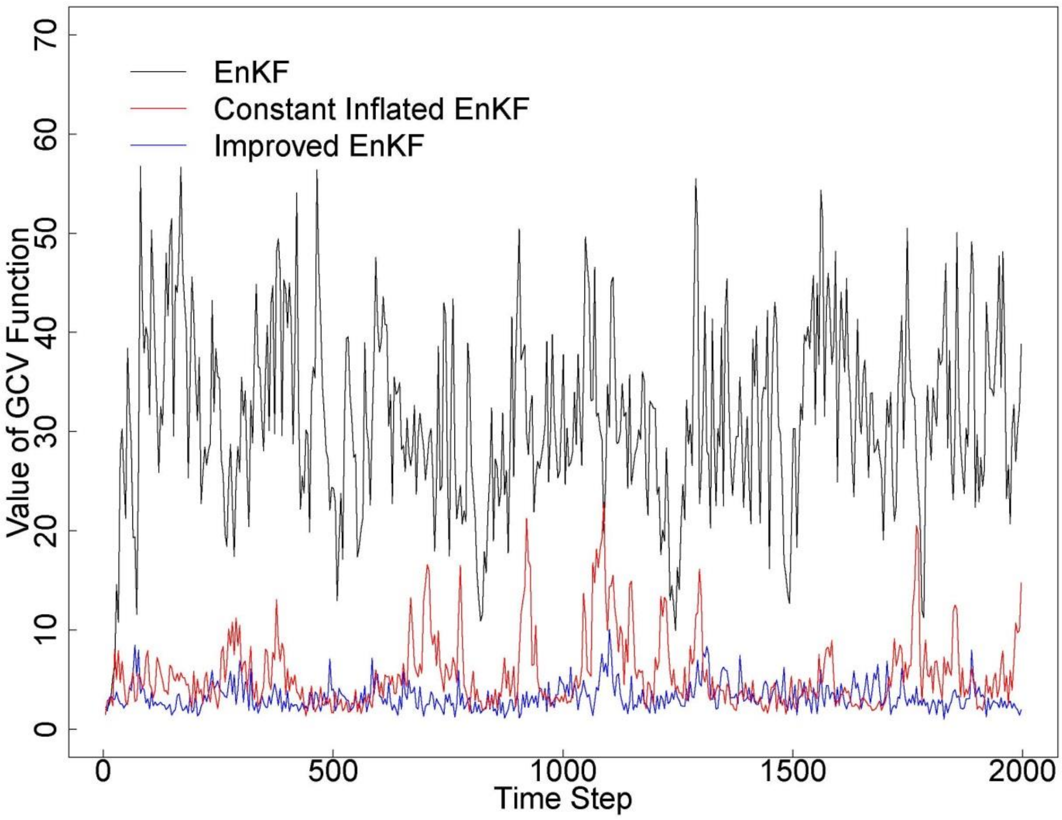

Figure 6. GCV function values of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line).

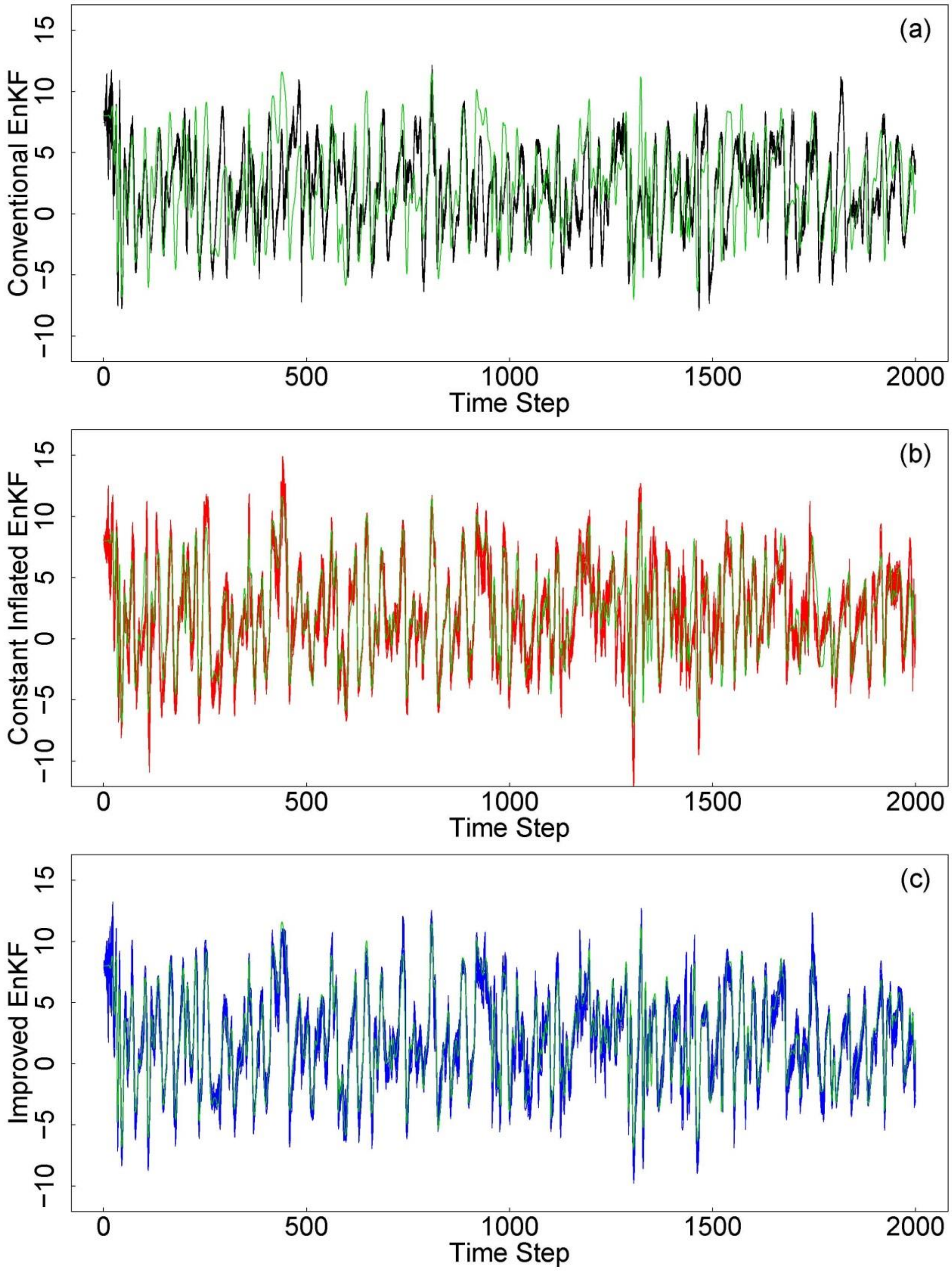

Figure 7. Ensemble analysis state members of the conventional EnKF (black line), the improved EnKF (blue line) and the constant inflated EnKF (red line). The green line refers to the true trajectory obtained by the numerical solution.

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.

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.