Accelerating the spin-up of Ensemble Kalman Filtering

Eugenia Kalnay* and Shu-Chih Yang
University of Maryland

Abstract

Ensemble Kalman Filter (EnKF) has that disadvantage that the spin-up time needed to reach its asymptotic level of accuracy is longer than the corresponding spin-up time in variational methods (3D-Var or 4D-Var). This is because the ensemble has to fulfill two independent requirements, namely that the mean be close to the true state, and that the ensemble perturbations represent the "errors of the day". As a result, there are cases such as radar observations of a severe storm, where EnKF may spin-up too slowly to be useful. A scheme is proposed to accelerate the spin-up of EnKF applying a no-cost Ensemble Kalman Smoother, and using the observations more than once in each assimilation window in order to maximize the initial extraction of information. The performance of this scheme is tested with the Local Ensemble Transform Kalman Filter (LETKF) implemented in a Quasi-geostrophic model, which requires a very long spin-up time when initialized from a cold start. Results show that with the new "running in place" scheme the LETKF spins-up and converges to the optimal level of error at least as fast as 3D-Var or 4D-Var. Additional computations (2-4 iterations for each window) are only required during the initial spin-up, since the scheme naturally returns to the original LETKF after spin-up is achieved.

1. Introduction

The relative advantages and disadvantages of 4-dimensional Variational Data Assimilation (4D-Var), already operational in several numerical forecasting centers, and Ensemble Kalman Filter (EnKF), a newer approach that does not require the adjoint of the model, are the focus of considerable current research (e.g., Lorenc, 2003, Kalnay et al, 2007a, Gustafson, 2007, Kalnay et al., 2007b, Miyoshi and Yamane, 2007).

One area where 4D-Var seems to have a clear advantage over EnKF is in the initial spin-up, since the evidence thus far is that 4D-Var converges faster than EnKF to its asymptotic level of accuracy. For example, Caya et al. (2005) compared 4D-Var and EnKF for a storm simulating the development in a sounding corresponding to 00UTC 25 May 1999. They found that "Overall, both assimilation schemes perform well and are able to recover the supercell with comparable accuracy, given radial-velocity and reflectivity observations where rain was present. 4DVAR produces generally better analyses than the EnKF given observations limited to a period of 10 min (or three volume scans), particularly for the wind components. In contrast, the EnKF typically produces better analyses than 4DVAR after several assimilation cycles, especially for model variables not functionally related to the observations." In other words, for the severe

* Corresponding author: 3431 Computer and Space Sciences Bldg., College Park, MD, 20742-2425.
ekalnay@atmos.umd.edu

storm problem the EnKF eventually yields better results than 4D-Var, presumably because of the assumptions made on the 4D-Var background error covariance, but during the crucial *initial time of storm development*, when radar data starts to become available, EnKF provides a worse analysis. For a global shallow water model, which is only mildly chaotic, Zupanski et al. (2006) found that initial perturbations that had horizontally correlated errors converged faster and to a lower level of error than perturbations created with white noise. In agreement with these results, Liu (2007) found using the SPEEDY global primitive equations model that perturbations obtained from differences between randomly chosen states (which are naturally balanced and have horizontal correlations of the order of the Rossby radius of deformation) converged faster than white noise perturbations.

Yang et al (2008a) compared 4D-Var and the Local Ensemble Transform Kalman Filter (LETKF, Hunt et al., 2007) within a quasi-geostrophic channel model. They found that if the LETKF is initialized from randomly chosen fields, it takes more than 100 days before it converges to the optimal level of error. If, on the other hand, the ensemble mean is initialized from an existent 3D-Var analysis, which is already close to the true state, the LETKF converges to its optimal level very quickly, within about 3-5 days. However, 3D-Var and 4D-Var converge fast without needing a good initial guess. This has also been observed for severe storm simulations (Caya et al., 2005), especially when using real radar observations (Jidong Gao, 2008, personal communication). It is not surprising that EnKF spins-up more slowly than 3D-Var or 4D-Var because in order to be optimal the ensemble has to satisfy two independent requirements, namely that the mean be close to the true state of the system, and that the ensemble perturbations represent the characteristics of the "errors of the day" in order to estimate the evolving background error covariance $\mathbf{B}$. In both 3D-Var and 4D-Var, by contrast, $\mathbf{B}$ is assumed to be constant.

The option of initializing the EnKF from a state close enough to the optimal analysis, such an existent 3D-Var analysis, with balanced perturbations having realistic horizontal correlations, is feasible within a global operational system, and as a result spin-up is not a serious problem for EnKF. However, there are other situations, such as the storm development discussed above, where radar information is not available before the storm starts, so that no information is available to guide the EnKF in the spin-up towards the optimal analysis. The system may start from an unperturbed state without precipitation, and if a severe storm develops within a few minutes and the EnKF takes considerable real time to spin-up from the observations, it will "miss the train" and give results that are less useful for severe storm forecasting than 4D-Var or even 3D-Var.

In this note we propose a new method to accelerate the spin-up of the EnKF by "running in place" during the spin-up phase and using the observations more than once in order to extract maximum information. We find that it is possible to accelerate the convergence of the EnKF so that (in terms of real time) it spins-up even faster than 3D or 4D-Var. Section 2 contains a brief theoretical motivation and discussion of the method, results are presented in Section 3 and a discussion is given in Section 4.

2. Spin-up, no-cost smoothing and "running in place" in EnKF

Hunt et al. (2007) provided a new derivation of the linear Kalman Filter equations by showing that in the cost function

$$J(\mathbf{x}) = \left[\mathbf{x} - \overline{\mathbf{x}}_n^b\right]^T \left(\mathbf{P}_n^b\right)^{-1} \left[\mathbf{x} - \overline{\mathbf{x}}_n^b\right] + \left[\mathbf{y}_n^o - \mathbf{H}_n\mathbf{x}\right]^T \left(\mathbf{R}_n^{-1}\right)\left[\mathbf{y}_n^o - \mathbf{H}_n\mathbf{x}\right], \qquad (1)$$

the background term represents the Gaussian distribution of a state with the maximum likelihood trajectory (history), i.e., the analysis/forecast trajectory that best fits the data from $t = t_1,...,t_{n-1}$. This state is obtained by using the forecast model $\mathbf{M}_{t_{n-1},t_n}$ to advance the previous maximum likelihood analysis $\overline{\mathbf{x}}_{n-1}^a$ and the corresponding analysis error covariance $\mathbf{P}_{n-1}^a$ to the new analysis time t_n. In other words, the following relationship is satisfied for some constant *c*:

$$\sum_{j=1}^{n-1}\left[\mathbf{y}_j^o - \mathbf{H}_j\mathbf{M}_{t_n,t_j}\mathbf{x}\right]^T \mathbf{R}_j^{-1}\left[\mathbf{y}_j^o - \mathbf{H}_j\mathbf{M}_{t_n,t_j}\mathbf{x}\right] = \left[\mathbf{x} - \overline{\mathbf{x}}_n^b\right]^T \left(\mathbf{P}_n^b\right)^{-1}\left[\mathbf{x} - \overline{\mathbf{x}}_n^b\right] + c \qquad (3)$$

After the cost function in (1) is minimized finding the analysis $\overline{\mathbf{x}}_n^a$ and its corresponding covariance $\mathbf{P}_n^a$, a similar relationship holds for the analysis at t_n for some constant *c*':

$$\left[\mathbf{x} - \overline{\mathbf{x}}_n^b\right]^T \left(\mathbf{P}_n^b\right)^{-1}\left[\mathbf{x} - \overline{\mathbf{x}}_n^b\right] + \left[\mathbf{y}_n^o - \mathbf{H}_n\mathbf{x}\right]^T \left(\mathbf{R}_n^{-1}\right)\left[\mathbf{y}_n^o - \mathbf{H}_n\mathbf{x}\right] = \left[\mathbf{x} - \overline{\mathbf{x}}_n^a\right]^T \left(\mathbf{P}_n^a\right)^{-1}\left[\mathbf{x} - \overline{\mathbf{x}}_n^a\right] + c' \qquad (4)$$

Equating the terms in (4) that are linear and quadratic in $\mathbf{x}$, the linear Kalman Filter equations for a perfect model are obtained.

This derivation makes clear that Kalman Filter yields the maximum likelihood estimate $\overline{\mathbf{x}}_n^a$ with the corresponding error covariance $\mathbf{P}_n^a$ at time t_n if the model is linear and perfect and *if the previous analysis* $\overline{\mathbf{x}}_{n-1}^a$ *at* t_{n-1} *is also the maximum likelihood state estimate at the previous analysis time*. Hunt et al. (2007) also indicate that a system can be initialized with a limited number of observations at the initial time t_1 by assuming that the initial background error covariance is large but not infinitely large. Although this introduces into the cost function an additional quadratic term, they point out that "with sufficient observations over time, the effect of this term [on the background error covariance] at time t_n decreases in significance as n increases". In other words, with sufficient observations, the Kalman Filter spins-up and eventually converges and yields the maximum likelihood solution and its error covariance.

The EnKF, like the Kalman Filter, also provides a maximum likelihood analysis, except that the background and analysis error covariances are estimated from an ensemble of K generally nonlinear forecasts:

$$\mathbf{P}_n^b \approx \frac{1}{K-1}\mathbf{X}_n^{bT}\mathbf{X}_n^b, \qquad (5)$$

where $\mathbf{X}_n^b$ is a matrix whose columns or the background (forecast) perturbations $\mathbf{x}_{n,k}^b - \overline{\mathbf{x}}_n^b$ and $\overline{\mathbf{x}}_n^b = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_n^b$ is the most likely forecast state, i.e., the ensemble average. Similar equations are valid for the analysis mean $\overline{\mathbf{x}}_n^a$ and the analysis error covariance $\mathbf{P}_n^a$. Thus, EnKF, like the original Kalman Filter, is a sequential data assimilation system where, after the new data is used at the analysis time it should be discarded (Ide et al., 1997), but this is true only if the previous analysis and the new background are the most likely states given the past observations. In other words, *if the system has converged after the initial spin-up, all the information from past observations is already included in the background*. In contrast, 4D-Var is a smoother that best fits all the observations (even asynoptic data) within an assimilation window. We note that EnKF can be also easily extended to 4-dimensions as in 4D-Var, allowing for the assimilation of asynoptic observations made between two analyses (e.g., Hunt et al., 2004). In EnKF only the observational increments that project on the subspace of the ensemble forecasts can be assimilated. Therefore the observational increments computed at the observation time, which are linear combinations of the ensemble forecasts, can be shifted either forward or backward to the analysis time by simply using the same linear combination of the ensemble forecasts obtained at the observation time.

In summary, after the initial spin-up, all the information from past observations is already included in the background field, so that the observations should be used only once and then discarded. However, there is no theoretical reason why this constraint should also be applied when EnKF is "cold-started", and the initial ensemble is not representative of the most likely state and its uncertainty. In practical applications, nevertheless, the rule of using the data only once is usually applied (e.g., Zupanski et al. 2006), and a slow EnKF spin-up observed. In this note we suggest that when a quick EnKF spin-up (in real time) is needed in order to make useful short-range forecasts for fast weather instabilities, the initial observations can be used more than once in order to extract more information from them, and that this procedure leads to a much faster spin-up of the initial ensemble in real time. This "running in place" algorithm is made possible by the use of a "no-cost" Ensemble Kalman Smoother (EnKS) (Kalnay et al., 2007b, Yang et al., 2008a).

The no-cost EnKS is easy to implement. Consider an assimilation window $[t_{n-1}, t_n]$ within a Square-Root type of EnKF (e.g., Tippett et al. 2003, Whitaker and Hamill, 2002, Ott et al., 2004). The analysis ensemble members at time t_n are each a weighted average (linear combination) of the ensemble forecasts at t_n (Hunt et al., 2007)[1]. Since the ensemble

[1] We note that Yang et al. (2008b) explored the characteristics of the analysis weight fields and found that they vary smoothly on large scales. As a result, if the analysis (i.e., the computation of the weights) is carried out on a very sparse analysis grid and then

analysis estimates the linear combination of the trajectories that best fits the observations within an assimilation window, not just at the end of the interval, the no-cost EnKS valid at the beginning of the window is obtained by simply applying the same weights obtained at analysis time t_n to the initial ensemble at t_{n-1}. Yang et al. (2008a) tested this scheme and found that indeed, the no-cost EnKS smoothed ensemble at t_{n-1} is more accurate than the analysis ensemble valid at t_{n-1}, as could be expected from the fact that the smoothed ensemble at the beginning of the window has benefited from the information provided by the "future" observations in the window $[t_{n-1}, t_n]$. Although the no-cost smoothing improves the initial analysis at t_{n-1}, it does not improve the final analysis at t_n, since the forecasts started from the new initial analysis ensemble will end as the final analysis ensemble (at least in a linear sense).

With the no-cost EnKS it is then possible to go backwards in time within an assimilation window, and then advance with the regular EnKF using the initial observations repeatedly in order to extract maximum information from them. This improves the quality (likelihood) of the initial ensemble mean faster, and leads the ensemble-based background error covariance to be more representative of the true forecast error statistics.

The algorithm that we have tested (not necessarily the best) is as follows: We start the EnKF from a randomly chosen initial ensemble mean and random perturbations at t_0, and integrate the initial ensemble to t_1. Then the "running in place" loop with $n = 1$, is:

a) Perform a standard EnKF analysis and obtain the analysis weights at t_n, saving the mean square observations minus forecast (OMF) computed by the EnKF.

b) Apply the no-cost smoother to obtain the smoothed analysis ensemble at t_{n-1} by using the same weights obtained at t_n.

c) Perturb the smoothed analysis ensemble with a small amount of random Gaussian perturbations, a method similar to additive inflation. These added perturbations have two purposes: they avoid the problem of otherwise reaching the same final analysis at t_n as in the previous iteration, and they allow the ensemble perturbations to evolve into fast growing directions that may not have been included in the unperturbed ensemble subspace.

d) Integrate the perturbed smoothed ensemble to t_n. If the forecast fit to the observations is smaller than in the previous iteration according to a criterion such as

$$\frac{OMF^2(iter) - OMF^2(iter+1)}{OMF^2(iter)} > \varepsilon, \qquad (6)$$

---

interpolated to the in-between grid points, the interpolated weight analysis is not only computationally more efficient, but the interpolation does not degrade and may actually improve upon the full resolution analysis.

go to a) and perform another iteration. If not, let $t_{n-1} \leftarrow t_n$ and proceed to the next assimilation window.

3. Results

Figure 1 shows the RMS error of the analysis obtained during spin-up, using several methods over 200 analysis cycles of 12 hours each (corresponding to a total of 100 days). All the methods started from the same a randomly chosen mean state and in the case of LETKF, from perturbations created as Gaussian noise. 3D-Var (dashed blue line) takes about 60 cycles to spin-up, and 4D-var (full blue line) takes about 80 cycles, but converges to a much lower RMS error than 3D-Var. The standard LETKF (black line) using the observations once and discarding them takes much longer, a total of 170 cycles. During the first 120 cycles the ensemble perturbations develop into the "errors of the day", and between 120 and 170 cycles the LETKF converges rather quickly to the optimal level of error. After they attain convergence, LETKF and 4D-Var RMS errors are similar.

A preliminary experiment with the LETKF "running in place" algorithm allowing for repeated use of the observations but fixing the number of iterations at 10 is shown with a dashed black line. The LETKF with 10 iterations spins-down even faster than 4D-Var and converges in only about 50 cycles but to a higher level of error, close to 3D-Var. This is not surprising, since once the system is close to the maximum likelihood solution, as indicated by the theoretical arguments discussed above, observations should be used only once and then discarded. By using 10 iterations after the spin-up, the EnKF analysis fits the data too closely and this increases the analysis errors.

The adaptive approach (6) tests whether the system is optimal by checking whether iterations reduce the ensemble forecast error, and stops iterating when the relative improvement is less than ε. A low value of $\varepsilon = 0.01$ (not shown) leads to a faster initial reduction of errors but requires a large number of iterations (Figure 2). Values of ε within a range of 0.02-0.05 give optimal results, leading to a spin-down of the initial errors similar to 3D-Var and faster than 4D-Var, and converging to and error level at least as good as that of 4D-Var (see red line in Figure 1 corresponding to $\varepsilon = 0.05$).

We also tested whether the use of additive inflation with perturbations that are horizontally correlated would accelerate the spin-up, as found by Zupanski et al. (2006) for the initial perturbations. Figure 1 shows with a green line the result of the LETKF with $\varepsilon = 0.05$, as in the red line, but with the additive perturbations chosen so that their background error covariance is the 3D-Var covariance, i.e., the columns of the matrix $\sqrt{\mathbf{B}_{3D-Var}}\mathbf{E}$, where $\mathbf{E}$ is a matrix whose columns are random Gaussian numbers such that $\mathbf{E}\mathbf{E}^T = \mathbf{I}$. Since $\mathbf{B}_{3D-Var}$ was obtained using the NMC method (Parrish and Derber, 1992, Yang et al., 2008a), the additive perturbations based on $\mathbf{B}_{3D-Var}$ have horizontal correlation

lengths with synoptic scales, whereas the additive Gaussian perturbations used for the other experiments discussed before have very small correlation lengths. The green line in Figure 1 shows that when the additive perturbations are horizontally correlated, convergence takes place faster than with the Gaussian additive perturbations, even when the same criterion $\varepsilon = 0.05$ is used for both. This agrees with the conclusion of Zupanski et al. (2006) that horizontal correlation of the perturbations accelerates spin-up. Nevertheless, once convergence has been achieved, the accuracy of the system with noisy perturbations (red) is slightly better than the system with $\mathbf{B}_{3D-Var}$ perturbations.

Figure 2 compares the number of iterations required by "running in place" schemes. It shows that with $\varepsilon = 0.01$ the number of iterations required starts at about 50, and remains at a range of 2-10 iterations even after convergence, suggesting that the criterion is too strict, leading to inefficient spin-up. With $\varepsilon = 0.05$ the system with synoptic scale ($\mathbf{B}_{3D-Var}$-based) additive perturbations converges faster, reaching 1-2 iterations after only about 30 data assimilation cycles, and then oscillates between 1 and 2 iterations. The system with uncorrelated Gaussian additive inflation (also with $\varepsilon = 0.05$) takes about 50 data assimilation cycles to reach a single iteration (i.e., using the data only once). During the spin-up period the number of iterations is 2-4, and after convergence it automatically returns to the regular LETKF.

4 Discussion

The results obtained are very encouraging: it is possible to spin-up the LETKF (and other EnKF algorithms) when a cold-start and fast convergence to the optimal level of error (in terms of real or physical time) are required, by simply using the initial observations several times rather than only once. The no-cost Ensemble Kalman Smoother, with the smoothed analysis ensemble at the beginning of an assimilation window given by using the analysis weights of the ensemble forecast at the end of the window enables this algorithm to extract the maximum information from the initial observations. It is necessary to add small perturbations to the ensemble, in a procedure akin to additive inflation. The number of iterations needed is estimated by checking whether the smoothed analysis reduces the forecast error (OMF). A level of relative reduction ε of about 2-5% was found to work well in this quasi-geostrophic model, leading to about 2-4 iterations during spin-up, and when the system converges it naturally returns to the original LETKF.

In the case of a developing storm, it would be possible to use the weight interpolation algorithm of Yang et al (2008b) to perform the additional iterations locally, "where the action is", rather than throughout the whole domain. We found that additive inflation with horizontal correlations accelerates the initial spin-up, in agreement with Zupanski et al. (2006), but later is slightly worse than uncorrelated errors. These exploratory experiments are encouraging, agree with a similar acceleration found by Anna Trevisan (personal communication, 2008) using initial bred vectors, and may be applicable to other problems such as ocean data assimilation where a fast spin-up is desirable.

Acknowledgements:

We are grateful to Dr. Jidong Gao who pointed out the difficulties of a slow spin-up for EnKF in severe storm prediction using radar data. The influence of the paper by B. Hunt, E. Kostelich and I. Szunyogh is also gratefully acknowledged. This research was supported by NASA grants NNG06GB77G, NNX08AD90G and NNG06GB77G, and DOE grant DEFG0207ER64437.

5. References:

Caya, C, J. Sun, and C. Snyder, 2005: A Comparison between the 4DVAR and the Ensemble Kalman Filter Techniques for Radar Data Assimilation, *Mon. Wea. Rev.*, **133**, 3081-3094.

Hunt, B. R., E. Kostelich, I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: a Local Ensemble Transform Kalman Filter. *Physica D,* **230**, 112-126.

Kalnay, E., H. Li, T. Miyoshi, S.-C. Yang and J. Ballabrera, 2007a: 4D-Var or Ensemble Kalman Filter? *Tellus A*, **59**, 758–773.

----------, 2007b: Response to Comments by N. Gustaffson. *Tellus A*, **59**, 778-780

Liu, J., 2008: Applications of the LETKF to adaptive observations, analysis sensitivity, observation impact and the assimilation of moisture. *PhD dissertation*, University of Maryland, 154 pages.

Lorenc, A. C., 2003: The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. Quart. J. Roy. Meteor. Soc., 129, 3183-3203.

Ott, E., B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415-428.

Parrish, D. and J. Derber, 1992: The National Meteorology Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747-1763.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble Square Root Filters. *Mon. Wea. Rev.*, **131**, 1485-1490.

Whitaker, J. S. and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations, Mon. Wea. Rev. 130, 1913–1924.

Yang, S-C, M. Corazza, A. Carrassi, E. Kalnay, and T. Miyoshi, 2008a: Comparison of ensemble-based and variational-based data assimilation schemes in a quasi-geostrophic model. *Mov. Wea. Rev., under revision*.

Yang, S-C, E. Kalnay, B. Hunt, N. Bowler, 2008b: Weight interpolation for efficient data assimilation with the Local Ensemble Transform Kalman Filter, *Quart. J. Roy. Meteor. Soc.*, *under revision*.

Zupanski, M., S. J. Fletcher, I. M. Navon, B. Uzunoglu, R. P. Heikes, D. A. Randall, T. D. Ringlee and D. Daescu, 2006: Initiation of ensemble data assimilation. Tellus, 58A, 159-170.

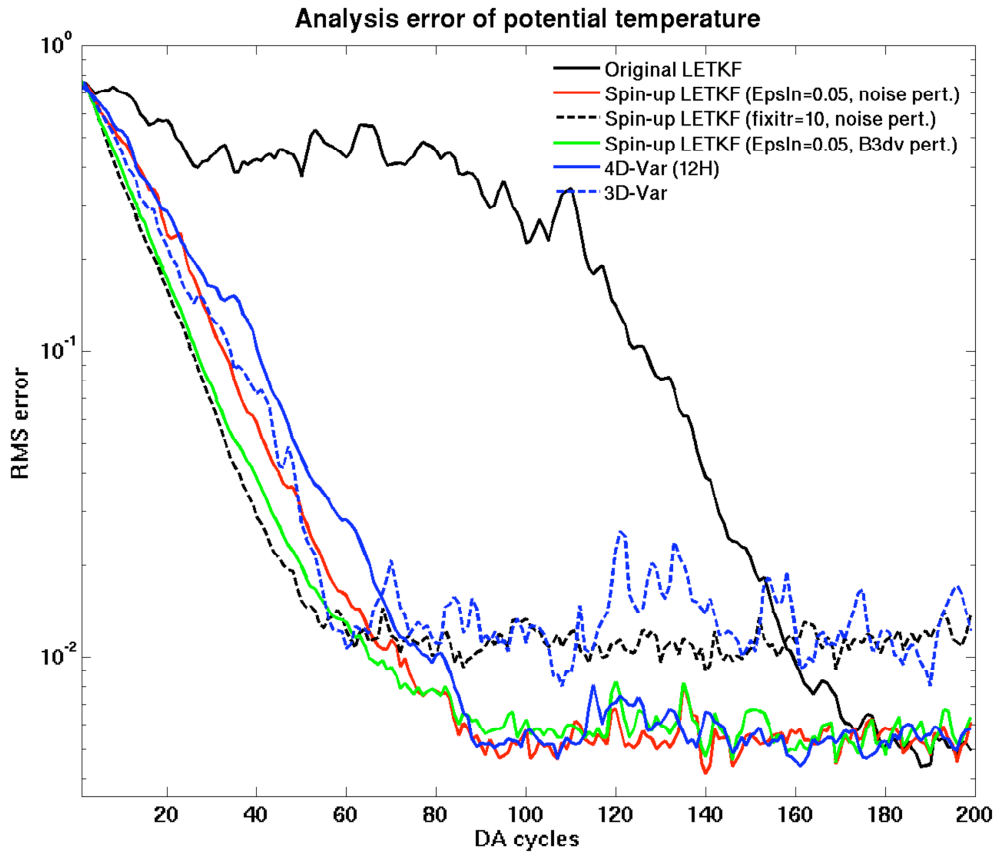

Figure 1 Time series of RMS analysis errors in potential temperature at the bottom level of the original LETKF (black line), 4D-Var (blue line) and 3D-Var (dashed blue line). The dashed black line represents the LETKF "running in place" algorithm with Gaussian additive inflation but with a fixed number of iterations (10). The red line is for an adaptive number of iterations with $\varepsilon = 0.05$ and Gaussian additive inflation, and the green line is as the red one, but with correlated additive inflation (see text).

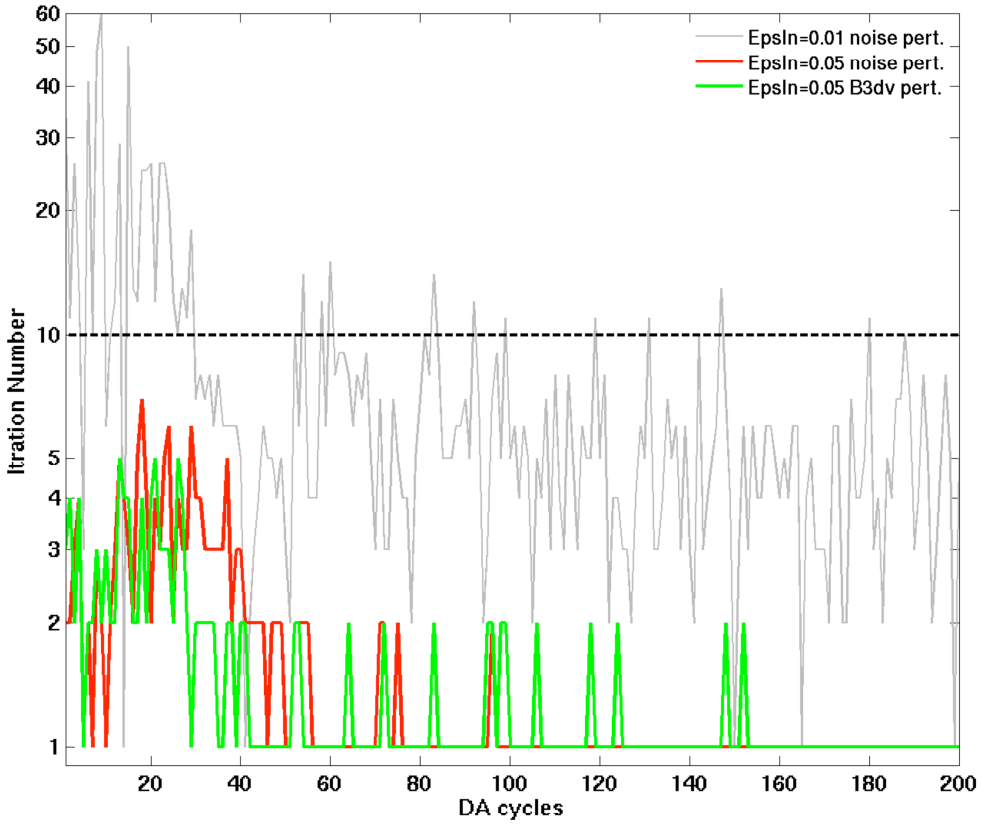

Figure 2 Number of iterations required by the spin-up LETKF with Gaussian additive inflation using $\varepsilon = 0.05$ (red line), $\varepsilon = 0.01$ (thin grey line), and $\varepsilon = 0.05$ (red line) but correlated additive (green line). The dashed black line is as the red line but fixing the number of iterations at 10.