Reducing Correlation Sampling Error in Ensemble Kalman Filter Data Assimilation

JEFFREY L. ANDERSON

Data Assimilation Research Section, National Center for Atmospheric Research,* Boulder, Colorado

(Manuscript received 6 February 2015, in final form 12 November 2015)

ABSTRACT

Ensemble Kalman filters are widely used for data assimilation in large geophysical models. Good results with affordable ensemble sizes require enhancements to the basic algorithms to deal with insufficient ensemble variance and spurious ensemble correlations between observations and state variables. These challenges are often dealt with by using inflation and localization algorithms. A new method for understanding and reducing some ensemble filter errors is introduced and tested. The method assumes that sampling error due to small ensemble size is the primary source of error. Sampling error in the ensemble correlations between observations and state variables is reduced by estimating the distribution of correlations as part of the ensemble filter algorithm. This correlation error reduction (CER) algorithm can produce high-quality ensemble assimilations in low-order models without using any a priori localization like a specified localization function. The method is also applied in an observing system simulation experiment with a very coarse resolution dry atmospheric general circulation model. This demonstrates that the algorithm provides insight into the need for localization in large geophysical applications, suggesting that sampling error may be a primary cause in some cases.

1. Introduction

Ensemble Kalman filters are widely used for data assimilation in numerical weather prediction on both global (Houtekamer et al. 2014; Whitaker et al. 2008) and regional scales (Zhu et al. 2013; Cavallo et al. 2013), ocean prediction (Keppenne et al. 2008; Karspeck et al. 2013), land surface (Rosolem et al. 2014), and hydrology prediction (Reichle et al. 2002), as well as many other types of applications (Chen and Oliver 2010; Emerick and Reynolds 2011; Shaman and Karspeck 2012). State-of-the-art prediction models for many of these applications have evolved to use all available computational resources. Therefore, there is a natural desire to minimize ensemble sizes while still providing reasonably high-quality ensemble analyses.

Many variants of ensemble filters provide the exact Kalman filter solution for linear prediction models with Gaussian observational error and sufficiently large ensembles (Anderson 2009b; Tippett et al. 2003). However, all of these prerequisites are violated for large geophysical applications. Consequently, basic ensemble filters without ad hoc adjunct algorithms like inflation and localization and with affordable ensemble sizes generally diverge from the observed system. Other filter variants (Burgers et al. 1998; Ott et al. 2004) also tend to diverge without additional algorithmic enhancements.

Two fundamental problems associated with small ensembles for large geophysical applications are insufficient prior ensemble variance and spurious prior correlations between observations and state variables. Heuristic methods addressing these problems have been developed, in particular inflation (Anderson and Anderson 1999) and localization (Houtekamer and Mitchell 1998; Hamill et al. 2001; Furrer and Bengtsson 2007). Inflation algorithms reduce the loss of variance during the assimilation (Zhang et al. 2004) or restore variance when it is lost (Anderson 2009a). Localization algorithms attempt to reduce errors in the correlations. In most cases, small ensembles are implicitly assumed to overestimate the magnitude of correlation and localization algorithms reduce ensemble correlations (Anderson 2012, hereafter A12). Adaptive algorithms

---

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

**Corresponding author address:** Jeffrey L. Anderson, NCAR, 1850 Table Mesa Dr., Boulder, CO 80305.
E-mail: jla@ucar.edu

DOI: 10.1175/MWR-D-15-0052.1

© 2016 American Meteorological Society
that estimate the required inflation (Whitaker and Hamill 2012) and localization (Bishop and Hodyss 2007, 2009; Zhang and Oliver 2010) as adjunct parts of the ensemble assimilation have been developed.

Localization and inflation often correct for more than just errors from small ensembles. In large geophysical applications, inflation may correct primarily for systematic model errors (Li et al. 2009a,b) and localization may also correct for model errors that produce incorrect correlations.

A12 ignored these other aspects of localization and inflation and explicitly assumed that ensemble sampling error was the only source of correlation errors. By assuming that ensembles were random draws from a specified prior distribution of correlations, it was possible to correct for some sampling errors in the correlations between observations and state variables. Applying this sampling error correction (SEC) algorithm led to ensemble filters that still required traditional fixed localization, but were less sensitive to the tuning of traditional localization width.

This manuscript extends the approach of A12 by explicitly estimating the distribution of correlations between observations and state variables. The correlation distributions are estimated for subsets of pairs of observations and state variables as in Anderson and Lei (2013, hereafter AL13). This distribution is then used to calculate a correction for sampling error in the computation of correlation in the ensemble filter. This correlation error reduction (CER) algorithm is described in section 2. Section 3 describes low-order perfect-model experiments used to test the algorithm. Sections 4 and 5 present results from linear and nonlinear low-order models with various observing system configurations. Section 6 presents results using a low-order dry atmospheric general circulation model. Sections 7 and 8 discuss details of the algorithmic performance and present conclusions.

2. Reducing correlation sampling error

For observations with uncorrelated observational errors (or after an appropriate rotation for correlated errors), the assimilation step in many sequential ensemble Kalman filters is defined without loss of generality by the impact of a single observation $y$ on a single state vector component $x$ (Anderson 2003):

$$
\Delta x_n = \hat{b} \Delta y_n, \quad n = 1, \ldots, N.
$$

Here $n$ indexes the $N$ ensemble members, $\Delta y_n$ is the $n$th ensemble estimate of the increment for the observed quantity, $\Delta x_n$ is the corresponding increment for the $n$th ensemble estimate of the state variable component, and $\hat{b}$ is the sample regression coefficient:

$$
\hat{b} = \hat{r} \hat{\sigma}_x/\hat{\sigma}_y,
$$

where $\hat{r}$ is the sample correlation and $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the sample standard deviations of $x$ and $y$, respectively.

Following A12, sampling error is assumed to occur in the computation of $\hat{b}$ due to finite ensemble size $N$. Sampling error in the standard deviations is assumed to be small as justified in the appendix of A12, so only sampling error in the correlation $\hat{r}$ is considered. The impact of also considering sampling error in the standard deviations is discussed in section 7.

The algorithm described here estimates the prior probability distribution for the correlation $r(y, x)$ between $y$ and $x$ using information from previous state variable updates. Let the prior distribution for $r(y, x)$ be $P^{-}(y, x) = P[r(y, x) \mid \{Y, X\}^{-}]$, where $\{Y, X\}^{-}$ is the set of all pairs of observations and state variables that have been previously used in the assimilation. The likelihood of the true correlation being $\hat{r}$ given that the ensemble sample correlation is $\hat{r}$, $P[r(y, x) \mid \hat{r}(y, x)]$ is also required. A posterior distribution, $P^{+}(y, x)$, that combines information from the prior and likelihood is computed using Bayes’s formula:

$$
P^{+}(y, x) = P[r(y, x) \mid \{Y, X\}^{-}, (y, x)]
= P[r(y, x) \mid \hat{r}(y, x)]P^{-}(y, x)/\text{Norm},
$$

where Norm is a scalar that makes the posterior a PDF. The mean value of $P^{+}(y, x)$, $\tau^{+}(y, x)$ is assumed to be the best available estimate and is used to compute increments for $x$ instead of the sample value $\hat{r}(y, x)$,

$$
\Delta x_n = \tau^{+}(y, x)(\hat{\sigma}_x/\hat{\sigma}_y)\Delta y_n.
$$

Figure 1 illustrates the process of computing $\tau^{+}(y, x)$. For diagnostic purposes, it is possible to write (4) in the form of a localization,

$$
\Delta x_n = \alpha r(y, x)(\hat{\sigma}_x/\hat{\sigma}_y)\Delta y_n,
$$

where

$$
\alpha = \tau^{+}(y, x)/\hat{r}(y, x).
$$

Implementing this algorithm requires a discrete representation of the prior and posterior distributions of a correlation and the likelihood. The range $[-1, 1]$ of the correlation is divided into $S$ congruent subintervals and the mean probability density in each
subinterval is defined as $d(i), i = 1, \ldots, 5$. The product in (3) is computed by taking the products of the mean density in each subinterval. A total of 200 subintervals of width 0.01 ($S = 200$) are used in all results presented here (see section 7).

The discrete likelihood for ensemble size $N$ is computed by an offline Monte Carlo computation like the one used in A12. It is assumed that before the assimilation starts, the correlation between an observation and a state variable is uniformly distributed on $[-1, 1]$. A set of $K$ correlation values ($K = 10^8$ here) is defined that equally partitions the interval $[-1, 1]$ with

$$r_k = -1 + 2(k-1)/(K-1), \quad k = 1, \ldots, K.$$ (7)

For each $r_k$, a pseudorandom sample of size $N$ from a bivariate normal distribution with covariance

$$
\begin{bmatrix}
1 & r_k \\
r_k & 1
\end{bmatrix}
$$

is generated and the sample correlation coefficient $\hat{r}_k$ is computed. The set of all values of $r_k$ that have sample correlations $\hat{r}_k$ in a given subinterval are viewed as a random sample of the likelihood for this range of values of $\hat{r}_k$. More formally, let $I(r)$ be the function that converts a correlation into the index $(1, \ldots, 200)$ of the corresponding subinterval:

$$I(r) = \left[\frac{100(r + 1) + 1}{200}\right], \quad -1 \leq r < 1$$

$$I(r) = 200, \quad r = 1.$$ (8)

The set of pairs of true and sample correlation intervals \{(I(r_k), I(\hat{r}_k))\} for each $I(\hat{r}_k)$ are discrete estimates of the likelihood for $\hat{r}_k$. An example of a discrete likelihood estimate, $P[r|\hat{r} = 0.2]$ for $N = 20$ is shown by the red curve in Fig. 1. This likelihood is approximately Gaussian with a maximum at 0.2 but is skewed to the left.

The normalized product in (3) of the prior and the likelihood (blue curve in Fig. 1) is the best available estimate of the PDF of correlations for the update of state variable $x$ by observation $y$.

Optimal prior distributions for $r(y, x)$ depend strongly on the physical relation between $y$ and $x$. AL13 developed a method to compute empirical localization functions (ELFs) for subsets of pairs of state variables and observations. The ELF algorithm is not based on sampling error arguments, but rather does an a posteriori optimization using the output of an observing system simulation experiment to produce a localization coefficient for each subset. That use of such subsets is adopted here with an independent estimate of the correlation distribution being computed for each subset, but it is important to note that the algorithm here is an extension of A12, not AL13. In the simple examples in subsequent sections, the subsets are distinguished only by the distance between $y$ and $x$ unless otherwise noted in sections 5c, 6, and 7. AL13, Lei and Anderson (2014a), and Lei and Anderson (2014b) discuss the selection of subsets for more complex applications.

Starting an assimilation experiment requires an initial estimate of the prior correlation distribution for each subset. All experiments here use a uniform distribution on $[-1, 1]$; sensitivity to this choice is discussed in section 7.

If all true correlations for a given ($y, x$) subset were the same throughout the assimilation experiment, the posterior distribution for the correlation for a given subset in (3) would be used as the prior for the next ($y, x$) pair from that subset. However, the true ($y, x$) distribution cannot be assumed to be stationary in this way for real applications. As the synoptic situation changes, the expected correlation between an observation 200 km from a state variable might change significantly (e.g., depending on whether a front existed between the two). Instead, for potentially nonstationary correlation distributions, the posterior distribution in (3) is treated as an independent estimate of distributions that can occur for this subset. The subsequent prior for the subset is computed as a weighted average of the current prior and the posterior

$$P^+_{next} = (P^-(y, x) + \beta P^+(y, x))/(1 + \beta).$$ (9)

The relative weight $\beta$ given to the posterior distribution controls how quickly the prior PDF evolves as pairs are assimilated; a larger value for beta means that the estimate evolves more quickly but is more subject to noise. Results here use $\beta = 1/10000$ and sensitivity to $\beta$ is discussed in section 7.

A summary of the steps in the CER algorithm for observation $y$ impacting state variable $x$ is as follows:
1) Use the distance between \( x \) and \( y \) to determine which subset to use.
2) Use the sample correlation \( \hat{r}(y, x) \) (red dashed value in Fig. 1) to select the appropriate prior likelihood (red curve in Fig. 1) using (8).
3) Compute the posterior correlation distribution (blue curve in Fig. 1) as the product of the prior distribution for the subset (green curve in Fig. 1) and the likelihood using (3).
4) Use the mean of the posterior (blue dashed value in Fig. 1) to compute increments for \( x \) using (4).
5) Compute an updated prior estimate for this subset by adding a small fraction times the posterior PDF to the prior (red curve + fraction of blue curve) using (9).

There can be instances when \( \hat{r}(y, x) \) is very close to 0. The posterior mean \( \tau^+ (y, x) \) can be much larger so that \( \alpha \) in (6) could be very large. Rare large values for \( \alpha \) can dominate the time average value of localization that will be compared to previously documented localization methods. Therefore, if the absolute value of the sample correlation \( \hat{r}(y, x) \) is less than a threshold value \( r_{\text{crit}} \) and \( |\tau^+ (y, x)| > |\hat{r}(y, x)| \), the value of \( \tau^+ (y, x) \) is changed to \( \hat{r}(y, x) \) in (4). For all results shown, \( r_{\text{crit}} \) was set to 0.1. Sensitivity of results to the value of \( r_{\text{crit}} \) is discussed in section 7.

3. Evaluation of CER algorithm

Perfect-model assimilation experiments are used to evaluate the CER algorithm. Forward observation operators are applied to the state vector from a single long run of a forecast model, the “truth” run, and random samples from a specified observational error distribution are added to generate synthetic observations. These observations are then assimilated by an ensemble filter using the same forecast model.

The initial truth for all experiments except those with the GCM in section 6 is a state that is close to the model attractor (just zero for the linear model in section 4). A truth integration of 110 000 assimilation steps is started from the initial truth. Initial ensemble members are generated by adding independent draws from Normal(0, 0.2) to each state variable for each ensemble.

A sequential EAKF (Anderson 2001, 2003) using the parallel implementation of Anderson and Collins (2007) is used to assimilate the observations. The first 10 000 assimilation steps are discarded as spin up and diagnostics are reported for the subsequent 100 000 steps. Spatially and temporally varying adaptive inflation (Anderson 2009a) is applied in all low-order model cases, but the parameters of the inflation vary to facilitate comparison to previously published results. The root-mean-square error of the ensemble mean from the truth,

\[
\text{RMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\bar{x}_m - x_{m,T})^2}
\]

is used to evaluate results where \( \bar{x}_m \) is the ensemble mean, \( x_{m,T} \) is the true value, and the subscript \( m \) indexes the \( M \) model state variables. All results are for prior estimates but no qualitative differences were found for posterior estimates. No traditional localization is applied with the CER algorithm. Comparison cases that do not use CER use standard Gaspari–Cohn (GC) localization (Gaspari and Cohn 1999) with a specified half-width.

4. Simple linear model

The first model is the simple linear model from A12,

\[
x_{i,2} = 1.05x_{i,1}, \quad i = 1, \ldots, 200,
\]

where the first subscript indexes the state vector component and the second indexes the time. Each of the 200 components evolves independently and is observed every time step with a simulated observational error drawn from Normal(0, 1). For computational convenience without loss of generality, the true value of each of the 200 variables is simply zero at all times. The model and observation operator are linear, and the observational error is Gaussian, so the Kalman filter gives the optimal solution with time mean RMSE of approximately 0.318. For ensemble sizes \( N > 200 \) the standard EAKF has no sampling error and reproduces the Kalman filter solution to machine precision, but for \( N \leq 200 \), the EAKF diverges. If observations are localized with a delta function so that the \( i \)th observation impacts only the \( i \)th state variable, the EAKF converges to the Kalman filter solution for any ensemble size \( N > 1 \).

The CER algorithm is applied with adaptive inflation standard deviation fixed at 0.6 (Anderson 2009a) and the inflation value damped 10% of the distance to 1:

\[
\sqrt{\lambda_p} = 0.9(\sqrt{\lambda_n} - 1) + 1
\]

before each assimilation step as in A12, where \( \lambda \) is the variance inflation for a given state variable, subscript \( u \) refers to the posterior at the previous assimilation time, and subscript \( p \) is the prior at the current assimilation time. Figure 2 shows the time mean RMSE and ensemble spread as a function of ensemble size for the CER. Ensemble size 5 gave the smallest RMSE (0.324) and most consistent spread. Ensemble sizes of 20 and
larger gave nearly identical values of RMSE (0.331) and spread (0.368). An ensemble size of 201 that would give the exact answer without CER and adaptive inflation also has RMSE of 0.331. Ensemble size 4 periodically diverged and reconverged to the truth while ensemble size 3 diverged.

Figure 2 can be compared to Fig. 2 from A12, which shows results using the SEC algorithm applied with no GC localization and the same adaptive inflation as used here. For ensemble sizes of 160 and 200 the SEC gives nearly identical results to CER, but for smaller ensemble sizes the SEC error is larger and the SEC diverged for ensemble sizes less than 50. In this simple case where the only source of error is sampling error, the CER algorithm does nearly as well as the exact solution for ensemble sizes greater than 4 and required no tuning of localization.

5. Lorenz-96 40-variable model

The second model examined is the 40-variable configuration of the Lorenz-96 model (L96; Lorenz and Emanuel 1998) with standard parameter settings of forcing $F = 8.0$, a time step of 0.05 units, and fourth-order Runge–Kutta time differencing. To facilitate comparison to the ELF results in AL13, the same three observation distributions examined there are discussed here.

a. Frequent low-quality observations

For the first test, all 40 state variables are observed every model time step with an observational error variance of 16. Figure 3 shows the time mean RMSE for ensemble sizes of 5, 10, 20, and 40 for a variety of half-widths using standard GC localization and for the CER algorithm (horizontal lines). The CER RMSE is slightly larger than for the best GC half-width for all ensemble sizes. The GC results for the 20-member ensemble are identical to those shown in AL13’s Fig. 2, while the 10- and 40-member ensemble results are the same as in AL13’s Fig. 3. The RMSE for the CER for the 10-, 20-, and 40-member ensemble cases is very similar to the RMSE for the best ELF from AL13.

Figure 4 plots the time mean equivalent localization $\alpha$ in (6) as a function of the distance between observations and state variables for the four ensemble sizes along with the GC function that gave the smallest RMSE (note that AL13 used separation, a vector function of two locations, ranging from −19 to 20 for subsets while distance, the magnitude of the separation ranging from 0 to 20, is used here). The CER localization curves are approximately Gaussian except for a distinct dip for distances of one grid interval. This dip has been seen in
previous adaptive localization studies like A12 and is a function of the dynamics of the L96 model. The prior correlation distribution for a separation of one grid interval has roughly the same probability mass for small correlation values as the distribution for two intervals (Fig. 8). This means that there are more small correlation values with accompanying larger sampling error so that a separation of one requires localization approximately as small a separation of two.

For five members the dip is a local minimum with the localization for distance one smaller than for distance two. For all ensemble sizes, the localization from the CER is smaller than the best GC whenever the GC localization is greater than 0.1. This suggests that assuming that the ensembles act as random samples from an underlying distribution of correlations is not a very good choice in this case. In other words, the problem is fairly close to one in which a sufficiently large ensemble gives the correct result without sampling error. Therefore, the CER always localizes too aggressively resulting in larger RMSE than the best GC.

The CER localizations can be compared to the ELF localizations in AL13’s Fig. 1 for 20 members and Fig. 5 for 10 members. The shapes are very similar with the same kink at distance one. It is unclear why the ELF RMSE was not smaller than that from the best GC for this case. However, the similarity between the ELF and the CER localization patterns suggests that sampling error, the only thing used to construct the CER, is a major error source in this case.

b. Infrequent high-quality observations

The second case has observations of all state variables every 12 model time steps with an observational error variance of 1. For this observation system, the prior RMSE is about twice as large as for the previous case, but the posterior RMSE is much smaller. The fact that the prior error is larger suggests that this case may be further from the linear limit where the Kalman filter is exact so that the sampling error approximation may be more appropriate than in more linear cases.

Figure 5 shows RMSE for 10-, 20-, and 40-member ensembles for a selection of GC half-widths and for the CER. The CER RMSE is smaller than that for the best GC half-width for all ensemble sizes and is again very similar to that for the best ELFs (AL13’s Fig. 7); the CER is better than the best ELF for 10- and 40-member ensembles and slightly worse for 20.

Figure 6 shows the time mean localization for the CER along with the best GC for 10, 20, and 40 members. In this case, there is a local minimum in localization for distance one grid interval for all ensemble sizes. The CER localizations are smaller than the best GC for small distances between observation and state variable, but larger for distances greater than about three grid intervals. This suggests that the CER does better than the best GC in this case for two reasons. First, because the prior errors are much larger, the sampling error that the CER corrects is a more dominant source of error. Second, because the best localization is not as similar to a Gaussian as in the previous case, the baseline GC localization is further from optimal so it is easier for the CER to do better. The CER can have a more aggressive localization for small distances and less aggressive for larger distances than the best GC. Again, the CER localizations are very similar in shape to the best ELFs show in AL13’s Fig. 8 implying that correlation sampling error is the dominant error source in the filter for this case.
Figure 7 shows the evolution of the prior correlation PDF in the 20-member CER algorithm for the subset with the closest observations and state variables, one grid interval apart. The initial distribution (not shown) is Uniform(−1, 1). By assimilation time 20, the probability that the correlation is close to either 1 or −1 has been reduced [note that there are 80 (y, x) pairs in each subset for each assimilation time, except for distances of 0 and 0.5 with 40 pairs]. As the assimilation proceeds, a clear mode develops around correlation 0.25 and the probability of correlations close to 1 or −1 becomes negligible. By 400 assimilation times, the distribution has nearly equilibrated and it varies little over the remaining 109600 assimilations. The equilibrated distribution is strongly skewed toward negative correlations with a hint of bimodal behavior.

Figure 7b shows the evolution of the 20-member correlation distribution for observations and state variables that are relatively far apart with a distance of 8 grid intervals. The distribution evolves to have large probability of correlation close to 0 and negligible probability that the absolute value of correlation is greater than 0.5. The mode of the distribution is almost always in one of the two discrete intervals bounded above or below by zero.

Figure 8 shows the equilibrated 20-member correlation distributions for a selection of distances between the observation and state variable. The one and eight grid interval distances were already discussed. The two grid interval distance has a mode of about −0.4 with a skewed tail extending to large positive correlations. It does not, however, appear to have a second mode as for the one interval case. As discussed above, this may explain why the localization for one grid interval is a local minimum for the 20-member ensemble in Fig. 6. As the distance increases past two intervals, the correlation modes in Fig. 8 are close to 0 with increasingly narrow distributions. Correlation PDFs for distances greater than 8 are similar to the distance 8 case.

Although the prior correlation PDFs approximately converge as the assimilation proceeds, the localizations computed with (6) for subsets with small distance between observation and state variable do not. Figure 9 shows the localization for different distance subsets as sequences of pairs of observations and state variables are updated for the 20-member case. For distances of 3 grid intervals or less, localization varies from more than 1 to less than 0.1 over a sequence of 500 pairs. The
variation is smaller for larger distances. Localization for the 6 grid interval subset is almost always between 0.3 and 0.5 and for the 16 grid interval subset is almost always between 0.03 and 0.05. This time variation is a result of the fact that the correlations for a subset vary in time and the correlation error reduction equivalent localizations are functions of the current sample correlation in the ensemble and the pdf of correlations for the subset.

c. Many observations of sums of state variables

Following AL13, observations with error variance 1 are constructed as the sum of 17 consecutive state variables. The location of the middle (9 out of 17) summed variable is the observation location. For example, the observation located at gridpoint 20 is drawn from the following distribution:

\[
\text{Normal} \left( \sum_{j=12}^{28} x_{i,j}, 1 \right).
\]

These observations mimic satellite radiance observations whose forward operators are weighted averages of variables in an atmospheric volume. With observational error variance of 1, each observation has small information content compared to those in previous examples, again similar to some remote sensing observations. To constrain the prior RMSE to be much less than the L96 climatological spread, there are eight observations at each location every model time step, a total of 320 observations per assimilation time. These observations should be weakly correlated with many state variables and, hence, particularly prone to sampling error (A12).

Naive application of the CER algorithm with 20-member ensembles gave time mean RMSE that was much larger than for the best GC localization or the ELFs from AL13. This is because the subsets of \((y, x)\) pairs were distinguished only by the distance between \(y\) and \(x\). However, the parallel sequential assimilation algorithm used here (Anderson and Collins 2007) works in a joint phase space where the forward operators for all observations are computed at the beginning of each assimilation step. The model state vector becomes the set consisting of the standard 40 state variables plus the (320 in this case) observed variables. When an observation is assimilated, increments are computed and applied not only to the 40 original state variables but also to all of the 320 observed variables that have not yet been assimilated. In this case, that means that a given distance subset includes two distinct types of \((y, x)\) pairs. One type contains an observation \(y\), the sum of 17 adjacent state variables, and a state variable that is \(k\) grid intervals from the center of these 17. The second type contains an observation \(y\) and another observation with a center that is \(k\) grid intervals from the center of \(y\). The correlation PDFs for these two types of pairs are quite different. When the two types of pairs are combined to estimate a single correlation PDF, the result is not very similar to either alone.

To improve the performance of the CER, correlation PDFs are estimated for 41 subsets of \((y, x)\) pairs. A total of 21 of the subsets are pairs of observations and the original state variables with distances from 0 to 20 grid intervals. The second 20 subsets are pairs of observations with other observations with distances from 1 to 20 grid intervals.

The RMSE for a 20-member ensemble for this CER application and a variety of GC localization half-widths are shown in Fig. 10. The CER produces time-averaged RMSE of less than 1.3 while the best GC is just less than 1.7. The best ELF from AL13 gives an RMSE of about 1.45, but only a single ELF was computed rather than one for pairs of observations with state variables and one for observations with observations as discussed in the previous paragraph.

The time mean localization for this case is shown in Fig. 11 along with the best GC function. Note that there are two curves for the CER localization: one for state variables and one for other observed variables. The state variable localization is somewhat reminiscent of ELFs for this case with values of about 0.5 for small distance and largest localization values at 7 grid intervals. However, the state localization at distances larger than seven intervals is larger and has more structure than for the
ELFs. The localization for observations has values greater than 1 for distances between 2 and 6 and is larger than the localization for state until distances of 15. The best GC function is bracketed between the two localizations for distances out to six intervals. The small RMSE for the CER suggests that sampling error explains a significant portion of the error in this case. Since the CER localizations are distinctly non-Gaussian, the GC is unable to compete. The ELF is also not as good presumably because it did not work with separate \((y, x)\) subsets for the state and observed variables.

The localization values greater than 1 for the observed variables suggest that in some cases, correlation sampling error can cause correlations to be too small, rather than too large. AL13 also found some instances of ELFs with values larger than one for small distances. It was argued that the ELFs were acting as an empirical inflation in those cases.

6. Low-order dry dynamical core

The low-resolution GFDL AM2 B-grid dynamical core (GFDL Global Atmospheric Model Development Team 2004) with 30 latitudes, 60 longitudes, and 5 levels that was used in A12 and Anderson et al. (2005) with forcing from Held and Suarez (1994) is used next. As in A12, surface pressure along with wind components and temperature at all five levels are observed every 12 h at 180 approximately regularly spaced latitudes and longitudes. Simulated observational errors are drawn from Normal(0, 200²) for surface pressure (PS, units are Pa), and Normal(0, 3²) for wind components \((U, V)\) and temperature \(T\) (units are \(\text{m s}^{-1}\) and \(\text{K}\), respectively).

The initial truth and ensemble for the experiment are identical to those in A12. Observations are generated from a 200-day truth integration and a 20-member ensemble assimilation is then applied. The first 50 days are discarded, and the final 150 days (300 assimilation times) are used to compute statistics. Baseline cases are run with GC localizations of 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0, and 1.6 rad (the last case results in very large RMSE and is not included in Fig. 12).

Corresponding assimilations with CER are also run. The CER cases have no GC localization, but they only allow observations to impact state variables that are within the region where the GC case had nonzero localization (i.e., observations impact state variables within twice the half-width). For the 1.6 rad case, this means that each observation in the CER impacts all state variables.

Subsets of observation–state pairs are a function of the horizontal separation (60 categories), vertical separation (4 categories), observation type (4 categories for PS, \(U, V, T\), and state variable type (4 categories) for a total of 3840 subsets. This is similar to the subsets used for the ELF method in Lei and Anderson (2014a).

Unlike in A12, no inflation is applied to either the baseline or CER cases in order to highlight the impact of localization as opposed to adaptive inflation. As noted in Anderson et al. (2005), this low-order GCM is unusually robust when used with no inflation. This makes it a convenient tool for looking at the impacts of the CER method without the complications of interaction with the adaptive inflation that is required for general good performance in the L96 applications.

The assimilation quality is evaluated by the spatial and temporal mean of the RMSE for the ensemble mean
of the model PS variables [as in (10) with $m$ indexing only the $M = 1800$ surface pressure variables from the model]. Results are qualitatively similar for any of the other model variables. Figure 12 shows the time mean PS RMSE as a function of background horizontal GC localization radius half-width for the baseline and CER cases (cf. to Fig. 6 in A12, which has inflation). The CER RMSE is always smaller than that for the corresponding GC case, and the CER values change very little as the impact region increases.

Figure 13 shows the time mean equivalent localization from the CER for the impact of a north–south wind component observation on the model’s middle level on east–west wind state variables on the same level (cf. to Fig. 10 in A12 for sampling error correction and A12’s Fig. 12 for a group filter; Anderson 2007) for the 1.6 half-width case. The maximum localization values are significantly less than one with non-Gaussian structure apparent close to the observation. For larger separations, the CER algorithm has very effectively determined that localization should be close to 0.

7. Discussion

The CER algorithm estimates the correlation distribution for different subsets of $(y, x)$ pairs as part of the assimilation process. With uniform initial PDFs, the CER ensemble filter converged to a good estimate of the true model state for all ensemble sizes greater than 4 for all low-order model assimilation cases discussed. When the prior correlation PDF for a subset is uniform, the CER localization is just the mean of the likelihood and closely approximates the behavior of the SEC algorithm in A12. As the correlation estimates evolve during the assimilation, the CER RMSE decreases. However, as shown by Fig. 7, it can take many assimilation steps for the correlation estimates to improve. All experiments were repeated starting with initial correlation PDFs that were Normal(0, 0.05), so that all $(y, x)$ pairs are expected initially to be nearly uncorrelated. All experiments also converged with this initial correlation PDF but most took longer to equilibrate. The initial correlation distribution can be viewed as similar to a background error correlation estimate in variational or hybrid assimilation. If one had good a priori estimates of expected correlation, for instance that $(y, x)$ pairs with large separation were uncorrelated, this information could be incorporated in the initial estimates to reduce convergence time. Once the prior correlation distributions have converged for applications like the ones examined here, there is no longer a need to update them. The converged correlation PDFs are essentially climatological distributions of the correlation and in some cases it might be possible to compute these offline.

The CER algorithm has several free parameters that impact performance. The relative weight, $\beta$, given to the latest posterior estimate in (9) was set to $1/10000$ here. Equilibrated RMSE was largely unaffected by varying this parameter between 1/100 and 1/100000. The algorithm equilibrated faster with larger $\beta$. However, larger values led to noisier estimates of prior correlation after
many assimilation cycles. Cases with 1/1000 that were run for 200000 assimilation steps began to have larger RMSE; lightly smoothing the discrete correlation PDF estimates every 10000 assimilation steps eliminated this increase in RMSE.

The second free parameter is \( r_{\text{crit}} \), used to avoid large equivalent localizations. Values between 0.01 and 0.2 produced nearly identical RMSE and had negligible impact on time mean localization. RMSE increased for values outside this range.

A third free parameter is the number of subintervals, \( S \), in the discrete representation of the correlation PDFs and likelihoods. The computational cost is a linear function of \( S \), so smaller is cheaper. Reducing \( S \) to 40 (subinterval width 0.05) had a negligible impact on any of the CER assimilations. Reducing \( S \) to 20 led to small increases in RMSE for all experiments.

The number of subsets for which correlation is estimated clearly has significant impact on the results as demonstrated by the case discussed in section 5c. In that case, the correlation distributions required for state variables and observed variables at the same distance were very different. However, it may be possible to combine subsets that do not have such large differences in correlation distributions. To explore this, the experiment in section 5b was repeated with fewer subsets. Cases were done with 2, 3, 4, 5, 7, 10, and 20 pairs of consecutive \((y, x)\) distances in each subset (a total of 10, 7, 5, 4, 3, 2, and 1 subsets). Figure 14 shows the time mean RMSE and spread for the original case and these additional cases with the best GC results as a baseline. RMSE and spread increase as the number of subsets is decreased, but cases with up to 4 distances per subset (a total of 5 subsets) are still better than the best GC. As shown in Fig. 8, the equilibrated correlation distributions for the first five distances are quite different. This suggests that application of CER in larger models might be able to proceed with a reasonable number of subsets, similar to the ELF results in Lei and Anderson (2014b). However, the fact that the CER did poorly for the case in section 5c when subsets were only a function of distance is informative. The ELF, which does an a posteriori fit to find a localization that minimizes the analysis RMSE, performed much better in this case. This suggests that the correlation PDF for the impact of summed observations on state variables is sufficiently different from the PDF for the impact on other observations, thus, combining them in one subset is not effective.

The CER algorithm only considers sampling error in the prior correlation, not in the sample standard deviations. Sample estimates of standard deviation have a small bias, but sample estimates of the quotient of standard deviations in (2) can be more significantly biased. An algorithm similar to the CER can correct for sampling error in the quotient. This algorithm was tested for all cases here, but had a negligible impact for ensemble sizes of 10 or greater. For smaller ensemble sizes, very small reductions in RMSE were found for most cases.

The CER algorithm requires \( O(S) \) multiplications and several sums of \( S \) elements to normalize PDFs for each \((y, x)\) pair where \( S \) is the number of subintervals used in the discrete representation of PDFs. The cost of the standard filter is \( O(N) \) for each \((y, x)\) pair. However, other parts of the assimilation and the model advance are identical. For \( N = 20 \), the CER algorithm implemented in the Data Assimilation Research Testbed (Anderson et al. 2009) required approximately 1.5 times as much computation as the basic ensemble filter. This ratio can be reduced by noting that the likelihood functions (red curve in Fig. 1) and correlation prior PDFs (Fig. 8) are significantly nonzero only for a fraction of the \([-1, 1]\) correlation range. More sophisticated implementations could only perform multiplications and sums for subintervals in which both are significantly nonzero. Since the correlation PDFs converge fairly rapidly during the assimilation, practical applications, especially for large ensemble sizes, could estimate PDFs for an initial spinup period. After that, a simple lookup table could return the value of the CER correlation given the ensemble sample correlation for a given subset. For subsets where the underlying correlation is always small (e.g., subsets at large horizontal distance in the low-order GCM), it takes fewer than 100 observation state variable pairs for a 10-member ensemble (and even fewer pairs for larger ensembles) to be able to state confidently that the correlation is very small. An efficient implementation could very quickly identify such subsets and cease to compute updates to reduce computation time.
The ELF approach has been used as a benchmark here since it does an a posteriori computation of a good localization. However, as implemented in work to date, the ELF requires an iterative process using a sequence of long observing system simulation experiments, so it is much more expensive and complex than the CER algorithm used here. In addition, the CER provides the capability to introduce a priori information about background correlation distributions; there is no similar capability for the ELF.

8. Conclusions and next steps

An algorithm that allows ensemble data assimilation without tuning localization functions has been developed and applied to low-order models. The algorithm assumes that ensemble sampling error in the computation of correlations is the primary source of error in an assimilation. This assumption is clearly false for some applications (like the example in section 4) since the standard ensemble adjustment Kalman filter with a large enough ensemble and no localization is the optimal solution. The fact that the new algorithm is competitive with other empirical methods for computing localization suggests, however, that the assumption may be approximately valid for many applications. Further research in larger models is required to determine if the new algorithm will be effective. For instance, issues related to the interaction of localization and model balance (Greybush et al. 2011; Kepert 2009) are not addressed by the results here (Oke et al. 2007). In addition, all experiments presented here are in situations with no model systematic errors. Realistic applications will include bias and it is possible that tuned a priori localization will be more effective in such cases. Applying the correlation error reduction algorithm to the large atmospheric model applications explored with empirical localization functions in Lei and Anderson (2014a,b) will be the next step.

Acknowledgments. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. Thanks to the DART team for support of the code and to Lili Lei, Abhishek Chatterjee, and three anonymous reviewers for constructive comments on earlier drafts.

REFERENCES


