Comparisons of Empirical Localization Techniques for Serial Ensemble Kalman Filters in a Simple Atmospheric General Circulation Model

LILI LEI AND JEFFREY L. ANDERSON
National Center for Atmospheric Research, Boulder, * Colorado

(Manuscript received 9 May 2013, in final form 7 October 2013)

ABSTRACT

Two techniques for estimating good localization functions for serial ensemble Kalman filters are compared in observing system simulation experiments (OSSEs) conducted with the dynamical core of an atmospheric general circulation model. The first technique, the global group filter (GGF), minimizes the root-mean-square (RMS) difference between the estimated regression coefficients using a hierarchical ensemble filter. The second, the empirical localization function (ELF), minimizes the RMS difference between the true values of the state variables and the posterior ensemble mean. Both techniques provide an estimate of the localization function for an observation’s impact on a state variable with few a priori assumptions about the localization function. The ELF localizations can have values larger than 1.0 at small distances, indicating that this technique addresses localization but also can correct the prior ensemble spread in the same way as a variance inflation when needed. OSSEs using ELF localizations generally have smaller root-mean-square error (RMSE) than the optimal Gaspari and Cohn (GC) localization function obtained by empirically tuning the GC width. The localization functions estimated by the GGF are broader than those from the ELF, and the OSSEs with the GGF localization generally have larger RMSE than the optimal GC localization function. The GGFs are too broad because of spurious correlation biases that occur in the OSSEs. These errors can be reduced by using a stochastic EnKF with perturbed observations instead of a deterministic EAKF.

1. Introduction

Since the ensemble Kalman Filter (EnKF) was first proposed by Evensen (1994) in an oceanographic application, it has subsequently been widely used in atmospheric applications (e.g., Burgers et al. 1998; Houtekamer and Mitchell 1998; Whitaker et al. 2004). However, the EnKF is subject to sampling error when it is applied in large geophysical applications with affordable ensemble sizes. Sampling error can lead to filter divergence where the filter no longer responds to the observations owing to the underestimated ensemble spread. In many cases, filter divergence due to insufficient spread can be avoided using covariance inflation (Anderson and Anderson 1999), which can be implemented in a number of ways (e.g., Anderson 2007b; Whitaker et al. 2008; Anderson 2009).

A related problem with using a limited number of ensemble members is the occurrence of spuriously large magnitudes for estimated covariances between observations and state variables. Houtekamer and Mitchell (1998) found that EnKF performance could be improved by excluding the impacts of distant observations on a state variable to reduce the impact of spurious correlations. The efficacy of a more general localization that limits the impact of an observation on remote state variables was demonstrated by Houtekamer and Mitchell (2001), Hamill et al. (2001), and Furrer and Bengtsson (2007).

Localization is often implemented by multiplying the regression coefficient (equivalently the component of the Kalman gain) by a distance-dependent function that traditionally varies from 1.0 at the observation location to 0.0 at a threshold distance. It can be implicitly implemented in some filters like the local ensemble transform Kalman filter (Ott et al. 2004; Miyoshi et al. 2007). The most commonly used localization function is the approximately Gaussian fifth-order piecewise continuous polynomial function from Gaspari and Cohn (1999, subsequently called the GC function). The GC function has a single parameter that defines its width, and this...
ically computed localization functions require additional a function of ensemble size and the correlation between an developed a related algorithm to compute localization as correlation lengths of the model. Anderson (2012) de-
localization based on sensitivity matrices and principal correlation. Emerick and Reynolds (2010) computed the error correlation and adapts to the width of the true error
function as a power of the ensemble cor-
discrete scales. Bishop and Hodyss (2009a,b) proposed
developed a multiscale filter, which replaces the sample covariance with a multiscale tree that is composed of nodes distributed over a relatively small number of discrete scales. Bishop and Hodyss (2009a,b) proposed a localization function as a power of the ensemble cor-
This localization function moves with the true error correlation and adapts to the width of the true error correlation. Emerick and Reynolds (2010) computed the localization based on sensitivity matrices and principal correlation lengths of the model. Anderson (2012) de-
veloped a related algorithm to compute localization as a function of ensemble size and the correlation between an observation and a state variable. However, these dynam-
computed localization functions require additional computations and increase the complexity of the filter.

A method to estimate the localization while making few a priori assumptions for the shape of the localization function, an empirical localization function (ELF), was proposed by Anderson and Lei (2013). The ELF takes sampling error and other potential errors into account and can automatically provide an estimate of the localization function for any possible observation type with a given kind of state variable. Anderson and Lei (2013) demonstrated that in a low-order model for an observation type whose forward operator involves a sum of state variables (similar to a satellite radiance) the ELF produced a non-Gaussian localization function and out-
performed the best GC localization function.

As a proof-of-concept, the ELF was first tested in the Lorenz96 model (Lorenz and Emanuel 1998). This study further explores the use of ELFs for serial EnKFs in the dynamical core of the Geophysical Fluid Dynamics Laboratory (GFDL) B-grid climate model (Anderson et al. 2004). The intermediate B-grid model has more realistic dynamics compared to the simple Lorenz96 model, and it has less complexity compared to a full-
physics numerical weather prediction (NWP) model. Thus the exploration of the ELF in the B-grid model helps to further understand the ELF and also provides a foundation for applying the ELF in real NWP models in subsequent work.

In addition to providing the localization function, the ELF can help to understand why localization is needed and what factors should be considered when creating a localization function. To further investigate these issues, the hierarchical filter proposed by Anderson (2007a) is also applied. These two localization algorithms are de-
scribed in section 2 and results from a simple ideal sim-
ulation are compared in section 3. This purely statistical simulation demonstrates the need for localization and examines the advantages of the ELF compared to the hierarchical filter when the prior ensemble sample cor-
relation is biased. The two localization functions are then applied and compared in an observing system simulation experiment (OSSE) with the B-grid model. The experi-
dmental design is described in section 4, and the results are discussed in section 5. The localization functions computed from the hierarchical filter are further discussed in
section 6 where different kinds of filters are applied. The conclusions are presented in section 7.

2. Algorithm

a. Computation of localization functions

Two kinds of localization functions are described here. The global group filter (GGF) minimizes the expected RMS difference between the regression coefficients using an extension of the hierarchical ensemble filter method (Anderson 2007a). The empirical localization function (ELF; Anderson and Lei 2013) is defined to minimize the root-mean-square error (RMSE) of the posterior en-
semble mean compared to the true value given the results of an OSSE.

Given an ensemble simulation experiment, define \(X\) as the set of all model-state variables that could be modified by assimilating observations in the simulation. Let \(x_m\) be an element of \(X\), \(m \in \{1, \ldots, M\}\), where \(M\) is the total number of state variables. Define \(Y\) as the set of all potential observations that could be assimilated in the simulation. Let \(y_l\) be an element of \(Y\), \(l \in \{1, \ldots, L\}\), where \(L\) is the total number of observations. Define the pair \((y, x)\), where \(y \in Y\) and \(x \in X\); then the set of all pairs consists of the domain for the localization function.

Here, the localization function is defined as a function of subsets of the domain. For instance, to find the localization value for a radiosonde temperature observation and a state variable \(u\)-wind component with a separation...
of 900–1000 km, the pairs of radiosonde temperature observations and model grid point values of u-wind component having separation between 900 and 1000 km can constitute a subset \( \{(y_k, x_k)\}, k \in \{1, \ldots, K\} \), where \( K \) is the total number of pairs in the subset. The localization value is computed from the ensemble information about the observations and state variables in the subset.

For each pair in the subset, a group of estimates of the regression coefficient between the observation and state variable can be provided by a hierarchical ensemble filter in which a set (group) of ensemble assimilations starting with different ensemble initial conditions are performed. The GGF is computed by minimizing the RMS difference between the estimated regression coefficients for all the pairs in the subset. Let \( b_{k,g} \) be the estimate of the regression coefficient for the \( k \)th pair in the subset from the \( g \)th group of ensemble assimilation, \( k \in \{1, \ldots, K\} \) and \( g \in \{1, \ldots, G\} \), where \( K \) is the total number of pairs in the subset, and \( G \) is the group size of the hierarchical ensemble filter. The function to be minimized for the subset is

\[
J = \sqrt{\sum_{k=1}^{K} \sum_{i=1}^{G} \sum_{j \neq i}^{G} (\alpha b_{k,j} - \hat{b}_{k,j})^2}, \tag{1}
\]

where \( \hat{b}_{k,j} \) is the regression coefficient for the \( k \)th pair from the \( i \)th group, and \( \alpha \) is the localization value. Equation (1) is similar to Eq. (2) in Anderson (2007a), but with an additional summation over all the pairs in the subset. Taking a derivative with respect to \( \alpha \) and seeking a minimum gives the localization value as

\[
\alpha = \left( \frac{\sum_{k=1}^{K} \left( \sum_{i=1}^{G} \hat{b}_{k,i} \right)^2}{\sum_{k=1}^{K} \sum_{i=1}^{G} \sum_{j \neq i}^{G} \sigma_{k,j}^2} - 1 \right) / (G - 1). \tag{2}
\]

While the GGF uses a hierarchical ensemble filter to compute the localization, the ELF uses the results from an OSSE to find the localization that minimizes the RMS difference between the posterior ensemble mean and the true value of the model-state variables in the subset (Anderson and Lei 2013). For the \( k \)th pair in the subset, the posterior ensemble mean of state variable \( x_k \) after assimilating observation \( y_k \)

\[
x_k^p = \bar{x}_k + \alpha \hat{b}_k \Delta y_k, \tag{3}
\]

where superscripts \( p \) and \( u \) denote prior and posterior, overbar is the ensemble mean, \( \Delta y \) is the mean increment of the observed quantity, \( \hat{b} \) is the sample regression coefficient, and \( \alpha \) is the localization value. In the subsequent derivation, the \( x, y, \) and \( b \) variables are obtained from the output of the original OSSE. To find the ELF localization, the cost function

\[
J = \sqrt{\sum_{k=1}^{K} (\bar{x}_k^t - x_k^t)^2} = \sqrt{\sum_{k=1}^{K} (\bar{x}_k^t + \alpha \hat{b}_k \Delta y_k - x_k^t)^2} \tag{4}
\]

is minimized where superscript \( t \) denotes the true value, and \( K \) is the total number of pairs in the subset. The cost function Eq. (4) is a minimum when its derivative with respect to \( \alpha \) is 0, so that

\[
\alpha = \sum_{k=1}^{K} (x_k^t - \bar{x}_k^t) \hat{b}_k \Delta y_k / \sum_{k=1}^{K} (\hat{b}_k \Delta y_k)^2. \tag{5}
\]

The synthetic observations \( Y \) in an OSSE are generated by random draws from a normal distribution \( N(y_k^o, \sigma_{o,k}^2) \), where \( y_k^o \) is the true observation value, and \( \sigma_{o,k}^2 \) is the observation error variance. The ELF can be computed with the true observation values by integrating over all the possibilities of the synthetic observations (Anderson and Lei 2013). This leads to a smoother ELF because the sampling error of the observations is reduced. The localization value \( \alpha \) with integral observations for Eq. (5) can be written as

\[
\alpha = \sum_{k=1}^{K} (x_k^t - \bar{x}_k^t) \hat{b}_k (A_k + B_k y_k^o) / \sum_{k=1}^{K} \{[\hat{b}_k (A_k + B_k y_k^o)]^2 + (\hat{b}_k B_k)^2 \sigma_{o,k}^2\}, \tag{6}
\]

where

\[
A_k = \left( \frac{\sigma_{u,k}^2}{\sigma_{o,k}^2} - 1 \right) y_k^o, \quad \text{and} \tag{7}
\]

\[
B_k = \frac{\sigma_{u,k}^2}{\sigma_{o,k}^2}. \tag{8}
\]

The \( y_k^o \) is the prior ensemble mean of the observed quantity of the \( k \)th pair in the subset \( \sigma_{p,k}^2 \), and \( \sigma_{o,k}^2 \) are the prior and posterior ensemble sample variances for the observation of the \( k \)th pair in the subset. The posterior ensemble sample variance \( \sigma_{u,k}^2 \) is given by

\[
\sigma_{u,k}^2 = \left[ \frac{\sigma_{o,k}^2}{\sigma_{o,k}^2} + \left( \frac{\sigma_{u,k}^2}{\sigma_{o,k}^2} \right)^{-1} \right]^{-1}. \tag{9}
\]

### b. Statistical test of the localization functions

The uncertainty in the two types of localization functions can be assessed by applying standard statistical tests. The statistical tests help to eliminate spurious nonzero empirical localization values, especially at large separations. A Student’s \( t \) test is used for the GGF, while a \( z \) test is used...
for the ELF because the standard error of the localization value $a$ can be obtained from the least squares fit [Eq. (5)] and the sample size $K$ is usually very large.

To apply the $t$ test for the GGF, the subset that contains $K$ pairs is first divided into $D$ new subsets in which there are $K/D$ pairs. The localization value $a$ is then computed for each of these new $D$ subsets using Eq. (2). With the null hypothesis that $a = 0$ and $D-1$ degrees of freedom, a $t$ value is computed, and the null hypothesis is accepted if the $t$ value is smaller than the critical value given by a confidence level.

For the ELF, the $z$ test is more appropriate than the Student’s $t$ test. Given Eq. (5), the ELF is the slope of the least squares fit, so the standard error of the slope can be computed. Given the slope and its standard error, a $z$ test can be applied with a null hypothesis that $a = 0$, and the null hypothesis is accepted if the $z$ value is out of the critical region given by a confidence level. The confidence level for both the Student $t$ test and $z$ test is $95\%$ in the following discussions.

3. Ideal simulation

The two kinds of localization function are first explored and compared in an ideal simulation. In the ideal simulation, the true correlation is known, and the simulated state variables and observations are random draws from a bivariate normal distribution. The statistical simulation demonstrates the need for localization and examines different impacts on the two localization functions from spurious correlations.

A set of $M$ true correlation values that are evenly distributed from 0 to 1 is selected

$$r_m' = (m - 1)/(M - 1), m = 1, \ldots, M,$$  

where superscript $t$ denotes the truth. For results shown in this section, $M$ is set to 101. For each $r_m'$, $K$ random samples with ensemble size $N$ drawn from a bivariate normal distribution with covariance matrix $\begin{bmatrix} 1 & r_m' \\ r_m' & 1 \end{bmatrix}$ are generated. Let $(y_{k,n}, x_{k,n}), n = 1, \ldots, N$ be the ensemble members of the $k$th sample, where $x$ is the state variable and $y$ is the observed quantity. For each sample, the true value $(y_{k,n}', x_{k,n}')$ is also a draw from the same bivariate normal distribution. Given an observation error variance $R$, the ELF localization for each true correlation value $r_m'$ can be computed using Eq. (6).

For each $r_m'$, $G$ groups of $K$ random samples with ensemble size $N$ from a bivariate normal distribution with covariance $\begin{bmatrix} 1 & r_m' \\ r_m' & 1 \end{bmatrix}$ are generated to simulate the hierarchical filter and compute the GGF. For each sample, there are $G$ estimated regression coefficients. The GGF localization value for each true correlation value $r_m'$ can be computed using Eq. (2).

By setting $K = 100000$, $G = 2$, and $R = 1.0$, the ELF and GGF localization functions for ensemble sizes ranging from 16 to 128 are shown in Fig. 1. The localization values are close to 1 (no localization) when the true correlation is close to 1 or when the ensemble size is large. On the other hand, the localization values are close to 0 for small ensemble sizes or small correlations (Bishop and Hodyss 2009a,b), indicating that the sampling error is relatively large. These results are consistent with the sampling error correction method (Anderson 2012), and demonstrate the need for localization with limited ensemble size and small correlations.

The ELF and GGF in Fig. 1 are similar to each other, especially with large ensembles when the true correlation values are known and the simulated state variables and observations are drawn from the same bivariate normal distribution. The GGFs with larger groups ($G = 4, 8, 16$) are similar to those for $G = 2$. The ELFs with a variety of observation error variances ($R = 0.1, 4.0, not shown$) are similar to those for $R = 1.0$.

Next, a simulated correlation bias $\Delta r$ is added into the true correlation $r_m'$ to examine the ability of GGF and ELF to detect spurious correlations in the presence of model bias. Let $\hat{r}_m$ be the true correlation with added bias. For each $\hat{r}_m, (y_{k,n}', x_{k,n}')$, the ensemble members are drawn from a bivariate normal distribution with covariance matrix $\begin{bmatrix} 1 & \hat{r}_m \\ \hat{r}_m & 1 \end{bmatrix}$, while the true value $(y_{k,n}', x_{k,n}')$ is still drawn from the bivariate normal distribution with unbiased covariance matrix $\begin{bmatrix} 1 & r_m' \\ r_m' & 1 \end{bmatrix}$. Using the random samples, the GGF and ELF are computed by Eqs. (2) and (6).
The ELF is narrower than the GGF.

The GGF and ELF with a variety of correlation biases are shown in Fig. 2, for $K = 100000$, $N = 32$, $G = 2$, and $R = 1.0$. The abscissa in Fig. 2 ends at 0.8 to avoid $r_m$ being larger than 1.0. The ELF has localization values around 0.0 when $r_m$ is close to 0.0 for all values of correlation bias. The ELF localization decreases as the correlation bias increases when $r_m$ is larger than 0.1. This indicates that the ELF is able to correct the spurious correlations and also demonstrates the need for localization. However, the GGF has localization values larger than 0 when $r_m$ is close to 0, with larger localization values as the correlation bias increases. Not surprisingly the GGF cannot correct the spurious correlations when correlation bias occurs, because the estimated regression coefficients used to compute the GGF are based on the random samples drawn from biased correlations. But the ELF uses the true value $(y^t, x^t)$ that is drawn from a bivariate normal distribution with true correlation $r_m^t$ and computes the localization from the error of the biased prior ensemble mean against the true value. This allows the ELF to detect the spurious correlations. The ELF and GGF are similar to each other when the correlation is unbiased, but when the correlation bias is introduced the ELF is narrower than the GGF.

4. Experimental design

To further explore the localization functions, an OSSE is conducted using the dynamical core of the GFDL B-grid atmospheric model (Anderson et al. 2004). The B-grid model state variables are surface pressure, temperature, and zonal and meridional wind components. The horizontal model grid points are 30 latitudes $\times$ 60 longitudes and there are 5 vertical levels; the model time step is 1 h.

The true initial condition is obtained from a long integration of the model. Small perturbations drawn from a normal distribution with mean 0 and standard deviation of $10^{-4}$ are added to each state variable of the true initial condition to generate $N$ perturbed states. The true initial condition and $N$ ensemble members are then advanced for 10 years. The result is a set of $N$ ensemble members that can be seen as random draws from the model’s climatological distribution and are the ensemble initial conditions for assimilation experiments. For a hierarchical ensemble filter, the ensemble initial conditions for the groups of ensemble simulation experiment are also generated with this procedure. The default ensemble size is 32, and the default group size $G$ for the hierarchical ensemble filter is 2.

The true model trajectory is integrated for an additional 30 yr and synthetic observations of temperature, zonal and meridional wind components, and surface pressure are generated by adding random draws from a normal distribution with mean 0 and specified observation error variances to spatially interpolated values from the gridded true state. The observation error variances are 2000 hPa$^2$ for surface pressure, 2 K$^2$ for temperature, and 8 m$^2$ s$^{-2}$ for zonal and meridional wind components. The spatial distribution of the synthetic observations is 300 profiles approximately evenly distributed on the sphere. Each profile has observations of temperature and wind components on every model vertical level plus surface pressure. Synthetic observations are available once per day.

Assimilation experiments starting from the ensemble initial conditions and assimilating the synthetic observations are conducted with the Data Assimilation Research Testbed (DART; Anderson et al. 2009) using the ensemble Kalman filter with perturbed observation (e.g., Burgers et al. 1998; Houtekamer and Mitchell 1998). The use of the ensemble adjustment Kalman filter (EAKF; Anderson 2001) will be discussed in section 6. To avoid filter divergence, time-varying but spatially uniform state-space adaptive inflation (Anderson 2009) is applied.

An assimilation experiment with a GC localization function is first conducted for 30 yr and the first 10 yr are discarded to eliminate initial errors from starting with a climatological ensemble distribution. The second 10 yr are used to compute the localization functions described in section 5. When computing the localization functions, the observation set $Y$ is comprised of the true gridded state values rather than the synthetic observations. As discussed in section 2, the ELF can be computed with the true observation value by integrating over all possible values of the synthetic observations [Eq. (6)]. This procedure reduces the sampling error from observational noise and results in a smoother ELF. Also the sample
size $K$ for each subset is increased by using the true gridded state values rather than the synthetic observations in set $\mathbf{Y}$. For instance, the size of the localization function domain $(\mathbf{Y}, \mathbf{X})$ for the surface pressure observations with the surface pressure variables at an assimilation time increases from $300 \times 1800$ to $1800 \times 1800$, because there are 300 synthetic surface pressure observations but 1800 true surface pressure gridded state values.

Additional assimilation experiments using the computed localization functions also start from the ensemble initial conditions and the first 10 yr are also discarded. The second (third) 10 yr are used for dependent (independent) verification of the different localization experiments. For each state variable at every model vertical level, the RMSE averaged over the horizontal domain and the second or third 10 yr is computed for verification.

5. Results

An assimilation experiment is first conducted from the climatological ensemble initial conditions with the default ensemble size 32 for 30 yr. The GC localization function (Gaspari and Cohn 1999) with half width of 0.8 rad (5096 km), the first guess for the GC localization half width, is applied in the assimilation experiment. The GGF and ELF localizations are computed from the second 10 yr for each observation type at every model vertical level with each kind of model state variable at every model vertical level. For instance, there is one empirical localization function for temperature observations at model level 3 with zonal wind variables at model level 4 as a function of horizontal separation. Thus there are a total of 256 (16 squared) empirical localization functions, given 16 state variables at every horizontal grid point (surface pressure, and temperature, zonal, and meridional winds at 5 vertical levels). There are 64 distance bins each with width 0.05 rad (319 km) for the empirical localization functions.

The ELFs for a temperature observation on the bottom level with state variables also on the model bottom level are shown in Fig. 3. The blue solid lines represent the raw ELFs, and the red dashed lines display the ELFs after setting values with $z$-test confidence level less than 95% to 0. The $z$ test helps to eliminate noisy estimates at larger separations. Note that the effect of the $z$ test is small here because of the large sample sizes for computing the ELF. The ELFs with state variables zonal and meridional winds (Figs. 3c and 3d) have no values at 0 distance because the wind components are on staggered model grid points. These four ELFs are similar to each other except for the detailed structure at small separations ($<0.2$ rad). The ELFs are narrower than the 0.8-rad GC function, decreasing to 0 at around 0.7 rad compared to 1.6 rad for the GC. This suggests that the first guess of GC half width is too broad. The 0.6-rad GC function has better agreement with the tails of the ELFs than the 0.8-rad GC function.

The ELFs have values larger than 1.0 at small distances ($<0.3$ rad). This suggests that the prior ensemble mean needs to move further toward the observations than in the GC assimilation, consistent with a prior ensemble spread that is too small. For an extreme case in which the observation type is the same as the kind of state variable and the distance between the observation and state variable is 0, the localization is simply an inflation. In this case, the sample regression coefficient $b$ in Eq. (3) is 1, and the localization value $\alpha$ in Eq. (3) scales the mean increment of the observed quantity $\Delta y$.

The green solid and dashed lines in Fig. 3 show the correlation coefficient and standard error of the slope from the least squares fit [Eq. (5)]. The correlation coefficients of temperature observations with nontemperature-state variables are smaller than for temperature observations with temperature, and the correlation coefficients become small for separations greater than approximately 0.5 rad. The standard error is small for all ELFs, because the sample size is large ($\sim 10^6$). The other ELFs that are not shown have similar patterns to those in Fig. 3.

To eliminate the impact of the ELF acting as an inflation, a normalized ELF (ELFN) is computed. The localization values for the subset in which the observation has the same type as the state variable and their separation is 0 are first computed. As noted above, these values can be regarded as empirical inflation values. The prior ensemble spreads are first inflated by the empirical inflation values and then used to compute the ELFNs by Eq. (6). The ELFNs for temperature observations on the bottom level with the state variables on the same level are shown in Fig. 4. Again the $z$ test helps to remove noise for large separations. Compared to the ELFs in Fig. 3, the ELFNs have similar structures but smaller localization values. The ELFN of temperature with temperature (Fig. 4b) has $\alpha$ of 1.0 at 0 separation by construction. The correlation coefficients and standard errors of ELFNs are similar to those of ELFs in Fig. 3.

The 256 GGF localizations with group size 2 are also computed using the second decade of the assimilation period. Figure 5 displays the GGFs of the temperature observation on the bottom level with the state variables on the same level (blue lines in Fig. 5). To estimate confidence, the training data are divided into $D = 8$ equal pieces and a GGF is computed for each. These eight subsample results (green lines in Fig. 5) are very similar to each other except at large separations ($>3.0$ rad). Based on these eight estimates of each GGF, the Student’s $t$ test
for the null hypothesis of $\alpha = 0$ is applied. Values that do not reject the null hypothesis at the 95% confidence level are set to 0 (red dashed in Fig. 5). Similar to the $z$ test for the ELFs, the effect of the Student’s $t$ test is small because of large sample sizes in the training data. There would be more obvious impact from the Student’s $t$ test when the sample size is smaller. The four GGFs with different state variables in Fig. 5 are similar to each other except for the detailed structure at small separation (<0.2 rad). Compared to the ELFs (Fig. 3), the GGFs have values either close to or smaller than 1.0 at 0 distance, thus the GGF is unable to detect the insufficient prior ensemble spread. The GGFs are broader than the ELFs and ELFN, although it is slightly narrower than the GC half width of 0.8 rad. The GGFs have smaller magnitudes than the 0.6-rad GC function at small separations, but they have broader tails than the 0.6 rad GC function. The other GGFs (not shown) have similar patterns to those in Fig. 5.

The ELFs, ELFNs, and GGFs are applied in new assimilation experiments starting from the climatological ensemble initial conditions. Figure 6 shows the RMSE with these three kinds of localization functions and various half widths of the GC localization functions from the second 10 yr.

On average the GC half width of 0.6 rad gives the smallest RMSE, although the GC half width of 0.5 rad has smallest RMSE for surface pressure and the GC half width of 0.55 rad has smallest RMSE for temperature at levels 2, 4, and 5, for zonal wind at level 4 and for meridional wind at level 5. This optimal GC half width of 0.6 is referred to as the GCO experiment.

The ELF (blue line in Fig. 6) has smaller RMSE than the GCO for most state variables and vertical levels; it has slightly larger RMSE for zonal wind at levels 3 and 4 and for meridional wind at level 4. The ELFN (red line) has smaller RMSE than the GC half-width 0.8 from which the localization functions are computed, while it has larger RMSE than the GCO and ELF. Thus the ELF has advantages over the ELFN because the ELF can correct the underestimated prior ensemble spread.
spread. The GGF (cyan line) has smaller RMSE than the GC half-width 0.8 for zonal and meridional winds in general, but it has larger RMSE than the GC half-width 0.8 for surface pressure and temperature, and the GGF has larger RMSE than the GCO, ELF, and ELFN. The localization functions provided by the GGF are broader than those from the ELF and ELFN (Figs. 3–5) and cannot correct the underestimated prior ensemble spread.

The adaptive inflation values averaged over the second 10 yr for the assimilation experiments are shown in Fig. 7. As expected, the inflation values increase as the GC half width increases because less localization results in larger errors from remote observation impacts and further reduced ensemble spread. The ELF has slightly larger inflation than the GCO. The ELFN has smaller inflation than the ELF because it has smaller localization values at small separations so it does not decrease the ensemble spread as much as the ELF. The GGF has larger inflation than the ELF and ELFN and smaller inflation than the GC half-width 0.8, which is consistent with being broader than the ELF and ELFN and narrower than the GC half-width 0.8. In an assimilation experiment with model error, the inflation values would be expected to increase.

The RMSE and inflation of the assimilation experiments from the third 10 yr are shown in Figs. 8 and 9. The third 10-yr simulation provides independent data for verification because the ELF, ELFN, and GGF are computed from the second 10-yr simulation with GC half-width 0.8. The ELF has similar RMSE to the GCO for surface pressure and temperature, and it has smaller RMSE for zonal and meridional winds except for the zonal wind at the model top. Since the B-grid model has heavy damping at the model top, the results at the top level can be misleading. Thus with independent data verification, the ELF can still outperform the GCO, but the ELF does not require intensive tuning of the localization parameter. Similar to the results from the second 10 yr, the ELFN has smaller RMSE than the GC half-width 0.8, while it has larger RMSE than the GCO and ELF. The GGF has larger RMSE than the GC half-width 0.8, ELFN, GCO and ELF, because the GGF is
broader than the ELF and ELFN and the GGF cannot
detect the underestimated prior ensemble spread.

Consistent with the inflation values from the second
10 yr of assimilation, the inflation values increase with
larger GC half widths in the independent data period.
The ELF has slightly larger inflation than the GCO. The
ELFN has smaller inflation than the ELF and GCO. The
GGF has larger inflation than the ELF and ELFN and
smaller inflation than the GC with half-width 0.8, which
indicates that the GGF is too broad.

Anderson and Lei (2013) show that the ELFs appear
to converge to a solution if the construction process is
iterated. Thus the output from the second 10 yr of as-
simulation with the ELF is used to construct a second
ELF. A z test with 95% confidence level and null hy-
pothesis that the localization value is 0 is also applied to
the second ELF. The second ELFs for temperature
observations at the bottom level with the state variables
at the same level after the z test are shown by the red
dash-dotted lines in Fig. 3. The second ELFs are very
similar to the first ELFs, except that the second ELFs
have larger values for separations between 0.5 and
1.0 rad. The RMSE of the second ELF in the second
(third) 10-yr simulation is shown in Fig. 6 (Fig. 8). In
general, the second ELF has similar RMSE to the first
ELF, and it also has smaller RMSE than the GCO. The
inflation value of the second ELF in the second (third)
10-yr simulation is shown in Fig. 7 (Fig. 9). Compared to
the first ELF, the second ELF has a larger inflation value,
which is consistent with the slightly larger localization
value between distance 0.5 and 1.0 rad.

The empirical localization functions shown in Figs.
3–5 are constructed from the second 10 yr of the sim-
ulation. However, similar results are obtained from the
empirical localization functions using 2- and 6-month
simulations. When even shorter simulation periods are
used to compute the empirical localization functions,
the difference between the empirical localization func-
tions and the empirical localization functions constructed
from a 10-yr simulation are larger (not shown), which is
consistent with Anderson and Lei (2013). Further re-
search is needed to understand how large a sample size
Fig. 6. The RMSE of the GC localization function with different half widths (black line), and the RMSE of the first ELF (blue line), the second ELF (green line), ELFN (red line), and GGF (cyan line) for each model state variable at every vertical level in the second 10 yr.
is required to produce a good empirical localization function.

For both dependent and independent data verification, the first and second ELFs generally compare favorably to the GCO (Figs. 6 and 8). The ELF does not assume a Gaussian-like localization function and does not require tuning the localization scale. Thus compared to the GC localization function, the ELF has more flexibility for the shape and may require less computational effort. The ELFs have smaller RMSE than the ELFN, because the ELFs are able to detect the underestimated prior ensemble spread and correct the underestimated spread with localization values $a$ larger than 1.0 (Figs. 3 and 4).

The GGFs with increased group size $G = 4$ are very similar to those with $G = 2$ (not shown). The localization functions computed by the GGF with increased group size are still too broad compared to the GCO. The RMSE and inflation of the GGFs with $G = 4$ are also similar to those of the GGFs with $G = 2$.

The broad localization functions provided by the GGF appear to result from spurious biased correlations between groups that occur due to assimilating the same observations for each group in the original OSSE. As discussed in an ideal simulation, the GGF is broader than the ELF when correlation biases exist because the GGF only uses estimated regression coefficients from a hierarchical filter. Therefore, in a subset, the pairs that are contaminated by spurious correlation biases give estimated regression coefficients that agree more than they should. These pairs then lead to larger GGF values, because the GGF is dominated by the pairs that have similar estimated regression coefficients from the groups [Eq. (2)]. For instance, for temperature observations with temperature variables at the bottom level, the ELFN has localization value around 0.017 at distance of 0.7 rad (Fig. 4b), while the GGF has localization value around 0.18 (Fig. 5b). The ELF does not use a hierarchical filter and uses the true observations and state variables that have correct correlations.

6. GGF and ELF computed from EAKF

To further explore why the GGF provides excessively broad localization functions, the GGFs and ELFs are next computed from an assimilation experiment using the deterministic ensemble adjustment Kalman filter (EAKF; Anderson 2001). The EnKF that uses perturbed observations has an additional source of noise by adding random errors to the observation priors during the assimilation. As discussed in section 5, the hypothesis for the broad GGFs is that spurious correlation biases between groups result from a group of ensemble simulations that assimilate the same observations. Therefore, the GGF computed from a group of ensemble simulations using the EAKF could have even broader GGFs because the deterministic assimilation does not add noise that acts to destroy the spurious correlation biases between groups.

Figure 10 shows the GGFs and ELFs of surface pressure observations with surface pressure variables for the EnKF and EAKF computed from an assimilation with GC half width of 0.4 rad, 32 ensemble members, and two groups. The last 200 days of the 1-yr assimilation are used to compute the GGFs and ELFs, and the statistical tests are not applied. The GGF for the EAKF is much broader than for the EnKF. For instance, the EnKF GGF for separation 1 rad is around 0.13, while the EAKF GGF is around 0.42. The ELF for the EAKF is similar to that for the EnKF; both of them are around 0.08 for separation 1 rad and decrease to 0 for separation 2 rad. The other 255 GGFs and ELFs (not shown) have similar patterns as in Fig. 10. By applying the 256 EAKF GGFs in a new simulation experiment starting from the climatological ensemble initial conditions, the GGF EAKF has larger RMSE and inflation than the GGF EnKF consistent with the broader localization functions.

To diagnose the differences between the GGFs and ELFs for the EnKF and EAKF, for each pair in the subset of surface pressure observations with surface pressure variables having separation between 0.45 and 0.5 rad and
FIG. 8. As in Fig. 6, but for the third 10 yr of the simulation.
from the last 20 days of the assimilations, the estimated regression coefficients with the EnKF and EAKF are shown in Figs. 11a and 11b. If the estimated regression coefficients from the two groups for each pair agree, the blue dots should align along the one-to-one line with slope 1 (i.e., the GGF localization is 1). On the other hand, if the estimated regression coefficients from the two groups are independent, the blue dots should be randomly distributed and the expected slope is zero (i.e., the GGF localization is 0). Figures 11a and 11b show that the estimated regression coefficients between the two groups from the EAKF are more correlated than those from the EnKF because there are larger spurious correlation biases between the two groups from the EAKF than those from the EnKF. Thus the slope for the EAKF is larger than for the EnKF, consistent with the broader GGFs for the EAKF (Fig. 10).

Similar to Figs. 11a and 11b, Figs. 11c and 11d show the prior error and mean increment of each pair in the subset from the first group of assimilations with the EnKF and EAKF. The slope of the least squares fit equals the ELF defined in Eq. (5). The EAKF has less increment variation than the EnKF, but the slope (i.e., the localization) for the EAKF is similar to that for the EnKF. Thus the ELF has less impact from the EAKF than the GGF. Moreover, the slopes based on the prior errors and increments are smaller than those based on the estimated regression coefficients between the two groups for the EAKF and EnKF, and this is consistent with Fig. 10.

7. Summary and conclusions

Two types of localization functions are discussed and compared in this manuscript. One minimizes the RMS difference between the estimated regression coefficients from a hierarchical ensemble filter (GGF), and the other minimizes the RMS difference between the true values and posterior ensemble mean (ELF). Both the GGF and ELF are able to automatically provide an estimate of the localization function of an observation type and a kind of state variable without assuming a Gaussian-like localization function and tuning the width of the GC localization. In an ideal simulation in which the true correlation is known, the need for localization is demonstrated. The GGF and ELF are similar to each other, especially when the ensemble size is large. When spurious correlation biases occur, the GGF provides broader localization functions than the ELF because the ELF uses the true values and the GGF relies on the estimated regression coefficients from a group of ensemble simulations that can be contaminated by spurious correlation biases. Thus the ELF is able to correct the spurious correlations more effectively than the GGF.

The GGF and ELF are compared in an OSSE that is conducted with the dynamical core of the GFDL B-grid climate model. The ELFs of an observation type with different kinds of state variables are generally similar to each other. Similar to the GC localization function, the ELFs decrease with increasing separation. However, unlike the GC localization function, the ELFs can have values larger than 1.0 at 0 separation indicating insufficient prior ensemble spread. The ELF can correct the underestimated spread and play the role of an inflation when the ensemble spread is too small. A normalized ELF (ELFN) is computed to avoid this inflation effect of the ELF. The ELFNs have smaller values of $\alpha$ and similar structures at small separations compared to the ELFs. The GGFs of an observation type with different kinds of
state variables are similar to each other in general. The GGFs have broader localization functions than the ELFNs and ELFs, and the GGFs have values either close to 1.0 or smaller than 1.0 at 0 distance. The GGF cannot detect the underestimated ensemble spread. Compared to the GC localization function, the empirical localization functions correct the localization function shape that has impact on how to spread the observation information, and they also optimize the distance beyond which the localization value is 0.0.

The ELFs, ELFNs, and GGFs are applied as the localization in additional assimilation experiments. In both dependent and independent data verification periods, the ELF generally has smaller RMSE than the best GC half width, thus the ELF can outperform the optimal GC localization function without assuming a Gaussian-like localization function or requiring intensive computational effort to tune the localization scale. The ELFN has larger RMSE than the optimal GC localization function because it does not correct for the insufficient ensemble spread. The ensemble spread could increase if a spatially and temporally varying adaptive inflation was used instead of the temporal varying but spatially uniform adaptive inflation used here. The GGF has larger RMSE than the ELF and ELFN because the GGF provides overly broad localization functions and it cannot detect the insufficient ensemble spread.

Similar results are obtained by varying the initial guess of the half width of the GC localization function, ensemble size, and adding model noise. The GGFs with

FIG. 11. The estimated regression coefficients of each pair in the subset of surface pressure observations with surface pressure variables having separation between 0.45 and 0.5 rad from two groups for (a) EnKF and (b) EAKF, and the prior error and mean increment of the pairs for (c) EnKF and (d) EAKF. The red line denotes the least squares fit of the data.
group size 4 are similar to those with the default group size 2. Thus the GGF tends to have excessively broad localization functions and results in large RMSE and inflation compared to the ELF and ELFN. The broad localization functions provided by the GGF are caused by spurious correlation biases between groups built up by the data assimilation in the B-grid model with a hierarchical ensemble filter. The spurious correlation biases between groups lead to larger localization values than those without spurious correlation biases between groups as demonstrated in an ideal simulation. It is further demonstrated by computing the GGF in an OSSE using the deterministic EAKF. The GGF from a group of ensemble assimilations using the EAKF is even broader, because the EnKF with perturbed observations adds additional noise to the simulations compared to the deterministic EAKF, and this additional noise destroys the spurious correlation biases between groups that are built up by the data assimilation. The GGF underestimates the localization in the dynamical core of the GFDL B-grid climate model, but the GGF for other models that are more chaotic needs further investigation.

Given the experimental design in this manuscript, the ELF and GGF may require less computational cost than finding an optimal half width for the GC localization function. The GGF requires $G$ (group size of a hierarchical ensemble filter) times more computational cost than the ELF. The empirical localization functions will have more advantages of computational expense over manually tuning the GC half width when multiple GC half widths need to be tuned for different subsets of observations and state variables.

Although the empirical localization functions are applied here only to serial ensemble Kalman filters, it may be possible to approximate the method in ensemble Kalman filters that simultaneously assimilate observations. First, an ELF can be computed by Eq. (6) from the output of any ensemble OSSE, serial or simultaneous. For ensemble Kalman filters that assimilate observations simultaneously, a localization matrix that has the same dimension as the background error covariance matrix can be constructed that approximates the empirical localization function, and the Schur product of the background error covariance matrix and the localization matrix provides the localized background error covariance matrix.

The ELF and GGF are potentially scalable to full-physics operational NWP models. Further investigations of the ELF and GGF in NWP models are needed, since the dynamical core of the GFDL B-grid climate model is a simple atmospheric general circulation model. Implementation of ELFs in the Community Atmospheric Model version 5.0 (CAM; Neale et al. 2010) and the Weather Research and Forecasting Model (WRF; Skamarock et al. 2008) is underway.

In this manuscript, the ELF uses the true values; however, ELFs can also be computed from the output of real ensemble assimilation experiments without knowing the truth (Anderson and Lei 2013). Also the GGF and ELF could be computed on the fly during an assimilation rather than using archived output from a separate assimilation experiment. This procedure will be explored in a future report.

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 Doug Nychka and Zhengyuan Zhu for helpful discussions and Nancy Collins and Tim Hoar for technical support and to two anonymous reviewers who helped to significantly improve an earlier version.

REFERENCES


