

Data assimilation using a climatologically augmented local ensemble transform Kalman filter

By MATTHEW KRETSCHMER^{1*}, BRIAN R. HUNT² and EDWARD OTT^{1,3}, ¹*Department of Physics, University of Maryland, College Park, MD, USA*; ²*Department of Mathematics, University of Maryland, College Park, MD, USA*; ³*Department of Electrical Engineering, University of Maryland, College Park, MD, USA*

(Manuscript received 11 November 2014; in final form 20 April 2015)

ABSTRACT

Ensemble data assimilation methods are potentially attractive because they provide a computationally affordable (and computationally parallel) means of obtaining flow-dependent background-error statistics. However, a limitation of these methods is that the rank of their flow-dependent background-error covariance estimate, and hence the space of possible analysis increments, is limited by the number of forecast ensemble members. To overcome this deficiency ensemble methods typically use empirical localisation, which allows more degrees of freedom for the analysis increment by suppressing spatially distant background correlations. The method presented here improves the performance of an Ensemble Kalman filter by increasing the size of the ensemble at analysis time in order to boost the rank of its background-error covariance estimate. The additional ensemble members added to the forecast ensemble at analysis time are created by adding a collection of ‘climatological’ perturbations to the forecast ensemble mean. These perturbations are constant in time and provide state space directions, possibly missed by the dynamically forecasted background ensemble, in which the analysis increment can correct the forecast mean based on observations. As the climatological perturbations are calculated once, there is negligible computational cost in obtaining the additional ensemble members at each analysis cycle. Included here are a formulation of the method, results of numerical experiments conducted with a spatiotemporally chaotic model in one spatial dimension and discussion of possible future extensions and applications. The numerical tests indicate that the method presented here has significant potential for improving analyses and forecasts.

Keywords: ensemble Kalman filter, climatological covariance, data assimilation

1. Introduction

Data assimilation aims to optimally combine a best-guess forecast of a system state with observations of the true system state. This best-guess is known as the background state and is typically made with a computer model. Modern data assimilation includes variational and ensemble methods, which update the background state vector using either static (for variational methods) or flow-dependent (for ensemble methods) background-error covariance models. Due to the size of atmospheric models, static covariances between atmospheric variables are typically approximated as spatially homogeneous and isotropic (Wang et al., 2007). Static covariance matrices have high rank and can effectively encode

important relationships and balance constraints. In contrast, flow-dependent covariance matrices, such as those approximated using ensemble methods, can distinguish physical spatially varying and time-dependent relationships that may not be accounted for by static covariance matrices. However, such ensemble-derived covariance estimates are typically restricted to much lower rank than the covariance estimates used in variational methods. The empirical technique of localisation (Houtekamer and Mitchell, 1998, 2001; Hamill et al., 2001; Ott et al., 2004) is the most often used method for compensating for the rank deficiency of ensemble-derived covariance estimates (see Section 2 for further discussion).

The aim of this paper is to propose a method for improving the performance of ensemble methods by effectively increasing the rank of their estimates of background-error covariance relationships. The way in which this is accomplished is through the inclusion of information from a static,

*Corresponding author.
email: mkretsch@umd.edu

climatologically derived background-error covariance estimate, such as that which is often used in variational methods. One of the simplest and most common ways to combine flow-dependent and static covariance estimates is through a linear combination of their respective covariance matrices, which is essential to the workings of hybrid data assimilation methods (Hamill and Snyder, 2000; Lorenc, 2003). In contrast, here we consider remaining entirely within the ensemble data assimilation framework and achieve higher rank by using climatological information to construct additional ensemble members. This approach achieves a higher-rank estimate of the background-error covariance matrix without simultaneously increasing the number of forecasted ensemble members.

Referring to the ensemble members corresponding to computed forecasts as the ‘dynamic’ ensemble, at analysis time we augment the ensemble by adding supplementary static ‘climatological’ ensemble members. These climatological ensemble members are formed by adding constant-in-time perturbations to the background dynamic ensemble mean at each analysis time. In our method, these climatological perturbations are chosen to be approximately parallel to the directions of the leading eigenvectors of a static, climatological background-error covariance matrix, thus potentially enabling the analysis to allow for additional error directions that may not be well-represented by the dynamic ensemble. After these new ensemble members have been created and the ensemble has been enlarged, assimilation is performed on the collection of both dynamic and climatological ensemble members. The analysis ensemble members which correspond to the updated background (dynamic) ensemble members are then forecasted to the next analysis time, and the cycle is repeated. The accuracy of ensemble data assimilation methods can strongly depend on the number of ensemble members used, and we believe that our method will obtain benefits from increased ensemble size, without correspondingly increasing the number of forecasts carried out.

A related but fundamentally different technique to enhance ensemble perturbations is additive covariance inflation (Hamill and Whitaker, 2005; Wang et al., 2009). This approach adds perturbations randomly sampled from a climatological error distribution to each ensemble member during each analysis cycle. A key difference between this approach and the climatologically augmented Local Ensemble Transform Kalman Filter (caLETKF) is that the caLETKF increases the size of the ensemble, and hence the rank of the background covariance.

The rest of this paper is organised as follows. Section 2 provides background on ensemble data assimilation methods, while Section 3 formulates our proposed method. Section 4 describes numerical experiments testing the effectiveness of

our approach, and Section 5 contains a review and discussion of the results. Section 6 contains our conclusions and additional discussion.

2. Background: the ensemble Kalman filter

Ensemble data assimilation methods are newer than variational methods and show great potential for many geophysical applications. Specifically, we focus here on the ensemble Kalman filter method (Evensen, 1994; Burgers et al., 1998). These filters estimate the time-evolving background covariance matrix, $\mathbf{P}^b$, as the sample covariance of an ensemble of k_d model forecasts. Each of these forecasts can be done independently of the others, allowing for a naturally parallel computational method for forecasting background-error covariances. However, due to computational limitations, the ensemble size k_d is typically much smaller than the size N of the model state vector, $k_d \ll N$. Thus, despite gaining flow-dependence, the estimate of $\mathbf{P}^b$ is severely rank deficient. Empirical techniques such as localisation are one way to remedy this rank deficiency. Localisation works by suppressing correlations between model variables beyond some spatial distance determined by an empirically tuned localisation radius. Though localisation can substantially increase the effective rank of background covariances, practical limitations on the ensemble size may still result in information being missed by the ensemble. Decreasing the localisation radius in such a situation is not necessarily a solution to this problem, as too much localisation can lead to deleterious effects, such as increased imbalances in the analyses (Greybush et al., 2011). On the contrary, larger localisation radii may improve state estimates at the cost of requiring a larger ensemble.

There are several variants of the ensemble Kalman filter (e.g. Houtekamer and Mitchell, 1998; Anderson, 2001; Bishop et al., 2001; Whitaker and Hamill, 2002; Ott et al., 2004; Wang et al., 2004). A class of ensemble filters, known as square root filters, work by finding transformations that, when applied to the background ensemble members, produce a new collection of ensemble members whose mean and sample covariance obey the Kalman filter equations. Though we apply our method to one of these Kalman filter variants known as the Local Ensemble Transform Kalman Filter (LETKF) (Hunt et al., 2007), it can also be applied to other ensemble Kalman filter formulations.

The LETKF finds an ensemble of analysis state vectors whose mean and sample covariance match those given by the Kalman filter equations, in the case of a linear observation operator H . It represents the k_d member background ensemble by its mean $\bar{\mathbf{x}}^b$, a N -dimensional vector,

and a $N \times k_d$ dimensional matrix of perturbations $\mathbf{X}^b$, obtained by subtracting the mean from each ensemble member. The LETKF finds the analysis ensemble mean $\bar{\mathbf{x}}^a$ and ensemble perturbations $\mathbf{X}^a$ by transforming $\bar{\mathbf{x}}^b$ and $\mathbf{X}^b$ via

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^b + \mathbf{X}^b \mathbf{w}, \quad (1)$$

$$\mathbf{X}^a = \mathbf{X}^b \mathbf{W}. \quad (2)$$

In the LETKF, the transformations $\mathbf{w}$ and $\mathbf{W}$ are found locally at each model grid point and depend on the observations (and their error covariance matrix $\mathbf{R}$) contained within a local analysis region centred at each model grid point. The transformations in eqs. (1) and (2) are applied at each grid point to the ensemble of state vectors at that grid point. The global analysis ensemble is then formed by combining the results of each local analysis.

3. A climatologically augmented ensemble Kalman filter

As mentioned above, we present here a method which incorporates climatologically derived error covariance information into a pure ensemble data assimilation framework. In ensemble Kalman filters, a collection of k_d model forecasts is used to estimate the most likely system state and its uncertainty. Each of these forecasts is integrated using the full non-linear dynamics from one analysis time to the next, and is represented by an N -dimensional state vector, which we denote as $\tilde{\mathbf{x}}_d^i$. The ensemble mean, given by

$$\bar{\mathbf{x}}_d = \frac{1}{k_d} \sum_{i=0}^{k_d} \tilde{\mathbf{x}}_d^i,$$

is interpreted as the best estimate of the system state. Its uncertainty is estimated by the dynamic ensemble sample covariance matrix,

$$\mathbf{P}_d^b = \frac{1}{k_d - 1} \sum_{i=0}^{k_d} (\tilde{\mathbf{x}}_d^i - \bar{\mathbf{x}}_d)(\tilde{\mathbf{x}}_d^i - \bar{\mathbf{x}}_d)^T = \frac{1}{k_d - 1} \mathbf{X}_d \mathbf{X}_d^T,$$

where $\mathbf{X}_d$ is the $N \times k_d$ dimensional matrix of dynamic ensemble perturbations, $\mathbf{x}_d^i = \tilde{\mathbf{x}}_d^i - \bar{\mathbf{x}}_d$, and the superscript T denotes the transpose operation. We incorporate climatological covariance information into this framework by increasing the size of the ensemble at analysis time, by adding to the dynamic ensemble a collection of k_c ‘climatological’ ensemble members, $\tilde{\mathbf{x}}_c^j$, for $k_d + 1 \leq j \leq k_d + k_c$. In practice, the size k_c of this climatological ensemble is determined by a number of factors including convenience, empirical tests of forecast effectiveness and the user’s computational resources. Each of these N -dimensional climatological ensemble members are created by adding a

climatologically derived perturbation $\mathbf{x}_c^j$ to the dynamic ensemble mean:

$$\tilde{\mathbf{x}}_c^j = \bar{\mathbf{x}}_d + \mathbf{x}_c^j.$$

Thus, the ensemble on which the analysis is performed may be represented as a mean $\bar{\mathbf{x}}_d$ and a $N \times (k_d + k_c)$ dimensional perturbation matrix $\mathbf{X} = [\mathbf{X}_d \ \mathbf{X}_c]$, where the columns of $\mathbf{X}_d$ are the dynamic ensemble perturbations, and the columns of $\mathbf{X}_c$ are the climatological ensemble perturbations.

Our particular choice of climatological perturbations $\mathbf{X}_c$, described below, is made under two conditions. First, in order that the mean of the combined ensemble be the same as the mean of the dynamical ensemble, we require that the mean of the columns of $\mathbf{X}_c$ be 0 (note that the mean of the columns of $\mathbf{X}_d$ is 0 by definition). Second, we want the population covariance $\mathbf{P}_c^b = \frac{1}{k_c} \mathbf{X}_c \mathbf{X}_c^T$ of the climatological ensemble to approximate the true climatological background-error covariance matrix $\mathbf{B}$. The sample covariance $\mathbf{P}^b = \frac{1}{k_d + k_c - 1} \mathbf{X} \mathbf{X}^T$ of the combined ensemble may then be written as

$$\begin{aligned} \mathbf{P}^b &= \frac{1}{k_d + k_c - 1} [(k_d - 1) \mathbf{P}_d^b + k_c \mathbf{P}_c^b] \\ &= \frac{k_d - 1}{k_d + k_c - 1} \mathbf{P}_d^b + \frac{k_c}{k_d + k_c - 1} \mathbf{P}_c^b. \end{aligned} \quad (3)$$

Thus, as with many hybrid methods (Hamill and Snyder, 2000; Lorenc, 2003), the implicit background covariance $\mathbf{P}^b$ is a linear combination of a dynamical (flow-dependent) covariance $\mathbf{P}_d^b$ and a climatological covariance $\mathbf{P}_c^b$, with coefficients whose sum is 1.

In most applications, the true climatological background-error covariance matrix, $\mathbf{B}$, is not known a priori, and is estimated using various physical arguments and statistical techniques (we denote this estimate by $\mathbf{B}_{\text{est}}$). The climatological perturbations $\mathbf{x}_c^j$, which are the columns of $\mathbf{X}_c$, are derived from the orthogonal eigenvectors of the background-error covariance estimate $\mathbf{B}_{\text{est}}$. Specifically, they are derived from k_c columns of the matrix $\mathbf{A} = \mathbf{V} \mathbf{D}^{1/2}$, where $\mathbf{V}$ is a $N \times N$ dimensional matrix whose columns are the orthonormal eigenvectors of $\mathbf{B}_{\text{est}}$, and $\mathbf{D}$ is a $N \times N$ dimensional diagonal matrix of the eigenvalues of $\mathbf{B}_{\text{est}}$, so that $\mathbf{B}_{\text{est}} = \mathbf{A} \mathbf{A}^T$. The collection of k_c scaled orthonormal eigenvectors, where $k_c \ll N$, is made of the eigenvectors which correspond to the k_c largest eigenvalues of $\mathbf{B}_{\text{est}}$, and we interpret these climatological vectors as representing the k_c directions of greatest climatological error variability. Once these k_c columns of $\mathbf{A}$ have been chosen, we multiply each of them by $\sqrt{k_c}$, to make their covariance matrix (normalised by k_c) approximate $\mathbf{B}_{\text{est}}$. We then subtract from these k_c vectors their mean, to make their sum equal zero.

We store the resulting climatological perturbation vectors in the columns of the $N \times k_c$ dimensional matrix X_c .

In our experiments, we use as an analysis algorithm the LETKF (Ott et al., 2004; Hunt et al., 2007). We perform the analysis on the collection of $k_d + k_c$ dynamic and climatological ensemble members, and upon completion of the analysis procedure, the ensemble mean is updated according to eq. (1). The analysis also produces a collection of $k_d + k_c$ analysis ensemble perturbations. The ‘dynamic’ analysis ensemble perturbations are created from the first k_d analysis ensemble perturbations, which for the LETKF are the ones that are closest to the dynamic background ensemble perturbations [Ott et al. (2004), in particular, see Appendix A]. The mean of these k_d perturbations is subtracted from each perturbation to yield the dynamic analysis ensemble perturbations. The dynamic analysis ensemble members are then calculated by adding the analysis ensemble mean to each of these dynamic analysis perturbations. This k_d member analysis ensemble is then forecasted forward in time to the next analysis time. Especially for other Ensemble Kalman filters, it may be fruitful to consider other methods of selecting k_d analysis perturbations from a $(k_d + k_c)$ -member analysis ensemble. Our choice of the first k_d perturbations is appropriate for filters like the LETKF and perturbed observations EnKF (Burgers et al., 1998; Houtekamer and Mitchell, 1998), for which there is a natural correspondence between pairs of background and analysis ensemble members.

We think of the new climatological members as representing potential error directions that might not be captured by the dynamically forecasted ensemble members. In this sense, the data assimilation algorithm can search in a higher dimensional space for corrections to the dynamic ensemble. Though the addition of ensemble members increases the cost of the analysis computation, we show in Section 5 that a major benefit of the caLETKF is that greater analysis accuracy can be achieved with fewer forecasts. We note that the static, climatological perturbations only need to be calculated once, so that the increased computational cost of our method comes purely from carrying out the analysis in a higher dimensional space.

4. Set-up of our numerical experiments

We apply our climatological ensemble augmentation method in a series of experiments that use a one-dimensional chaotic model which we call Lorenz Model II (Lorenz, 2005). Lorenz Model II represents the flow of an ‘atmospheric-like’ quantity Z around a circle of constant ‘latitude’. Specifically, Z is defined on a lattice of N grid points with periodic boundary conditions. The value of Z at a given grid

point n is denoted by Z^n , and its temporal behaviour is given by

$$\frac{dZ^n}{dt} = [Z, Z]^{K,n} - Z^n + F. \quad (4)$$

The terms on the right-hand side of eq. (4) are meant to roughly model quantities analogous to spatially averaged non-linear advection, linear dissipation, and constant forcing, respectively. For our experiments, we set the parameter F to have a value of 15, and take $N = 240$. The first term on the right-hand side of eq. (4) is given by

$$[Z, Z]^{K,n} = \frac{1}{K^2} \sum_{j=-J}^J \sum_{l=-J}^J (Z^{n-K+j-l} Z^{n+K+j} - Z^{n-2K-l} Z^{n-K-j}), \quad (5)$$

where K is an integer-valued parameter, $J = K/2$ if K is even and $J = (K+1)/2$ if K is odd. If K is even, the primed summation notation in eq. (5) denotes an ordinary sum with the first and last terms in the sum each multiplied by $1/2$. For our experiments, $K = 8$. The result of the spatial averaging present in eq. (5) is that the Model II states are characterised by spatially smooth waves.

We construct an estimate, B_{est} , of the background-error covariance matrix through the National Meteorological Center (NMC) Method of Parrish and Derber (1992). This method approximates the background-error covariance matrix using a large set (for our experiments ~ 50000) of differences between 6- and 24-hour forecasts that verify at the same time. These forecasts were started from initial conditions generated using a 40-member LETKF. To account for magnitude differences between the estimate of the background-error covariance matrix found using the NMC method and the true climatological background-error covariance matrix, the estimate of the background-error covariance matrix estimated using the NMC method is typically multiplied by a constant scalar factor that is tuned for the particular application (e.g. 3D-Var). To accomplish this tuning within the caLETKF, the climatological perturbations X_c , generated as described in Section 3, are multiplied by a scalar factor α . For the cases explored here, we found that the analysis error was insensitive to the precise value of α , and that the value $\alpha = 1$ produced a near-minimum error across a range of values of k_c and k_d . Thus, we used $\alpha = 1$ for all of the results reported here.

A key step in our proposed method is the generation of the leading eigenvectors of the static, climatological background-error covariance matrix. For the experiments described here, small system size allowed all of the eigenvectors of the climatological background-error covariance matrix to be easily calculated. However, for large, operational meteorological applications, explicit diagonalisation of the complete B matrix for systems of order $N \sim 10^8 - 10^9$ is

problematic, since B is too large to be stored in a computer. On the contrary, the approximate effect of multiplying vectors by such matrices is available and is, for example, widely used for preconditioning in variational data assimilation. Furthermore, we only require the leading eigenvectors, not all of them, and methods such as the Lanczos algorithm are capable of calculating the leading $m \ll N$ eigenvectors numerically while requiring only m evaluations of the action of B on a vector (Golub and Van Loan, 1996). Exploiting special structure of the background-error covariance matrix can make these numerical procedures even more efficient and tractable. We again emphasise that these operations must only be completed once and can be done offline, as they are the same at each analysis cycle.

We employ a ‘perfect model’ set-up in our numerical experiments, so that the truth against which our state estimate is compared is generated from a free model integration that uses the same parameter values and dynamics as are used to forecast the ensemble. Observations are generated by adding Gaussian white noise, with mean 0 and standard deviation 1, to the truth state vector at the observation times and locations. We chose a spatially homogeneous, static observation network. Specifically, we assimilate 12 observations at each analysis time, with one observation every 20 model grid points. We perform assimilations every 0.05 model time units, analogous to approximately every 6 hours for the atmosphere (Lorenz, 2005). The initial ensembles used in our numerical experiments are found by randomly sampling widely time-separated states from a long run of the forecast model states.

As a benchmark for comparison, we use results produced when analyses are performed using the standard LETKF algorithm (i.e. without a climatological supplement to the ensemble). Benchmark runs have the same ensemble size, assimilate the same set of observations, and compare against the same truth run as experiments with the climatologically augmented algorithm. In both sets of experiments, the localisation radius used by the analysis algorithm is constant at 20 model grid points, with no tapering of the observation influence (Hunt et al., 2007). In addition, we utilize 3% multiplicative covariance inflation (Anderson and Anderson, 1999) in all experiments involving the standard LETKF and 2.75% multiplicative covariance inflation for experiments using the caLETKF. These inflation factors were tuned to minimise analysis root-mean-square error (RMSE) over the interval of 0–7% inflation.

To compare the performance of the caLETKF and the LETKF, we measure the RMSE between the truth and forecast ensemble mean for several forecast lead-times f . We denote the difference at location n between the ensemble mean of the f -hour forecast ensemble verifying at time r and the truth at the same time as $\epsilon(r, n, f)$. The RMSE of the

f -hour forecast ensemble mean verifying at analysis time r is then expressed as

$$RMSE(r, f) = \left\{ \sum_{n=1}^N (\epsilon(r, n, f))^2 / N \right\}^{1/2}. \quad (6)$$

The temporally averaged RMSE of the forecast ensemble mean is then found by averaging eq. (6) over all c times during which statistics are calculated, to yield the average RMS error for a given forecast lead-time f ,

$$\langle RMSE(f) \rangle = \sum_{r=1}^c \left\{ \sum_{n=1}^N (\epsilon(r, n, f))^2 / N \right\}^{1/2} / c, \quad (7)$$

where $r=1$ is chosen to correspond to an analysis time occurring after a sufficiently long spin-up time. We compare the LETKF and caLETKF through the analysis accuracy [$f=0$ in eqs. (6) and (7)] and for ensemble forecasts of various lead-times [$0 < f \leq 72$ in eq. (7)]. Though the present work focuses on the usage of the RMSE as a metric of analysis and forecast performance, we recognise that other diagnostics (e.g. the spread/skill relationship) will be useful and necessary for further explorations of the properties of the caLETKF.

5. Experimental results

The effectiveness of our method was first examined through time series of the analysis RMSE. The RMS error of the analysis ensemble mean was recorded at each analysis time, the results of which are shown in Fig. 1, for both the caLETKF and standard LETKF. Here the dynamic ensemble size of both the standard LETKF and the caLETKF are both equal to 10, while the caLETKF supplements the dynamic ensemble with 10 additional static, climatological ensemble members at each analysis time. We note that the LETKF case with $k_d = 10$ dynamic ensemble members does not converge. Once our method (red curve) has converged, it produces analysis errors that are significantly smaller than those produced using the standard LETKF with $k_d = 10$ dynamic ensemble members (blue curve). The standard LETKF, albeit with a substantially larger dynamic ensemble size ($k_d = 20$, black curve), can achieve analysis accuracies similar to those of our method shown in Fig. 1. Comparing the results for the standard LETKF with $k_d = 10$ dynamic ensemble members with results for the caLETKF for $k_d = 10$ dynamic and $k_c = 10$ climatological ensemble members, one can see that the inclusion of the additional climatological ensemble members helps to stabilise the filter (i.e. fluctuations in the errors are substantially reduced). The behaviour shown in Fig. 1 suggests that the climatological members account for realistic errors in directions not captured by the dynamic ensemble members,

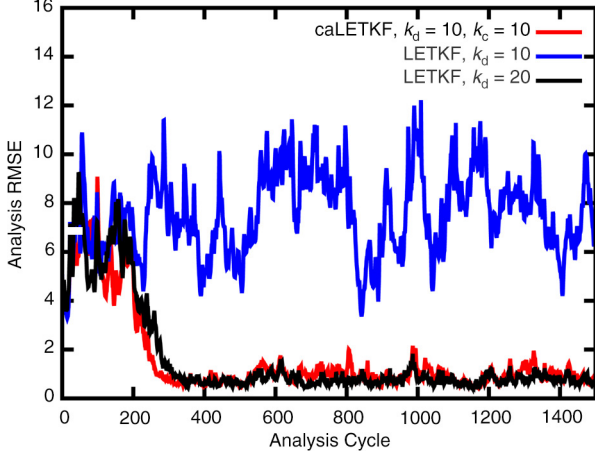


Fig. 1. An example in which a standard LETKF analysis with insufficient ensemble size (10 dynamic ensemble members, blue curve) is stabilised by augmenting the ensemble with 10 additional climatological ensemble members (red curve). The RMS error of the analysis ensemble mean [eq. (6)] is plotted at each analysis cycle over a 1500 analysis cycle period. Both experiments assimilate the same observations on the same observation network. For comparison, results from an experiment where the standard LETKF has $k_d = 20$ dynamic ensemble members (black curve) is also included.

and are providing value to the overall assimilation without the expense of being forecasted themselves.

To better quantify how much of an advantage is gained by climatological augmentation of the ensemble at analysis time, we also compare how RMS analysis error changes with increasing dynamic ensemble size, for both the augmented and standard analysis methods. At each dynamic ensemble size, the RMS analysis error of the ensemble mean is averaged over 50 000 analysis cycles, after an initial 1000 spin-up cycles. For each dynamic ensemble size listed on the x-axis of Fig. 2, the caLETKF was computed with $k_c = 10$ climatological ensemble members. As Fig. 2 shows, the advantage of the caLETKF is shown in its convergence, at smaller dynamic ensemble size, to the error level reached by the traditional LETKF at larger dynamic ensemble size. Specifically, our method needs approximately 13 dynamic ensemble members to achieve the same error that the LETKF achieves with approximately 28 dynamic ensemble members. Figure 2 further suggests that, if limited by computational resources for calculating ensemble member forecasts, it can be beneficial to consider using climatological ensemble members during the analysis, rather than striving to increase total ensemble size.

In the experiments detailed above, comparisons between the caLETKF and LETKF analyses are made when the caLETKF uses $k_c = 10$ climatological ensemble members. Figure 3 explores how many climatological ensemble members the caLETKF needs by changing the number of

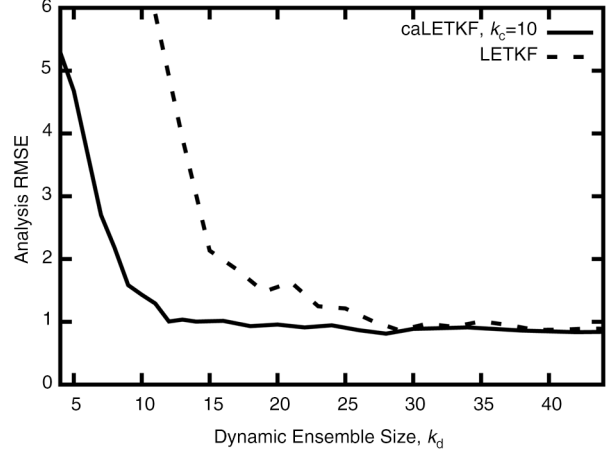


Fig. 2. A comparison of the RMS error of analysis ensemble means, eq. (7), between both the standard LETKF (dashed curve) and the caLETKF (solid curve). After an initial spin-up of 1000 analysis cycles, RMS error is averaged over 50 000 analysis cycles and is plotted as a function of dynamic ensemble size. For the experiments shown here, the climatologically augmented method uses $k_c = 10$ climatological ensemble members. Below $k_d = 4$ dynamic ensemble members, we find that the caLETKF is susceptible to filter divergence, while the standard LETKF is susceptible to filter divergence below $k_d = 10$ dynamic ensemble members.

climatological ensemble members while keeping the size of the dynamic ensemble constant at $k_d = 15$. The plot of Fig. 3 shows time-averaged RMS error of the analysis ensemble mean versus climatological ensemble size, and each data point shown here is averaged over 50 000 analysis cycles. Figure 3 shows that as few as about eight climatological ensemble members can be used without a significant

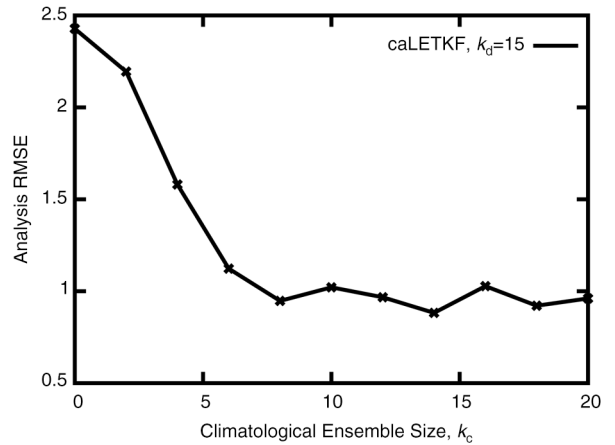


Fig. 3. Here we compare RMS error of the analysis ensemble mean for the caLETKF, eq. (7), as a function of climatological ensemble size k_c . Fifteen dynamic ensemble members ($k_d = 15$) are used, and each trial is averaged over 50 000 analysis cycles, discarding the first 1000 cycles as spin-up.

loss of analysis accuracy, implying that the caLETKF curve of Fig. 2 would look very similar for any $k_d \geq 8$. This suggests that much of the information missed by the dynamic ensemble is found in a relatively small number of state space directions.

The experiments just described measure analysis accuracy of the caLETKF while keeping forecasting costs (i.e. k_d) constant. We next present analysis accuracy results from an experiment which keeps analysis cost constant, by varying the size of the climatological ensemble while keeping constant the sum of the number of dynamic and climatological ensemble members, $k_d + k_c$. Specifically, the total ensemble size on which the analysis is performed is kept constant at $k_d + k_c = 30$ ensemble members. The value $k_d + k_c = 30$ is chosen because, as shown in Fig. 2, if a dynamic ensemble and no contributing static members ($k_c = 0$) are used, the LETKF analysis accuracy (dashed curve in Fig. 2) quickly degrades when the number of dynamic ensemble members falls below 30, $k_d < 30$. Figure 4 shows the time-averaged RMSE of the analysis ensemble mean, averaged over 50000 analysis cycles, as a function of climatological ensemble size. We see that the inclusion of climatological ensemble members allows the caLETKF to achieve the same accuracy with 12 dynamic ensemble members (and $k_c = 18$ climatological members) as that achieved with the standard LETKF with 30 dynamic ensemble members (and $k_c = 0$ climatological ensemble members). For reference, the horizontal line represents the analysis accuracy of the standard LETKF when $k_d = 30$ dynamic ensemble members are used, which is the smallest

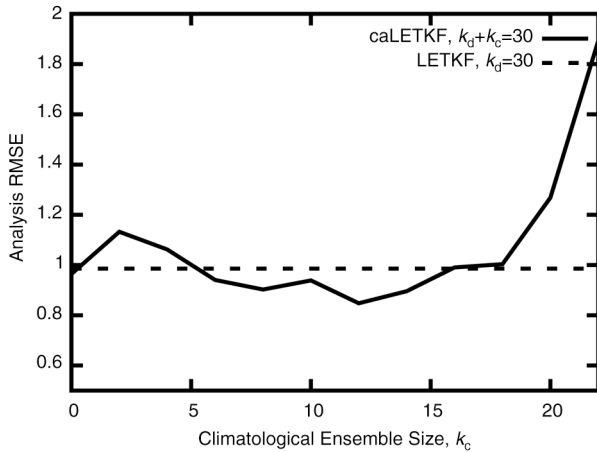


Fig. 4. Analysis accuracy of the caLETKF at constant analysis cost. Here, the sum of the dynamic and climatological ensemble sizes is kept constant at $k_d + k_c = 30$, and the climatological ensemble size is plotted versus analysis RMS error, eq. (7), averaged over 50000 analysis cycles, after discarding 1000 initial cycles. For small dynamic ensemble sizes, $k_d < 8$ and $k_c > 22$, the caLETKF was susceptible to filter divergence.

dynamic ensemble size at which the LETKF converges. Here, the caLETKF is advantageous over the LETKF, as it achieves comparable analysis errors with many fewer (as low as $k_d = 12$) dynamic ensemble members.

To measure ensemble forecast accuracy, we average the accuracy of the forecast ensemble mean over a series of 50000 ensemble forecasts. Forecast accuracy information is stored for a maximum lead-time of three model days. The forecast accuracy, as measured by RMSE [eq. (7)] is plotted in Fig. 5 as a function of forecast lead-time. For this experiment, the caLETKF uses $k_d = 20$ dynamic ensemble members and $k_c = 10$ climatological ensemble members. Forecast results from the caLETKF are compared against results from the LETKF with $k_d = 20$ and $k_d = 30$ dynamic ensemble members. Forecast results from the LETKF with $k_d = 30$ dynamic ensemble members outperform those of the LETKF with $k_d = 20$ dynamic ensemble members, and are on average comparable to the caLETKF with $k_d = 20$ and $k_c = 10$. Figure 5 shows that gains made from more accurate caLETKF analyses persist, and lead to more accurate forecasts. In fact, the caLETKF corresponds to gains in forecast lead-time of 24 to 17 hours: the 1-d forecasts initialised with the 20 dynamic member caLETKF analysis

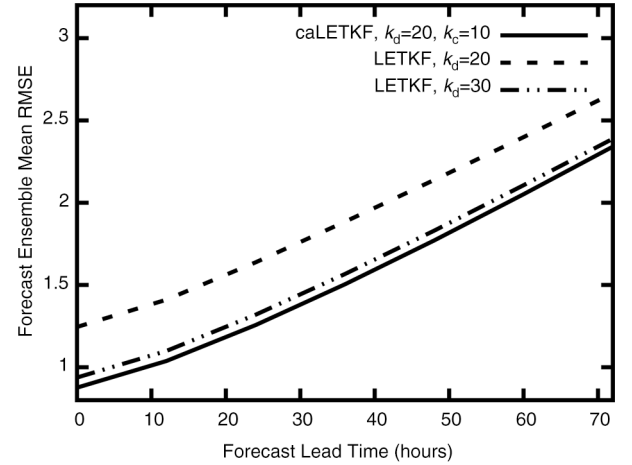


Fig. 5. Ensemble forecast accuracy as a function of lead-time f , for forecasts initialised from caLETKF and LETKF analysis ensembles. Ensembles were forecasted forward in time, and the mean of the forecast ensemble was compared against the truth. The resulting errors were averaged over a sample of 50000 forecasts at each lead-time, using eq. (7). Forecast results initialised from caLETKF analysis ensembles with $k_d = 20$ dynamic and $k_c = 10$ climatological ensemble members are shown as a solid black curve. For comparison, results of forecasts initialised from LETKF analysis ensembles with $k_d = 20$ and $k_d = 30$ dynamic ensemble members are shown as dashed and dot-dashed curves, respectively. The small difference (0.06 in forecast RMSE) between the LETKF $k_d = 30$ result and the caLETKF result is within the level of statistical fluctuations seen in our experimental system (for example, see the variation of the solid and dashed curves in Fig. 2 for $k_d \geq 30$).

ensemble are as accurate as the 20 dynamic member LETKF analysis ensemble. Though these gains diminish slightly with time, at constant dynamic ensemble size ($k_d = 20$), 48-hour caLETKF-initialised forecasts are as accurate as 30-hour LETKF-initialised forecasts, and 72-hour caLETKF-initialised forecasts are as accurate as 55-hour LETKF-initialised forecasts.

6. Summary and conclusions

The techniques of data assimilation can broadly be categorised as either variational or ensemble methods. While variational methods have been around for longer than ensemble methods, and are in widespread use in both operational and research settings, ensemble methods have advantages of their own, such as flow-dependent covariance estimates and natively computationally parallel forecasting. Recently, there has been a push to combine these methods to take advantage of their mutual advantages. The result has been the ‘hybrid’ methods, which combine the flow-dependent covariance estimates of ensemble methods with the climatological covariance estimates used by variational methods. Inspired by the success of these methods, we present here an ensemble method that aims to take advantage of climatological covariance information while staying in a purely ensemble framework.

The method we present here computes the analysis in a higher dimensional space than standard ensemble Kalman filters, by considering an augmented ensemble during the analysis. At the start of each analysis, the background dynamic ensemble mean is computed, along with the background dynamic ensemble perturbations. Additional ensemble members are created by adding to the background dynamic ensemble mean perturbations derived from a climatological background-error covariance matrix. We generate these climatological perturbations from the eigenvectors that correspond to the largest eigenvalues of the climatologically estimated background-error covariance matrix. These chosen eigenvectors correspond to the directions in model space that climatologically account for the most forecast error variability. Preliminary results (not shown) indicate that generating climatological perturbations from the eigenvectors of a static covariance estimate $\mathbf{B}_{\text{est}}$ is advantageous over the simpler approach of using perturbations that are samples taken at each analysis cycle from the archive of forecast errors used to generate $\mathbf{B}_{\text{est}}$ and scaled by a factor independent of k_c . For the experimental results reported here, we estimate the background-error covariance matrix through the NMC method of Parrish and Derber (1992). We anticipate that other methods of generating the climatological background-error covariance matrix might be called for in other settings.

We conducted a series of numerical experiments using a one-dimensional chaotic model of Lorenz (Lorenz, 2005). In these experiments, we compared the caLETKF against the standard LETKF analysis algorithm. Initial experiments compared time-series analysis errors generated by both methods at a constant dynamic ensemble size (Fig. 1). These experiments showed that the caLETKF with $k_c = 10$ added climatological ensemble members can have a clear advantage over the standard LETKF.

Our next series of experiments investigated how well the new method performs as a function of dynamic ensemble size when the climatological ensemble size is fixed at $k_c = 10$ (Fig. 2). This experiment showed that the caLETKF converged and achieved a consistent level of analysis accuracy at approximately 13 dynamic ensemble members, while the LETKF needed approximately 28 dynamic ensemble members to perform as accurately. In this experiment, we observed that both the caLETKF and LETKF produce similar results at large enough dynamic ensemble sizes. In more complex, highly flow-dependent systems, we imagine that a large number of climatological ensemble members could reduce the relative weight given to the dynamical portion of the covariance enough to potentially degrade performance. In these scenarios, one might introduce an additional tunable scaling factor that, in conjunction with α , would independently control the weights of dynamic and climatological perturbations in the estimation of background covariances.

A third experiment explored the effect on analysis accuracy of varying the climatological ensemble size (Fig. 3). This experiment found that fewer than the 10 climatological ensemble members used in the previous experiments could be used during the analysis with negligible loss of analysis accuracy. In addition to comparing our method to the standard LETKF at constant forecasting cost, as just described, the accuracy of the caLETKF was also tested against the standard LETKF at constant analysis cost. Here, the total size of the ensemble is kept constant (i.e. the sum of dynamic and climatological ensemble sizes), while the proportion of the ensemble members that are climatological ensemble members is varied (Fig. 4). Results from these experiments indicate that analysis accuracy can be maintained by replacing a significant number of dynamic ensemble members with climatological ensemble members. In applications where the same result applies, this could substantially reduce forecast costs.

Our last series of experiments investigated the accuracy of 1-, 2- and 3- d ensemble forecasts (Fig. 5). This experiment found that the caLETKF analysis gains discussed earlier were retained during the forecasts, as ensemble forecasts initialised from caLETKF analyses were more accurate than forecasts initialised from LETKF analyses. We find the results of these numerical experiments with the

caLETKF to be encouraging, and suggest that it be tested on larger and more realistic atmospheric models. Numerical experiments with larger, more complex models will necessitate other, more advanced diagnostics (e.g. the spread-error relationship) to measure and quantify the benefits of the caLETKF. The caLETKF could potentially be very useful in these settings, as fewer dynamic ensemble members can be used without loss of analysis accuracy.

7. Acknowledgements

This work was supported by ONR grant N000141210785. We thank Kayo Ide for helpful comments.

References

- Anderson, J. L. 2001. An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **129**, 2884–2903.
- Anderson, J. L. and Anderson, S. L. 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.* **127**, 2741–2758.
- Bishop, C. H., Etherton, B. J. and Majumdar, S. J. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Mon. Weather Rev.* **129**, 420–436.
- Burgers, G., Jan van Leeuwen, P. and Evensen, G. 1998. Analysis scheme in the ensemble Kalman filter. *Mon. Weather Rev.* **126**, 1719–1724.
- Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**, C5, 10143–10162.
- Golub, G. H. and Van Loan, C. F. 1996. *Matrix Computations*, Vol. 3. JHU Press, Baltimore, Maryland.
- Greybush, S. J., Kalnay, E., Miyoshi, T., Ide, K. and Hunt, B. R. 2011. Balance and ensemble Kalman filter localization techniques. *Mon. Weather Rev.* **139**, 511–522.
- Hamill, T. M. and Snyder, C. 2000. A hybrid ensemble Kalman filter-3D variational analysis scheme. *Mon. Weather Rev.* **128**, 2905–2919.
- Hamill, T. M. and Whitaker, J. S. 2005. Accounting for the error due to unresolved scales in ensemble data assimilation: a comparison of different approaches. *Mon. Weather Rev.* **133**, 3132–3147.
- Hamill, T. M., Whitaker, J. S. and Snyder, C. 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Weather Rev.* **129**, 2776–2790.
- Houtekamer, P. and Mitchell, H. 1998. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **126**, 796–811.
- Houtekamer, P. L. and Mitchell, H. L. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **129**, 123–137.
- Hunt, B. R., Kostelich, E. J. and Szunyogh, I. 2007. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica D*. **230**, 112–126.
- Lorenc, A. C. 2003. The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Q. J. Roy. Meteorol. Soc.* **129**, 3183–3203.
- Lorenz, E. N. 2005. Designing chaotic models. *J. Atmos. Sci.* **62**, 1574–1587.
- Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J. and co-authors. 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A*. **56**, 415–428.
- Parrish, D. F. and Derber, J. C. 1992. The National Meteorological Center’s spectral statistical-interpolation analysis system. *Mon. Weather Rev.* **120**, 1747–1763.
- Wang, X., Bishop, C. H. and Julier, S. J. 2004. Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble? *Mon. Weather Rev.* **132**, 1590–1605.
- Wang, X., Hamill, T. M., Whitaker, J. S. and Bishop, C. H. 2009. A comparison of the hybrid and EnSRF analysis schemes in the presence of model errors due to unresolved scales. *Mon. Weather Rev.* **137**, 3219–3232.
- Wang, X., Snyder, C. and Hamill, T. M. 2007. On the theoretical equivalence of differently proposed ensemble-3DVAR hybrid analysis schemes. *Mon. Weather Rev.* **135**, 222–227.
- Whitaker, J. S. and Hamill, T. M. 2002. Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.* **130**, 1913–1924.