

## On the Propagation of Information and the Use of Localization in Ensemble Kalman Filtering

YOUNG-NOH YOON AND EDWARD OTT

*Department of Physics, University of Maryland, College Park, College Park, Maryland*

ISTVAN SZUNYOGH

*Department of Atmospheric Sciences, Texas A&M University, College Station, Texas*

(Manuscript received 29 January 2010, in final form 10 June 2010)

### ABSTRACT

Several localized versions of the ensemble Kalman filter have been proposed. Although tests applying such schemes have proven them to be extremely promising, a full basic understanding of the rationale and limitations of localization is currently lacking. It is one of the goals of this paper to contribute toward addressing this issue. The second goal is to elucidate the role played by chaotic wave dynamics in the propagation of information and the resulting impact on forecasts. To accomplish these goals, the principal tool used here will be analysis and interpretation of numerical experiments on a toy atmospheric model introduced by Lorenz in 2005. Propagation of the wave packets of this model is shown. It is found that, when an ensemble Kalman filter scheme is employed, the spatial correlation function obtained at each forecast cycle by averaging over the background ensemble members is short ranged, and this is in strong contrast to the much longer range correlation function obtained by averaging over states from free evolution of the model. Propagation of the effects of observations made in one region on forecasts in other regions is studied. The error covariance matrices from the analyses with localization and without localization are compared. From this study, major characteristics of the localization process and information propagation are extracted and summarized.

### 1. Introduction

In an ensemble Kalman filter (EnKF), at any given analysis time  $t = nT$ , the estimated state of the atmosphere and the corresponding error covariance reflecting uncertainty in the state estimation are represented by a finite ensemble of global system states (e.g., Evensen 1994; Burgers et al. 1998; Houtekamer and Mitchell 1998). Each ensemble member is then integrated forward in time using a physical model of the atmosphere, thus creating an ensemble of forecasts at the next analysis time,  $t = (n + 1)T$ . By suitable incorporation (assimilation) of new measurements, the data assimilation process creates a new ensemble of system states that reflects the most probable atmospheric state and its uncertainty based on the available combined knowledge contained in the forecast ensemble at time  $(n + 1)T$  and

the new measurement data. The process then repeats at the  $t = (n + 2)T$  cycle, and so on.

Several localized versions of EnKF have been proposed (e.g., Houtekamer and Mitchell 2001; Hamill et al. 2001; Anderson 2001, 2007; Ott et al. 2004; Hunt et al. 2007). In this paper, we concentrate on one localization scheme called the local ensemble Kalman filter (LEKF; Ott et al. 2004) and a subsequent computationally more efficient version, the local ensemble transform Kalman filter (LETKF; Hunt et al. 2007). We expect that similar results would be obtained using any other of the available EnKF localization schemes. In LEKF and LETKF, local regions surrounding each gridpoint location are defined, the analysis is performed in each local region, and the results are combined to form a global analysis ensemble. The motivation for using localization is that it is expected to work well with drastically fewer ensemble members than would be the case if localization were not employed. In fact, without localization, the computational requirements for a meaningful EnKF analysis are so vast that they make the approach completely infeasible in typical weather forecasting settings. On the

---

*Corresponding author address:* Young-noh Yoon, Box 234, Dept. of Physics, University of Maryland, 82 Regents Dr., College Park, MD 20742.  
E-mail: mystyle@umd.edu

other hand, the use of localization has been shown to be successful and eminently feasible for both toy models and real atmospheric models (e.g., Houtekamer and Mitchell 2005; Houtekamer et al. 2005; Whitaker et al. 2008; Szunyogh et al. 2008).

One goal of this paper is to study the basic properties and rationale for localization. To do this, we will employ numerical experiments on a simple model system of equations introduced by Lorenz (Lorenz 2005). The model is simple and small enough that (unlike a real operational model) we can employ an EnKF without localization. We will compare the results obtained with the localized LETKF scheme with those obtained without localization [henceforth referred to as ensemble transform Kalman filter (ETKF); ETKF was introduced in Bishop et al. (2001)]. By doing this, we will be able to study some aspects relevant to the issues of localization, such as what determines a good size of the localization region, what dynamics justifies (or not) localization, and so on.

The second goal of this paper is to study the propagation of information in a spatiotemporally chaotic system with wave-like dynamics that are similar to that of weather. In particular, we use the traditional approach of discussing the propagation of information based on the study of phase and group velocities of the waves in the model (e.g., Charney 1949; Persson 2000; Szunyogh et al. 2002; Zimin et al. 2003). An alternative approach that has recently been gaining increasing attention is based on using tools of probability theory—most important, measures based on relative entropy of information (e.g., Kleeman 2007). The relation between localization and the advection of information from the observation location by the model dynamics has been studied using the hierarchical ensemble filter by Anderson (2007).

The organization of this paper is as follows: In section 2, we introduce the model system (Lorenz 2005) that we employ, and we investigate some of its salient properties. In section 3, we investigate correlations of the model. In section 4, we explain the forecast and localization procedures used in the ETKF and LETKF schemes. In section 5, we study correlations of the background ensemble of an ETKF as a function of spatial distance. In section 6, we compare covariance matrices obtained using an ETKF and an LETKF. In section 7, we show how observations taken in a small region at a certain time affect the forecast at other spatial points in the future. Section 8 provides further discussions and summarizes our main conclusions.

## 2. Model

Lorenz (2005) discusses three one-dimensional toy models that incorporate many features shown in real

atmospheric dynamics and in global numerical weather prediction models. The first model (Lorenz model 1) was originally introduced in Lorenz (1996) and Lorenz and Emanuel (1998). This model has become the standard model of choice for the initial testing of EnKF schemes. The popularity of the model is in part due to the similarity between the propagation of uncertainties (forecast errors) in Lorenz model 1 and global models in the midlatitude storm-track regions. In particular, the errors are propagated by dispersive waves whose behavior is similar to that of synoptic-scale Rossby waves, and the magnitude of the errors has a doubling time of about 1.5 days (where the dimensionless model time has been converted to dimensional time by assuming that the characteristic dissipation time scale in the real atmosphere is 5 days; see Lorenz 1996). Lorenz model 2 adds the feature of a smooth spatial variation of the model variables that resembles the smooth variation of the geopotential height (streamfunction) at the synoptic and large scales in the atmosphere. Lorenz model 3, the most refined and “realistic” of the three models in Lorenz (2005), adds a rapidly varying small-amplitude component to the smooth large-scale flow, mimicking the effects of small-scale atmospheric processes. In our following simulations, we use Lorenz model 3.

The equation of evolution of the scalar state variable  $Z_n$  at position  $n$  is the following:

$$\begin{aligned} dZ_n/dt = & [\mathbf{X}, \mathbf{X}]_{k,n} + b^2[\mathbf{Y}, \mathbf{Y}]_{1,n} + c[\mathbf{Y}, \mathbf{X}]_{1,n} \\ & - X_n - b Y_n + F, \end{aligned} \quad (1)$$

where  $n$  is a spatial variable ( $n = 1, 2, \dots, N$ ), periodic boundary conditions are employed ( $Z_{N+1} = Z_1$ ), and the  $\mathbf{X}$  and  $\mathbf{Y}$  vectors are defined as

$$\mathbf{X}_n = \sum_{i=-I}^I' (\alpha - \beta|i|) Z_{n+i} \quad \text{and} \quad (2)$$

$$\mathbf{Y}_n = Z_n - \mathbf{X}_n, \quad (3)$$

$$\text{with } \alpha = (3I^2 + 3)/(2I^3 + 4I) \quad \text{and}$$

$$\beta = (2I^2 + 1)/(I^4 + 2I^2).$$

The prime notation on  $\Sigma'$  signifies that the first and the last terms in the summation are divided by 2. The bracket of any two vectors  $\mathbf{X}$  and  $\mathbf{Y}$  is defined as

$$\begin{aligned} [\mathbf{X}, \mathbf{Y}]_{K,n} = & \sum_{j=-J}^J' \sum_{i=-J}^J' (-X_{n-2K-i} Y_{n-K-j} \\ & + X_{n-K+j-i} Y_{n+K+j})/K^2 \end{aligned} \quad (4)$$

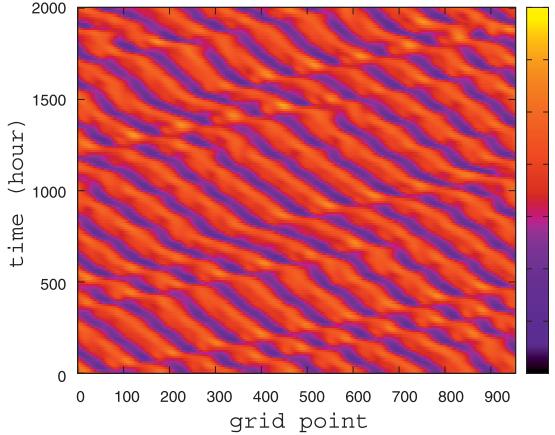


FIG. 1. Time evolution of state  $Z_n$  of Lorenz model 3. In this figure, 6 h are defined to be 0.05 time units in the model equations considering the damping time of the model.

when  $K$  is even, and  $\Sigma'$  is replaced by  $\Sigma$  when  $K$  is odd;  $J = K/2$  when  $K$  is even, and  $J = (K - 1)/2$  when  $K$  is odd. The parameter values used throughout this paper are  $N = 960$ ,  $K = 32$ ,  $b = 10$ ,  $c = 2.5$ ,  $F = 15$ , and  $I = 12$ . We solve Eqs. (1)–(4) using a Runge–Kutta scheme with a time step sufficiently small to resolve the fast time scale of the model. We find that these equations lead to a state profile that shows small-scale activity  $\mathbf{Y}_n$  superposed on a large-scale smooth component  $\mathbf{X}_n$ , thus crudely mimicking the multiscale dynamics of a real atmospheric system.

Figure 1 obtained from a numerical solution of Eq. (1) shows how states evolve in time. The horizontal axis is the spatial location, and the vertical axis is the time. We define 6 h to be 0.05 time units in the time evolution equation following Lorenz [Lorenz justifies this correspondence between  $t$  in Eq. (1) and pseudohours by considering the damping time of the model]. The colors in Fig. 1 represent the values of the state variables  $Z_n$ . Red corresponds to high values, and blue corresponds to low values. The figure shows that waves of approximately 7 spatial periods propagate to the left. This observed structure mimics what is seen at large scale in the atmosphere where Rossby waves play a prominent role.

The structure found in Fig. 1 also suggests that an insight can be gained by representing the field as a modulated sinusoidal wave. For this purpose, modulating envelopes were extracted from the states at each time using the method introduced in the paper by Zimin et al. (2003). The method of Zimin et al. for extracting modulating envelopes is explained briefly in appendix A. Based on the observed structure in Fig. 1, we computed envelopes using wavenumbers of only 7 and 8 to show smooth flow of the envelope (see appendix A). Figure 2 depicts the envelope amplitude in time and space for the same

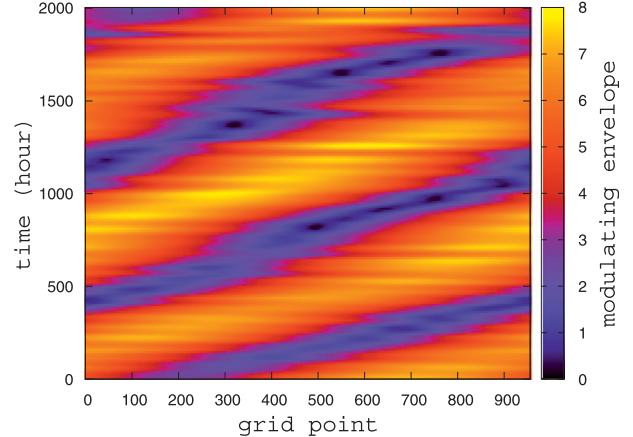


FIG. 2. Time evolution of the wave packet envelopes extracted by the envelope extraction method introduced in Zimin et al. (2003). Wave numbers of only 7 and 8 were used to smooth out the envelopes.

numerical run as in Fig. 1. From Fig. 2, we see that wave packets propagate to the right, in contrast to the leftward propagation of the individual troughs and ridges seen in Fig. 1. That is, in some appropriate sense, the wave turbulent state consists of waves that have phase velocities that move to the left (as in Fig. 1) and group velocities that move to the right (as in Fig. 2). This situation is similar to that of Rossby waves in the atmosphere, whose phase velocity is always westward relative to the mean zonal flow and whose group velocity can be either eastward or westward.

### 3. Correlation structure of the model

We calculated the correlations  $C_m(\tau)$  between values at two different locations separated by  $m$  grid points in space and by  $\tau$  in time as follows:

$$C_m(\tau) = \frac{\langle [Z_n(t) - \langle Z_n(t) \rangle][Z_{n+m}(t + \tau) - \langle Z_n(t) \rangle] \rangle}{\langle [Z_n(t) - \langle Z_n(t) \rangle]^2 \rangle}, \quad (5)$$

where angle brackets denote an average over  $n$  and  $t$ . The resulting plots,  $C_m(0)$  (Fig. 3) and  $C_m(\tau)$  (Fig. 4), show that wave dynamics plays a central role in the spatiotemporal evolution of the correlation, as demonstrated through three properties: 1) the spatial correlation has the structure of a wave packet with carrier wavenumber 7 and a packet envelope that decreases monotonically with increasing  $|m|$  (Fig. 3), 2) the wave structure shifts to the left with a phase velocity of about  $-0.77$  grid points per hour (Fig. 4), and 3) the packet envelope shifts to the right at a rate that we identify as a group velocity (e.g., see the bright yellow spots along

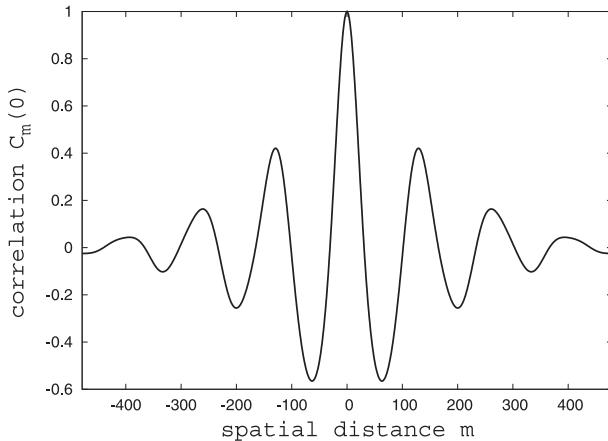


FIG. 3. Correlations  $C_m(0)$ ; i.e., the correlation values of Fig. 4 with time difference  $\tau$  set to 0.

the red troughs in Fig. 4). Under the assumption of ergodicity, the correlation in Eq. (5) can be regarded as describing an ensemble drawn from the distribution that defines the climate of the model. In section 5, we will consider another type of correlation function that can be regarded as describing an ensemble drawn from a distribution of short-term forecast uncertainties associated with a fairly dense observational network.

To make property 3 more transparent, we extracted the envelopes at each temporal distance  $\tau$  from Fig. 4 using the envelope extraction method in Zimin et al. (2003) with wavenumbers of 6, 7, and 8. We obtained the results depicted in Fig. 5. It is seen from Fig. 5 that, similar to Fig. 2, the envelope in  $m-\tau$  space propagates to the right at a group velocity of about 1.37 grid points per hour.

#### 4. Forecasting and localization

In our numerical assimilation experiments, we will employ “perfect model” simulations of forecast/analysis cycles. That is, we carry out a time evolution run of our model, Eqs. (1)–(4), which we regard as simulating a “true” atmospheric state evolution that we wish to analyze and forecast. We then take simulated measurements of this true state  $Z_n$  by adding uncorrelated noise of Gaussian distribution to the true state variable  $Z_n$  at each grid point and at each analysis time. The mean of the noise is 0 and the standard deviation is 0.3, which is about the size of the small-scale activity  $\mathbf{Y}$  in the model. The standard deviation of the noise can be also compared with the climatological standard deviation of the state variable  $Z_n$ , which is 4.67 (a noise standard deviation of 0.3 is used throughout this paper). Using these simulated measurements, we do analyses employing the same model,

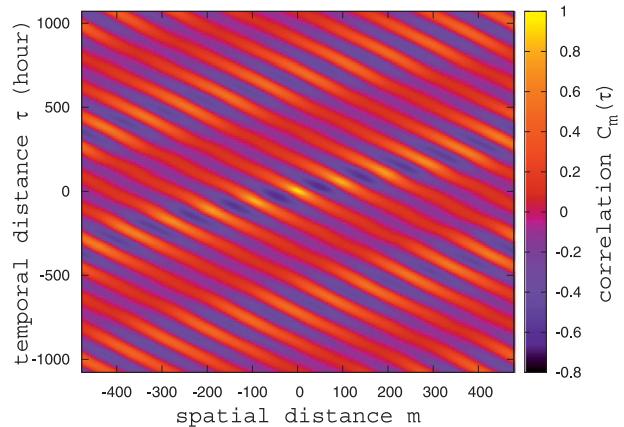


FIG. 4. Correlations  $C_m(\tau)$  between state variables at two different points,  $(n, t)$  and  $(n + m, t + \tau)$  in Fig. 1. The values were averaged over all of the space-time locations  $(n, t)$ .

Eqs. (1)–(4), as we use to generate the true state evolutions [i.e., Eqs. (1)–(4)].

In our ETKF implementation, we use the same technique as described in Hunt et al. (2007), but without employing localization. To faithfully represent the full system state and its covariance in the absence of localization, the number of ensemble members  $k$  that we use in the ETKF must be at least on the order of the number of growing local Lyapunov exponents, which increases as the number of grid points of the model increases in general. Thus, we need many ensemble members when there are many grid points in the model. We integrate each ensemble member during the evolution phase of the forecast cycle, and the required computing power to do this scales like  $k$ . In carrying out our ETKF procedure, it is also necessary to invert a  $k \times k$  matrix, which requires a number of computer operations on the order

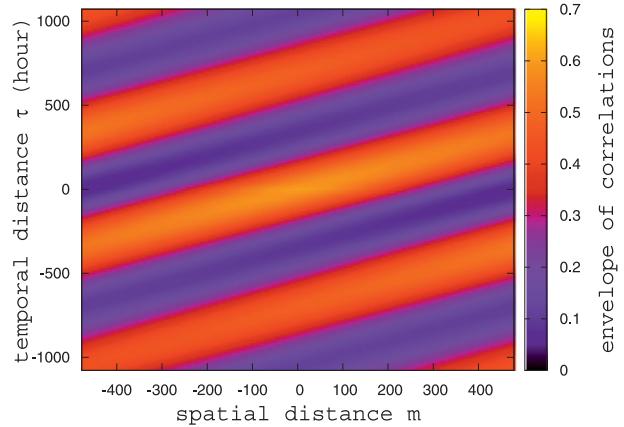


FIG. 5. Envelopes extracted from correlations in Fig. 4 at each time  $\tau$  using wave numbers only from 6 to 8 to smooth out the result.

of  $k^3$ . Hence, if we can do computations in local regions with many fewer grid points than in the global region, and thus with correspondingly many fewer ensemble members, then the required computational power is greatly reduced. For a more detailed discussion of the computational cost of LETKF, see Hunt et al. (2007) and Szunyogh et al. (2008).

If the errors in the state estimates at two different grid points that are far from each other are independent, then we might be able to compute analysis ensembles for each location while ignoring the other. This qualitative type of consideration is what motivates localization of the analysis. A brief description of the localization process used in LETKF and implemented on our one-dimensional model, Eq. (1), is the following: First, choose an appropriate spatial size  $L$  for the localization. At each grid point  $i$ , a local patch is assigned as consisting of the  $2L + 1$  grid points  $(i - L, i - L + 1, \dots, i + L - 1, i + L)$  for  $i = 1, 2, \dots, N$ . We then compute the analysis ensemble in each local patch using observations in the local patch. Then, we construct the global analysis ensemble that combines all of the local analysis ensembles by taking weighted averages of analysis values from each local patch. Ott et al. (2004), Hunt et al. (2007), and appendix B explain how to do the steps in this procedure in detail. How accurate is this localization procedure? The supposed main requirement is that observations outside the local patch should not affect the information from the local patch that is used in the evolution of the ensemble to the next analysis cycle. We discuss this issue for Lorenz model 3 in what follows. Note that, because we investigate a univariate model, the issue of balance does not arise. The issue of balance and its interaction with localization has been discussed by Kepert (2009).

## 5. Spatial correlations of the background ensemble

As we have pointed out, Fig. 3 shows a long-range correlation essentially extending over the entire length of the simulation region. Thus, one might think that we will lose a lot of information that comes through correlations from observations outside the local patch when we localize the analysis scheme. As we shall subsequently argue, this is not the case. In particular, we emphasize that it is the covariance matrix of the background ensemble around the ensemble mean that is used in the analysis, not the covariance matrix of the climatological distribution of true states themselves.

To formulate a correlation relevant to what is used in our analyses, we first carry out a perfect-model simulation of a forecasting situation. In particular, we begin by preparing a background ensemble  $Z_n^{(k)}(t)$ ,  $k = 1, 2, \dots, 960$ , by running 1000 forecast cycles (we use  $k$  and  $t$  to

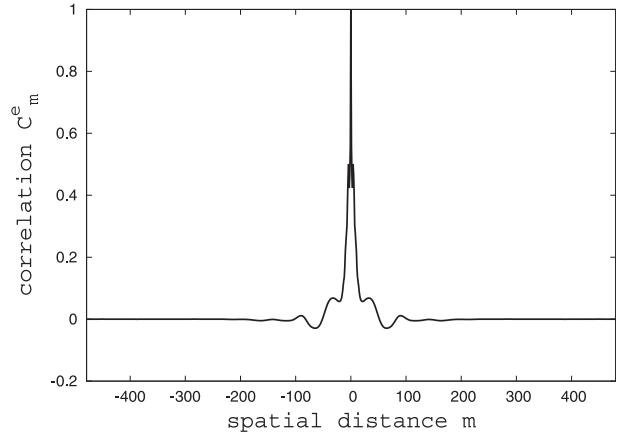


FIG. 6. Correlation  $C_m^e$  of the background ensemble of ETKF with 6-h cycle time vs spatial distance  $m$ . The values were averaged over 1000 forecast cycles.

denote an ensemble member and cycle time, respectively). During each cycle, we saved the background ensemble  $Z_n^{(k)}(t)$  and took simulated observations with noise of uncorrelated Gaussian distribution at each grid point at the analysis time, and then we updated the ensemble with ETKF using a multiplicative covariance inflation factor of 1.13 [see step 5 of appendix B; the factor 1.13 gives the minimum rms error of the resulting state estimate—see Anderson and Anderson (1999) for a general explanation of covariance inflation], and we then evolved the true state and the ensemble from  $t$  to  $t + 6$  h (except where otherwise stated, a 6-h cycle time and 1.13 multiplicative covariance inflation factor are used throughout this paper). Using these parameters, we calculated the following ensemble-based spatial correlation in the background:

$$C_m^e = \left\langle \frac{A}{\sqrt{B} \sqrt{C}} \right\rangle_{n,t}, \quad (6)$$

with  $A = \langle [Z_n^{(k)}(t) - \bar{Z}_n(t)][Z_{n+m}^{(k)}(t) - \bar{Z}_{n+m}(t)] \rangle_k$ ,

$B = \langle [Z_n^{(k)}(t) - \bar{Z}_n(t)]^2 \rangle_k$ , and

$C = \langle [Z_{n+m}^{(k)}(t) - \bar{Z}_{n+m}(t)]^2 \rangle_k$ ,

where  $\langle \cdot \rangle_k$  denotes an average over ensemble members  $k$ ,  $\langle \cdot \rangle_{n,t}$  denotes an average over space  $n$  and time  $t$ , and  $\bar{Z}_n = \langle Z_n^{(k)} \rangle_k$ . Figure 6 shows the result. As can be seen, the magnitude of the correlation is less than 3% of its peak value for  $|m| > 46$ , and the half-width of the strong peak around zero distance is roughly on the order of 10 for 6-h cycle time simulation. Thus, the correlations of deviations of ensemble members from the ensemble mean will be negligible when 2 grid points are separated by more than 46 grid points.

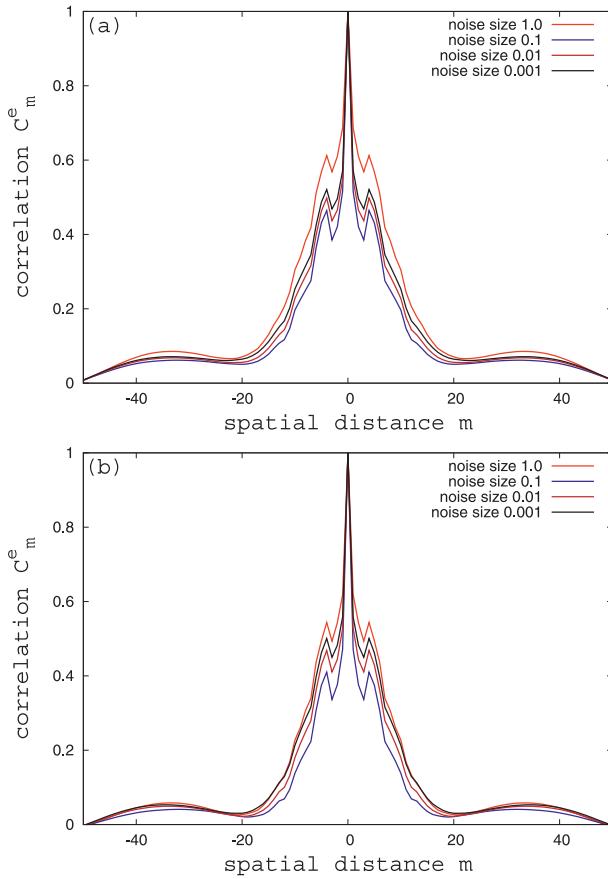


FIG. 7. Correlations  $C_m^e$  (a) of the background ensembles and (b) of the analysis ensembles of ETKF with several different noise sizes. The values were averaged over 1000 forecast cycles. The red, blue, brown, and black lines are for standard deviations of the noise: 1.0, 0.1, 0.01, and 0.001, respectively.

To analyze further the result shown in Fig. 6, we note that the background correlation is determined by the structure (correlation) of the analysis error and the model dynamics that spreads the effects of the analysis error in space and time. The shape of the analysis correlation function is very similar to that of the background correlation function and is only weakly dependent on the magnitude of the observation noise (Figs. 7a,b). A reasonable assumption is that the spatial expansion of the correlation pattern of the analysis ensemble during the model integration is controlled by the group velocity of the waves that compose the envelope of the analysis error pattern. Although the group velocity associated with these waves is not necessarily identical to the group velocity of the waves observed in the free run of the model, the group velocity observed for the free run can be considered to be an estimate of the maximum group velocity for the given dynamical system. Thus, the comparatively short range of the correlations is natural if

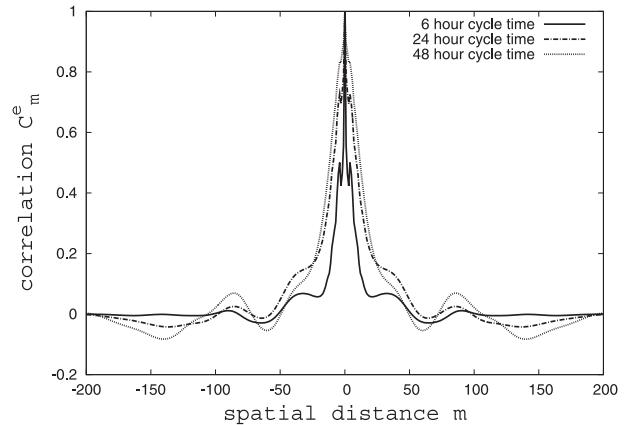


FIG. 8. Correlation  $C_m^e$  of the background ensembles of ETKF with various cycle times vs spatial distance  $m$ . The values were averaged over 1000 forecast cycles. The solid, dash-dotted, and dotted lines are for 6-, 24-, and 48-h cycle times, respectively.

one assumes that the expansion of the observed half-width of the correlation pattern of the analysis ensemble during the model integration cannot be larger than the maximum group velocity multiplied by the forecast time used in calculating the values  $Z_n^{(k)}(t)$  in Eq. (6) and identifies the maximum group velocity (maximum information propagation speed) to be on the order of the speeds found in Figs. 1–5 (i.e.,  $\sim 1$  grid point per hour).

In Fig. 8, we show results for the same quantity as plotted in Fig. 6, but for the cases in which the assimilation cycle times are 6 (as in Fig. 6), 24, and 48 h. The observations were taken only at analysis times with corresponding intervals of 6, 24, and 48 h. Consistent with the above interpretation, we observe that the correlation function spreads as the cycle time is increased.

## 6. ETKF and LETKF covariance matrices

Now we compare the covariance matrices obtained from ETKF analyses with those obtained from LETKF analyses. We ran forecast cycles keeping two sets of ensembles, one for ETKF and one for LETKF. The number of ensemble members used in our ETKF implementation was 960, which is the total number of grid points of the model. Our LETKF implementation used a local patch size of  $2 \times 50 + 1$  and number of ensemble members equal to 101, the same as the number of grid points in the local patch. The multiplicative covariance inflation factor was 1.13 for both ETKF and LETKF, which gives the minimum rms state estimate errors for both. We computed the covariance matrix of the ETKF analysis ensemble and the covariance matrix of the LETKF analysis ensemble at several analysis times. Because they have nonnegligible values only near the diagonal, we chose the elements  $P_{ij}$  of the covariance matrices with  $|j - i| \leq 7$

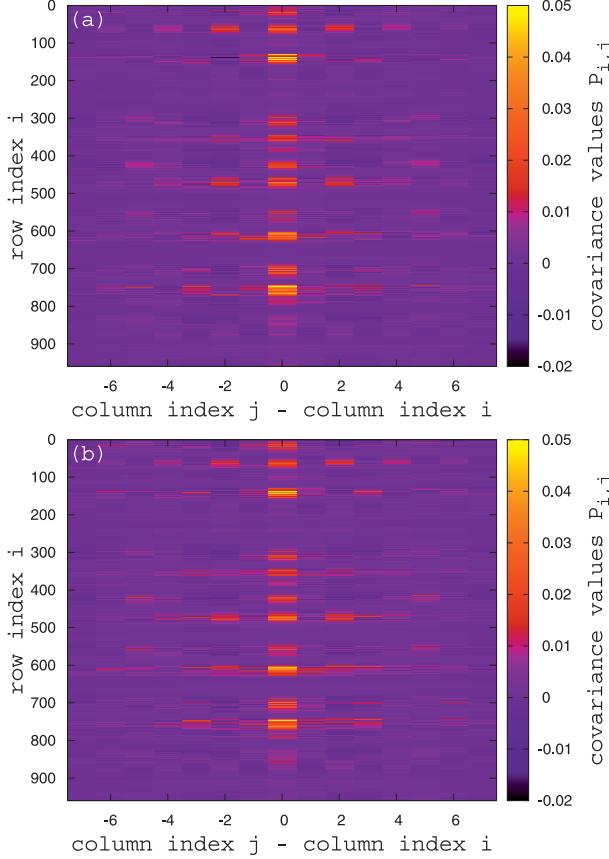


FIG. 9. Covariance matrix values of analysis ensemble at a certain forecast time of (a) ETKF and (b) LETKF. The  $y$  axis is the row index  $i$ , and the  $x$  axis is the difference between the column indices  $j$  and  $i$ .

and stacked them row by row so that elements with the same separation of indices  $j - i$  are aligned vertically. Figures 9a,b show the ETKF and LETKF covariance matrices at a representative analysis time. The vertical axis is index  $i$ , the horizontal axis is  $j - i$ , and the color represents the covariance values between grid points  $i$  and  $j$ :  $P_{i,j}$ . Red corresponds to high values, and blue corresponds to low values. These figures show similar patterns for ETKF and LETKF. Figure 10 is a plot of  $\bar{P}_i$  versus  $i$ , where  $\bar{P}_i$  is the average of  $P_{i+l,i+l}$  over the 11 values of  $l$  for  $-5 \leq l \leq 5$ . The red color is for ETKF, and the blue color is for LETKF. This figure also shows remarkably similar patterns for ETKF and LETKF. The mean rms errors of the state estimates [defined as  $\sqrt{\langle (\bar{Z}_n^a - Z_n)^2 \rangle_n}_{\text{cycle}}$ , where  $\bar{Z}_n^a$  is the mean of the analysis ensemble and  $\langle \dots \rangle_n$  and  $\langle \dots \rangle_{\text{cycle}}$  are averages over grid points and over forecast cycles, respectively] obtained from ETKF and LETKF are 0.0979 and 0.0997.

We also studied the patch size and cycle time dependence of two global measures of the goodness of

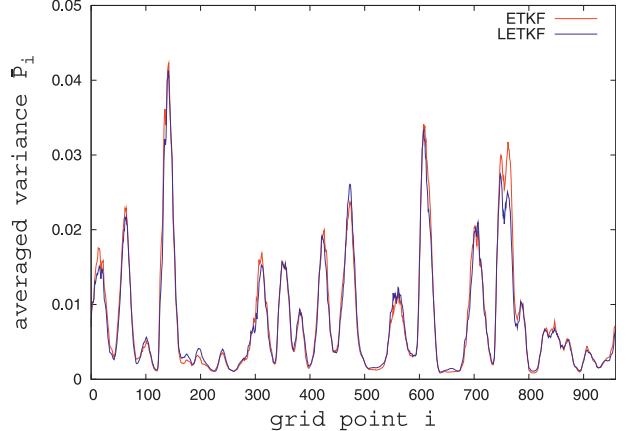


FIG. 10. Variance values of analysis ensemble at a certain forecast time. The values were averaged over 11 neighboring grid points. Red plot is for ETKF, and the blue plot is for LETKF.

LETKF assimilations for 6-, 12-, 24-, and 48-h cycle times. One measure is the time-averaged rms error of the state estimate by averaged ensemble as defined above. The other measure is

$$\delta = \left\langle \frac{\langle |\bar{P}_i^{\text{LETKF}} - \bar{P}_i^{\text{ETKF}}| \rangle_i}{\langle \bar{P}_i^{\text{ETKF}} \rangle_i} \right\rangle_{\text{cycle}}, \quad (7)$$

where  $\langle \dots \rangle_i$  denotes an average over all of the grid points. We used the same number of ensemble members as the number of grid points in a local patch for LETKF. Figures 11 and 12 show that the rms error and the quantity  $\delta$  become smaller as we increase the local patch size. Using the rms error as our figure of merit, we see

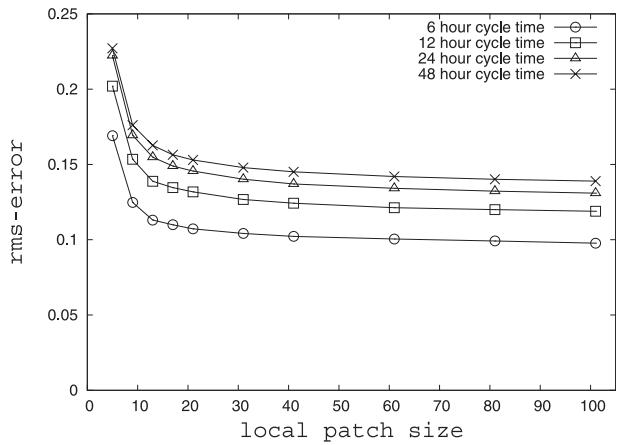


FIG. 11. Averaged rms errors of the state estimates by the means of the analysis ensembles vs local patch size. The values were averaged over 2000 forecast cycles. The lines with the circles, squares, triangles, and times signs correspond to 6-, 12-, 24-, and 48-h cycle times.

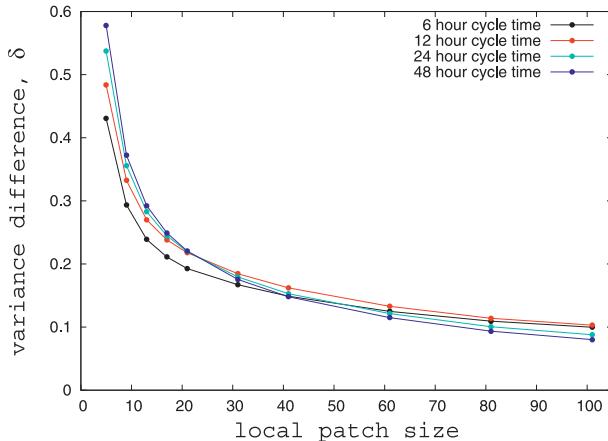


FIG. 12. Values of  $\delta$  (normalized difference between variances of analysis ensembles of ETKF and LETKF) vs local patch size. The values were averaged over 2000 forecast cycles. The black, red, cyan, and blue colors correspond to 6-, 12-, 24-, and 48-h cycle times.

that there is relatively little benefit to increasing the patch size past about 15, although there is more substantial improvement in  $\delta$  as the patch size is increased past 15. Thus, we see that the improvement in the estimate of the mean of the ensemble saturates more quickly with increase in patch size than the improvement in the estimate of the variance of the ensemble. From Fig. 12, we can also see that longer forecast cycle time makes  $\delta$  larger when the local patch size is smaller than about 20. However, we cannot see a clear pattern when the local patch size is large. Figure 13 shows the difference between the rms errors of LETKF and ETKF divided by the rms error of ETKF, a normalized rms error difference.

## 7. Demonstration of the localized influence of observations

If the local patch size is large enough to encompass most of the correlation seen in Fig. 6, then it is to be expected that the analysis at grid points around the middle of the patch does not lose too much information by ignoring observations outside the local patch. In what follows, we report evidence supporting this intuition.

We now examine how observations taken at time  $t$  and in a small region around spatial location  $i$  affect a forecast at space-time point  $(t + \tau, i + m)$ . To do this, we first ran 100 forecast cycles using the ETKF and noisy observations at all the grid points (as done in section 5), thus bringing the ensemble members close to the true state. Then, at current time  $t$ , we took observations at only five fixed contiguous locations,  $i - 2, i - 1, i, i + 1$ , and  $i + 2$  (taking observations at only one location did not produce a clear result, possibly because the impact

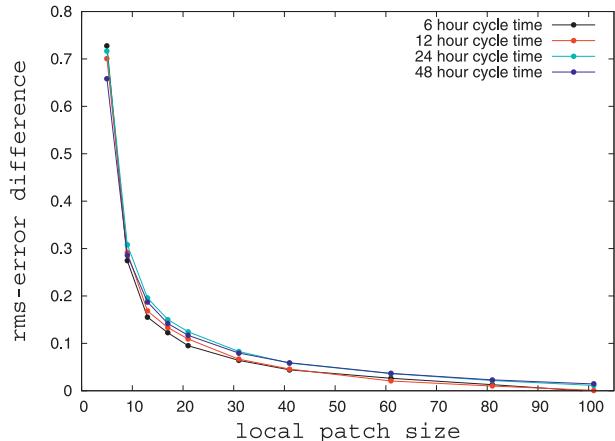


FIG. 13. Normalized difference between rms errors of state estimates with ETKF and LETKF vs local patch size. The values were averaged over 2000 forecast cycles. The black, red, cyan, and blue colors correspond to 6-, 12-, 24-, and 48-h cycle times.

of one observation was not strong enough). We then used these observations to calculate an analysis ensemble. Then, we calculated the mean of the background ensemble and the mean of the analysis ensemble and evolved the two means from the current time  $t$  to the time  $t + \tau$ . We denote the values of these two evolved means at each grid point  $n$  by  $\bar{Z}_n^b(\tau)$  and  $\bar{Z}_n^a(\tau)$  ( $b$  and  $a$  denote background and analysis, respectively). We did this for values of  $\tau$  ranging from  $\tau = 0$  h to  $\tau = 150$  h in 6-h intervals, calculating and saving the quantities

$$D_m(\tau) = [\bar{Z}_{i+m}^a(\tau) - \bar{Z}_{i+m}^b(\tau)]^2. \quad (8)$$

We then proceed to the next time cycle  $t \rightarrow t + 6$  h and repeat this process: i.e., at time  $t$  we now take observations at all the grid points, calculate the analysis ensemble, evolve it and the true state forward to create a new background ensemble at the new time  $t \rightarrow t + 6$  h, and repeat the previously described five-observation-point calculation of  $D_m(\tau)$ . Figure 14a shows a plot of  $\langle D_m(\tau) \rangle_{\text{cycle}}$ , the average of  $D_m(\tau)$  over 4000 cycles. The horizontal axis is  $m$ , and the vertical axis is  $\tau$ . The color represents  $\langle D_m(\tau) \rangle_{\text{cycle}}$ . Red corresponds to high values, and blue corresponds to low values. The white color represents values higher than the upper limit of the color bar. The figure shows that observations do not affect points that are far from the observation points within up to about 30 h. However, all of the points become affected by the observations after about 30 h.

To investigate why all of the points eventually become affected after some time, we plotted the time evolution of the differences between the two means,

$$d_m(\tau_i) = \bar{Z}_{i+m}^a(\tau_i) - \bar{Z}_{i+m}^b(\tau_i), \quad (9)$$

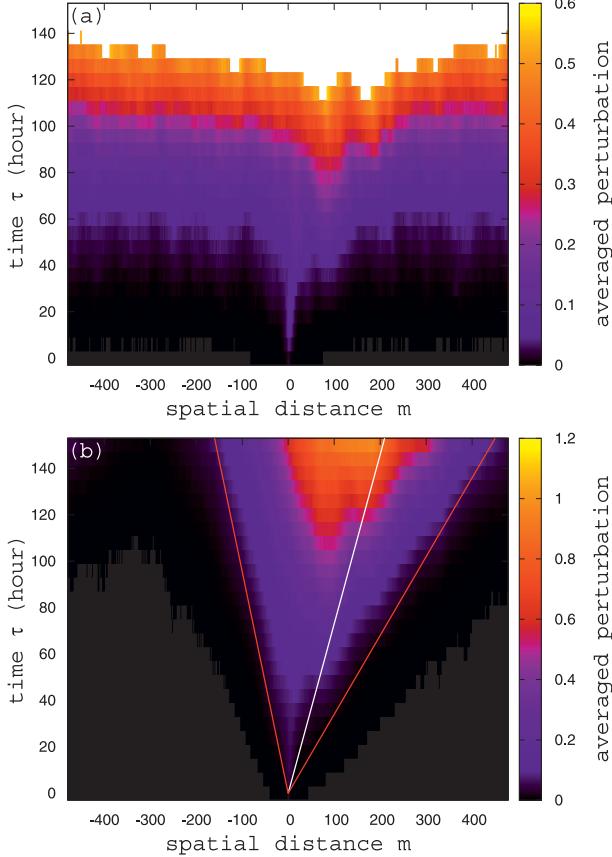


FIG. 14. Values of  $\langle D_m(\tau) \rangle_{\text{cycle}}$  (evolution of the perturbation to the ensemble mean caused by five localized observations). The values were averaged over 4000 cycles, (a) with ETKF and (b) with LETKF. The white line shows the group velocity of Lorenz model 3, and the two red lines show the boundaries of the wedgelike region of the perturbations.

at  $\tau_i = i \times 6 \text{ h}$  ( $i = 0, 1, \dots, 7$ ) at a certain cycle time (see Fig. 15a). Figure 15a shows  $d_m(\tau_i)$  plotted versus  $m$  at 6-h intervals from  $\tau = 0 \text{ h}$  to  $\tau = 42 \text{ h}$ . It is seen that small differences at points far from the observation points at  $\tau = 0 \text{ h}$  become amplified and dominate the  $d_m$  values at later times. At small  $\tau$ , these differences are close to zero but are not exactly zero. We computed the rms values of  $d_m(0)$  averaged over grid points  $m$  and cycles varying the number of ensemble members, where the average over  $m$  is taken only for  $m < -300$  and  $m > 300$ . Figure 16 shows that the impact of the localized observation on points far from the observation point decreases as the ensemble size increases. So, these small impacts are likely to be caused by the finite size of the ensemble [see Hamill et al. (2001) and Anderson (2007) for detailed discussions of noisy covariance values]. As  $\tau$  increases, the small perturbations far from the observation points are chaotically amplified and become large. We repeated this experiment using LETKF in place of ETKF. The

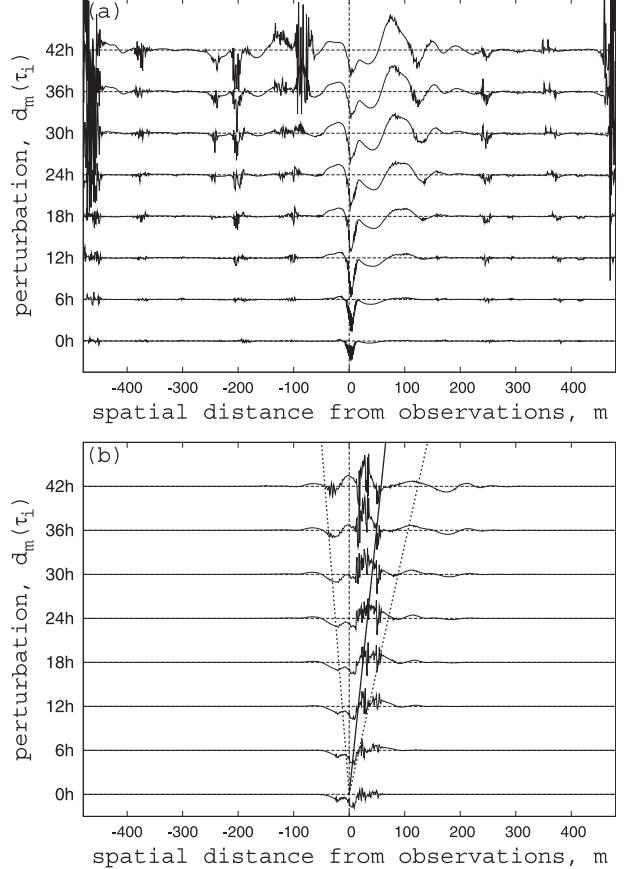


FIG. 15. Evolution of the perturbation to the ensemble mean caused by five localized observations,  $d_m(\tau_i)$ , in 6-h time intervals from  $\tau = 0 \text{ h}$  to  $\tau = 42 \text{ h}$ , at a certain cycle time, (a) with ETKF and (b) with LETKF. The solid line shows the group velocity of Lorenz model 3, and the two dotted lines show the same boundaries in Fig. 14b.

use of localization in LETKF has the effect of eliminating the randomlike small initial differences  $d_m(0)$  at large  $|m|$ . Figures 14b and 15b show the results. It is seen that the effect of the localized observations spreads to other points linearly, creating a wedgelike region in our  $\tau$ -versus- $m$  diagram indicated by two red lines in Fig. 14b and two dotted lines in Fig. 15b. The propagation speed corresponding to the right boundary of the wedge is roughly 3 grid points per hour, and the propagation speed corresponding to the left boundary of the wedge is roughly 1 grid point per hour. Thus, it appears that, during a 6-h cycle time, the effect of an observation will, on average, reach up to about 18 grid points to the right and up to about 6 grid points to the left. The maximum difference between the background mean and the analysis mean propagates at a speed that is close to the group velocity of wave packets (1.37 grid points per hour, shown by the white line in Fig. 14b and the solid line in Fig. 15b) determined from Fig. 5.

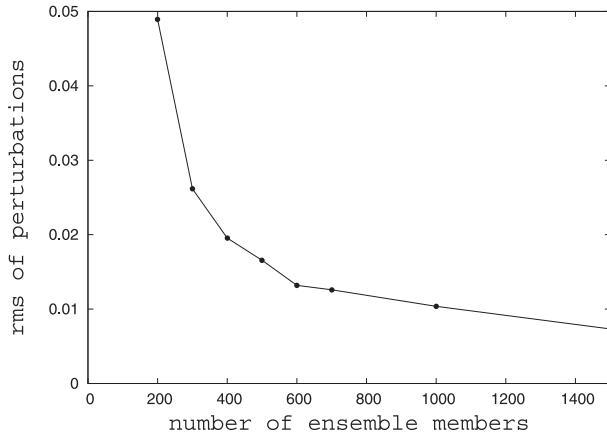


FIG. 16. The rms values of  $d_m(0)$  divided by the size of the noise of observations vs ensemble size of ETKF. The values were averaged over 2000 cycles and over grid points that are more than 300 grid points away from the middle of the observation points.

To see the effect of observations at a space–time point  $(t, n)$  on a forecast at  $(t + \tau, n + m)$  when observations are taken at all of the grid points, we did an experiment similar to the one described above. This time, we took observations at all of the grid points (instead of at 5 adjoining grid points) and calculated the value  $\tilde{C}_m(\tau)$  instead of  $D_m(\tau)$  defined in the above experiment:

$$\tilde{C}_m(\tau) = \left\langle \frac{[\bar{Z}_{n+m}^a(\tau) - \bar{Z}_{n+m}^b(\tau)][\bar{Z}_n^a(0) - \bar{Z}_n^b(0)]}{\sigma_{\text{noise}}^2} \right\rangle_{n, \text{cycle}}, \quad (10)$$

where  $\sigma_{\text{noise}}$  denotes the standard deviation of the observation noise. Figure 17 shows  $\tilde{C}_m(\tau)$ . The horizontal axis is  $m$ , and the vertical axis is  $\tau$ . The color represents  $\tilde{C}_m(\tau)$ . Red corresponds to high values, and blue corresponds to low values. The white color represents values that are higher than the upper limit of the color bar. This plot also shows that the effect of observations propagates globally to the right at a speed that is close to the group velocity of wave packets (shown by a long white line). The orientation of major axes of the blue blobs slopes to the left, and this slope is on the order of the phase velocity ( $-0.77$  grid points per hour, shown by short white lines) determined from Fig. 4.

## 8. Conclusions

The correlation function obtained from the time evolution of the state of Lorenz model 3 and the envelopes extracted from it both show that there is predominant wave packet propagating to the right (sections 2 and 3). We found that the correlation length of the deviations from the ensemble mean of the background ensemble

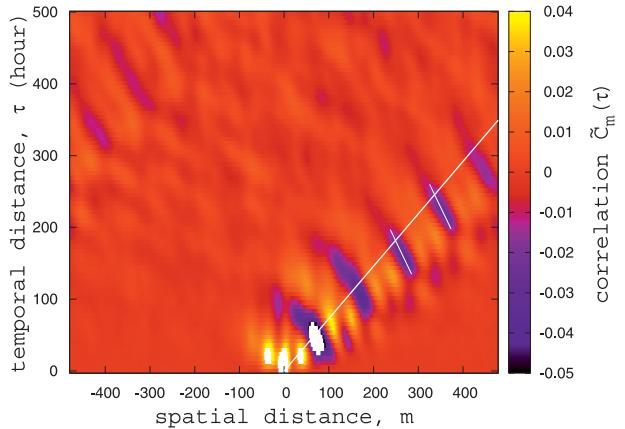


FIG. 17. Values of  $\tilde{C}_m(\tau)$  [correlations between perturbations at  $(0, n)$  and at  $(\tau, n + m)$  caused by observations]. The observations were taken at all the grid points. The values were averaged over grid points  $n$  and forecast cycles. The long white line shows the group velocity of the Lorenz model 3, and the short white lines show phase velocity of the model.

of the ETKF at each forecast cycle with Lorenz model 3 is very much shorter than the correlation length of the climatological distribution of model states (section 5). Thus, we argued that we do not lose much information from localization of the analysis if the size of the local patch is big enough to cover the correlation length of the deviations of the background ensemble. The comparison of the covariance matrices of the analysis ensembles with and without localization shows that they have similar patterns (section 6), thus providing strong support for the achievable accuracy of localization. The effect of an observation at space–time point  $(t, n)$  on a forecast at  $(t + \tau, n + m)$  was found to be local. In addition, the information obtained from the observations propagates both forward (to the right) and backward (Fig. 14b). However, the forward propagation, which is in the direction of the group velocity (to the right), is faster than the backward propagation (to the left), and the maximum effect propagates at a speed that is close to the group velocity (section 7).

**Acknowledgments.** This work was supported by ONR Contract N00014-07-1-0734 and NSF Contract ATM-0935538. Craig Bishop and an anonymous reviewer made several useful suggestions to improve the paper.

## APPENDIX A

### Extracting an Envelope

See Zimin et al. (2003) for the original introduction of the method presented here. Let us say that we have a function

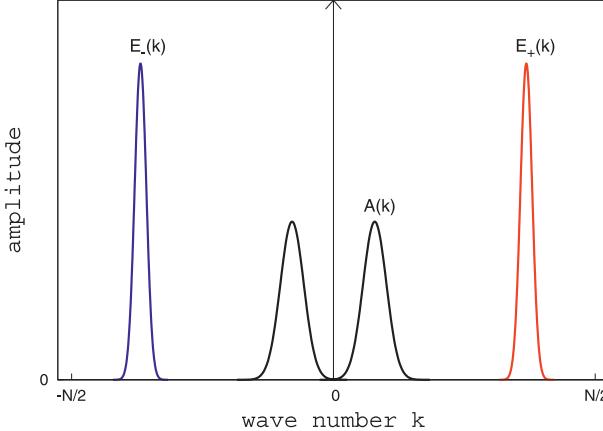


FIG. A1. DFT of  $a(n)$ ,  $e^{i\phi(n)}$ , and  $e^{-i\phi(n)}$ . The terms  $A(k)$ ,  $E_+(k)$ , and  $E_-(k)$  are DFTs of  $a(n)$ ,  $e^{i\phi(n)}$ , and  $e^{-i\phi(n)}$ , respectively, and  $f(n) = a(n) \cos[\phi(n)]$ ,  $0 \leq n < N$ .

$$f(n) = a(n) \cos[\phi(n)], \quad 0 \leq n < N, \quad (\text{A1})$$

where  $n$  is an integer,  $a(n)$  is an envelope, and  $\cos[\phi(n)]$  is an oscillating part with  $\phi(n)$  monotonically increasing. We want to extract  $a(n)$  given  $f(n)$ . First,  $f(n)$  can be expressed as

$$f(n) = \frac{1}{2}a(n)e^{i\phi(n)} + \frac{1}{2}a(n)e^{-i\phi(n)}. \quad (\text{A2})$$

Let us define  $F(k)$ ,  $A(k)$ ,  $E_+(k)$ , and  $E_-(k)$  to be discrete Fourier transforms (DFTs) of  $f(n)$ ,  $a(n)$ ,  $e^{i\phi(n)}$ , and  $e^{-i\phi(n)}$ , respectively. Assuming each function is  $N$ -periodically extended, we have the following:

$$F(k) = \frac{1}{2}\text{DFT}\{a(n)e^{i\phi(n)}\} + \frac{1}{2}\text{DFT}\{a(n)e^{-i\phi(n)}\} \quad (\text{A3})$$

$$= \frac{1}{2N}A(k)*E_+(k) + \frac{1}{2N}A(k)*E_-(k) \quad (\text{A4})$$

$$\begin{aligned} &= \frac{1}{2N} \sum_{l=0}^{N-1} A(l)E_+(k-l) + \frac{1}{2N} \sum_{l=0}^{N-1} A(l)E_-(k-l), \end{aligned} \quad (\text{A5})$$

where the asterisk denotes circular convolution and the curly braces denote the DFT function. The term  $E_+(k)$  is centered around a positive frequency, and  $E_-(k)$  is centered around a negative frequency. If we assume that  $a(n)$  changes much more slowly than  $\phi(n)$ , then  $A(k)$  is well separated from  $E_+(k)$  and  $E_-(k)$ . See Fig. A1, where the horizontal axis is the wavenumber  $k$  and the vertical axis is the amplitude of the DFT values. The red, blue, and black plots are for DFTs of  $e^{i\phi(n)}$ ,  $e^{-i\phi(n)}$ , and  $a(n)$ , respectively. Therefore,  $A(k)*E_+(k)$  will be on the

positive frequency side and  $A(k)*E_-(k)$  will be on the negative frequency side. So, taking inverse discrete Fourier transform of  $F(k)$  only with positive frequencies,  $\text{DFT}_+^{-1}\{F(k)\}$ , is equivalent to taking inverse discrete Fourier transform of  $(2N)^{-1}A(k)*E_+(k)$ . Therefore, we have the following:

$$\text{DFT}_+^{-1}\{F(k)\} = \text{DFT}^{-1}\left\{\frac{1}{2N}A(k)*E_+(k)\right\} \quad (\text{A6})$$

$$= \frac{1}{2}a(n)e^{i\phi(n)}. \quad (\text{A7})$$

Thus,

$$a(n)e^{i\phi(n)} = 2\text{DFT}_+^{-1}\{F(k)\}. \quad (\text{A8})$$

Taking the absolute values of both sides, we obtain

$$a(n) = 2|\text{DFT}_+^{-1}\{F(k)\}|. \quad (\text{A9})$$

In the above equation, only several frequencies that have large DFT values can be used to approximate the envelope. This has the effect of smoothing out the envelope by sacrificing accuracy. As more and more frequency components are added, more and more wiggles are present in the envelope.

## APPENDIX B

### LETKF Algorithm

We present a summary of the LETKF algorithm of Hunt et al. (2007) used in our study with Lorenz model 3. We assume a discrete model that has grid points. Each grid point has a variable. A local region associated with grid point  $n$  is defined as a region consisting of grid points  $i$  with  $n - L \leq i \leq n + L$  for a certain integer  $L$ . The input is a global background ensemble of  $m_g$ -dimensional model state vectors  $[\mathbf{x}_g^{b(i)}, i = 1, 2, \dots, k]$ , an  $l_g$ -dimensional vector  $\mathbf{y}_g^o$  of observations, a function  $H$  that maps the  $m_g$ -dimensional model space to the  $l_g$ -dimensional observation space, and an  $l_g \times l_g$  observation error covariance matrix  $\mathbf{R}_g$ . Here, the subscript  $g$  denotes global. The output is a global analysis ensemble of  $m_g$ -dimensional model state vectors  $[\mathbf{x}_g^{a(i)}, i = 1, 2, \dots, k]$ . The algorithm consists of nine steps:

- 1) Apply  $H_g$  to each  $\mathbf{x}_g^{b(i)}$  to form the global background observation ensemble  $[\mathbf{y}_g^{b(i)}]$  and average the latter vectors to get the  $l_g$ -dimensional column vector  $\bar{\mathbf{y}}_g^b$ . Subtract this vector from each  $\mathbf{y}_g^{b(i)}$  to form the columns of the  $l_g \times k$  matrix  $\mathbf{Y}_g^b$ .

- 2) Average the vectors  $[\mathbf{x}_g^{b(i)}]$  to get the  $m_g$ -dimensional column vector  $\bar{\mathbf{x}}_g^b$  and subtract this vector from each  $[\mathbf{x}_g^{b(i)}]$  to form the columns of the  $m_g \times k$  matrix  $\mathbf{X}_g^b$ . For each grid point, do the following steps 3–8.
- 3) Select the rows of  $\bar{\mathbf{x}}_g^b$  and  $\mathbf{X}_g^b$  corresponding to the local region of  $m$  grid points associated with the given grid point, forming the  $m$ -dimensional vector  $\bar{\mathbf{x}}^b$  and the  $m \times k$  matrix  $\mathbf{X}^b$ . Select the rows of  $\bar{\mathbf{y}}_g^b$  and  $\mathbf{Y}_g^b$  corresponding to the  $l$  observations for the local region, forming the  $l$ -dimensional vector  $\bar{\mathbf{y}}^b$  and the  $l \times k$  matrix  $\mathbf{Y}^b$ . Select the corresponding rows of  $\mathbf{y}_g^o$ , forming the  $l$ -dimensional vector  $\mathbf{y}^o$ . Select the corresponding rows and columns of  $\mathbf{R}_g$ , forming the  $l \times l$  matrix  $\mathbf{R}$ .
- 4) Compute the  $k \times l$  matrix  $\mathbf{C} = (\mathbf{Y}^b)^T \mathbf{R}^{-1}$ .
- 5) Compute the  $k \times k$  matrix  $\tilde{\mathbf{P}}^a = [(k-1)\mathbf{I}/\rho + \mathbf{C}\mathbf{Y}^b]^{-1}$ , where  $\rho > 1$  is a covariance inflation factor.
- 6) Compute the  $k \times k$  matrix  $\mathbf{W}^a = [(k-1)\tilde{\mathbf{P}}^a]^{1/2}$ , where the power  $1/2$  means the symmetric square root.
- 7) Compute the  $k$ -dimensional vector  $\bar{\mathbf{w}}^a = \tilde{\mathbf{P}}^a \mathbf{C}(\mathbf{y}^o - \bar{\mathbf{y}}^b)$  and add it to each column of  $\mathbf{W}^a$ , forming a  $k \times k$  matrix whose columns are the analysis vectors  $[\mathbf{w}^{a(i)}]$ .
- 8) Multiply  $\mathbf{X}^b$  by each  $\mathbf{w}^{a(i)}$  and add  $\bar{\mathbf{x}}^b$  to get the local analysis ensemble members  $[\mathbf{x}^{a(i)}]$  for the given grid point.
- 9) After performing steps 3–8 for each grid point, form the global analysis ensemble  $[\mathbf{x}_g^{a(i)}]$  by combining  $[\mathbf{x}^{a(i)}]$  from each local region by taking weighted averages,

$$x_{n,g}^{a(i)} = \sum_{j=n-L}^{n+L} x_{n,l(j)}^{a(i)} f(n-j),$$

where the subscript  $l(j)$  denotes the local region associated with the grid point  $j$  and  $f(\cdot)$  is a weighting function defined on the integers from  $-L$  to  $L$  for a certain value  $L$ .

## REFERENCES

- Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903.
- , 2007: Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D*, **230**, 99–111.
- , and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758.
- Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436.
- Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, **126**, 1719–1724.
- Charney, J. G., 1949: On a physical basis for numerical prediction of large-scale motions in the atmosphere. *J. Meteor.*, **6**, 371–385.
- Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10 143–10 162.
- Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790.
- Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811.
- , and —, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137.
- , and —, 2005: Ensemble Kalman filtering. *Quart. J. Roy. Meteor. Soc.*, **131**, 3269–3289.
- , —, G. Pellerin, M. Buehner, M. Charron, L. Speck, and B. Hansen, 2005: Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Wea. Rev.*, **133**, 604–620.
- Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230** (1–2), 112–126.
- Kepert, J. D., 2009: Covariance localisation and balance in an ensemble Kalman filter. *Quart. J. Roy. Meteor. Soc.*, **135**, 1157–1176.
- Kleeman, R., 2007: Information flow in ensemble weather predictions. *J. Atmos. Sci.*, **64**, 1005–1016.
- Lorenz, E. N., 1996: Predictability—A problem partly solved. *Proc. Seminar on Predictability*, Vol. 1, Reading, United Kingdom, ECMWF, 1–18.
- , 2005: Designing chaotic models. *J. Atmos. Sci.*, **62**, 1574–1587.
- , and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, **55**, 399–414.
- Ott, E., and Coauthors, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428.
- Persson, A., 2000: Synoptic-dynamic diagnosis of medium range weather forecast systems. *Proc. Seminar on Diagnosis of Models and Data Assimilation Systems*, Reading, United Kingdom, ECMWF, 123–137.
- Szunyogh, I., Z. Toth, A. V. Zimin, S. J. Majumdar, and A. Persson, 2002: Propagation of the effect of targeted observations: The 2000 Winter Storm Reconnaissance program. *Mon. Wea. Rev.*, **130**, 1144–1165.
- , E. J. Kostelich, G. Gyarmati, E. Kalnay, B. R. Hunt, E. Ott, and E. Satterfield, 2008: A local ensemble transform Kalman filter data assimilation system for the NCEP global model. *Tellus*, **60A**, 113–130.
- Whitaker, J. S., T. M. Hamill, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, **136**, 463–482.
- Zimin, A. V., I. Szunyogh, D. J. Patil, B. R. Hunt, and E. Ott, 2003: Extracting envelopes of Rossby wave packets. *Mon. Wea. Rev.*, **131**, 1011–1017.