ABSTRACT

Title of dissertation: INFORMATION FLOW
IN AN ATMOSPHERIC MODEL
AND DATA ASSIMILATION

Young-noh Yoon, Doctor of Philosophy, 2011

Dissertation directed by: Professor Edward Ott
Department of Physics

Weather forecasting consists of two processes, model integration and analysis (data assimilation). During the model integration, the state estimate produced by the analysis evolves to the next cycle time according to the atmospheric model to become the background estimate. The analysis then produces a new state estimate by combining the background state estimate with new observations, and the cycle repeats. In an ensemble Kalman filter, the probability distribution of the state estimate is represented by an ensemble of sample states, and the covariance matrix is calculated using the ensemble of sample states. We perform numerical experiments on toy atmospheric models introduced by Lorenz in 2005 to study the information flow in an atmospheric model in conjunction with ensemble Kalman filtering for data assimilation. This dissertation consists of two parts.

The first part of this dissertation is about the propagation of information and the use of localization in ensemble Kalman filtering. If we can perform data assimilation locally by considering the observations and the state variables only near

each grid point, then we can reduce the number of ensemble members necessary to cover the probability distribution of the state estimate, reducing the computational cost for the data assimilation and the model integration. 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. We address these issues and elucidate the role played by chaotic wave dynamics in the propagation of information and the resulting impact on forecasts.

The second part of this dissertation is about ensemble regional data assimilation using joint states. Assuming that we have a global model and a regional model of higher accuracy defined in a subregion inside the global region, we propose a data assimilation scheme that produces the analyses for the global and the regional model simultaneously, considering forecast information from both models. We show that our new data assimilation scheme produces better results both in the subregion and the global region than the data assimilation scheme that produces the analyses for the global and the regional model separately.

INFORMATION FLOW IN AN ATMOSPHERIC MODEL AND DATA ASSIMILATION

by

Young-noh Yoon

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Edward Ott, Chair/Advisor
Professor Brian R. Hunt
Professor Thomas M. Antonsen
Professor Michelle Girvan
Professor Kayo Ide

Dedication

To my parents Minja Kim and past Yongwan Yoon.

Acknowledgments

First, I would like to thank my advisor Edward Ott for his invaluable guidance and advice. I also would like to thank Istvan Szunyogh for his detailed advice. I also would like to thank Brian Hunt for refreshing the project by introducing new ideas when the project was a little bit at a loss.

I also thank Jonghee Lee for helping me find a research area that I can enjoy. He helped me make the brave decision of choosing the right research group when I did not have enough courage to make a big decision.

Finally, I thank my father past Yongwan Yoon and my mother Minja Kim for their love and everything they gave me. They always encouraged me to keep my spirits high during the long journey of my study.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Weather forecast cycle

Weather forecasting employs two cyclically repeated processes, model integration and analysis (data assimilation). During the forecast phase by the model integration $[nT \leq t \leq (n+1)T]$, the state estimate produced by the analysis at time $t = nT$ evolves to the next cycle time $t = (n+1)T$ according to the atmospheric model to become the 'background' estimate at time $t = (n+1)T$. The analysis at time $t = (n+1)T$ then produces a new state estimate by combining the background state estimate with new observations, and the cycle repeats. The analysis state estimate is more accurate than a state estimate that could be obtained by considering only the new observations, because the observations in a real setting have uncertainties, and observations of all relevant variables are not taken everywhere. This dissertation focuses on the application of the ensemble Kalman filtering method to data assimilation.

1.2 Kalman filter

In Kalman filters and probabilistic weather forecasting, in general, one deals with distribution functions of the system state, rather than with a single estimate of

the system state. Consider a state vector $\mathbf{x}$ that evolves according to the equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}). \tag{1.1}$$

We assume that we have a background probability distribution of the state estimate $f_b(\mathbf{x})$ and the new observations $\mathbf{y} = \mathbf{H}(\mathbf{x}) + (\text{noise})$ at a certain analysis time. We further assume that we know the probability distribution of the observation $f_o(\mathbf{y}|\mathbf{x})$ when the true state is $\mathbf{x}$. We want to find a new probability distribution $f_a(\mathbf{x})$ that takes into account both the known background distribution $f_b(\mathbf{x})$ and the information from the new observations. First, we identify the sample space to be $(\mathbf{x}, \mathbf{y})$. We denote the joint probability distribution, the marginal probability distribution of $\mathbf{x}$, the conditional probability distribution of $\mathbf{x}$ conditioned on $\mathbf{y}$, and the conditional probability distribution of $\mathbf{y}$ conditioned on $\mathbf{x}$ by $f(\mathbf{x}, \mathbf{y})$, $f(\mathbf{x})$, $f(\mathbf{x}|\mathbf{y})$, and $f(\mathbf{y}|\mathbf{x})$, respectively. Then, we have

$$f_b(\mathbf{x}) = f(\mathbf{x}), \tag{1.2}$$

$$f_a(\mathbf{x}) = f(\mathbf{x}|\mathbf{y}), \tag{1.3}$$

$$f_o(\mathbf{y}|\mathbf{x}) = f(\mathbf{y}|\mathbf{x}). \tag{1.4}$$

The analysis probability distribution $f_a(\mathbf{x})$ can be calculated using Bayes' theorem as

$$f_a(\mathbf{x}) = f(\mathbf{x}|\mathbf{y}) = \frac{f(\mathbf{x}, \mathbf{y})}{\int f(\mathbf{x}, \mathbf{y})\, d\mathbf{x}} = \frac{f(\mathbf{x}) f(\mathbf{y}|\mathbf{x})}{\int f(\mathbf{x}, \mathbf{y})\, d\mathbf{x}} = \frac{f_b(\mathbf{x}) f_o(\mathbf{y}|\mathbf{x})}{\int f(\mathbf{x}, \mathbf{y})\, d\mathbf{x}}. \tag{1.5}$$

Thus, if we fix $\mathbf{y}$ and take $\mathbf{x}$ as a variable, then we have

$$f_a(\mathbf{x}) \propto f_b(\mathbf{x}) f_o(\mathbf{y}|\mathbf{x}). \tag{1.6}$$

If the observation operator is linear such that $\mathbf{H}(\mathbf{x}) = \mathbf{H}\mathbf{x}$ and all the probability distributions are Gaussian such that

$$f_b(\mathbf{x}) \propto \exp\left[-(1/2)(\mathbf{x}-\bar{\mathbf{x}}_b)^T\mathbf{P}_b^{-1}(\mathbf{x}-\bar{\mathbf{x}}_b)\right], \tag{1.7}$$

$$f_o(\mathbf{y}|\mathbf{x}) \propto \exp\left[-(1/2)(\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right], \tag{1.8}$$

where the covariance matrix of the background distribution is denoted by $\mathbf{P}_b$, and the covariance matrix of the observation noise is denoted by $\mathbf{R}$, then from Eq. (1.6), we have

$$\begin{aligned} f_a(\mathbf{x}) &\propto \exp\left[-(1/2)(\mathbf{x}-\bar{\mathbf{x}}_b)^T\mathbf{P}_b^{-1}(\mathbf{x}-\bar{\mathbf{x}}_b)\right] \\ &\quad \times \exp\left[-(1/2)(\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right] \qquad (1.9) \\ &\propto \exp\left[-(1/2)\left\{(\mathbf{x}-\bar{\mathbf{x}}_b)^T\mathbf{P}_b^{-1}(\mathbf{x}-\bar{\mathbf{x}}_b)\right.\right. \\ &\quad \left.\left.+(\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right\}\right] \qquad (1.10) \\ &\propto \exp\left[-(1/2)(\mathbf{x}-\bar{\mathbf{x}}_a)^T\mathbf{P}_a^{-1}(\mathbf{x}-\bar{\mathbf{x}}_a)\right] \qquad (1.11) \end{aligned}$$

with

$$\mathbf{P}_a = (\mathbf{P}_b^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}, \tag{1.12}$$

$$\bar{\mathbf{x}}_a = \mathbf{P}_a\,(\mathbf{P}_b^{-1}\bar{\mathbf{x}}_b + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}). \tag{1.13}$$

If the time evolution is linear such that

$$\mathbf{x}_{n+1} = \mathbf{M}_n\mathbf{x}_n, \tag{1.14}$$

where $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$ are the state vectors at the cycle time n and $n+1$, and $\mathbf{M}_n$ is time evolution operator that maps $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$, then once we know the mean of the

analysis state estimate $\bar{\mathbf{x}}_{n,a}$ and the analysis error covariance matrix $\mathbf{P}_{n,a}$ at the cycle time n, the background at time $n+1$ is again Gaussian, and we can calculate the mean of the background state estimate $\bar{\mathbf{x}}_{n+1,b}$ and the background error covariance matrix $\mathbf{P}_{n+1,b}$ at the next cycle time $n+1$ as

$$\bar{\mathbf{x}}_{n+1,b} = \mathbf{M}_n \bar{\mathbf{x}}_{n,a}, \tag{1.15}$$

$$\mathbf{P}_{n+1,b} = \mathbf{M}_n \mathbf{P}_{n,a} \mathbf{M}_n^T. \tag{1.16}$$

1.3 Ensemble Kalman filter

In real situations, the evolution is nonlinear. Even if we wanted to approximate the evolution over a cycle time as linear, a suitable linear approximation $\mathbf{M}_n$ for use in Eq. (1.14) is hard to obtain. In this case, an ensemble Kalman filter provides an attractive alternative. In an ensemble Kalman filter, the probability distribution of the state estimate is represented by an ensemble of sample states, and the covariance matrix is calculated using the ensemble of sample states. In principle, the number of sample states must be large enough to yield a good representation of the state distribution function. At the analysis time, a new analysis ensemble of samples is calculated incorporating the information from the new observations and assuming the probability distribution of the estimate is Gaussian. During the forecast phase, each ensemble member is integrated forward in time using the atmospheric model. At the next analysis time, this yields a new set of sample system states, which can then be used to obtain new estimates of $\mathbf{P}_b$ and $\bar{\mathbf{x}}_b$ for use in Eq. (1.7). Note that this procedure essentially supposes that assuming Gaussianity provides useful results, in

spite of the fact that nonlinearity of the state evolution (and also possibly of the observation operator $\mathbf{H}(\mathbf{x})$) will cause this not to be strictly true. Nevertheless, the effectiveness of this assumption seems to be substantial, as can be assessed by seeing how well it works in practice; e.g., by seeing if reasonably good forecasts are produced.

1.4 Localization

While the ensemble Kalman filter approach (Sec. 1.3) provides a useful framework, its *direct* application to weather forecasting using current global atmospheric models is computationally impossible. This is because such a model evolves a state consisting of the atmospheric variables (fluid velocity, temperature, etc.) evaluated on all the grid points. Thus, the state vector can typically have a dimensionality of more than 10^6. Moreover, a very large number of sample states would be necessary to effectively span all the significant covariances necessary for making an effective estimate of the global state distribution function. Thus, we require some way of circumventing this problem. To do this, workers in the field have suggested and implemented the idea of 'localization', which is explained in Chapter 2.

1.5 Organization of the thesis

This dissertation consists of two parts, chapter 2: "On the propagation of information and the use of localization in ensemble Kalman filtering", and chapter 3: "Ensemble regional data assimilation using joint states". In chapter 2, we show

how localization in data assimilation is justified by studying the correlation length of the background ensemble and how the information obtained from new observations propagates in time and space. In chapter 3, we propose a new data assimilation scheme for situations in which we have access to a global model and a regional model of higher accuracy defined in a subregion within the global domain. Using numerical experiments on toy models, we show that our proposed new scheme can be substantially more accurate than schemes currently in use.

Chapter 2

On the propagation of information and the use of localization in ensemble Kalman filtering

2.1 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., [8, 5, 10]. 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., [11, 9, 1, 3, 21, 14]. In this chapter, we concentrate on one localization scheme called the local ensemble Kalman filter (LEKF) [21] and a subsequent computationally more efficient version, the local ensemble transform Kalman filter (LETKF) [14]. We expect that simi-

lar 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 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., [12, 13, 27, 25].

One goal of this chapter 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 [18]. 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 [4]). 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 chapter 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., [6, 22, 26, 29]. 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., [16]. 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 [3].

The organization of this chapter is as follows: In section 2.2, we introduce the model system [18] that we employ, and we investigate some of its salient properties. In section 2.3, we investigate correlations of the model. In section 2.4, we explain the forecast and localization procedures used in the ETKF and LETKF schemes. In section 2.5, we study correlations of the background ensemble of an ETKF as a function of spatial distance. In section 2.6, we compare covariance matrices obtained using an ETKF and an LETKF. In section 2.7, we show how observations taken in a small region at a certain time affect the forecast at other spatial points in the future. Section 2.8 provides further discussions and summarizes our main conclusions.

2.2 Model

Lorenz 2005[18] 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 [17] and Lorenz and Emanuel 1998 [19]. 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 [17]). 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 [18], 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:

$$dZ_n/dt = [X, X]_{K,n} + b^2[Y, Y]_{1,n} + c[Y, X]_{1,n} - X_n - bY_n + F, \tag{2.1}$$

where n is a spatial variable ($n = 0, 1, \ldots, N-1$), periodic boundary conditions are

employed ($Z_N = Z_0$), and the X and Y vectors are defined as

$$X_n = \sum_{i=-I}^{I}{}'(\alpha - \beta|i|)Z_{n+i}, \tag{2.2}$$

$$Y_n = Z_n - X_n, \tag{2.3}$$

$$\alpha = (3I^2 + 3)/(2I^3 + 4I), \tag{2.4}$$

$$\beta = (2I^2 + 1)/(I^4 + 2I^2), \tag{2.5}$$

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

$$[X, Y]_{K,n} = \sum_{j=-J}^{J}{}' \sum_{i=-J}^{J}{}'(-X_{n-2K-i}Y_{n-K-j} + X_{n-K+j-i}Y_{n+K+j})/K^2 \tag{2.6}$$

when K is even, and Σ' is replaced by Σ 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 chapter are $N = 960, K = 32, b = 10, c = 2.5, F = 15$, and $I = 12$. We solve Eqs. (2.1)–(2.6) 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 Y_n superposed on a large-scale smooth component X_n, thus crudely mimicking the multiscale dynamics of a real atmospheric system.

Figure 2.1 obtained from a numerical solution of Eq. (2.1) 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 2005 [Lorenz justifies this correspondence between t in Eq. (2.1) and pseudohours by considering the damping time of the model]. The colors in Fig. 2.1

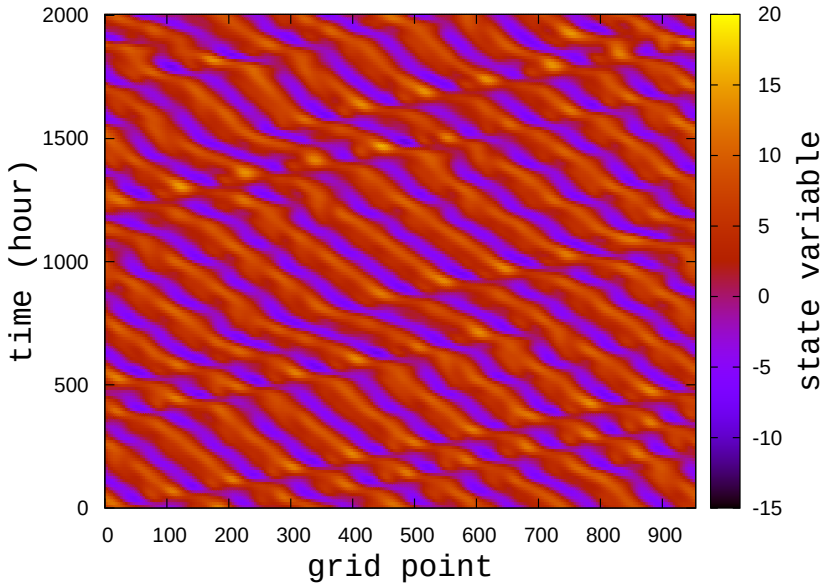

Figure 2.1: Time evolution of state Z_n of Lorenz model 3. In this figure, 6 h are defined to be 0.05 time units in the model equations considering the damping time of the model.

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. 2.1 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 [29]. The method of [29] for extracting modulating envelopes is explained briefly in appendix A. Based on the observed structure in Fig. 2.1, we computed envelopes using wavenumbers of only 7 and 8 to show smooth flow of the envelope (see appendix A). Figure 2.2 depicts the envelope amplitude in

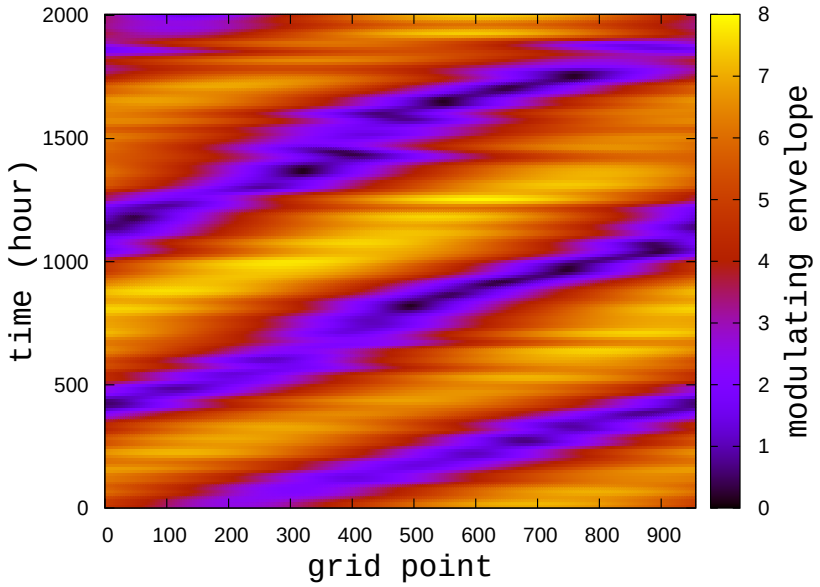

Figure 2.2: Time evolution of the wave packet envelopes extracted by the envelope extraction method introduced in Zimin et al. 2003 [29]. Wave numbers of only 7 and 8 were used to smooth out the envelopes.

time and space for the same numerical run as in Fig. 2.1. From Fig. 2.2, we see that wave packets propagate to the right, in contrast to the leftward propagation of the individual troughs and ridges seen in Fig. 2.1. 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. 2.1) and group velocities that move to the right (as in Fig. 2.2). 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. [24, 23, 7]

2.3 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 τ in time as follows:

$$C_m(\tau) = \frac{\langle\, [\, Z_n(t) - \langle Z_n(t)\rangle\,]\,[\, Z_{n+m}(t+\tau) - \langle Z_n(t)\rangle\,]\,\rangle}{\langle\, [\, Z_n(t) - \langle Z_n(t)\rangle\,]^2\,\rangle}, \tag{2.7}$$

where angle brackets denote an average over n and t. The resulting plots, $C_m(\tau)$ (Fig. 2.3) and $C_m(0)$ (Fig. 2.4), 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. 2.4), (2) the wave structure shifts to the left with a phase velocity of about -0.77 grid points per hour (Fig. 2.3), 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 the red troughs in Fig. 2.3). Under the assumption of ergodicity, the correlation in Eq. (2.7) can be regarded as describing an ensemble drawn from the distribution that defines the climate of the model. In section 2.5, 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 τ from Fig. 2.3 using the envelope extraction method in [29] with wavenumbers of 6, 7, and 8. We obtained the results depicted in Fig. 2.5. It is seen from Fig. 2.5 that, similar to Fig. 2.2, the envelope in m-τ space propagates to the

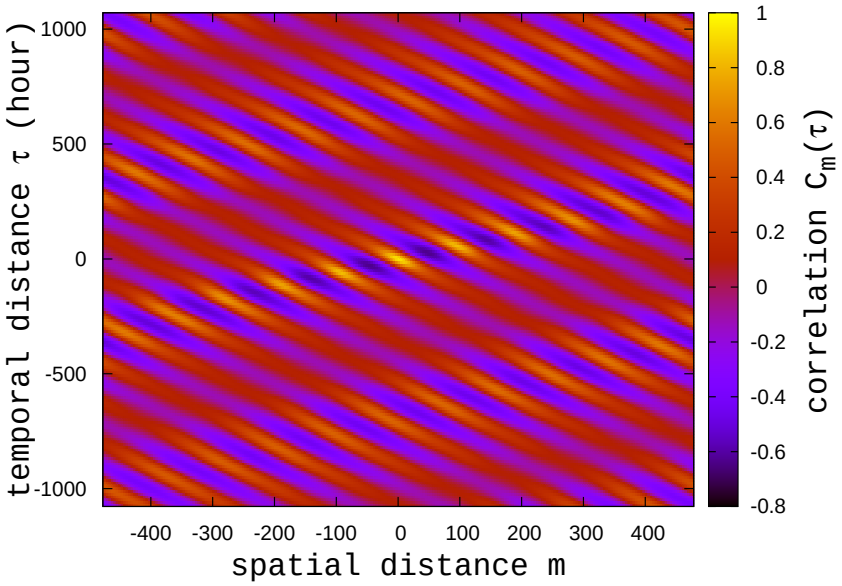

Figure 2.3: Correlations $C_m(\tau)$ between state variables at two different points, (n, t) and $(n + m, t + \tau)$ in Fig. 2.1. The values were averaged over all of the space-time locations (n, t).

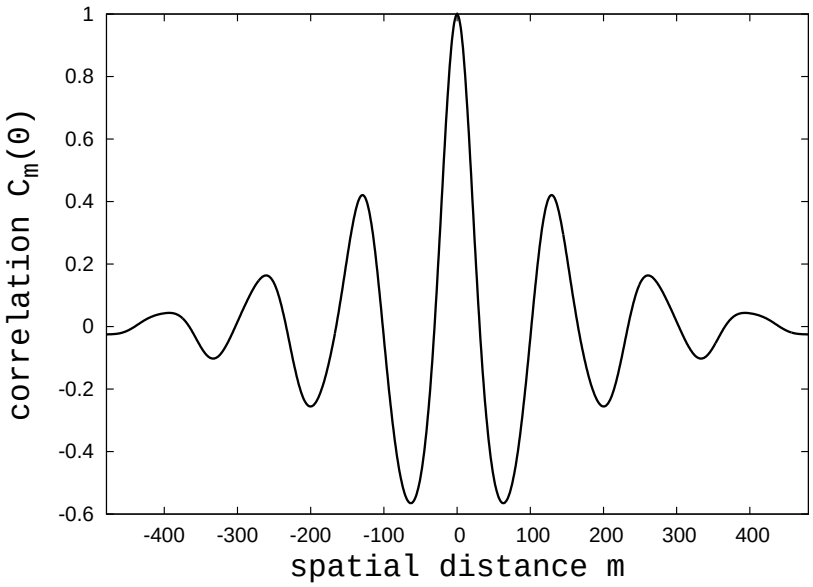

Figure 2.4: Correlations $C_m(0)$; i.e., the correlation values of Fig. 2.3 with time difference τ set to 0.

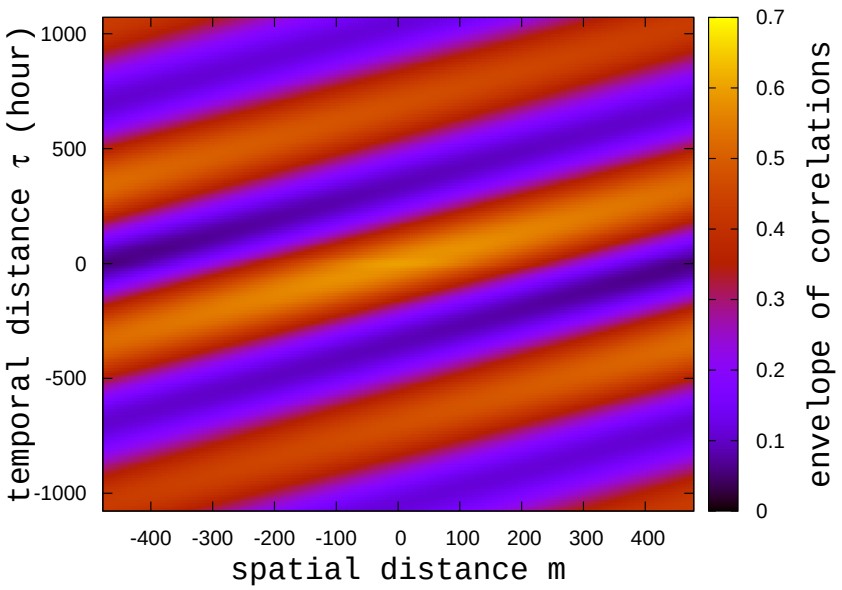

Figure 2.5: Envelopes extracted from correlations in Fig. 2.3 at each time τ using wave numbers only from 6 to 8 to smooth out the result.

right at a group velocity of about 1.37 grid points per hour.

2.4 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. (2.1)–(2.6), 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 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. Using these simulated measurements, we do analyses employing the same model, Eqs. (2.1)–(2.6), as we use to generate the true state evolutions [i.e., Eqs. (2.1)–(2.6)].

In our ETKF implementation, we use the same technique as described in Hunt et al. 2007 [14], 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 of K^3. Hence, if we can do computations in local regions with much fewer grid points than in the global region, and thus with correspondingly much 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 [14] and Szunyogh et al. 2008 [25].

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. (2.1), 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 $(i-L, i-L+1, \ldots, i+L-1, i+L)$ for $i = 0, 1, \ldots, N-1$. 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 [21], Hunt et al. 2007 [14], 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 [15].

2.5 Spatial correlations of the background ensemble

As we have pointed out, Fig. 2.4 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 $Z_n^{(k)}(t), k = 1, 2, \ldots, 960$, by running 1000 forecast cycles (we use k and t to 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 [2] 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 chapter). Using these parameters, we calculated the following ensemble-based spatial correlation in the background:

$$C_m^e = \left\langle \frac{\langle\, [\, Z_n^{(k)}(t) - \bar{Z}_n(t)\,]\,[\, Z_{n+m}^{(k)}(t) - \bar{Z}_{n+m}(t)\,]\,\rangle_k}{\sqrt{\langle\, [\, Z_n^{(k)}(t) - \bar{Z}_n(t)\,]^2\,\rangle_k}\sqrt{\langle\, [\, Z_{n+m}^{(k)}(t) - \bar{Z}_{n+m}(t)\,]^2\,\rangle_k}} \right\rangle_{n,t}, \tag{2.8}$$

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 2.6 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

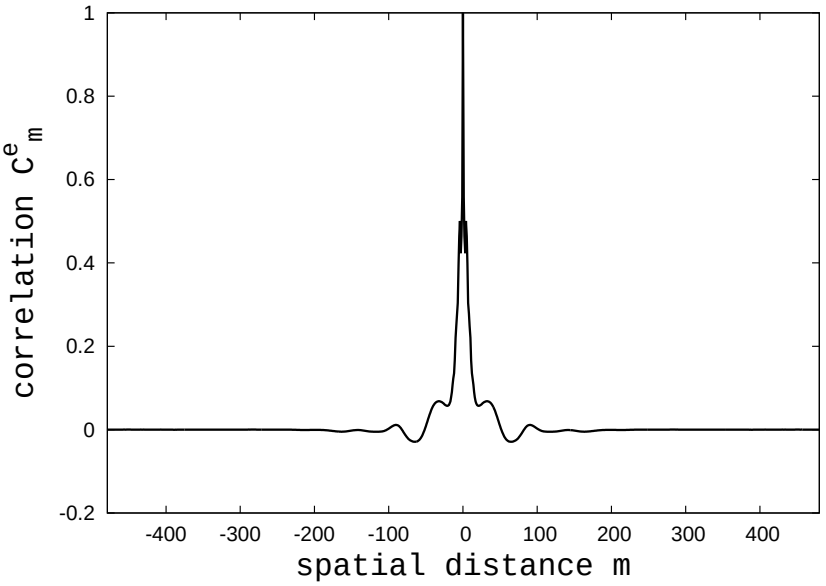

Figure 2.6: Correlation C^e_m of the background ensemble of ETKF with 6-h cycle time vs spatial distance m. The values were averaged over 1000 forecast cycles.

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.

To analyze further the result shown in Fig. 2.6, 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. 2.7a,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

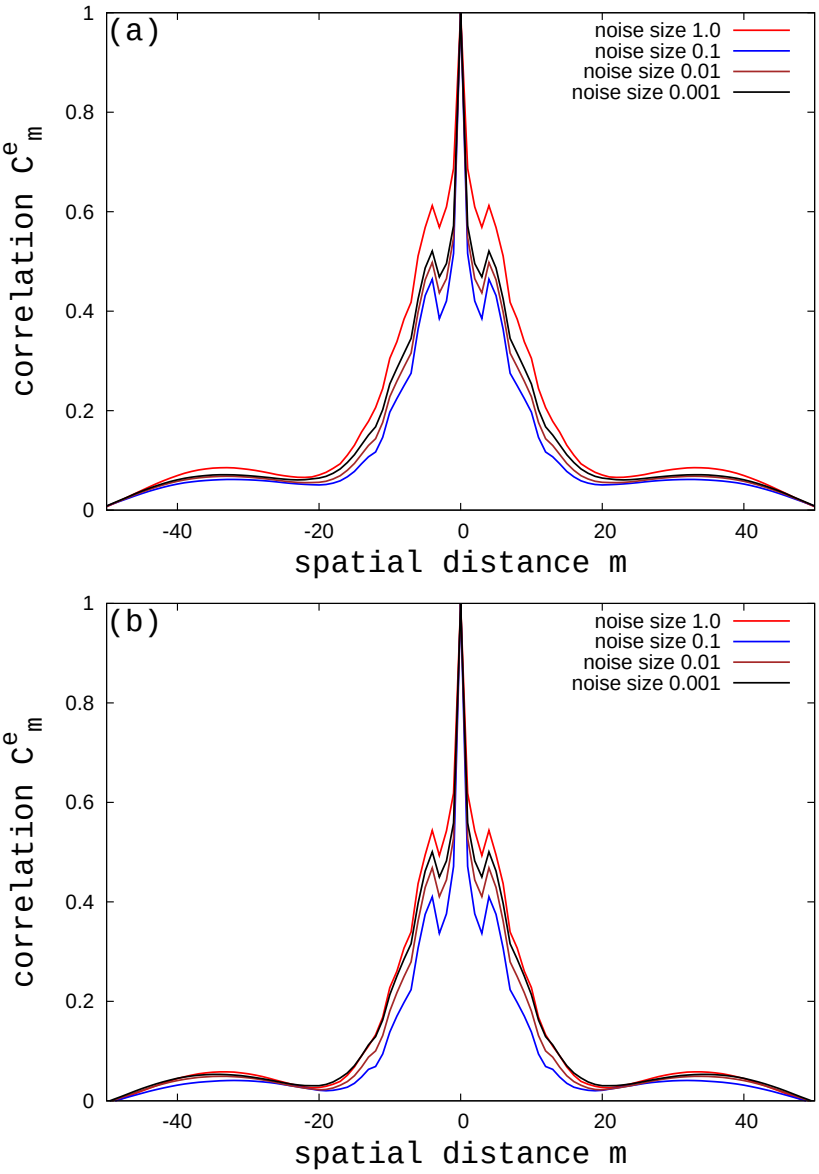

Figure 2.7: Correlations C_m^e (a) of the background ensembles and (b) of the analysis ensembles of ETKF with several different noise sizes. The values were averaged over 1000 forecast cycles. The red, blue, brown, and black lines are for standard deviations of the noise: 1.0, 0.1, 0.01, and 0.001, respectively.

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 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. (2.8) and identifies the maximum group velocity (maximum information propagation speed) to be on the order of the speeds found in Figs. 2.1–2.5 (i.e., $\sim$1 grid point per hour).

In Fig. 2.8, we show results for the same quantity as plotted in Fig. 2.6, but for the cases in which the assimilation cycle times are 6 (as in Fig. 2.6), 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.

2.6 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

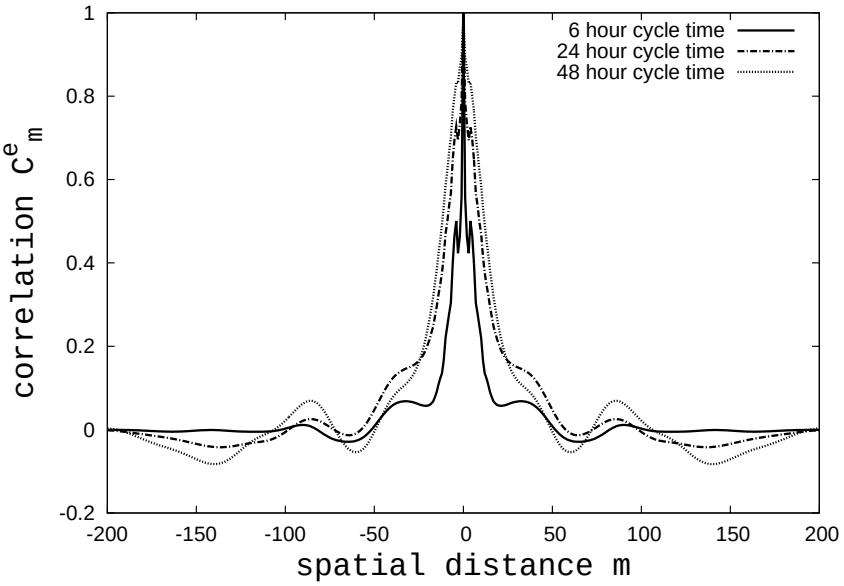

Figure 2.8: Correlation C^e_m of the background ensembles of ETKF with various cycle times vs spatial distance m. The values were averaged over 1000 forecast cycles. The solid, dash-dotted, and dotted lines are for 6-, 24-, and 48-h cycle times, respectively.

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 $P_{i,j}$ of the covariance matrices with $|j - i| \leq 7$ and stacked them row by row so that elements with the same separation of indices $j - i$ are aligned vertically. Figures 2.9a,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 2.10 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 $\langle \sqrt{\langle (\bar{Z}_n^a - Z_n)^2 \rangle_n} \rangle_{cycle}$, where $\bar{Z}_n^a$ is the mean of the analysis ensemble and $\langle \cdots \rangle_n$ and $\langle \cdots \rangle_{cycle}$ 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 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

$$\delta = \left\langle \frac{\left\langle |\bar{P}_i^{LETKF} - \bar{P}_i^{ETKF}| \right\rangle_i}{\left\langle \bar{P}_i^{ETKF} \right\rangle_i} \right\rangle_{cycle}, \tag{2.9}$$

where $\langle \cdots \rangle_i$ 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 2.11 and 2.12 show that the rms error and the quantity δ become smaller as we increase the local patch size. Using the rms error as our figure of merit, we see that there is relatively little benefit to increasing the patch size past about 15, although there is more substantial improvement in δ as the patch size

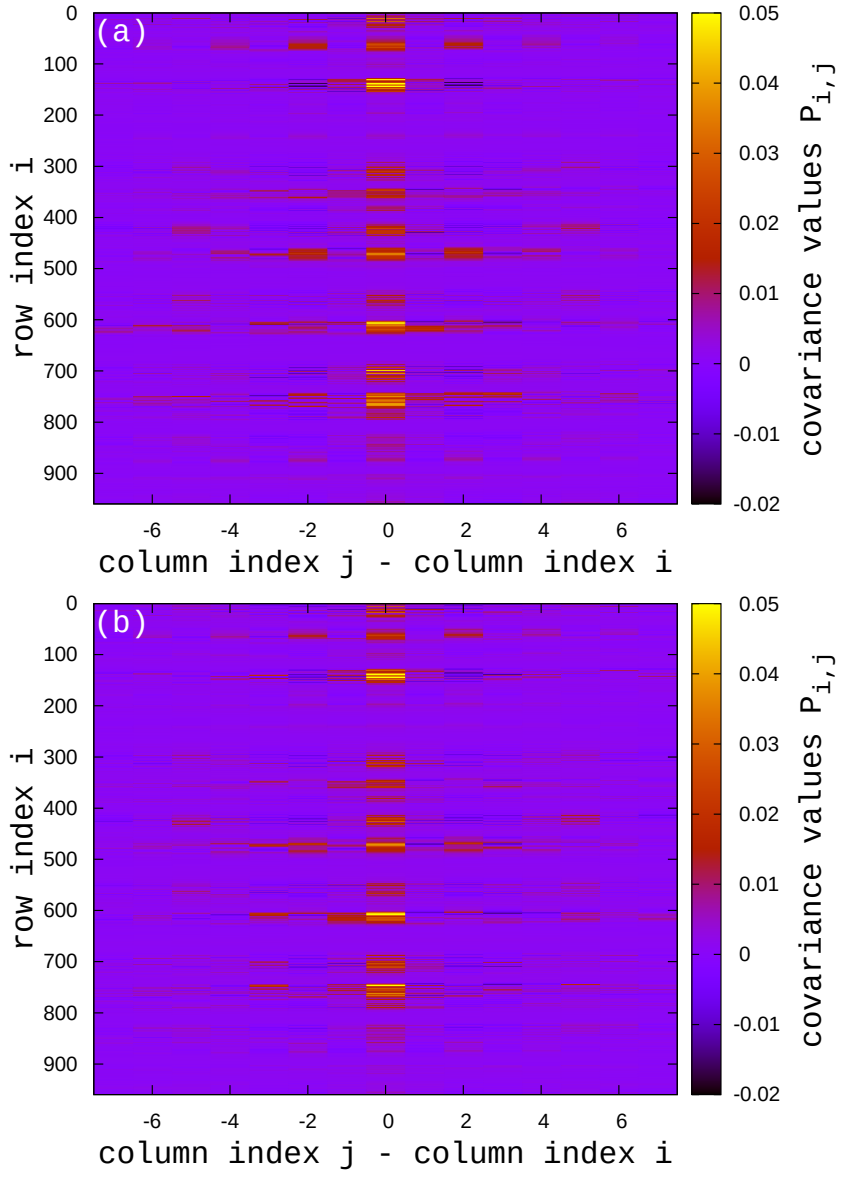

Figure 2.9: Covariance matrix values of analysis ensemble at a certain forecast time of (a) ETKF and (b) LETKF. The y axis is the row index i, and the x axis is the difference between the column indices j and i.

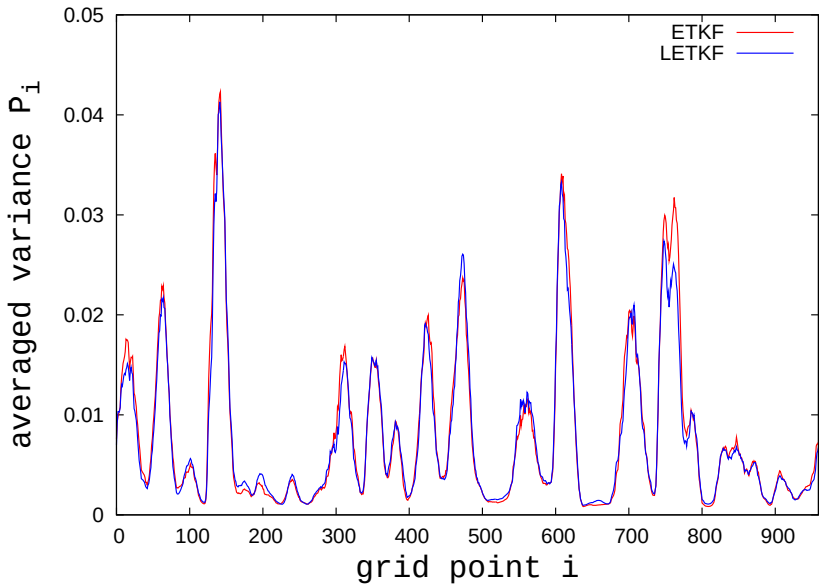

Figure 2.10: Variance values of analysis ensemble at a certain forecast time. The values were averaged over 11 neighboring grid points. Red plot is for ETKF, and the blue plot is for LETKF.

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. 2.12, we can also see that longer forecast cycle time makes δ 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 2.13 shows the difference between the rms errors of LETKF and ETKF divided by the rms error of ETKF, a normalized rms error difference.

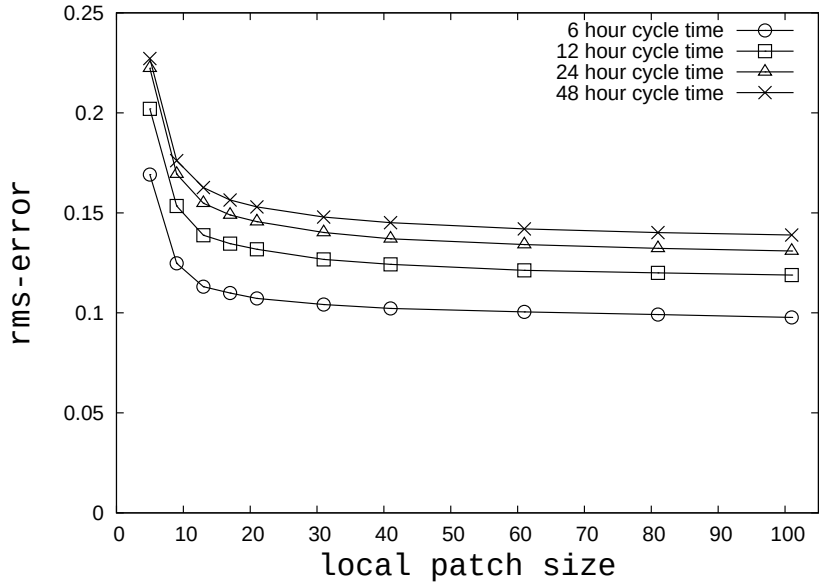

Figure 2.11: Averaged rms errors of the state estimates by the means of the analysis ensembles vs local patch size. The values were averaged over 2000 forecast cycles. The lines with the circles, squares, triangles, and times signs correspond to 6-, 12-, 24-, and 48-h cycle times.

2.7 Demonstration of the localized influence of observations

If the local patch size is large enough to encompass most of the correlation seen in Fig. 2.6, 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 2.5), thus bringing the ensemble members close to the true

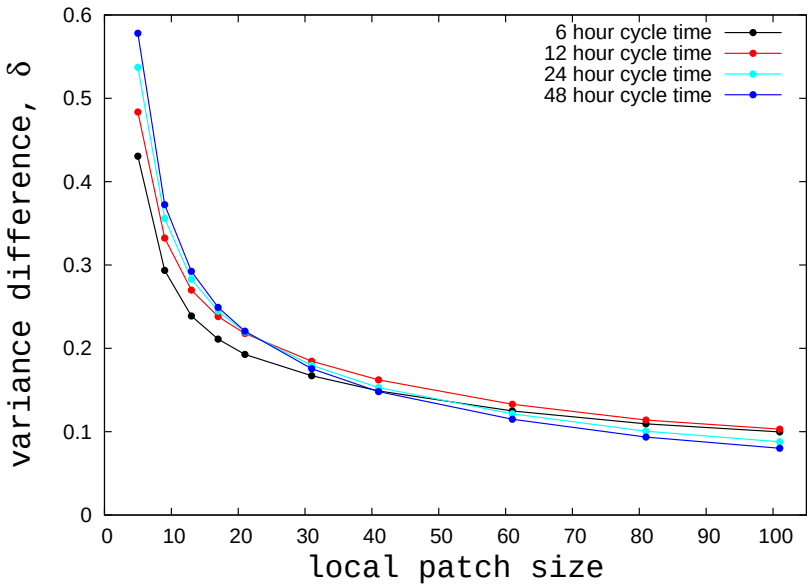

Figure 2.12: Values of δ (normalized difference between variances of analysis ensembles of ETKF and LETKF) vs local patch size. The values were averaged over 2000 forecast cycles. The black, red, cyan, and blue colors correspond to 6-, 12-, 24-, and 48-h cycle times.

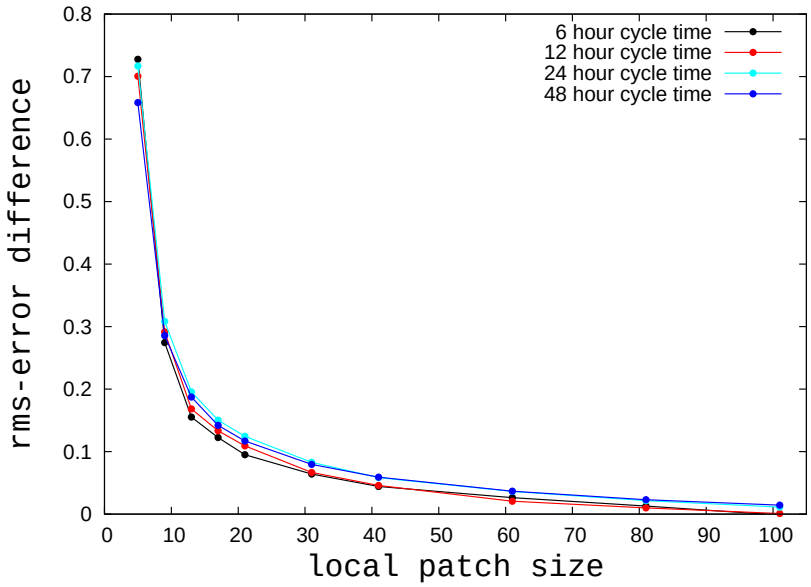

Figure 2.13: Normalized difference between rms errors of state estimates with ETKF and LETKF vs local patch size. The values were averaged over 2000 forecast cycles. The black, red, cyan, and blue colors correspond to 6-, 12-, 24-, and 48-h cycle times.

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 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 $\bar{Z}^b_n(\tau)$ and $\bar{Z}^a_n(\tau)$ (b and a denote background and analysis, respectively). We did this for values of τ ranging from $\tau = 0$ h to $\tau = 150$ h in 6-h intervals, calculating and saving the quantities

$$D_m(\tau) = (\, \bar{Z}^a_{i+m}(\tau) - \bar{Z}^b_{i+m}(\tau)\,)^2. \tag{2.10}$$

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 2.14a shows a plot of $\langle D_m(\tau)\rangle_{cycle}$, the average of $D_m(\tau)$ over 4000 cycles. The horizontal axis is m, and the vertical axis is τ. The color represents $\langle D_m(\tau)\rangle_{cycle}$. 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.

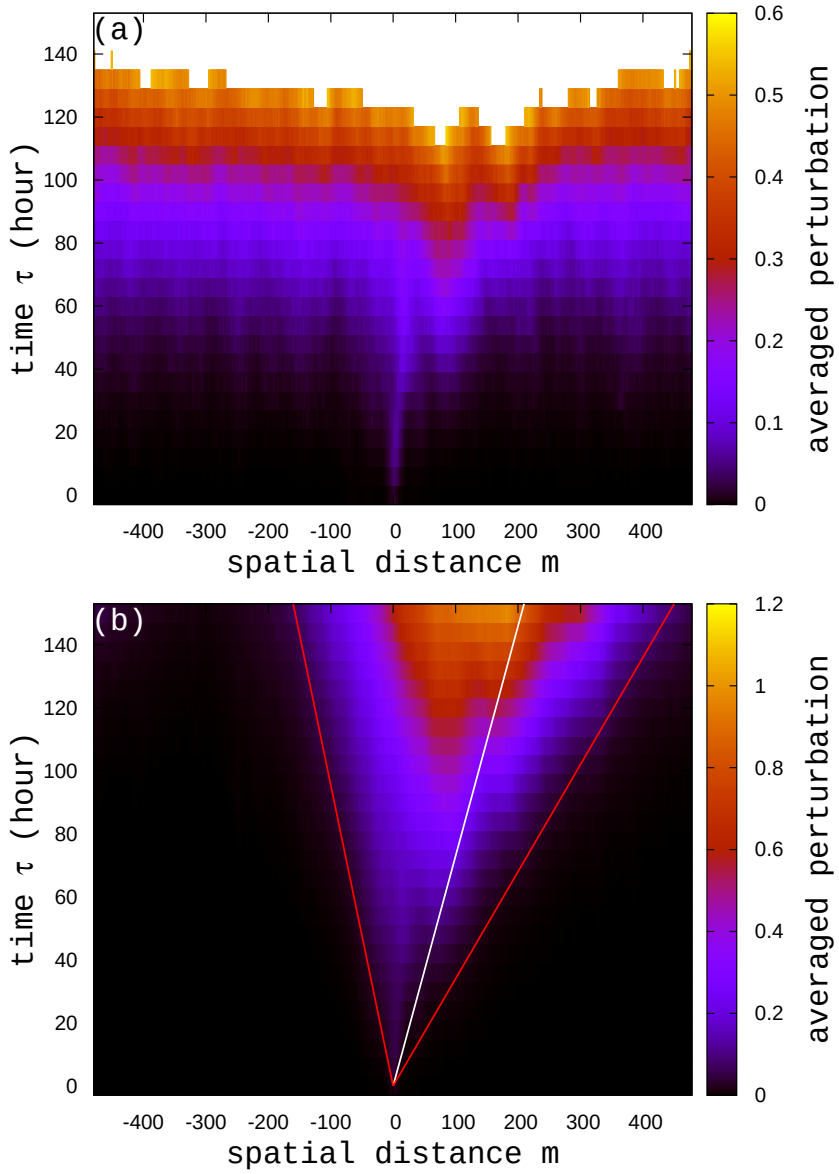

Figure 2.14: Values of $\langle D_m(\tau)\rangle_{cycle}$ (evolution of the perturbation to the ensemble mean caused by five localized observations). The values were averaged over 4000 cycles, (a) with ETKF and (b) with LETKF. The white line shows the group velocity of Lorenz model 3, and the two red lines show the boundaries of the wedgelike region of the perturbations.

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,

$$d_m(\tau_i) = \bar{Z}^a_{i+m}(\tau_i) - \bar{Z}^b_{i+m}(\tau_i), \tag{2.11}$$

at $\tau_i = i \times 6$ h ($i = 0, 1, \ldots, 7$) at a certain cycle time (see Fig. 2.15a). Figure 2.15a shows $d_m(\tau_i)$ plotted versus m at 6-h intervals from $\tau = 0$ h to $\tau = 42$ h. It is seen that small differences at points far from the observation points at $\tau = 0$ h become amplified and dominate the d_m values at later times. At small τ, 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 2.16 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 [9] and Anderson 2007 [3] for detailed discussions of noisy covariance values). As τ 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 use of localization in LETKF has the effect of eliminating the randomlike small initial differences $d_m(0)$ at large $|m|$. Figures 2.14b and 2.15b show the results. It is seen that the effect of the localized observations spreads to other points linearly, creating a wedgelike region in our τ-versus-m diagram indicated by two red lines in Fig. 2.14b and two dotted lines in Fig. 2.15b. The propagation speed corresponding to the right boundary of the wedge is roughly 3 grid points per hour, and the propagation speed

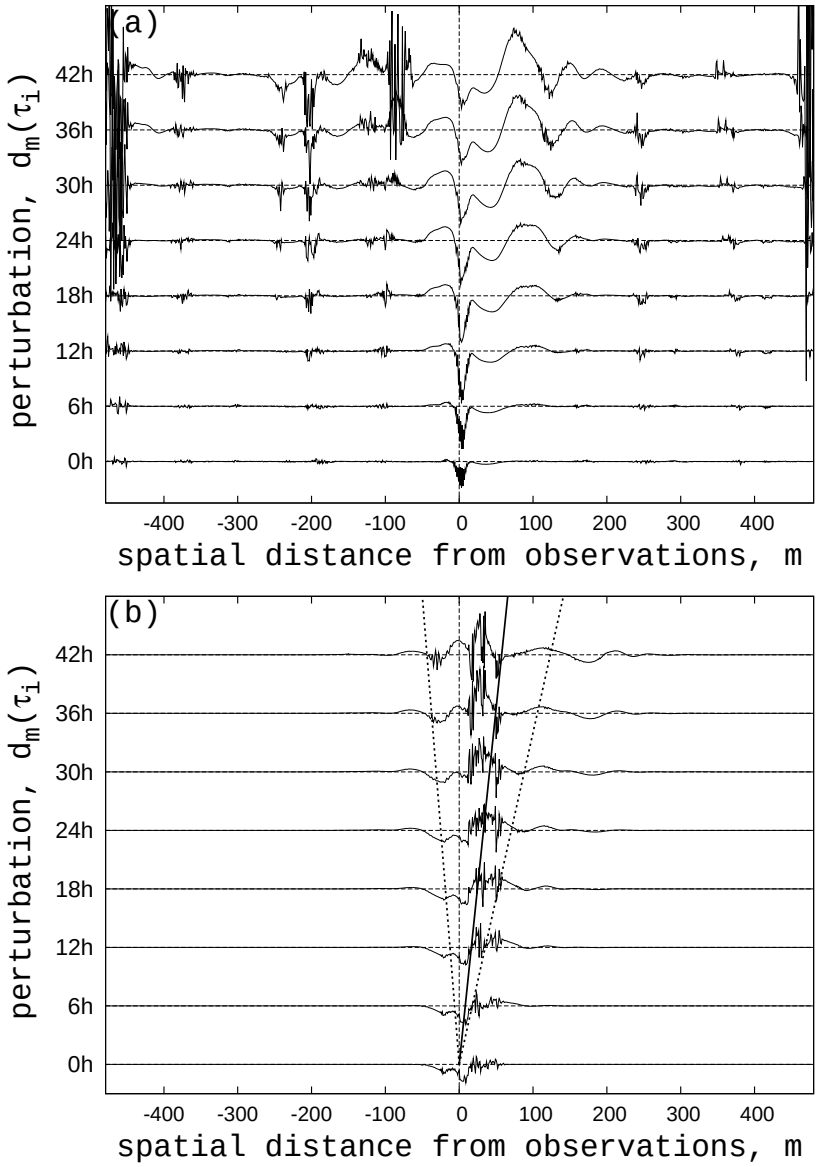

Figure 2.15: Evolution of the perturbation to the ensemble mean caused by five localized observations, $d_m(\tau_i)$, in 6-h time intervals from $\tau = 0$ h to $\tau = 42$ h, at a certain cycle time, (a) with ETKF and (b) with LETKF. The solid line shows the group velocity of Lorenz model 3, and the two dotted lines show the same boundaries in Fig. 2.14b.

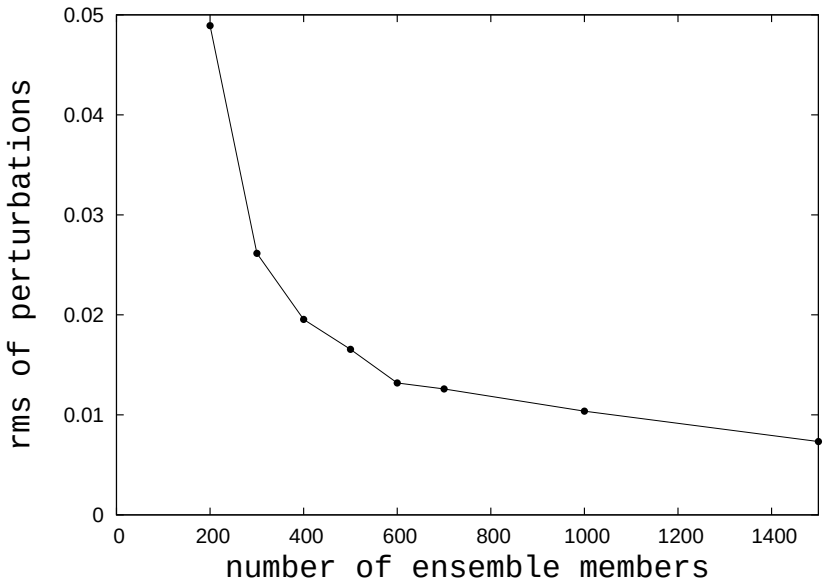

Figure 2.16: The rms values of $d_m(0)$ divided by the size of the noise of observations vs ensemble size of ETKF. The values were averaged over 2000 cycles and over grid points that are more than 300 grid points away from the middle of the observation points.

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. 2.14b and the solid line in Fig. 2.15b) determined from Fig. 2.5.

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:

$$\tilde{C}_m(\tau) = \left\langle \frac{[\bar{Z}^a_{n+m}(\tau) - \bar{Z}^b_{n+m}(\tau)][\bar{Z}^a_n(0) - \bar{Z}^b_n(0)]}{\sigma^2_{noise}} \right\rangle_{n,cycle}, \tag{2.12}$$

where σ_{noise} denotes the standard deviation of the observation noise. Figure 2.17 shows $\tilde{C}_m(\tau)$. The horizontal axis is m, and the vertical axis is τ. The color

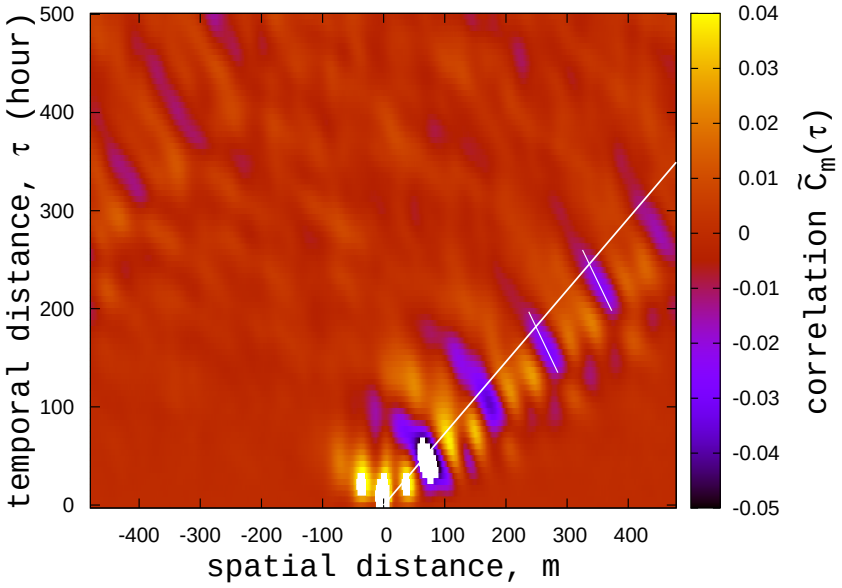

Figure 2.17: Values of $\tilde{C}_m(\tau)$ [correlations between perturbations at $(0, n)$ and at $(\tau, n + m)$ caused by observations]. The observations were taken at all the grid points. The values were averaged over grid points n and forecast cycles. The long white line shows the group velocity of the Lorenz model 3, and the short white lines show phase velocity of the model.

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. 2.3.

2.8 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 2.2 and 2.3). We found that the correlation length of the deviations from the ensemble mean of the background ensemble 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 2.5). 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 2.6), 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. 2.14b). However, the forward propagation, which is in the di-

rection 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 2.7).

Chapter 3

Ensemble regional data assimilation using joint states

3.1 Introduction

Assuming that we have a global model and a regional model of higher accuracy defined in a subregion inside the global region, we aim to produce a forecast which is better than the one from each model by using information from both models. We test two data assimilation methods. The first method is based on techniques most commonly used in current practice and has recently been tested in Merkova et al. 2011 [20]. In this method, the global and the regional data assimilations are done separately, and the regional model receives information from the global model through the boundaries during the integration phase, but the global model does not receive information from the regional model. The second method, which we call the joint state method, is proposed in this chapter. In this method, the global and regional data assimilations are coupled simultaneously using information contained in both the global and the regional forecast states, and the regional model receives information from the global model through the boundaries during the integration phase as in the separate analysis method. We use the Local Ensemble Transform Kalman Filter (LETKF) algorithm for data assimilation. This algorithm allows efficient implementation of the localization technique proposed by Ott et al. 2004 [21]. In order to test our global/regional assimilation techniques we use numerical

experiments based on simple atmospheric 'toy' models proposed in Lorenz 2005 [18] in conjunction with simulated observations. We compare results of our joint state method and results of the separate analysis method. We find that better forecasts are produced by using the joint state method than by using the separate analysis method. We note that our proposed scheme would most likely be of potential interest for centers, where both global and limited area forecasts and analyses are prepared.

The organization of the chapter is the following. Section 3.2 introduces the atmospheric toy models that we use. Section 3.3 describes the data assimilation schemes by the joint state method and by the separate analysis method. Section 3.4 describes how the regional model is coupled to the global model at the boundaries of the subregion during the integration phases of forecast cycles. Section 3.5 compares the results of our joint state method to those of the separate analysis method. Section 3.6 gives further discussion and summarizes our conclusions.

3.2 True model, global model, and regional model

Lorenz 2005 [18] introduced three simple, spatially discrete, 1-dimensional models that have been proven to be useful for testing weather data assimilation methods. Here we will use Lorenz's model 2 (which shows smooth propagating waves) and the more refined Lorenz model 3 (which shows small scale activities on top of smooth waves). Lorenz model 3 reduces to Lorenz model 2 when $I = 1$. In particular, for $I = 1$, Eq. (2.2) yields $X_n = Z_n$, which by Eq. (2.3) implies that

$Y_n = 0$. Thus, after changing notation, $n \to m$ and $Z_n \to Z_m$, we obtain

$$dZ_m/dt = [Z, Z]_{K,m} - Z_m + F, \tag{3.1}$$

where m is used to denote a point on the coarser grid of the global model.

We use Lorenz model 3 with parameter values $N = 960, K = 32, b = 10, c = 0.6, F = 15, I = 12$ to generate our simulated true dynamics, and Lorenz model 2 with $N = 240, K = 8, F = 15$ for the global model defined at every fourth grid point of the true model ($n = 0, 4, 8, \ldots, 956$). Thus the grid points for Eq. (3.1) occur at $m = n/4$, where $n = 0, 4, 8, \ldots, 956$. We assume that, between analyses, Eq. (3.1) for Z_m gives an approximation of the dynamical evolution of $Z_n(t)$ at the grid points $n = 4m$. When referring to locations or lengths of regions, we use the coordinate system of the true model throughout this chapter ($n = 0, 1, \ldots, 959$). For the regional model, we define a subregion extending from $n = n_0 = 240$ to $n = n_1 = 720$ grid, and use Lorenz model 3 with the same parameter values as the true model. In order to integrate this regional model, we must evaluate the bracket quantities on the right hand side of Eq. (2.1) defined by Eq. (2.6). For n too close to n_0 (n_1) this involves X, Y, and Z values at grid points outside the subregion, $n < n_0$ ($n > n_1$). Also, from Eq. (2.2), X_n in the regional model (and hence also Y_n) depends on $Z_{n'}$ values in $n' < n_0$ ($n' > n_1$) if n is within a distance I of n_0 (n_1). To evaluate these quantities, we use estimates of the required values of $Z_{n'}$ obtained from linear interpolation of the global values Z_m onto the n-grid. These interpolations essentially play the role of boundary conditions for the regional model.

3.3 Data assimilation

We selected 15 evenly spaced observation points starting from $n = 0$ ($n = 0, 64, 128, \ldots, 896$). Notice that all the observation points are at grid points defined in the global model. We construct simulated observations by adding random noise drawn from independent Gaussian distributions of standard deviation 1 to the true state values at the observation points.

We compare two data assimilation methods. The first method does data assimilation for the global model and the regional model separately, while the second method, which we call the joint state method, forms a combined state from the global model and the regional model and does data assimilation on the combined state. The intuition motivating our second method is that we expect the global and the regional estimates will both benefit from information exchange between them. We use LETKF for both methods. See Hunt et al. 2007 [14] for an explanation of LETKF.

For the separate analysis method, we use LETKF without much modification. For the global analysis, at each grid point $n = 4m$ defined in the global model, we define a local patch $[n - s, n + s]$ of size $2s + 1$ with $s = 40$, use the Ensemble Transform Kalman Filter (ETKF) to obtain an analysis for the $(2s/4 + 1)$ state values in each patch. This yields local patch analyses for each ensemble member. As done by others, e.g., Hunt et al. 2007 [14], we then use these patch analyses to form the global analysis states for each ensemble member by defining the value of the global ensemble field at each point $m = n/4$ to be the analysis state value of

that ensemble member in the center of patch $n = 4m$. For the regional analysis, at each grid point n defined in the regional model, we define a local patch, limiting the size near the two boundaries of the subregion so that the local patch is defined only inside the subregion, use ETKF, and take the patch analysis value at grid point n. Thus the global local patches always have size $2s + 1$, but the regional local patches have variable sizes depending on n. For n located in the subregion and also far away from the boundaries, the regional local patch has size $2s + 1$, while for n near the boundaries ($n + s > n_1$ or $n - s < n_0$), the regional local patch is the intersection, $[n - s, n + s] \cap [n_0, n_1]$, and has a size less than $2s + 1$.

For the joint state method, we use the same local patch size, $s = 40$. For each grid point n defined either in the global model or in the regional model, we define a global local patch and a regional local patch (where, as before, the regional patch is the intersection, $[n - s, n + s] \cap [n_0, n_1]$, which for some $n = 4m$ will be empty). For each such grid point n, we define a vector $\mathbf{x}_g^{(n)}$ by taking state values of the global local patch, and $\mathbf{x}_r^{(n)}$ by taking state values of the regional local patch, and we then form a local joint state vector $\mathbf{x}^{(n)}$ by concatenating $\mathbf{x}_g^{(n)}$ and $\mathbf{x}_r^{(n)}$, i.e.,

$$\mathbf{x}^{(n)} = \begin{pmatrix} \mathbf{x}_g^{(n)} \\ \\ \mathbf{x}_r^{(n)} \end{pmatrix}. \tag{3.2}$$

We also form a local observation vector $\mathbf{y}_o^{(n)}$ by taking observations in the local patches (from grid point $n - s$ to $n + s$). We define a local cost function $J^{(n)}(\mathbf{x}^{(n)})$

for grid point n as follows,

$$
\begin{aligned}
J^{(n)}(\mathbf{x}^{(n)}) &= (\mathbf{x}^{(n)} - \bar{\mathbf{x}}_b^{(n)})^T (\mathsf{P}_b^{(n)})^{-1} (\mathbf{x}^{(n)} - \bar{\mathbf{x}}_b^{(n)}) \\
&+ \left[\mathbf{y}_o^{(n)} - \mathbf{H}^{(n)}(\mathbf{x}^{(n)})\right]^T \mathsf{R}^{-1} \left[\mathbf{y}_o^{(n)} - \mathbf{H}^{(n)}(\mathbf{x}^{(n)})\right] \\
&+ \kappa \left[\mathbf{G}_g^{(n)}(\mathbf{x}_g^{(n)}) - \mathbf{G}_r^{(n)}(\mathbf{x}_r^{(n)})\right]^T \left[\mathbf{G}_g^{(n)}(\mathbf{x}_g^{(n)}) - \mathbf{G}_r^{(n)}(\mathbf{x}_r^{(n)})\right],
\end{aligned}
\tag{3.3}
$$

where $\bar{\mathbf{x}}_b^{(n)}$ and $\mathsf{P}_b^{(n)}$ are the local mean and the covariance matrix of the background ensemble, respectively, and R is the covariance matrix of the observation noise. $\mathbf{H}^{(n)}(\mathbf{x}^{(n)})$ is a local observation operator defined as

$$
H_i^{(n)}(\mathbf{x}^{(n)}) = \begin{cases} (1-\lambda)\, x_{g,j(i)} + \lambda\, x_{r,j(i)}, & \text{if } n_0 \le j(i) \le n_1; \\ \\ x_{g,j(i)}, & \text{otherwise,} \end{cases}
\tag{3.4}
$$

where $j(i)$ is the observation location of the i^{th} observation in the local patch, $x_{g,j(i)}$ is the global state value at location $j(i)$, and $x_{r,j(i)}$ is the regional state value at location $j(i)$. $\mathbf{G}_g^{(n)}(\mathbf{x}_g^{(n)})$ is a vector that consists of the state values of the global state at the grid points defined both in the global and the regional local patches. Similarly, $\mathbf{G}_r^{(n)}(\mathbf{x}_r^{(n)})$ is a vector that consists of the state values of the regional state at the grid points defined both in the global and the regional local patches. κ and λ are parameters. The third term is a constraint term that penalizes large differences between the estimates of the global and regional model states. We determine the value of x that minimizes the cost function $J^{(n)}(\mathbf{x}^{(n)})$ with the LETKF algorithm Hunt et al. 2007 [14].

In general, if our technique were to be applied in an operational setting, the grid points of the global and the regional models within the subregion will not

coincide. In that case, to calculate the third term in $J^{(n)}(\mathbf{x}^{(n)})$, an interpolation from the grid points of the regional model to the grid points of the global model or vice versa could be employed before the values of the regional and the global models are subtracted. Similarly, in an operational setting the observations are not at grid points, and $\mathbf{H}^{(n)}$ would then include interpolation.

3.4 Model integration

We define a smoothed regional state for the initial condition of the regional model for integration between analysis times as follows. After the analysis phase, we define spatial transition intervals of length 10 starting from the boundaries and ending inside the subregion. We then modify the regional analysis values in the transition intervals by taking weighted linear averages of the global analysis values and the regional analysis values. We do this in order to make the transition between the global model and the regional model smooth at the boundaries. For n such that $0 \leq n < 10$, we modify the regional ensemble members by

$$X^r_{k,n_0+n} \leftarrow (n/10)\, X^r_{k,n_0+n} + (1 - n/10)\, X^g_{k,n_0+n}, \tag{3.5}$$

$$X^r_{k,n_1-n} \leftarrow (n/10)\, X^r_{k,n_1-n} + (1 - n/10)\, X^g_{k,n_1-n}, \tag{3.6}$$

where $X^g_{k,n}$ and $X^r_{k,n}$ are the values of the k^{th} global and regional ensemble members at grid point n, respectively, and the subregion for the regional model is $[n_0, n_1] = [240, 720]$.

After performing the above smoothing process, we integrate each global and regional ensemble members for 6 hours using a fourth-order Runge-Kutta method,

dividing 6 hours into 24 time steps. We integrate the global ensemble members independent of the regional ensemble members. For the integration of the regional ensemble members, we use the necessary interpolated values of the corresponding global ensemble members outside the subregion at each Runge-Kutta time step to synchronize the global and the regional model at the boundaries.

3.5 Results

Before we tested the joint state method and the separate analysis method, we ran forecast cycles with 40 ensemble members using the global and the regional models separately and found that multiplicative covariance inflation factors of 0.024 and 0.02 for the global and the regional analyses, respectively, produce the lowest rms state estimate errors. We henceforth use these values in our data assimilations. For the joint state method, we found that $\lambda = 0.9$ and $\kappa = 0.04$ in Eqs. (3.3) and (3.4) give the lowest rms state estimate errors, and we use these values in all of our subsequent applications of the joint state method.

We first tested the separate analysis method and the joint state method without boundaries. That is, we used the whole region for both the global and regional models. Thus, there is no coupling between the global model and the regional model at the boundaries during the integration phases. In this setup, aside from the correlations induced by common observations in their assimilations, the separate analysis method corresponds to having independent global and regional forecasts. For the joint state method, the coupling between the global and regional models occurs only

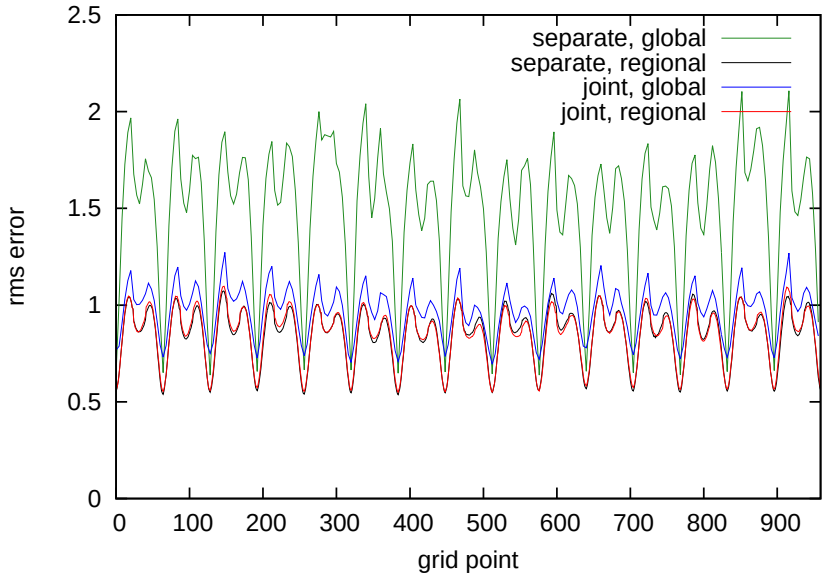

Figure 3.1: Rms errors of the state estimates of the separate analysis and the joint state analysis using the whole region for both the global and the regional models. The rms-error values were averaged over 10000 forecast cycles, discarding the values of 1000 initial cycles. The green and the black colors correspond to the global and the regional values obtained using the separate analysis method. The blue and the red colors correspond to the global and the regional values obtained using the joint state method.

at the analysis phases. Figure 3.1 shows the rms errors of state estimates given by the means of the ensemble members as a function of the grid point. The values were averaged over 10000 forecast cycles, discarding the values of 1000 initial cycles. The green and the black colors correspond to the global and the regional values obtained from the separate analysis method. The blue and the red colors correspond to the global and the regional values obtained from the joint state method. Error

minima occur at the observation points. The figure shows that the two regional rms errors are almost the same, while the global rms errors from the joint state method are much lower than the global rms errors of the separate analysis case indicating that, as one would expect, the information from the regional model substantially improved the estimate of the global model.

Now, we take a subregion $[n_0, n_1] = [240, 720]$, and introduce coupling between the global model and the regional model at the two boundaries during the integration phase. Figures 3.2(a) and 3.2(b) show the rms errors of the analysis and of a 1 day forecast, respectively, using the same color scheme as in Fig. 3.1. The two vertical dashed lines at grid points 240 and 720 indicate the boundaries of the subregion. The additional purple curves show the rms-error values in the perfect model scenario in which the forecast model was the true model (Lorenz model 3) which was used globally throughout the entire space. We view this as setting a standard for the best that could ever be done. These figures show that the joint state method performs better than the separate analysis method for both the global prediction and the regional prediction. We note that the global forecast obtained from the joint state method is better than the corresponding one from the separate analysis method even outside the subregion. This can be explained by the fact that the better global state estimates inside the subregion at the analysis phases can make better forecasts outside the subregion during the integration phases, and these better forecasts outside the subregion can make the regional forecasts better inside the subregion by providing better information at the boundaries during the integration phases. We also note that the global analysis improvements that result

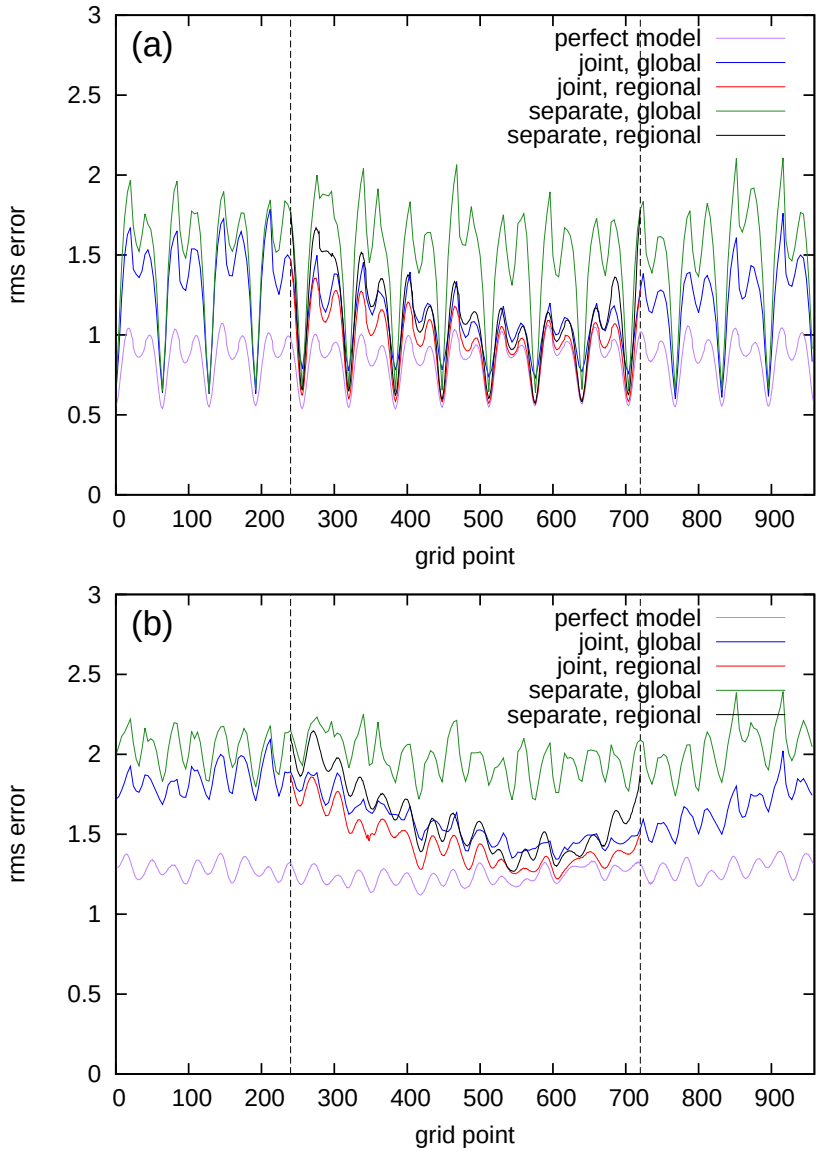

Figure 3.2: Rms errors of (a) the state estimates (b) 1 day forecasts of the separate analysis and the joint state analysis. The color scheme is the same as in Fig. 3.1. The additional purple curves show the rms-error values when assimilations were done globally using the true model (Lorenz model 3). The two vertical dashed lines at grid points 240 and 720 indicate the boundaries of the subregion.

from use of the joint state method are greater to the right of the subregion than to its left. This is consistent with the fact (Lorenz 2005 [18], Yoon et al. 2010 [28]) that, for these models, waves (and hence the information they carry) have group velocities that are predominantly rightward.

3.6 Discussion and conclusion

In this chapter we formulated a joint state method for regional forecasting. Using simulations employing simple models, we have numerically tested our method by comparing analysis and forecast results obtained using our method with results obtained using a separate analysis method. We found that the global forecast in the whole region and the regional forecast in the subregion are both noticeably improved when the joint state method is used compared to when the separate analysis method is used.

This work suggests several topics for future work. Most importantly, will the encouraging results from experiments using our Lorenz model set-up continue to apply when tests on real situations are done? What is the effect of regional model error? What are the benefits of applying our coupled analysis scheme to situations with multiple (perhaps overlapping) regional analyses?

Appendix A

Extracting an envelope

See [29] for the original introduction of the method presented here. Let us say that we have a function

$$f(n) = a(n)\cos[\,\phi(n)\,],\ 0 \le n < N, \tag{A.1}$$

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

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

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:

$$F(k) = \frac{1}{2}\,\mathrm{DFT}\{a(n)e^{i\phi(n)}\} + \frac{1}{2}\,\mathrm{DFT}\{a(n)e^{-i\phi(n)}\} \tag{A.3}$$

$$= \frac{1}{2N}A(k) * E_+(k) + \frac{1}{2N}A(k) * E_-(k) \tag{A.4}$$

$$\equiv \frac{1}{2N}\sum_{l=0}^{N-1} A(l)E_+(k-l) + \frac{1}{2N}\sum_{l=0}^{N-1} A(l)E_-(k-l), \tag{A.5}$$

where the asterisk denotes circular convolution. 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. A.1, 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, $\mathrm{DFT}_+^{-1}\{F(k)\}$, is equivalent to taking inverse discrete Fourier transform of $(2N)^{-1}A(k) * E_+(k)$. Therefore, we have the following:

$$\mathrm{DFT}_+^{-1}\{F(k)\} = \mathrm{DFT}^{-1}\left\{\frac{1}{2N}A(k) * E_+(k)\right\} \tag{A.6}$$

$$= \frac{1}{2}a(n)e^{i\phi(n)} \tag{A.7}$$

Thus,

$$a(n)e^{i\phi(n)} = 2\,\mathrm{DFT}_+^{-1}\{F(k)\} \tag{A.8}$$

Taking the absolute values of both sides, we obtain

$$a(n) = 2\,|\,\mathrm{DFT}_+^{-1}\{F(k)\}|. \tag{A.9}$$

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.

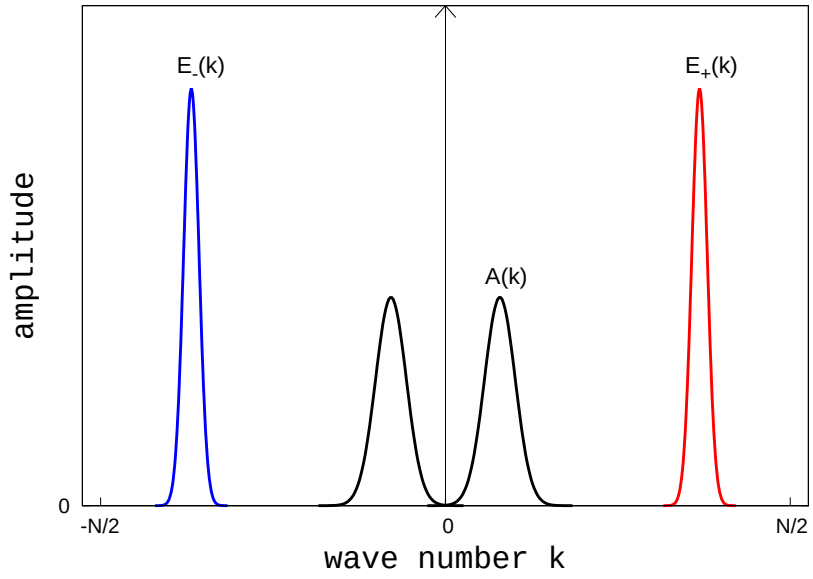

Figure A.1: DFT of $a(n)$, $e^{i\phi(n)}$, and $e^{-i\phi(n)}$. The terms $A(k)$, $E_+(k)$, and $E_-(k)$ are DFTs of $a(n)$, $e^{i\phi(n)}$, and $e^{-i\phi(n)}$, respectively, and $f(n) = a(n)\cos[\,\phi(n)\,]$, $0 \leq n < N$.

Appendix B

LETKF algorithm

We present a summary of the LETKF algorithm of [14] 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.

Input: a global background ensemble of m_g-dimensional model state vectors $\{\mathbf{x}_g^{b(i)}, i = 1, 2, \ldots, k\}$, an l_g-dimensional vector $\mathbf{y}_g^o$ of observations, a function $\mathbf{H}$ 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 R_g. Here the subscript g denotes global.

Output: a global analysis ensemble of m_g-dimensional model state vectors $\{\mathbf{x}_g^{a(i)}, i = 1, 2, \ldots, k\}$

1. Apply $\mathbf{H}$ 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 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 X_g^b.

For each grid point, do the following steps **3**–**8**.

3. Select the rows of $\bar{\mathbf{x}}_g^b$ and 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 X^b. Select the rows of $\bar{\mathbf{y}}_g^b$ and 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 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 R_g, forming the $l \times l$ matrix R.

4. Compute the $k \times l$ matrix $\mathsf{C} = (\mathsf{Y}^b)^T \mathsf{R}^{-1}$.

5. Compute the $k \times k$ matrix $\tilde{\mathsf{P}}^a = [(k-1)\mathsf{I}/\rho + \mathsf{C}\mathsf{Y}^b]^{-1}$, where $\rho > 1$ is a covariance inflation factor.

6. Compute the $k \times k$ matrix $\mathsf{W}^a = [(k-1)\tilde{\mathsf{P}}^a]^{1/2}$, where the power $1/2$ means the symmetric square root.

7. Compute the k-dimensional vector $\bar{\mathbf{w}}^a = \tilde{\mathsf{P}}^a \mathsf{C}(\mathbf{y}^o - \bar{\mathbf{y}}^b)$ and add it to each column of W^a, forming a $k \times k$ matrix whose columns are the analysis vectors $\{\mathbf{w}^{a(i)}\}$.

8. Multiply 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 $\{\mathbf{x}^{a(i)}\}$'s 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.

Bibliography

[1] Jefferey L. Anderson. An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, 129:2884–2903, 2001.

[2] Jefferey L. Anderson and Stephen L. Anderson. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, 127:2741–2758, 1999.

[3] Jeffrey L. Anderson. Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D*, 230:99–111, 2007.

[4] Craig H. Bishop, Brian J. Etherton, and Sharanya J. Majumdar. Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. *Mon. Wea. Rev.*, 129:420–436, 2001.

[5] Gerrit Burgers, Peter Jan van Leeuwen, and Geir Evensen. Analysis scheme in the ensemble kalman filter. *Mon. Wea. Rev.*, 126:1719–1924, 1998.

[6] J. G. Charney. On a physical basis for numerical prediction of large-scale motions in the atmosphere. *J. Meteor.*, 6:371–385, 1949.

[7] Robert E. Dickinson. Rossby waves – long-period oscillations of oceans and atmospheres. *Annu. Rev. Fluid Mech.*, 10:159–195, 1978.

[8] Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, 99(C5):10143–10162, 1994.

[9] Thomas M. Hamill, Jefferey S. Whitaker, and Chris Snyder. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, 129:2776–2790, 2001.

[10] P. L. Houtekamer and Herschel L. Mitchell. Data assimilation using an ensemble kalman filter technique. *Mon. Wea. Rev.*, 126:796–811, 1998.

[11] P. L. Houtekamer and Herschel L. Mitchell. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, 129:123–137, 2001.

[12] P. L. Houtekamer and Herschel L. Mitchell. Ensemble kalman filtering. *Q. J. R. Meteorol. Soc.*, 131:3269–3289, 2005.

[13] P. L. Houtekamer, Herschel L. Mitchell, Grard Pellerin, Mark Buehner, Martin Charron, Lubos Spacek, and Bjarne Hansen. Atmospheric data assimilation with an ensemble kalman filter: results with real observations. *Mon. Wea. Rev.*, 133:604–620, 2005.

[14] Brian R. Hunt, Eric J. Kostelich, and Istvan Szunyogh. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, 230(1-2):112–126, June 2007.

[15] Jeffrey D. Kepert. Covariance localisation and balance in an ensemble kalman filter. *Q. J. R. Meteorol. Soc.*, 135:1157–1176, 2009.

[16] Richard Kleeman. Information flow in ensemble weather predictions. *J. Atmos. Sci.*, 64:1005–1016, 2007.

[17] Edward N. Lorenz. Predictability - a problem partly solved. In *proc. seminar on predictability, vol 1*, pages 1–18, Reading, United Kingdom, 1996. ECMWF.

[18] Edward N. Lorenz. Designing chaotic models. *J. Atmos. Sci.*, 62:1574–1587, 2005.

[19] Edward N. Lorenz and Kerry A. Emanuel. Optimal sites for supplementary weather observations: simulation with a small model. *J. Atmos. Sci.*, 55:399–414, 1998.

[20] Dagmar Merkova, Istvan Szunyogh, and Edward Ott. Strategies for coupling global and limited-area ensemble Kalman filter assimilation. *Nonlin. Processes Geophys.*, 18:415–430, 2011.

[21] Edward Ott, Brian R. Hunt, Istvan Szunyogh, Aleksey V. Zimin, Eric J. Kostelich, Matteo Corazza, Eugenia Kalnay, D. J. Patil, and James A. Yorke. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, 56A:415–428, 2004.

[22] Anders Persson. Synoptic-dynamic diagnosis of medium range weather forecast systems. In *proc. seminar on diagnosis of models and data assimilation systems*, pages 123–137, Reading, United Kingdom, 2000. ECMWF.

[23] George W. Platzman. The Rossby wave. *Q. J. Roy. Meteorol. Soc.*, 94:225–248, 1968.

[24] Carl-Gustaf Rossby. Relation between variations in the intensity of the zonal circulation of the atmosphere and the displacements of the semi-permanent centers of action. *J. Mar. Res.*, 2:38–55, 1939.

[25] Istvan Szunyogh, Eric J. Kostelich, Gyorgyi Gyarmati, Eugenia Kalnay, Brian R. Hunt, Edward Ott, and Elizabeth Satterfield. A local ensemble transform kalman filter data assimilation system for the NCEP global model. *Tellus*, 60A:113–130, 2008.

[26] Istvan Szunyogh, Zoltan Toth, Aleksey V. Zimin, Sharanya J. Majumdar, and Anders Persson. Propagation of the effect of targeted observations: the 2000 winter storm reconnaissance program. *Mon. Wea. Rev.*, 130:1144–1165, 2002.

[27] Jeffrey S. Whitaker, Thomas M. Hamill, Xue Wei, Yucheng Song, and Zoltan Toth. Ensemble data assimilation with the ncep global forecast system. *Mon. Wea. Rev.*, 136:463–482, 2008.

[28] Young-noh Yoon, Edward Ott, and Istvan Szunyogh. On the propagation of information and the use of localization in ensemble kalman filtering. *J. Atmos. Sci.*, 67:3823–3834, 2010.

[29] Aleksey V. Zimin, Istvan Szunyogh, D. J. Patil, Brian R. Hunt, and Edward Ott. Extracting envelopes of Rossby wave packets. *Mon. Wea. Rev.*, 131(5):1011–1017, 2003.