A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts

JEFFREY L. ANDERSON

Geophysical Fluid Dynamics Laboratory, Princeton University, Princeton, New Jersey

STEPHEN L. ANDERSON

Metron, Inc., Reston, Virginia

(Manuscript received 28 August 1998, in final form 11 December 1998)

ABSTRACT

Knowledge of the probability distribution of initial conditions is central to almost all practical studies of predictability and to improvements in stochastic prediction of the atmosphere. Traditionally, data assimilation for atmospheric predictability or prediction experiments has attempted to find a single ''best'' estimate of the initial state. Additional information about the initial condition probability distribution is then obtained primarily through heuristic techniques that attempt to generate representative perturbations around the best estimate. However, a classical theory for generating an estimate of the complete probability distribution of an initial state given a set of observations exists. This nonlinear filtering theory can be applied to unify the data assimilation and ensemble generation problem and to produce superior estimates of the probability distribution of the initial state of the atmosphere (or ocean) on regional or global scales. A Monte Carlo implementation of the fully nonlinear filter has been developed and applied to several low-order models. The method is able to produce assimilations with small ensemble mean errors while also providing random samples of the initial condition probability distribution. The Monte Carlo method can be applied in models that traditionally require the application of initialization techniques without any explicit initialization. Initial application to larger models is promising, but a number of challenges remain before the method can be extended to large realistic forecast models.

1. Introduction

The production of ensemble forecasts of the state of the atmosphere has become common-place at the world's operational prediction centers during the past decade (Molteni et al. 1996; Tracton and Kalney 1993; Harrison et al. 1995). These ensemble forecasts are predicated on the notion that the state of the atmosphere as derived from all available observations is not known precisely, but can be represented in terms of a probability distribution. Operational ensemble forecast systems attempt to sample this initial state probability distribution and then produce samples of the resulting forecast probability distribution by integrating each individual member of the sample independently in a forecast model, usually a model developed for use in producing more traditional single discrete forecasts.

The operational prediction centers now routinely deliver a variety of ensemble-based products to forecasters. In general, reasonable interpretation of these products requires an assumption that individual ensemble forecasts are equally likely realizations of the forecast probability distribution. For example, the National Centers for Environmental Prediction produces a number of products that can be interpreted best in this context (Toth et al. 1997). The most obvious examples are charts of ''ensemble-based probability'' of a particular event.

One instance of such a chart would display a map of contours of the probability that precipitation amounts exceeding 1 mm will fall during a given time interval. These charts explicitly assume that the ensembles are random samples; the estimated probability of an event at a particular grid point is the number of ensemble members that produce the event divided by the total number of ensemble members.

Another example is the so-called spaghetti plot. An example of such a plot might display the 5640-m contour lines for 500-hPa heights from each member of the ensemble forecast. Although there is no longer an explicit assumption that the ensemble is a random sample of the probability distribution, such charts are nearly impossible to interpret if one does not make such an

Corresponding author address: Dr. Jeffrey L. Anderson, Geophysical Fluid Dynamics Laboratory, Princeton University, P.O. Box 308, Princeton, NJ 08542.
E-mail: jla@gfdl.gov

assumption. Both the published descriptions (Toth et al. 1997) of the spaghetti plots and discussions with producers and users of these products suggest an implicit assumption that the ensemble is a random sample of the forecast probability distribution.

A final example is the use of ensemble ''spread'' as a predictor of expected forecast skill. This is probably the most venerable of uses of forecast ensembles, and again, an explicit assumption that the ensemble is a random sample of the forecast distribution is required (Barker 1991; Houtekamer 1993; Kalnay and Dalcher 1987; Wobus and Kalnay 1995). Ensembles in which there is no notion of the relative likelihood of the different ensemble members (Brankovic et al. 1990; Buizza et al. 1993; Hoffman and Kalnay 1983; Toth and Kalnay 1993) cannot be interpreted easily for any of these applications of ensembles.

In what follows, a formal context for the stochastic prediction problem (Epstein 1969; Gleeson 1970), the nonlinear filtering/prediction problem, is described. In this context, the current and forecast states of the atmosphere are formally represented as probability distributions. Monte Carlo techniques are then applied to produce approximations to the solutions of the nonlinear filtering/prediction problem. These Monte Carlo techniques can be applied directly in low-order models; however, additional heuristic methods are required for application to large systems such as atmospheric forecast models. The result is a complete ensemble prediction system including an ensemble data assimilation that is designed to produce random samples of the initial state probability distribution, which can in turn be used to produce forecasts that are random samples of the forecast probability distribution. The ensemble mean of these analyses and forecasts should be competitive with analyses made by more traditional discrete data assimilation techniques.

Section 2 describes the nonlinear filtering/prediction problem. Section 3 develops Monte Carlo techniques to approximate the solution to the filtering/prediction problem. The result is an ensemble data assimilation/forecast system. Section 4 discusses techniques that can be used to evaluate performance of such a system. Section 5 presents results of the ensemble prediction system in a low-order, perfect model context; additional low-order results in section 6 address the system's ability to work in models that might require the application of ''initialization'' procedures to avoid spurious wave solutions. Section 7 discusses the extension of the method to realistic forecast models and observations, while section 8 presents conclusions.

2. The nonlinear filter

Observations of the atmosphere are sparse in both space and time and noisy. Traditional data assimilation systems for the atmosphere attempt to find a single representation of the atmospheric state that is the most likely given all the available observations (Parrish and Derber 1992). In the stochastic prediction problem, one seeks the probability distribution of the atmospheric state that is determined by all the available observations. Adopting a probabilistic approach, the state of the atmosphere, $\boldsymbol{\chi}_t$, at a particular time, t, has the conditional probability density function

$$\mathbf{p}(\boldsymbol{\chi}_t \mid \mathbf{Y}_t), \tag{1}$$

where $\mathbf{Y_t}$ is the set of all observations of the atmosphere that are taken at or before time t.

The rest of the development in this section is primarily drawn from the text of Jazwinski (1970). To proceed with the development of a nonlinear filter, it is necessary to introduce a discrete representation of the atmospheric state and stochastic model equations to represent the time evolution of this state:

$$d\mathbf{x}_t/dt = \mathbf{f}(\mathbf{x}_t, t) + \mathbf{G}(\mathbf{x}_t, t)\mathbf{w}_t. \tag{2}$$

Here, $\mathbf{x}_t$ is an n-dimensional vector that represents the state of the model system at time t, f is a deterministic forecast model, and $\mathbf{w}_t$ is a white Gaussian process of dimension r with mean 0 and covariance matrix $\mathbf{S}(t)$ while $\mathbf{G}$ is an $n \times r$ matrix. The second term on the right represents a stochastic component of the complete forecast model (2). For the time being, the stochastic term will be neglected; however, its inclusion is essential when one is attempting to apply a filter to data from a continuous system like the atmosphere. For the rest of this paper, the filter will be applied in a perfect model context in which

$$d\mathbf{x}_t/dt = \mathbf{f}(\mathbf{x}_t, t) \tag{3}$$

exactly represents the evolution of the system of interest.

Again, assume that a set of discrete observations, $\mathbf{Y}_t = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$, is available at time t. The observations are functions of the model state variables and include some observational error (noise), which is assumed to be Gaussian here (although the method developed can be extended to deal with non-Gaussian observational error distributions):

$$\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t, t) + \mathbf{v}_t. \tag{4}$$

Here, $\mathbf{h}$ is an m-vector function of the model state and time and $\mathbf{v}_t$ is an m-vector Gaussian observational noise; m, the number of observations, can itself vary with time. It is assumed below that the $\mathbf{v}_t$ for different times are uncorrelated. This is probably a reasonable assumption for many traditional ground-based observations although some newer types of observations, for instance, satellite radiances, may have significant temporal correlations in observational error.

As in the continuous case above, the conditional probability density of the model state at time t,

$$\mathbf{p}(\mathbf{x}_t \mid \mathbf{Y}_t), \tag{5}$$

is the complete solution to the filtering problem. The probability distribution (5) will be referred to as the *analysis probability distribution* or *initial condition probability distribution.* The forecast model (3) allows the computation of the conditional probability density at any time after the most recent observation time:

$$\mathbf{p}(\mathbf{x}_t\,|\,\mathbf{Y}_\tau) \qquad t > \tau. \tag{6}$$

This predicted conditional probability density can be used to make forecasts of the state of the model, or to provide the prior distribution at the time of the next available observations for the assimilation problem. The temporal evolution of this probability distribution is described by the Liouville equation as discussed in Ehrendorfer (1994). The probability distribution (6) will be referred to as the *first guess probability distribution* or *prior probability distribution* when it is being used to assimilate additional data, or the *forecast probability distribution* when a forecast is being made.

Assume that the conditional probability distribution (5) at time $t - 1$ is known and (6) is used to compute the distribution at time t, when a new observation $\mathbf{y}_t$ becomes available. The conditional distribution after making use of the new observation is

$$\mathbf{p}(\mathbf{x}_t\,|\,\mathbf{Y}_t) = \mathbf{p}(\mathbf{x}_t\,|\,\mathbf{y}_t,\ \mathbf{Y}_{t-1}). \tag{7}$$

Applying Bayes' rule gives

$$\mathbf{p}(\mathbf{x}_t\,|\,\mathbf{Y}_t) = \mathbf{p}(\mathbf{y}_t\,|\,\mathbf{x}_t,\ \mathbf{Y}_{t-1})\mathbf{p}(\mathbf{x}_t\,|\,\mathbf{Y}_{t-1})/\mathbf{p}(\mathbf{y}_t\,|\,\mathbf{Y}_{t-1}). \tag{8}$$

Since the observational noise $\mathbf{v}_k$ is assumed uncorrelated for different observational times, the observational error distribution at one time does not depend directly on the observational error at a previous time, so

$$\mathbf{p}(\mathbf{y}_t\,|\,\mathbf{x}_t,\ \mathbf{Y}_{t-1}) = \mathbf{p}(\mathbf{y}_t\,|\,\mathbf{x}_t). \tag{9}$$

It is also straightforward to compute the denominator in (8) as

$$\mathbf{p}(\mathbf{y}_t\,|\,\mathbf{Y}_{t-1}) = \int \mathbf{p}(\mathbf{y}_t\,|\,x)\mathbf{p}(\mathbf{x}\,|\,\mathbf{Y}_{t-1})\ d\mathbf{x}. \tag{10}$$

Incorporating (9) and (10) in (8) gives

$$\mathbf{p}(\mathbf{x}_t|\mathbf{Y}_t) = \mathbf{p}(\mathbf{y}_t|\mathbf{x}_t)\mathbf{p}(\mathbf{x}_t|\mathbf{Y}_{t-1}) \Big/ \int \mathbf{p}(\mathbf{y}_t|\boldsymbol{\xi})\mathbf{p}(\boldsymbol{\xi}|\mathbf{Y}_{t-1})\ d\boldsymbol{\xi}. \tag{11}$$

This equation expresses the way in which new observations are incorporated to modify the prior conditional probability distribution (first guess distribution) available from predictions based on earlier observations. The denominator is a normalization that guarantees that the total probability of all possible model states is 1. The numerator involves a product of two terms, the first representing new information from observations at time t and the second representing the prior constraints. The prior term gives the probability that a given model state, say $\mathbf{x}_t$, occurs at time t given information from all previous observations. The first term in the numerator of (11) then evaluates how likely it is that the observation $\mathbf{y}_t$ would be taken given that the state really was $\mathbf{x}_t$. If the relative probability of $\mathbf{y}_t$ being observed given that the true state is $\mathbf{x}_t$ is small, then the final probability of $\mathbf{x}_t$ being the truth is reduced. If the relative probability of $\mathbf{y}_t$ being observed given that the truth is $\mathbf{x}_t$ is high, then the final probability of $\mathbf{x}_t$ being the truth is increased. This computation is repeated for every possible value of the prior state in the (pointwise) product in the numerator of (11) to give the updated conditional distribution for the model state. This algorithm can be repeated recursively until the time of the latest observation, at which point (6) can be used to produce the forecast probability distribution at any desired time in the future.

3. Monte Carlo implementation of nonlinear filter

a. Representing the conditional distribution

Equations (2) and (11) of the previous section, in concert with some representation of the probability distribution at an initial time $\mathbf{p}(\mathbf{x}_0)$, define a nonlinear filter that can be used to assimilate data in atmospheric prediction models. However, a numerical implementation of the filter necessitates some discrete representation of probability distributions for the model state. The most straightforward approach to the problem is to assume that the probability distributions are approximately Gaussian, representing the probability distribution of a k-dimensional state by a k-vector of means and a $k \times k$ covariance matrix. As discussed below, using a Gaussian representation for the model state conditional probability distribution leads to data assimilation algorithms similar to the Kalman filter.

One could also choose to represent the conditional probability distribution in terms of the mean, covariance, and additional higher-order moments. While this might lead to increased accuracy, it is computationally challenging to perform the pointwise product in (11) using representations of this sort.

A fundamentally different approach, using a finite random sample of the conditional probability distribution as a discrete representation of the continuous distribution, is employed in the assimilation algorithm developed here. Such Monte Carlo algorithms can have a number of nice properties that allow the computation of approximate solutions to problems that may be intractable by other methods. Here, the Monte Carlo approach has a number of advantages. The most apparent is the ability of the method to represent probability distributions with non-Gaussian behavior. In addition, applying Monte Carlo solutions in conjunction with heuristic simplifications can make the filtering problem tractable in very large models.

b. Monte Carlo sample of prior distribution

The basic filtering algorithm derived in the preceding section is composed of two parts: advancing a conditional probability distribution from the end of one observation time to another to obtain the prior (first guess) probability density [Eq. (2)] and doing the pointwise product to incorporate the impact of new observations [Eq. (11)]. Mechanisms for implementing both operations are requisite for the Monte Carlo approach.

Advancing a conditional distribution in time is straightforward. Suppose one has a random sample, $\mathbf{x}_i$, $i = 1, \ldots, n$ of some (possibly vector) random variable $\mathbf{X}$ and a (possibly vector) function, $\mathbf{F}$. Then, if $\mathbf{Y} = \mathbf{F}(\mathbf{X})$, the set $\mathbf{y}_i = \mathbf{F}(\mathbf{x}_i)$ is a random sample of the random variable $\mathbf{Y}$ for all reasonable functions $\mathbf{F}$ (Cramer 1966). For example, $\mathbf{F}$ could be a forecast model in the form (2) or (3). Then, given a random sample of the model state conditional probability $\mathbf{p}(\mathbf{x}_t \mid \mathbf{Y}_t)$ at time t, one can generate a random sample of the prior distribution $\mathbf{p}(\mathbf{x}_{t+} \mid \mathbf{Y}_t)$ at a later time $t+$ by integrating each member of the sample in the forecast model. Similarly, a random sample of the forecast probability distribution can also be generated by integrating an ensemble of initial conditions that are a random sample of the conditional probability distribution at the time of the most recent observation.

Another interesting case is when the function $\mathbf{F}$ is a scalar function of $\mathbf{X}$, for instance, a function that returns the value of a single vector component, $\mathbf{x}_i$, of $\mathbf{X}$ or some more complex scalar quantity, say the rainfall interpolated to the location of a particular observing station. Here $\mathbf{F}$ could also be a composite function that returns a scalar function of a forecast field, or $\mathbf{F}$ could be the observational operator $\mathbf{h}$ in (4). In any of these cases, if one has a random sample of the model state conditional probability, applying $\mathbf{F}$ individually to the samples results in a random sample of the function.

c. Monte Carlo implementation of the pointwise product

The second step in the assimilation algorithm is to compute the product of the prior probability distribution with the observational error distribution. In the Monte Carlo implementation, the prior is represented by a random sample. For now, it is assumed that the observational error distribution is Gaussian, although it is possible to relax this constraint in a variant of the method being described. To simplify the development of the Monte Carlo algorithm, it is also assumed that the observational operator, $\mathbf{h}$ in (4), is the identity, that is, observations are available for all state variables at every observation time. Methods for relaxing this assumption are discussed in section 7. Given these assumptions, a pointwise product of a random sample and a Gaussian must be computed, resulting in a random sample of the product.

1) GAUSSIAN APPROXIMATION

One can proceed by using the Monte Carlo sample of the prior to construct a continuous approximation to the prior probability distribution. As noted in appendix A, the pointwise product of a pair of Gaussians is another Gaussian, and simple formulas exist for computing the mean, covariance, and area under the product. Given this, one way to compute the pointwise product is to compute a Gaussian distribution that is the best fit to the random sample

$$\mathbf{P}(\mathbf{x}) \approx \mathbf{N}(\mu, \Sigma), \tag{12}$$

where μ and Σ are the sample mean and covariance of the prior, respectively.

The mean and covariance of the pointwise product can then be computed by a single application of (A.2)–(A.3), and a random sample of this product can be created by standard methods. The denominator of (11) is simply a normalization factor and hence has no impact on the selection of the random sample of the product. A Monte Carlo method that uses this single Gaussian approximation of the prior probability distribution when performing the pointwise product will be referred to as a Gaussian filter in the following.

2) KERNEL APPROXIMATION

Using a Gaussian to represent the prior probability distribution for the pointwise product can be viewed as partially eliminating one of the fundamental advantages of a Monte Carlo approach, namely, the ability to represent non-Gaussian probability distributions. Instead of using a single Gaussian as a continuous representation of the prior probability distribution, one can use a standard kernel technique (Silverman 1986) in which a sum of Gaussian kernels is used to form a continuous representation from a random sample.

For reasons elaborated below, the method of Fukunaga (1972) is used to select the Gaussian kernels that are summed to form a continuous approximation to the random sample. For an n-member sample, a set of n Gaussian kernels is used,

$$\mathbf{P}(\mathbf{x}) \approx \sum_{i=1}^{n} \mathbf{K}_i(\mathbf{x}), \tag{13}$$

$$\mathbf{K}_i(\mathbf{x}) = \mathbf{N}(\mu_i, \alpha\Sigma), \tag{14}$$

where the mean of the ith kernel μ_i is the value of the ith member of the random sample and the covariance of all the kernels is the same, a constant factor times the covariance matrix Σ that would result if a single Gaussian were used as in (12). Figure 1 shows a schematic representation of the kernels and their relation to the single Gaussian approximation. A number of methods for computing the constant covariance reduction factor, α, have been developed (Silverman 1986). However, as noted in section 3e, the value of α is subsumed into a tuning constant and so does not need to be calculated explicitly.

With (13) used to generate a continuous approximation to the prior conditional probability distribution,

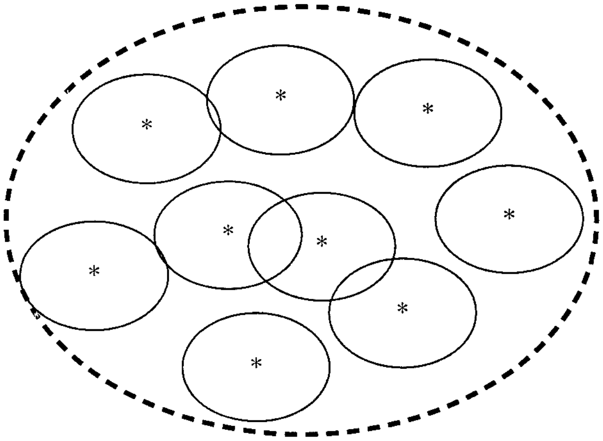

FIG. 1. Schematic representation of the kernels (solid) and their relation to the single Gaussian approximation (dashed).

a random sample of the pointwise product must be generated. Applying the distributive law, a continuous representation of the pointwise product probability distribution is

$$\mathbf{C}(\mathbf{x}) = \sum_{i=1}^{n} K_i(\mathbf{x})\mathbf{p}(\mathbf{y}\,|\,\mathbf{x}) = \sum_{i=1}^{n} c_i\mathbf{N}(\nu_i, \Sigma_{\text{new}}). \quad (15)$$

In (15), the product is computed by applying the formulas in appendix A to each Gaussian kernel in the summation for $\mathbf{P}(\mathbf{x})$ in turn. Each individual pointwise product is characterized by a mean, ν_i, an area, c_i, and a covariance, Σ_{new}; the covariance is the same for all the individual products since each kernel in $\mathbf{P}(\mathbf{x})$ has the same covariance and the product's covariance (A.2) depends only on the individual covariances. This fact also greatly reduces the work required to compute the n individual pointwise products.

Once the ν_i, c_i and Σ_{new} are computed, a new n-member random sample of $\mathbf{C}(\mathbf{x})$ must be generated. This is easily accomplished by noting that the c_i's define the relative probability that a sample should be taken from the ith member of the product sum. Selecting a random sample of the $\mathbf{C}(\mathbf{x})$ consists of repeating the following two steps n times. First, randomly select a kernel with probability

$$p_i = c_i \bigg/ \sum_{i=1}^{n} c_i. \quad (16)$$

Then, select a random sample from the Gaussian for the selected kernel using standard methods. This method is referred to as a kernel filter in the following.

This kernel filter can have distinct advantages over the Gaussian filter if the prior conditional probability distribution (or the product itself) has a distinctly non-Gaussian structure. Figure 2 shows a schematic demonstrating the additional power of the kernel method when assimilating observations from a low-order dynamical system. In this example, the attractor of the dynamical system being assimilated is confined to a limited region of the model's phase space. When a single Gaussian is used to represent the prior conditional distribution, the result can be a representation that has the majority of its probability density in regions of phase space that are never visited by the model. The kernel filter's ability to represent non-Gaussian prior distributions allows more of the density of the pointwise product to be placed in areas of phase space that are consistent with the model's dynamics. Comparisons of the single Gaussian and kernel methods are presented in more detail in sections 5 and 6.

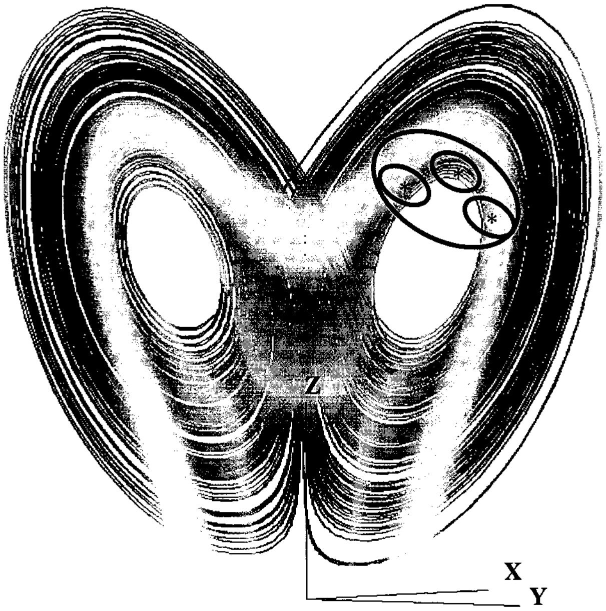

FIG. 2. Schematic representation of advantages of kernel over single Gaussian filter for a low-order model. The background is a projection of a trajectory from the Lorenz-63 model showing the attractor structure. Superimposed is an idealized image of the single Gaussian (outer curve) and kernel (inner curves) prior distributions for a three-member ensemble.

d. Comparison to other algorithms

A number of other approaches to solving the filtering problem (11) exist or could be developed. Perhaps the best known approximation for solving (11) is the Kalman filter. The Kalman filter assumes that the prior conditional probability distribution is Gaussian and represents this distribution explicitly in terms of its mean and covariance. Equation (3) is used to advance the mean of the prior distribution in time. The covariance matrix is advanced in time by linearizing (3) around the mean trajectory in phase space and applying the resulting linear operator to the covariance matrix. The pointwise product step can then be performed using the equations of appendix A, the result being the new mean and covariance for the state estimate. Miller et al. (1994) examine the application of a number of variants of the Kalman filter to the Lorenz-63 system.

The Gaussian filter described above is similar to the Kalman filter in that the prior conditional probability distribution is represented as a Gaussian. However, the Monte Carlo method does not require a linearized version of the model; it is also computationally cheaper in many cases since it requires only *n* integrations of the nonlinear model. On the other hand, if the ensemble size is smaller than the model phase space (as is always the case in realistic model problems of interest), the Monte Carlo method can at best sample the covariance of the prior distribution in an $n - 1$-dimensional subspace of the model phase space. Results later in section 7 discuss the efficacy of sampling only a subspace. In general, the Monte Carlo method naturally samples those directions in phase space with the most rapid perturbation growth, which can be argued to be the most important directions for the filtering problem.

A number of groups have combined traditional discrete data assimilation algorithms with ensemble forecasts. One method, the Observing System Simulation Experiment–Monte Carlo (OSSE–MC) is described in Houtekamer et al. (1996) and Houtekamer and Derome (1995). In these cases, a set of perturbations around the single discrete assimilated state is integrated in the nonlinear model. When new observations are available, this sample is converted to a Gaussian as in the Gaussian filter algorithm here and the pointwise product computed. A new sample is then selected from the product's probability distribution. However, the OSSE–MC method requires some independent assimilation mechanism to get the single control mean state. This mean estimate may be fundamentally inconsistent with the probability distribution being approximated by the ensemble of perturbations. In addition, constraining the perturbations to be distributed around the control implies that the ensemble does not represent a random sample of the forecast probability distribution (6). This is especially true when ensemble perturbations are required to be symmetric around the single control assimilation (Houtekamer et al. 1996) but is also true even without this additional constraint.

A method known as the ensemble Kalman filter (Evensen 1994; Evensen and Van Leeuwen 1996; Van Leeuwen and Evensen 1996; Evensen 1997) is functionally quite similar to the Gaussian filter developed here. The ensemble Kalman filter uses an ensemble to represent the covariance statistics of the assimilated state. The ensemble is integrated in the nonlinear model to get a sample of the prior distribution at the time of the next observation. At that point, a Gaussian is fit to the ensemble to get the covariance matrix of the prior distribution as in the Gaussian filter and the pointwise product is then performed in the context of the Kalman filter. A number of variants of this procedure are discussed in Evensen (1994). The results of applying the ensemble Kalman filter are probably quite similar to those obtained with the Gaussian filter discussed here. However, developing a partially nonlinear Monte Carlo method in the context of the linearized Kalman filter leads to a great deal of extra complication in the application and discussion of this method. It seems more natural to proceed directly to apply Monte Carlo methods to the fully nonlinear filter as is done here.

The kernel filter has all the advantages of the Gaussian filter but is potentially much more powerful. This method is somewhat similar to performing an ensemble of Kalman filters, and one could attempt to apply a kernel approach when doing the pointwise product in ensemble Kalman filter methods. Again, it seems much more natural to develop the kernel filter in the context of the fully nonlinear filter.

There are a variety of other variants of Monte Carlo techniques that could be applied to the filtering problem. A method that has received application in a variety of problems is the classical weighted Monte Carlo method with resampling. In algorithms of this class, the members of the random sample of the prior are assigned a weight at the time of the pointwise product corresponding to the relative probability given by (11). The same sample of points is then integrated to the next time when data become available and the procedure is repeated. Periodically, a resampling method similar to the basic Monte Carlo algorithm described above can be applied to generate a new set of points. The kernel method of section c above could be trivially modified to incorporate the weighting/resampling approach. Preliminary results have shown that when applying the method in low-dimensional models, this can be computationally somewhat more efficient. It can also help to reduce problems associated with the need for initialization (see section 6). However, when applied to higher-order models, the weighting method tends to be less effective because of the *empty space phenomenon* (e.g., in a 10-dimensional normal 99% of the mass of the distribution is at points more distant than 1.6 from the mean; Silverman 1986) and related difficulties with sampling probability densities in high-dimensional space. Future studies may further explore the efficacy of applying weighting/periodic resampling in the context of the Monte Carlo filter.

e. *Tuning the filter*

One of the common difficulties experienced when applying a variety of filtering techniques is filter divergence (Jazwinski 1970) in which the distribution produced by the filter drifts away from the truth. This normally occurs because the prior distribution becomes too narrow and the observations have progressively less impact on the pointwise product until the observations become essentially irrelevant. The most common approach to dealing with filter divergence is to add some (white) noise to the prior distribution to ‘‘broaden’’ this distribution and enhance the impact of the observations in the product.

The Gaussian and kernel filters described above are not immune to the filter divergence problem. In general, if the Gaussian filter is applied directly as described, it eventually diverges resulting in an increasingly tight prior distribution that is not influenced by new observations. A direct addition of noise to address the filter divergence may have undesirable consequences when applied in models of the type used to do atmospheric prediction (see discussion in section 6).

Filter divergence can also be fixed by a variety of other methods that attempt to avoid the unwarranted tightening of the prior distribution. Here, a particularly simple approach is taken: the covariance matrix, Σ, computed for the prior (12) is multiplied by a factor γ ($\gamma > 1$). By broadening the prior distribution artificially in this fashion, the divergence problem can be avoided while the implied prior distribution tends to remain on the local attractor (Anderson 1994, 1997). Obviously, making γ too large results in a filter in which the observations are given too much weight, so γ must be chosen with care. In general, the only viable method for choosing γ is experimentation. In the perfect model results presented in later sections, γ is chosen by trial and error to give an assimilation with the most favorable statistics. Tuning a filter for a real system is complicated by the limited number of observations, the lack of knowledge of the true state, and the presence of systematic model errors; all assimilation techniques have to deal with this same problem.

As noted above, when using the kernel filter, the sample covariance matrix computed from the prior distribution must be multiplied by a factor, α, to generate kernels of the appropriate size. While a number of heuristic methods for generating α exist (Silverman 1986), they should be regarded as at best rough approximations of the optimal value. When applying the kernel method, α can be computed in the same heuristic fashion as for γ in the preceding paragraph, resulting in a filter that does not experience filter divergence while giving a good representation of the model state distribution.

4. Evaluating ensemble analyses and forecasts for perfect model experiments

Before presenting some sample applications of the Monte Carlo filters, it is necessary to develop some tools for assessing the quality of analyses and forecasts. As noted in the introduction, the goal of the techniques developed here is to produce a random sample of the analysis probability distribution (5) and the forecast probability distribution (6). Since no analytic representation of these probability distributions is available, even in simple model applications with straightforward observational error covariances, it is a challenging problem to determine how well ensembles produced by a particular algorithm satisfy this criterion.

a. Partitioning by order statistics

Equation (5) describes all that can possibly be known about the truth given the available observations. If the filter worked perfectly, then the true state should be statistically indistinguishable from the random samples of the analysis distribution produced by the filter. Therefore, a necessary condition for a perfect assimilation is that the true state of the system and the random sample produced by the analysis are drawn from the same distribution; this condition is referred to as consistency between the ensemble analysis and the truth in what follows.

It is impossible to test this condition given a single sample of the truth and the analysis ensemble at a single time; it is also difficult to evaluate this condition for probability distributions in high-dimensional vector spaces. Anderson (1996b) described a technique, known as binning or Talagrand diagrams, that uses samples of scalar functions of the truth and the corresponding analysis ensemble at many different times to check for consistency between the truth and the analysis distribution. At each analysis time, this technique uses the analysis ensemble of a scalar quantity to partition the real line into $n + 1$ intervals (bins); the truth at the corresponding time falls into one of these $n + 1$ bins. As shown in Anderson (1996b), a necessary condition for the analysis ensemble to be a random sample of (5) is that the distribution of the truth into the $n + 1$ bins be uniform. In what follows, this is evaluated by applying a standard chi-square test to the distribution of the truth in the $n + 1$ bins. The null hypothesis here is that the truth and the analysis ensemble are drawn from the same distribution. Obviously, for large enough samples, the assimilation and the truth should be distinct since there are many approximations in the filtering algorithm. A very rough measure of the quality of the assimilation can be obtained by noting how large a sample must be used to demonstrate that the truth and assimilation are significantly different.

The binning technique can be applied only to scalars. However, one would like to have some information about higher-dimensional aspects of the analysis probability distribution. One very simple tool to evaluate this involves the ratio of the time-averaged rms error of the ensemble mean to the time-averaged mean rms error of the individual ensemble members. As shown by Murphy (1988, 1990), this ratio should be

$$R = \sqrt{(N + 1)/2N}$$

if the truth is statistically indistinguishable from a member of the analysis ensemble.

b. Minimizing rms

Consistency between the analysis ensemble and the truth is only a necessary condition for the truth and assimilation to be samples from the same distribution.

For instance, an ensemble analysis comprised of a random sample from a long model climatological run would be consistent with the truth. It is also necessary to minimize some measure of the average error between the analysis ensemble mean and the truth. The average rms difference between the ensemble mean and the truth is used here (Leith 1974).

5. Low-order model results: Lorenz-63 model

Tests in low-order models are essential to understanding the behavior of ensemble data assimilation and forecasting methods since the behavior of realistic forecast models is far too complicated to be readily understood. Here, results are first presented for the Lorenz-63 model (Lorenz 1963) (appendix B), which is a set of three coupled nonlinear partial differential equations in three variables (B1)–(B3). This model is chosen for a number of reasons: its low integration cost allows large numbers of comprehensive tests with large sample sizes, it is chaotic in its continuous form and has large sensitivity to initial conditions in its discretized form, it has an attractor with unusually simple structure, and it has been used in many previous studies of data assimilation and ensemble prediction (Anderson 1996a; Palmer 1993; Anderson and Hubeny 1997; Evensen 1997).

All results presented in this section are from perfect model experiments in which a very long control run of the Lorenz-63 model is assumed to represent the truth. Observations of the truth with a specified observational error are generated periodically by adding a random sample of a prescribed observational error distribution to the truth. Monte Carlo filters are then applied to these observations to produce random samples of the analysis probability distribution and forecast probability distributions for a range of forecast lead times. This perfect model framework has two advantages. First, it allows the verification of the assimilation and forecasts against the truth, which can never be known for a real system. Second, it eliminates the difficulties of dealing with systematic model error, which complicate the assimilation algorithm. Issues related to systematic error are discussed in section 7.

In the experiments in this section, the observational error distribution for a given experiment is fixed in time. To start the assimilation, an initial sample of the model state probability distribution is needed. Here, each member of the ensemble at the initial time is generated by adding an independent random sample of the observational error distribution to the truth. The first 5000 model time steps of the assimilation are discarded to avoid direct impacts of the early spinup portion of the assimilation.

a. Tuning the filter

As noted in section 3, it is necessary to select a value for the covariance inflation factors (γ for the single Gaussian representation and α for the kernel representation) in order to avoid filter divergence while producing the best assimilation and forecasts possible. Figure 3 shows the impact of varying the covariance inflation factor, α, for a 40-member ensemble assimilation of the Lorenz-63 model using a kernel filter. In this case, observations are available every 50th nondimensional time step with an observational error that has a standard deviation of 2.0 for each of *x, y,* and *z* with no covariance in the observational error. As noted in section 4, the goal of the assimilation is to produce a random sample of the conditional probability distribution that is consistent with the truth while minimizing the rms error of the ensemble mean from the truth. For values of α less

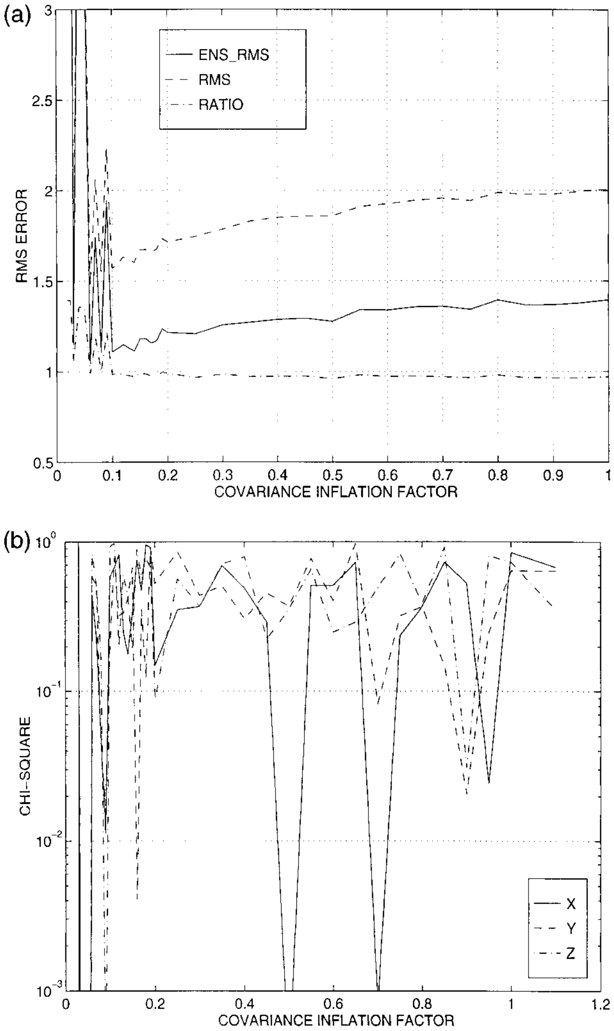

FIG. 3. Impact of covariance inflation factor, α, on the performance of a 40-member kernel filter assimilation applied to the Lorenz-63 model with observational error standard deviation 2.0 on each variable independently and observations available every 50 time steps. (a) The rms of the ensemble mean (solid), mean rms of the ensemble members (dashed), and the normalized ratio of the two (dash–dotted). (b) The significance of the chi-square statistic for the *x* (solid), *y* (dashed), and *z* (dash–dotted) variables.

TABLE 1. Mean rms error of the ensemble mean as a function of the observational interval for 40-member kernel and Gaussian filter assimilations of the Lorenz-63 system with observational error standard deviation 2.0 on each variable independently. Results are for time steps 5000–15 000 of a long assimilation run.

Observational interval	Kernel filter ensemble mean rms error	Gaussian filter ensemble mean rms error
1	0.365	0.348
10	0.600	0.666
20	0.805	0.940
50	1.04	1.28
100	1.51	1.70

than about 0.1, the filter may diverge completely from the truth and the rms error of the assimilation has approximately the same value as the difference between two randomly chosen states from a long integration (climate) of the model. Figure 3a shows that as α is increased, the problem of filter divergence begins to decrease. As α is increased past 0.1, the mean of the rms error and the ensemble mean error have minima at α approximately 0.15. As α is further increased, the rms error increases again as the assimilation begins to use less information from the prior distribution. For very large α, the rms error would be essentially the same as one would get by assuming that the model state distribution at observation times is equivalent to the observational distribution.

Figure 3b plots the significance of the chi-square statistic for the order statistic binning of x, y, and z as a function of α. For small α, the filter diverges and the truth almost always lies in one of the two outer bins, so the chi-square significance is very small. As α is increased, the value of chi-square significance increases until the test is unable to distinguish the distribution of the bins from uniform with the sample size (200 observation times) available. All three chi-square values are generally above the 10% confidence range for values of α greater than 0.1. The significance should be less than 10% one-tenth of the time by chance, even if the null hypothesis is true.

Figure 3a also shows the ratio of the mean of rms error of the ensemble mean to the individual rms errors from the ensemble, normalized by the expected value of approximately 0.716 for a 40-member ensemble. For small α, the ratio is much larger than 0.716. As α increases, the ratio decreases, passing through the value 0.716 for values of α for which the rms is slightly larger than its minimum. Since the chi-square statistics suggest that this is also a value of α for which the ensemble assimilation is relatively consistent with the truth, it would appear that values of α approximately 0.15 are most appropriate for this application. A similar tuning is undertaken for all the filters used throughout the rest of this study.

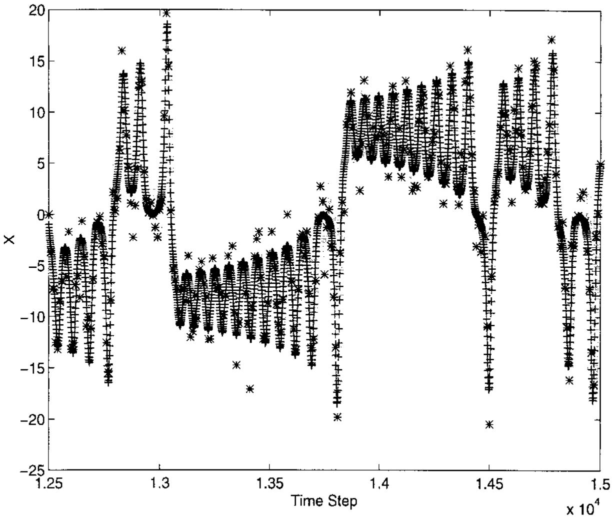

FIG. 4. Time sequence of values of x variable for truth (dark +), observed (dark *), and 10 members of the ensemble simulation (light +) from a 40-member kernel filter assimilation of the Lorenz-63 model with observations available every 10 steps with observational error standard deviation 2.0 on each variable independently.

b. *Kernel assimilation results for Lorenz-63*

The abilities of the kernel filter can be highlighted by examining the results of assimilations for the Lorenz-63 model. Table 1 shows the ensemble mean rms error of the assimilated ensemble as a function of observational frequency for 40-member ensembles. For frequent observations, the prior distributions are nearly Gaussian and the rms errors are relatively small. As an example, Fig. 4 shows the true, observed, and assimilated ensemble values for the x-variable of the Lorenz-63 system for observations available every 10 steps. The true trajectory demonstrates typical behavior for the Lorenz-63 system, switching from one attractor lobe (x positive) to the other (x negative) after some number of orbits. The assimilated ensemble (only 10 of the 40 members are shown to avoid even more clutter) generally tracks quite tightly along the true trajectory. Occasionally, the ensemble set spreads out, indicating that there is more uncertainty about the assimilated state. In some cases, some members of the ensemble follow trajectories into the other lobe of the attractor, for instance, around time step 13 000 or 13 750. At times like this, the prior distribution is bimodal and clearly not Gaussian.

As a control for the assimilation results, Fig. 5 shows the same sequence except that the assimilation is switched off at time 12 750. For the first orbit of the attractor lobe, all members of the ensemble (now technically a forecast) remain in the proper lobe. After the first orbit, the truth switches to the other lobe, but a few members of the ensemble fail to do so; even more ensemble members switch back to the wrong lobe after the next orbit. By about step 13 000, the ensemble members are nearly randomly distributed on the attractor.

Figure 6 shows the true, observed, and assimilated

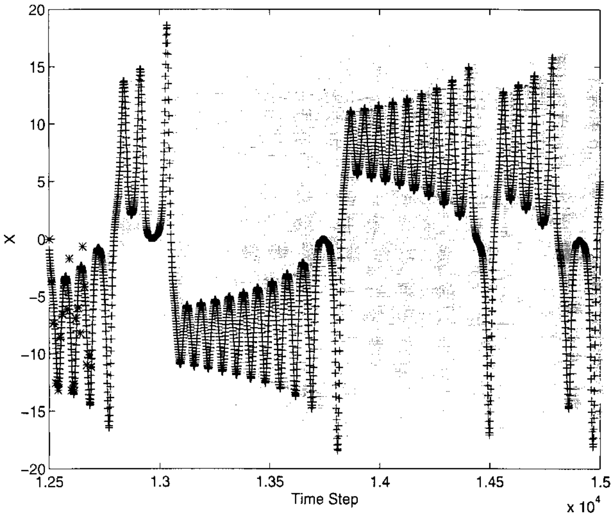

FIG. 5. Same as Fig. 4 except that all assimilation is stopped at time step 12 750 and the ensemble is allowed to evolve freely.

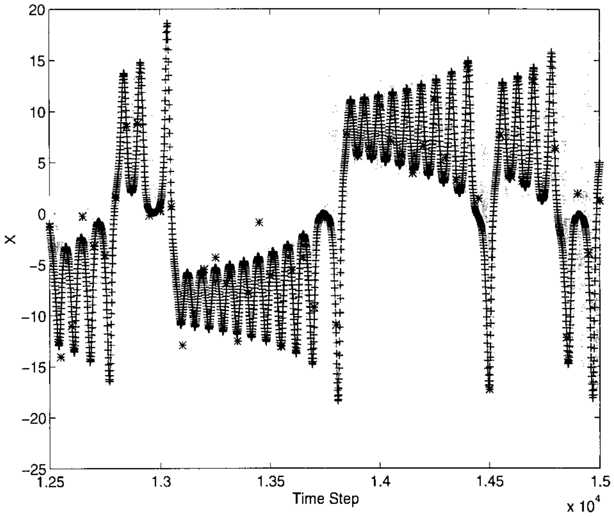

FIG. 6. Same as Fig. 4 except with observations available every 50 steps.

ensemble values for the *x*-variable for observations available only every 50 steps. The assimilation is clearly providing information about the truth, but the spread is much greater than for observations every 10 steps (Fig. 4). There are also a greatly increased number of incidents in which some of the ensemble trajectories move into the wrong lobe of the attractor.

For assimilations where the observations are available relatively infrequently, the prior distributions and even the assimilated distributions can become highly non-Gaussian. Figure 7 shows the probability density for the *x*-variable of the prior distribution at time step 12 750 of the assimilation shown in Fig. 6 (this continuous density plot is generated using the same kernel method that is used in the pointwise product for the assimilation). This prior distribution is significantly bimodal and is arguably trimodal, with about 25 of the ensemble points in two peaks in the positive *x* attractor lobe and the remaining 15 well separated in the other lobe. While one should be suspicious that a trimodal distribution with just 40 samples is simply a result of undersmoothing the data with the kernel summation, in this case, there is a physical reason to accept the trimodality. The peak for negative *x* is due to ensemble members that are in the negative *x* attractor lobe. The upper peak for positive *x* in Fig. 7 comes from a set of ensemble members that moved into the positive *x* attractor lobe at approximately time step 12 650. The lower positive *x* peak corresponds to points that moved into the positive lobe one orbit later at about time step 12 700. The ability of the kernel filter to resolve non-Gaussian behavior in both the prior and assimilated distributions appears to be of importance in this type of dynamical system with fairly infrequent observations (for a particular observational error distribution).

The kernel filter assimilation is able to perform well even when observations are extremely infrequent or when observational error is very large, situations that are generally challenging for more traditional methods like the Kalman filter. Figure 8 displays results for an assimilation with observations every 10 steps, but with an observational error standard deviation of 10.0 for *x*, *y*, and *z*. The size of the observational error relative to the variations of the model can be seen clearly in the figures. Despite the relatively low quality observations, the assimilation does not diverge from the truth. Figure 8a shows that the spread of the ensemble is generally large and occasionally the ensemble is spread out over almost the entire attractor. Despite this, the ensemble mean generally stays quite close to the truth as shown in Fig. 8b. On the occasions when the ensemble mean does diverge from the truth, for instance, around time

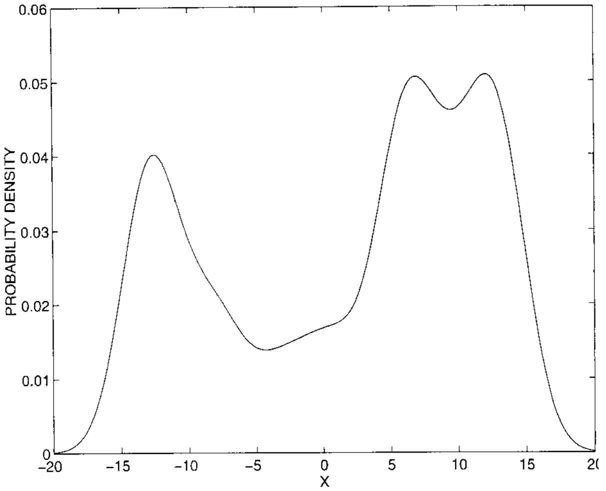

FIG. 7. Probability density distribution of *x* for the prior distribution at time step 12 750 for the same assimilation experiment as in Fig. 6.

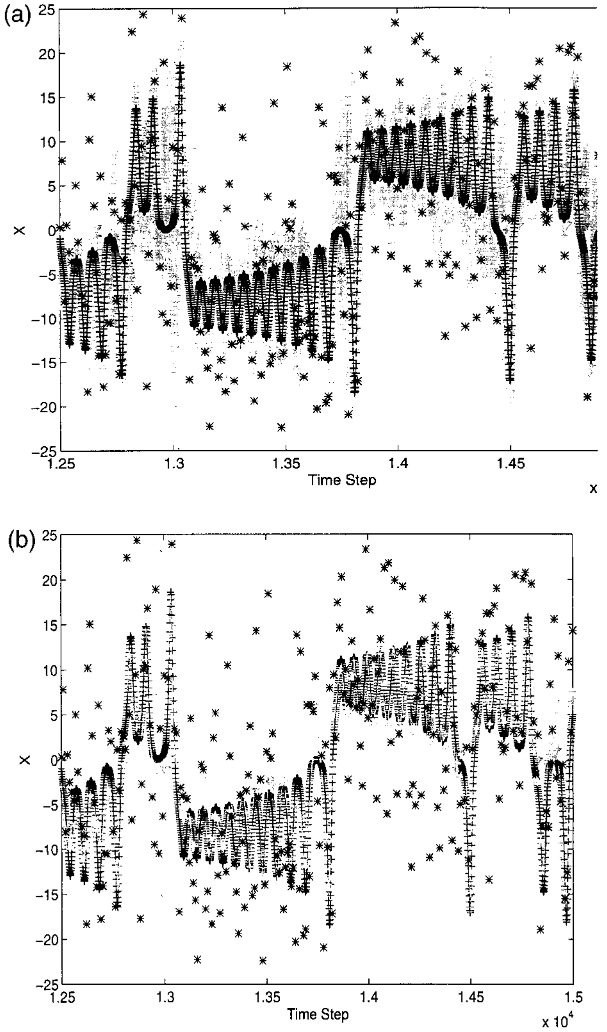

FIG. 8. (a) Same as Fig. 4 except with observational error standard deviation 10.0 on each variable independently. (b) Truth and observed as in (a) with ensemble mean of the assimilation (light +).

step 13 750, new observations are sufficient to pull it back toward the truth at later times.

All results displayed to this point are for 40-member ensembles. Table 2 shows the impact of varying the ensemble size on the rms of the assimilated ensemble mean for assimilations with observations available every 20 and every 50 model time steps. Increasing the ensemble size helps improve the assimilation in three ways. First, the estimates of the sample mean and covariance used to compute the kernels are significantly improved as the ensemble size becomes larger. Second, additional ensemble members allow better resolution of the details of the model's attractor. Third, as more kernels are available, the inflation factor, α, gets smaller so that the kernel distributions tend to spill into nonattractor areas to a lesser extent.

The quality of forecasts produced from the ensemble assimilations can also be evaluated. Figure 9 shows the rms error of the ensemble mean and the chi-square values as a function of forecast lead time for an assimilation with 40 ensemble members and observations available every 50 steps (shown in Fig. 6); results are shown out to 500 step lead forecasts (10 observation periods). As anticipated, the rms error increases with forecast lead time; however, the ensemble forecasts stay consistent with the true state of the system as demonstrated by the chi-square values. The chi-square significance only drops below 10%, which should happen by chance 10% of the time, in a handful of cases during the forecast set. This is the expected behavior for random samples of the forecast probability distribution; as the lead time increases, the distribution gets broader but remains consistent with the truth.

c. Single Gaussian versus kernel approximation

Section 3 presented two different methods for representing the continuous form of the prior distribution when performing the pointwise product. It was suggested there that the kernel filter approach should have advantages when applied to systems like the Lorenz-63 model. These advantages should be more pronounced when the prior (or assimilated) distribution is distinctly non-Gaussian, which should be more likely when the frequency of observations is reduced for a given observational error distribution.

Tables 1 and 2 contain comparisons of the Gaussian and kernel filter ensemble mean rms errors for a number of ensemble sizes and observation frequencies. In general, for frequent observations or very small ensembles, the two methods produce results that are quite similar,

TABLE 2. Mean rms error of the ensemble mean as a function of ensemble size for assimilations with observations available every 20 and every 50 model time steps for the Lorenz-63 model with observational error standard deviation 2.0 on each variable independently. Results are shown for kernel and Gaussian filter assimilations and are averaged over time steps 5000 to 15 000 of a long assimilation run.

Ensemble size	Kernel filter ens. mean rms obs. freq. 20	Gaussian filter ens. mean rms obs. freq. 20	Kernel filter ens. mean rms obs. freq. 50	Gaussian filter ens. mean rms obs. freq. 50
10	1.27	1.31	1.84	1.77
20	0.852	1.04	1.23	1.47
40	0.805	0.940	1.04	1.28
80	0.659	0.883	1.04	1.32
160	0.616	0.802	0.758	1.32

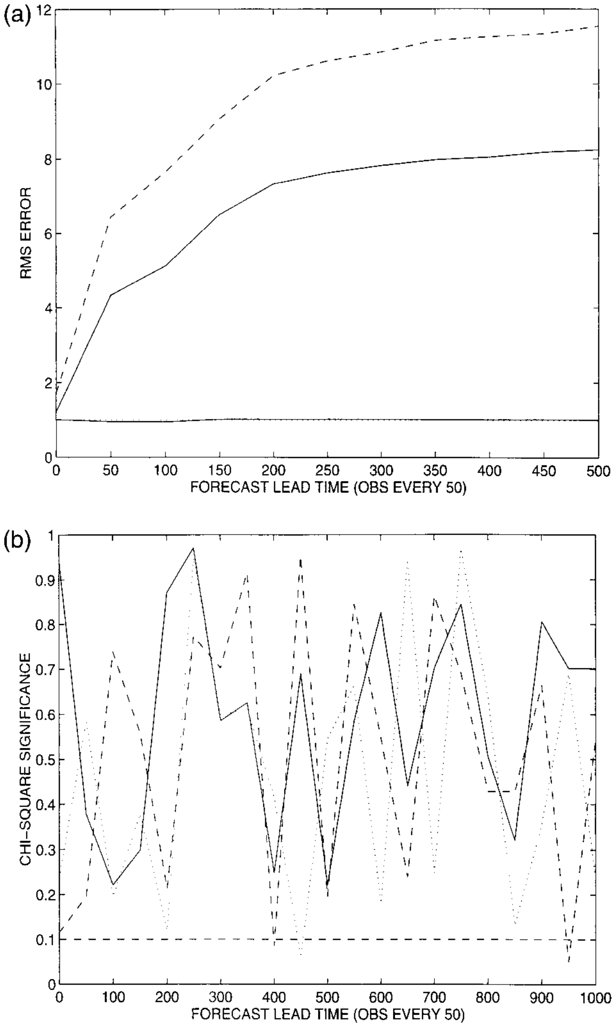

FIG. 9. Evaluation of quality of forecasts produced from same assimilation as in Fig. 6. (a) The rms of the ensemble mean (solid), the mean rms of the ensemble members (dash–dotted), and the normalized ratio of the two (dotted line nearly indistinguishable from the solid line at 1). (b) The significance of the chi-square statistic for the x (solid), y (dash–dotted), and z (dashed).

presumably because the prior distributions are almost always nearly Gaussian in these cases. As the frequency of observations decreases and/or the ensemble size gets larger, the advantages of the kernel filter can be realized and the Gaussian filter ensemble mean errors become larger than those for the kernel filter. Similar results were also found for even less frequent observations and when the observational error standard deviation was increased while holding the observational frequency constant. In all these cases, the prior distributions are becoming increasingly non-Gaussian, making the approximation used in the Gaussian filter increasingly inappropriate. As pointed out in the discussion of the previous subsection, the prior distributions for experiments like those shown in Figs. 6 and 7 are highly non-Gaussian at certain times. At these times, the kernel filter is not hampered by the assumption of a Gaussian prior and is able to produce superior assimilations. No instances were found for any of the low-order models examined in which the Gaussian filter performed significantly better than the kernel filter. In addition, the use of the kernel filter provides some additional benefits when the operator mapping from model variables to observations [**h** in Eq. (4)] is not easily invertible (recall that **h** is simply the identity for results here). The advantages of using the kernel filter can be even greater in higher-order systems as discussed in section 6.

6. Filtering and initialization

One of the major problems facing data assimilation systems in realistic forecast models is the need for initialization (Vautard and Legras 1986). The forecast models can be viewed as having attractors on which the dynamics is dominated by relatively slow, relatively balanced modes of some sort. However, perturbations off the attractor can result in large amplitude, relatively fast modes that may be unrealistic and generally have unfortunate numerical consequences (Warn and Menard 1986). The Lorenz-63 system discussed in the previous section does not have a serious need for initialization. Perturbations off the attractor generally decay exponentially toward nearby trajectories on the attractor (Anderson and Hubeny 1997).

Another low-order model, the nine-variable model (appendix C) of Lorenz (1980) can be used to evaluate the abilities of the filter assimilations in a system that can require initialization. This model is a truncated version of the primitive equations that has been used to study the behavior of gravity waves (Lorenz and Krishnamurthy 1987). Unlike the Lorenz-63 model, off-attractor perturbations in this model do not necessarily decay smoothly back to the attractor. Instead, most perturbations result in a transient period of high-amplitude, high-frequency gravity waves (Anderson and Hubeny 1997). If a data assimilation algorithm produces states that are not very close to the model attractor when new observations are combined with some prior estimate, the result is an assimilated state that is dominated by gravity wave noise that is unrelated to the true solution. Many traditional data assimilation algorithms are unable to incorporate information about the local structure of the attractor. These assimilation methods end up producing estimates of the state with larger errors than are found in the raw observations if they are applied directly to the nine-variable system. The standard solution is to apply some initialization algorithm to the assimilated state to enforce balance constraints that will reduce the resulting gravity wave amplitude.

Ostensibly, the analysis probability distribution produced by the kernel filter method knows about the constraints placed on the state by the model dynamics. However, the approximations involved in the solution method could lead to a loss of this information, leading

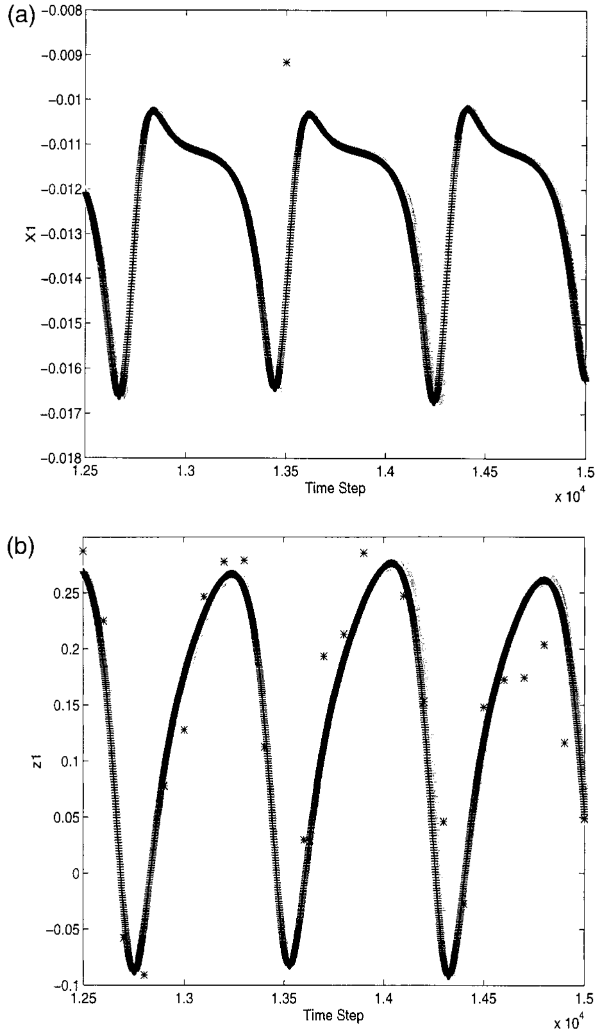

FIG. 10. (a) Time sequence of values of X_1 and (b) z_1 for truth (dark +), observed (dark *), and 10 members of the ensemble assimilation (light +) from a 40-member kernel assimilation of the nine-variable model with observations available every 100 steps with observational error standard deviation 0.05 on each variable independently.

to assimilated probability distributions that are not confined to be close to the model attractor. A kernel filter assimilation of the nine-variable model has been performed with observations available every 100 time steps and an observational error of 0.05 on each of the nine variables independently; a 40-member ensemble was used in this case. The assimilation has an average ensemble mean rms error of 5.7×10^{-3} and a mean rms error of 7.9×10^{-3} giving an rms ratio of 0.72, consistent with the truth being indistinguishable from a random sample of the ensemble. The chi-square statistics for the individual model components all indicate that the truth is indistinguishable from the ensemble with this sample size with a significance above 10%. Figure 10 shows a segment of the assimilation for variables X_1 and z_1, one of the divergence and height variables, respectively. Neither of these variables demonstrates any of the high frequency noise that is associated with gravity waves in this system. Particularly encouraging is the filter's ability to assimilate the **X** variables; the amplitude of these variables on the attractor is small (the system is nearly in balance) and the observational error is more than an order of magnitude larger than the variability of the variables (in fact, only one of the 25 observations during the period displayed in Fig. 10a even appears with the scale of this plot).

The attractor of the nine-variable model is quite similar to that for the Lorenz-63 model (Moritz and Sutera 1981). It is nearly flat locally and globally consists of a pair of nearly flat lobes that intersect at one edge. In this case, however, the attractor is embedded in a nine-dimensional phase space. A close examination of the prior distribution for the kernel filter shows that the ensemble distributions are generally able to represent this attractor structure, leading to the high quality assimilations.

The single Gaussian filter has been applied to the same assimilation problem and produces an ensemble mean rms error of 8.7×10^{-3} and a mean rms error of 1.2×10^{-2}. The single Gaussian filter appears to produce assimilated results that generate slightly more gravity wave noise in forecasts than do the kernel assimilation forecasts. This apparently leads to the kernel filter's advantage over the single Gaussian being larger in this nine-variable case than in most of the Lorenz-63 cases studied.

Increasing the observational frequency to every 10 time steps while simultaneously increasing the observational error standard deviations to 0.2 leads to an even more stringent test of the kernel filter. In this case, low quality observations are available frequently. Each time these poor observations are assimilated, the assimilation algorithm must avoid producing new ensemble members that are not close to the attractor. Figures 11a and 11b depict the ensemble assimilation for this case for the X_1 variable for the Gaussian and kernel filters, respectively. There is much more noise at all times in the Gaussian filter and orders of magnitude more noise at certain times. While the kernel filter is not noise-free in this case, the noise for the X_1 variable is still somewhat less than the natural variability of X_1. Figures 11c and 11d show the ensemble means for the same variables. Again, some high-frequency gravity wave noise is visible in both assimilations, but the noise is always significantly less in the kernel filter. The impacts of this noise are less noticeable for the height field due to its inherently larger natural variability, but again, the kernel filter produces a considerably better and less noisy assimilation of the z_1 variable as shown in Fig. 12.

It is encouraging that the kernel filter is able to produce nine-variable model assimilations that probably do

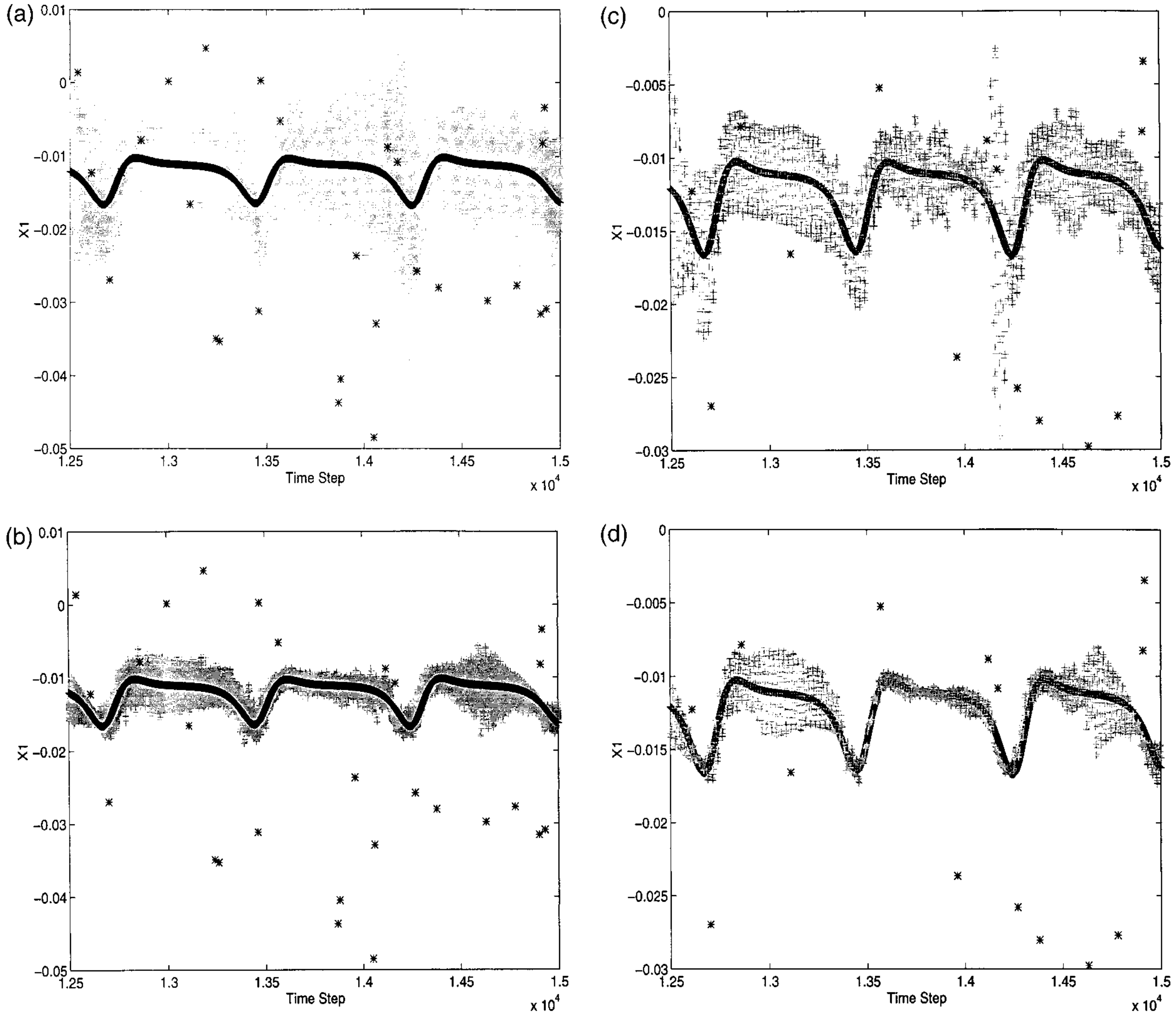

FIG. 11. Same as Fig. 10 except with observations every 10 steps and observational error standard deviation 0.2 on each variable independently: (a) and (c) from a Gaussian filter and (b) and (d) from a kernel filter assimilation. (a), (b) Ten members of the ensemble assimilation are shown by the light +. (c), (d) The ensemble mean is shown by the light +.

not require initialization, even with relatively frequent but very noisy observations. If the attractors of higher-dimensional models are confined predominantly to relatively low-dimensional manifolds locally, it is possible that these structures could also be represented by ensembles much smaller than the total phase space of the model. Initial results with higher-order models indicate that the advantages of the kernel filter become even greater in higher dimensions. Considerable further work will be needed to evaluate the potential for applying filters of this type in realistic forecast models that require initialization.

7. Application to more realistic systems

The results shown here demonstrate the capabilities of the kernel filter method in low-order systems with a number of simplifying assumptions. This section briefly discusses the potential for extending the method to higher-order models and more realistic observations while removing the most stringent simplifying assumptions.

a. Application to high-order models

The results of earlier sections have been for dynamical systems in which the number of phase space dimensions is small compared to the ensemble size. However, for realistic applications, the number of ensemble members would have to be very small compared to the number of phase space dimensions. One can still attempt to apply the kernel filter method with small ensembles to large models using a heuristic tech-

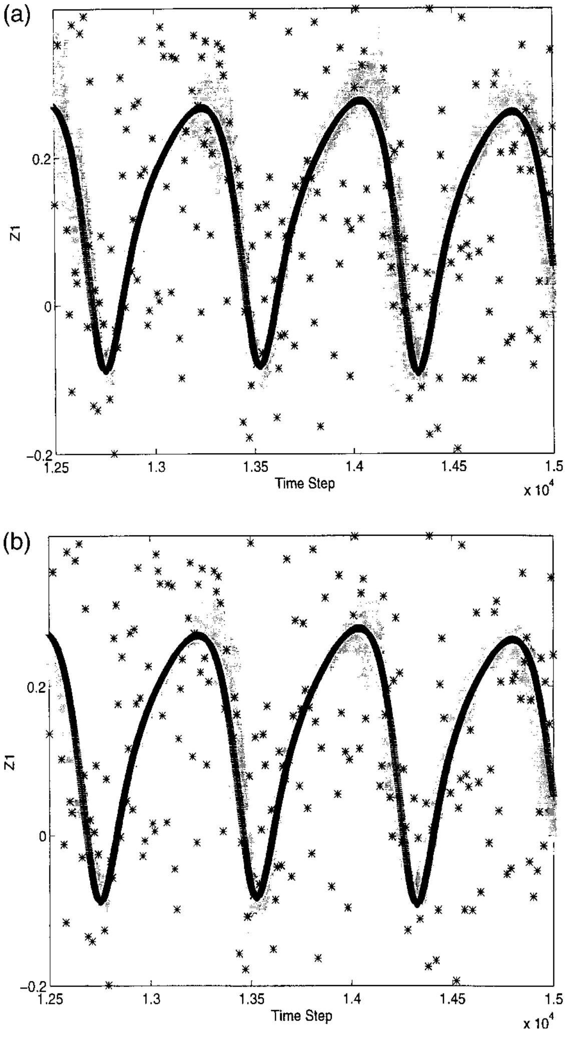

FIG. 12. Time sequence of values of z_1 with truth (dark +), observed (dark *) and 10 members of 40-member Gaussian (a) and kernel (b) filter assimilations (light +) of the nine-variable model with observations every 10 steps and observational error standard deviation 0.2 on each variable independently.

nique that is similar to classical regularization techniques.

Suppose an m-member ensemble is being used to do assimilation in a model with an n-dimensional phase space. Assume that, at time t_0, the m-member ensemble is a random sample of the analysis probability distribution (there is no contradiction in having a random sample of a probability distribution in a high-dimensional space with a small ensemble). One can proceed to generate a random sample of the prior distribution by advancing the ensemble in the model as in low-order applications. However, this prior distribution spans at most an $(m - 1)$-dimensional subspace of the n-dimensional phase space. One can calculate a basis for this subspace and project the observational error distribution onto this basis. The computation and random resampling of the pointwise product can then be completed in this subspace. This operation can be shown to produce a random sample of the analysis probability distribution in the $(m - 1)$-dimensional subspace but has no effect on the $(n - m + 1)$-dimensional portion of phase space, which is not spanned by the ensemble.

Initially, it seems unreasonable to assume that doing the pointwise product only in the ensemble subspace can result in an effective assimilation. However, most high-order dynamical systems used in atmospheric prediction are believed to have local attractor structures that are confined to very small submanifolds of the total phase space (Broomhead et al. 1991; Henderson and Wells 1988). In addition, ensemble members that lie near this local attractor will generally tend to be enriched in their projection on the attractor when integrated in the model [this is the underlying premise of some heuristic ensemble generation methods (Toth and Kalnay 1996)]. If the vast majority of the interesting dynamics takes place on a submanifold that is not of significantly larger size than the ensemble, this subspace assimilation technique may work.

A great deal of additional research is needed to better understand and test filter methods applied to high-dimensional systems. Initial tests, however, demonstrate that the method can work in models for which the phase space dimension is much larger than the ensemble size. An assimilation has been performed with a 40-member ensemble in a forced global barotropic spectral model at T42 resolution. Observations of the streamfunction are available at every grid point on a reduced grid (every fourth grid point in both latitude and longitude) every 12 h with an observational error standard deviation of 5×10^5 m^2 s^{-1} and the assimilation is performed in physical space, which has more than 3000 degrees of freedom. The kernel filter method was able to constrain successfully an assimilation in this case, giving rms errors for the ensemble mean nearly two orders of magnitude less than the observational errors and distributions for gridpoint streamfunction that were indistinguishable from random samples for assimilation sample sizes of 100.

b. Realistic observations

In the results presented, the operator **h** [Eq. (4)] mapping from the model state variables to the observations has been simply the identity. It is straightforward to extend the method to h operators that can be easily inverted to map from observations to model state. However, realistic observational operators generally do not have well-defined inverses.

A variety of possible extensions to the filter method could be developed to deal with more general **h** operators. As an example, the prior ensemble distribution can be mapped to the ''observational space'' by applying the **h** operator to each ensemble member.

The pointwise product can then be performed in observational space using the kernel method and a new random sample generated. Each member of this new sample is associated with a particular member of the prior ensemble. One can then solve an ensemble of constrained optimization problems to find points in the model state space that closely correspond to the newly generated observational space sample. This optimization may be considerably easier than similar operations required in some more conventional assimilation techniques for several reasons. First, a relatively good first guess is available since the kernel method resampling indicates which prior ensemble-member should be used as a first guess. Second, techniques such as adjoint methods can be readily applied to assist in gradient computations. The adjoints here are only for the **h** operator and may be much easier to compute than those required for adjoint applications of variational algorithms (Errico and Vukicevic 1992). Finally, the search need only be conducted in the phase space spanned by the ensemble, which is generally small compared to the phase space size of realistic models. Again, a great deal of additional work is needed to understand the best ways to apply filter algorithms with noninvertible **h** operators.

c. Systematic errors

All applications of the filter method described to date have been in a perfect model context for which the assimilating model is identical to that generating the truth. In applications to real observations, the assimilating model is expected to have systematic errors. Filter methods are traditionally able to adapt to the presence of systematic errors, which are a part of almost all real-world applications. However, systematic errors may present a greater challenge when Monte Carlo filters are applied in high-dimensional models. In this case, if the systematic error projects on directions in phase space that are not spanned by the ensemble prior, the assimilation will not affect that portion of the systematic error. Since there is no reason to expect that the systematic error will be predominantly confined to the attractor of the assimilating model, this is expected to be a serious problem. Possible solutions include applying a simpler assimilation method in the null-space of the ensemble prior. Imperfect model tests have suggested that even as simple a solution as damping the model state toward observations in the null space can be effective, but tests with real-world observations will be essential to develop solutions to the systematic error problem.

d. Stochastic models and parameterizations

In previous sections, the assimilating models have been assumed to be deterministic with the second term in Eq. (2) dropped. However, when applied to real systems, it becomes essential to account for the uncertainties in the model formulation (Houtekamer et al. 1996; Moritz and Sutera 1981; Buizza et al. 1998). This is particularly true for subgrid-scale parameterizations in atmospheric models. Traditionally, these parameterizations have been formulated as deterministic relations between the large-scale flow and the feedback of small-scale, unresolved features on the large scale. Clearly, this feedback is not deterministic but is instead stochastic. When applying filters to realistic systems, it is important to account for these stochastic terms in the assimilating model. This can be done by simply adding random samples from some rough estimate of the distribution, or by attempting to formulate stochastic subgrid-scale parameterizations. The importance of accounting for these stochastic model dynamics will need to be further investigated and appropriate solutions developed.

8. Conclusions

The kernel filter, a method to produce random samples of the analysis and forecast probability distributions of a dynamical system, has been described and validated in low-order model applications. The method is able to produce samples that are qualitatively very close to random in these applications while still producing ensemble mean assimilations and forecasts with small errors. The method has a number of advantages over methods previously described in the literature for the generation of ensemble members for predictions or predictability research.

A number of difficulties remain in extending the kernel filter method to more realistic applications. However, numerous applications exist for random samples of the analysis and forecast distributions even in low-order models. These samples can be used for studies of fundamental predictability questions in low-order models that have been performed traditionally using heuristically generated ensembles that are often poor approximations to the appropriate random samples. The kernel filter can also be used to do low-order model comparisons to techniques of ensemble generation that are applied in higher-order models.

A number of major obstacles must be surmounted in order to extend the assimilation method to high-order models with realistic observations. Initial tests of a technique for extending the kernel filter to models with phase spaces that are large compared to the ensemble size have been successful. Additional research is needed to see how these extend to prediction models. Additional work is also required to investigate extending the method to more realistic observation types and to deal with problems of systematic error. If successful, however, the application of the method in prediction models promises to enhance significantly the quality of ensemble forecasts over a range of spatial and temporal scales.

Acknowledgments. Reviews of earlier versions of this manuscript were significantly improved with the assistance of Tapio Schneider, David Baker, and John Lanzante. The authors are also indebted to Tom Hamill and an anonymous reviewer for thorough and helpful comments and suggestions.

APPENDIX A

Convolution of Two Gaussians

The convolution of two *m*-dimensional normals with means μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 is a normal (A1) with mean μ and covariance Σ defined in (A2)–(A4):

$$\mathbf{N}(\mu_1, \Sigma_1)N(\mu_2, \Sigma_2) = c\mathbf{N}(\mu, \Sigma), \tag{A1}$$

$$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}, \tag{A2}$$

$$\mu = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2), \tag{A3}$$

$$c = \frac{1}{(2\Pi)^{d/2}|\Sigma_1 + \Sigma_2|^{1/2}} \exp\left\{-\frac{1}{2}[(\mu_2 - \mu_1)^{\mathrm{T}}(\Sigma_1 + \Sigma_2)^{-1}(\mu_2 - \mu_1)]\right\}. \tag{A4}$$

APPENDIX B

Lorenz-63 Model

The Lorenz-63 model (Lorenz 1963) has become one of the mainstays for the study of chaotic systems (Palmer 1993). The model's three equations are

$$\dot{x} = -\sigma x + \sigma y, \tag{B1}$$

$$\dot{y} = -xz + rx - y, \tag{B2}$$

$$\dot{z} = xy - bz, \tag{B3}$$

where the dot represents a derivative with respect to time. The model is integrated using the standard values for the parameters *σ, b*, and *r* and the time step described in the original Lorenz paper resulting in a system with chaotic dynamics.

APPENDIX C

Nine-Variable Model

The nine-variable model is represented by the equations

$$\dot{X}_i = U_jU_k + V_jV_k - \nu_0a_iX_i + Y_i + a_iz, \tag{C1}$$

$$\dot{Y}_i = U_jY_k + Y_jV_k - X_i - \nu_0a_iY_i, \tag{C2}$$

$$\dot{z}_i = U_j(z_k - h_k) + (z_j - h_j)V_k - g_0X_i - K_0a_iz_i + F_i, \tag{C3}$$

$$U_i = -b_jx_i + cy_i, \tag{C4}$$

$$V_i = -b_kx_i - cy_i, \tag{C5}$$

$$X_i = -a_ix_i, \tag{C6}$$

$$Y_i = -a_iy_i, \tag{C7}$$

where each equation is defined for cyclic permutations of the indices (*i, j, k*) over the values (1, 2, 3). The *X, Y,* and *z* variables can be thought of as representing divergence, vorticity, and height, respectively, while the subscripts can be viewed as representing a zonal mean plus two wave components for each of the three fields. All the parameters in Eqs. (C1)–(C7) are selected as in Lorenz (1980) in order to produce a chaotic system.

REFERENCES

Anderson, J. L., 1994: Selection of initial conditions for ensemble forecasts in a simple perfect model framework. *J. Atmos. Sci.,* **53,** 22–36.

——, 1996a: Selection of initial conditions for ensemble forecasts in a simple perfect model framework. *J. Atmos. Sci.,* **53,** 22–36.

——, 1996b: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

——, 1997: The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: Low-order perfect model results. *Mon. Wea. Rev.,* **125,** 2969–2983.

——, and V. Hubeny, 1997: A reexamination of methods for evaluating the predictability of the atmosphere. *Non-linear Proc. Geosci.,* **4,** 157–166.

Barker, T. W., 1991: The relationship between spread and forecast error in extended range forecasts. *J. Climate,* **4,** 733–742.

Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. *Quart. J. Roy. Meteor. Soc.,* **116,** 867–912.

Broomhead, D. S., R. Indik, A. C. Newell, and D. A. Rand, 1991: Local adaptive Galerkin bases for large-dimensional dynamical systems. *Nonlinearity,* **4,** 159–197.

Buizza, R., J. Tribbia, F. Molteni, and T. Palmer, 1993: Computation of optimal unstable structures for a numerical weather prediction model. *Tellus,* **45A,** 388–407.

——, T. Petroliagis, T. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **124,** 1935–1960.

Cramer, H., 1996: Mathematical Methods of Statistics. Princeton University Press, 575 pp.

Ehrendorfer, M., 1994: The Liouville equation and its potential usefullness for the prediction of forecast skill. Part I: Theory. *Mon. Wea. Rev.,* **122,** 703–713.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

Errico, R. M., and T. Vukicevic, 1992: Sensitivity analysis using an adjoint of the PSU/NCAR mesoscale model. *Mon. Wea. Rev.,* **120,** 1644–1660.

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.

——, 1997: Advanced data assimilation for strongly nonlinear dynamics. *Mon. Wea. Rev.,* **125,** 1342–1354.

——, and P. J. van Leeuwen, 1996: Assimilation of Geosat altimeter data for the Agulhas Current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Wea. Rev.,* **124,** 85–96.

Fukunaga, K., 1972: *Introduction to Statistical Pattern Recognition.* Academic Press, 369 pp.

Gleeson, T. A., 1970: Statistical-dynamical prediction. *J. Appl. Meteor.,* **9,** 333–344.

Harrison, M. S. J., D. S. Richardson, K. Robertson, and A. Woodcock, 1995: Medium-range ensembles using both the ECMSF T63 and unified models—An initial report. UKMO Tech. Rep. 153, 25 pp. [Available from United Kingdom Meteorological Office, London Rd., Bracknell, Berkshire RG12 2S2 United Kingdom.]

Henderson, H. W., and R. Wells, 1988: Obtaining attractor dimensions from meteorological time series. *Advances in Geophysics,* Vol. 30, Academic Press, 205–236.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus,* **35a,** 100–118.

Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.,* **121,** 1834–1846.

——, and J. Derone, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.,* **123,** 2181–2196.

——, L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory.* Academic Press, 376 pp.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.,* **115,** 349–356.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1980: Attractor sets and quasi-geostrophic equilibrium. *J. Atmos. Sci.,* **37,** 1685–1699.

——, and V. Krishnamurthy, 1987: On the nonexistence of a slow manifold. *J. Atmos. Sci.,* **44,** 2940–2950.

Miller, R. N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.,* **51,** 1037–1056.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–120.

Moritz, R. E., and A. Sutera, 1981: The predictability problem: Effects of stochastic perturbations in multiequilibrium systems. *Advances in Geophysics,* Vol. 23, Academic Press, 345–383.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.,* **114,** 463–493.

——, 1990: Assessment of the practical utility of extended range ensemble forecasts. *Quart. J. Roy. Meteor. Soc.,* **116,** 89–125.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.,* **74,** 49–66.

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.,* **120,** 1747–1763.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 175 pp.

Tracton, S., and E. Kalnay, 1993: Operational ensemble forecasting at NMC: Practical aspects. *Wea. Forecasting,* **8,** 379–398.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

——, and ——, 1996: Ensemble forecasting at NMC and the breeding method. NMC Office Note 407, 34 pp. [Available from NMC Office, 5200 Auth Rd., Camp Springs, MD 20746.]

——, ——, S. M. Tracton, R. Wobus, J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting,* **12,** 140–153.

Van Leeuwen, P. J., and G. Evensen, 1996: Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Wea. Rev.,* **124,** 2898–2912.

Vautard, R., and B. Legras, 1986: Invariant manifolds, quasi-geostrophy and initialization. *J. Atmos. Sci.,* **43,** 565–584.

Warn, T., and R. Menard, 1986: Nonlinear balance and gravity-inertial wave saturation in a simple atmospheric model. *Tellus,* **38A,** 285–294.

Wobus, R. L., and E. Kalnay, 1995: Three years of operational prediction of forecast skill at NMC. *Mon. Wea. Rev.,* **123,** 2132–2148.