Accounting for Model Error from Unresolved Scales in Ensemble Kalman Filters by Stochastic Parameterization

FEI LU

Department of Mathematics, University of California, Berkeley, and Lawrence Berkeley National Laboratory, Berkeley, California

XUEMIN TU

Department of Mathematics, University of Kansas, Lawrence, Kansas

ALEXANDRE J. CHORIN

Department of Mathematics, University of California, Berkeley, and Lawrence Berkeley National Laboratory, Berkeley, California

(Manuscript received 20 December 2016, in final form 13 April 2017)

ABSTRACT

The use of discrete-time stochastic parameterization to account for model error due to unresolved scales in ensemble Kalman filters is investigated by numerical experiments. The parameterization quantifies the model error and produces an improved non-Markovian forecast model, which generates high quality forecast ensembles and improves filter performance. Results are compared with the methods of dealing with model error through covariance inflation and localization (IL), using as an example the two-layer Lorenz-96 system. The numerical results show that when the ensemble size is sufficiently large, the parameterization is more effective in accounting for the model error than IL; if the ensemble size is small, IL is needed to reduce sampling error, but the parameterization further improves the performance of the filter. This suggests that in real applications where the ensemble size is relatively small, the filter can achieve better performance than pure IL if stochastic parameterization methods are combined with IL.

1. Introduction

Model error due to unresolved scales can degrade the performance of data assimilation schemes. Such model error can arise from the failure to represent subgrid processes correctly, from computational resources that are too limited to resolve all scales, and from discretization and truncation errors.

Various methods have been proposed for taking model error into account. One can roughly divide them into direct and indirect approaches. In an indirect approach, one accounts for model error in ensemble data assimilation by correcting the ensemble during the assimilation step. The most widely used indirect methods are covariance inflation and localization (IL) algorithms, which correct the sample covariance (Houtekamer and Mitchell 1998; Anderson and Anderson 1999; Mitchell and Houtekamer 2000; Hamill et al. 2001). These algorithms were originally introduced to reduce sampling errors in the sample covariance as a result of insufficient ensemble size. Nevertheless, they have been found to compensate effectively for model errors and have been widely used for that purpose (see e.g., Mitchell and Houtekamer 2000; Hamill and Whitaker 2005; Anderson 2007a, 2009). Other examples of indirect techniques include covariance relaxation (Zhang et al. 2004) and bias correction methods that use innovations from data to remove bias in the forecast ensemble (Dee and Da Silva 1998). The drawbacks of these indirect methods include that they need empirical tuning, and more important, that the deficiency of the forecast model remains.

In a direct approach, one seeks a representation of the model error to augment and improve the forecast model, so that the forecast ensemble has correct

Corresponding author: Fei Lu, feilu@berkeley.edu

DOI: 10.1175/MWR-D-16-0478.1

statistics and dynamics. Examples include deterministic and stochastic parameterization methods (Palmer 2001; Meng and Zhang 2007; Berry and Harlim 2014; Mitchell and Carrassi 2015), additive random perturbations (Hamill and Whitaker 2005; Houtekamer et al. 2009), a low-dimensional method (Li et al. 2009), and averaging and homogenization methods (Pavliotis and Stuart 2008; Mitchell and Gottwald 2012; Gottwald and Harlim 2013). Representations of the model error can be derived either via data assimilation using the noisy observations, or before data assimilation using noiseless training data. In the latter case, numerous results demonstrate that stochastic parameterization is preferable to deterministic parameterization (Buizza et al. 1999; Palmer 2001; Pavliotis and Stuart 2008), and that a non-Markovian model is preferable to a Markovian model in the absence of scale separation (see, e.g., Wilks 2005; Crommelin and Vanden-Eijnden 2008; Danforth and Kalnay 2008; Chekroun et al. 2011; Majda and Harlim 2013; Kondrashov et al. 2015). These findings are consistent with the Mori–Zwanzig analysis (Zwanzig 1973, 2001; Chorin and Hald 2013; Chorin et al. 2000, 2002; Gottwald et al. 2015) in statistical physics, which shows that a closed system of equations for a subset of variables in a given problem consists of a Markovian term, a non-Markovian memory term, and a stochastic noise term. The abovementioned methods pose challenges when deriving an effective non-Markovian model, as a result of difficulties in inferring a continuous-time model from partial discrete data and then deriving an accurate discretization for it. A novel, efficient, discrete-time non-Markovian stochastic parameterization scheme for quantifying model error was introduced by Chorin and Lu (2015). This method is fully discrete, readily takes memory effects into account, simplifies the inference from discrete data, and requires no discretization. It leads to an improved non-Markovian forecast model that can capture key statistical and dynamical features of the resolved scales.

It is natural to ask whether the direct approach can be as good as or better than the methods of IL in accounting for model error in ensemble Kalman filters (EnKFs). Several direct methods have been studied for this purpose. Additive error representations were shown to improve the performance of the ensemble square root Kalman filter in Hamill and Whitaker (2005), bias removal methods augmented by additive noise were shown to outperform pure inflation schemes in the local ensemble transform Kalman filter in Li et al. (2009), and time-varying and time-constant model error representations were shown to reduce the tuning of IL in the ensemble transform Kalman filter in Mitchell and Carrassi (2015).

In the present study we examine the discrete-time parameterization and compare it with covariance inflation and localization in accounting for model error in the EnKF. We assume that offline noiseless training data of the resolved scales can be generated and used either to tune inflation and localization or to infer parameters in the parameterization. We examine both cases where the ensemble is large enough so that the sampling error is negligible, and where the sample is small and the sampling error needs to be reduced by IL. We carry out numerical tests on the two-layer Lorenz-96 system (Lorenz 1996), a simplified nonlinear model of atmospheric dynamics involving interacting resolved and unresolved scales of motion. A forecast model in the EnKF is a truncated model of the large scales alone, and its model error comes from the unresolved small scales. The parameterization directly accounts for the model error by constructing an improved forecast model for the filter and is compared with the IL approach.

The numerical results show that when the ensemble size is large, the parameterization outperforms IL in accounting for model error. To the best of our knowledge, this is the first comparison made in a case where the ensemble size is large enough for the sampling error to be negligible, so that both methods account exclusively for model error and their performance can be compared clearly. The numerical results also show that when the ensemble size is small, IL is needed to reduce sampling error, but the parameterization further improves the filter performance. This result is in line with the previous findings in work by Hamill and Whitaker (2005), Li et al. (2009), and Mitchell and Carrassi (2015) that show that with the combination of stochastic methods and IL the filter can achieve better performance than pure IL in small, practical ensemble sizes.

This paper is organized as follows. In section 2 we provide a quick review of the EnKF. In section 3 we review covariance inflation and localization algorithms, as well as discrete-time non-Markovian stochastic parameterization. We devote section 4 to a numerical study using the two-layer Lorenz-96 system, and conclude the paper with a discussion of the results in section 5.

2. The ensemble Kalman filter

The ensemble Kalman filter is a Monte Carlo implementation of Bayesian filtering with the Kalman filter update (Evensen 1994; Evensen and Van Leeuwen 1996; Houtekamer and Mitchell 1998; Burgers et al. 1998). This approach uses an ensemble of random samples,

also called particles, to approximate the forecast and analysis distributions by Gaussian distributions whose means and covariances are given by ensemble means and covariances. Among various EnKF algorithms, we choose to consider only the version with perturbed observations, introduced by Burgers et al. (1998) and Houtekamer and Mitchell (1998), and we refer to Lei et al. (2010) for a comparison of different versions of EnKF algorithms.

Suppose the filter uses a forecast model

$$\mathbf{x}_n = \mathbf{F}_n(\mathbf{x}_{n-l:n-1}), \tag{1}$$

where $\mathbf{x}_n \in \mathbb{R}^{d_x}$ is the state variable, $\mathbf{x}_{n-l:n-1} = (\mathbf{x}_{n-l}, \ldots, \mathbf{x}_{n-1})$, and $\mathbf{F}_n$ is a forecast operator at time n, which maps $\mathbb{R}^{l\times d_x}$ to $\mathbb{R}^{d_x}$ with $1 \le l \le n-1$. The forecast model can be either stochastic or deterministic, and either Markovian (e.g., $l=1$) or non-Markovian (e.g., $l>1$). The state variable is observed through a linear observation operator with Gaussian noise:

$$\mathbf{z}_n = \mathbf{H}\mathbf{x}_n + \boldsymbol{\varepsilon}_n,$$

where $\mathbf{H} \in \mathbb{R}^{d_z\times d_x}$ is the observation matrix and the $\boldsymbol{\varepsilon}_n \sim N(0, \mathbf{R})$ are independent Gaussian noises. In this study, we assume that the observation matrix $\mathbf{R}$ is known.

a. The standard EnKF

The EnKF iterates the following two steps, with an initial ensemble of particles $\{\mathbf{x}_0^{a,(i)}, i=1, \ldots, M\}$ sampled from the forecast distribution of the state variable $\mathbf{x}$ (e.g., the stationary distribution of the forecast model).

1) Forecast step: from the ensemble $\{\mathbf{x}_{1:n-1}^{a,(i)}\}$ at time $n-1$, generate a forecast ensemble $\{\mathbf{x}_n^{f,(i)}\}$ using the forecast model in (1); that is, $\mathbf{x}_n^{f,(i)} = \mathbf{F}_n(\mathbf{x}_{n-l:n-1}^{a,(i)})$. Here, the superscript in $\mathbf{x}_n^f$ denotes the ensemble from the forecast model, and the superscript in $\mathbf{x}_n^a$ denotes the ensemble of the posterior distribution after assimilating data in the following analysis step. If the forecast model is stochastic, independent realizations should be used at different times.
2) Analysis step: given a new observation $\mathbf{z}_n$, update the forecast ensemble to get a posterior ensemble of $\mathbf{x}_n$,

$$\mathbf{x}_n^{a,(i)} = \mathbf{x}_n^{f,(i)} + \mathbf{K}_n(\mathbf{z}_n^{(i)} - \mathbf{H}\mathbf{x}_n^{f,(i)}), \tag{2}$$

for $i=1, \ldots, M$, where the Kalman gain matrix is

$$\mathbf{K}_n = \mathbf{C}_n^f\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{C}_n^f\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}, \tag{3}$$

where the matrix $\mathbf{C}_n^f$ is the sample covariance of the forecast ensemble,

$$\mathbf{C}_n^f = \frac{1}{M-1}\sum_{i=1}^{M}(\mathbf{x}_n^{f,(i)} - \overline{\mathbf{x}_n^f})(\mathbf{x}_n^{f,(i)} - \overline{\mathbf{x}_n^f})^{\mathrm{T}},$$

where $\overline{\mathbf{x}_n^f} = (1/M)\sum_{i=1}^{M}\mathbf{x}_n^{f,(i)}$ and the $\mathbf{z}_n^{(i)}$ are obtained by adding random perturbations $\varepsilon_n^{(i)} \sim N(0, \mathbf{R})$ to $\mathbf{z}_n$,

$$\mathbf{z}_n^{(i)} = \mathbf{z}_n + \varepsilon_n^{(i)}.$$

b. A block update algorithm

At each time n, only the current state $\mathbf{x}_n^{(i)}$ of the ith particle is updated in the analysis step in the above standard EnKF, and the past trajectory $\mathbf{x}_{1:n-1}^{(i)}$ of the particle remains unchanged. Therefore, the time correlation between $\mathbf{x}_n$ and $\mathbf{x}_{1:n-1}$ is not properly represented by the ensemble. For a Markovian forecast model, this works fine, because the next state $\mathbf{x}_{n+1}$ depends only on the current state $\mathbf{x}_n$. For a non-Markovian model with lag l, however, the next state $\mathbf{x}_{n+1}$ depends directly on a block of the past trajectory $\mathbf{x}_{n-l+1:n}$. This requires the ensemble to properly represent the space–time correlation of $\mathbf{x}_{n-l+1:n}$, and therefore the states $\mathbf{x}_{n-l+1:n}$ should be updated as a whole at time n. Inspired by the block sampling algorithm of Doucet et al. (2006), we introduce the following block update algorithm that updates a block $\mathbf{x}_{n-L+1:n}$ with $L \ge l$ in the analysis step of the EnKF. This block update algorithm is akin to the fixed-lag smoother using EnKF (Khare et al. 2008), which is an implementation of the ensemble Kalman smoother (EnKS) discussed by Evensen and Van Leeuwen (2000) and Whitaker and Compo (2002).

We choose a block length $L \ge 1$, and define the augmented observation matrix $\tilde{\mathbf{H}} \in \mathbb{R}^{Ld_z\times Ld_x}$ and the augmented noise covariance $\tilde{\mathbf{R}} \in \mathbb{R}^{Ld_z\times Ld_z}$ as

$$\tilde{\mathbf{H}} = \mathrm{diag}(0, \ldots, 0, \mathbf{H}), \quad \tilde{\mathbf{R}} = \mathrm{diag}(0, \ldots, 0, \mathbf{R}). \tag{4}$$

For $n < L$, we use the above EnKF method. At time $n \ge L$, after obtaining the forecast ensemble $\{\mathbf{x}_n^{f,(i)}\}$, we update the ensemble of the block path $\mathbf{X}_n^{f,(i)} = (\mathbf{x}_{n-L+1:n-1}^{a,(i)}, \mathbf{x}_n^{f,(i)})$:

$$\mathbf{X}_n^{a,(i)} = \mathbf{X}_n^{f,(i)} + \tilde{\mathbf{K}}_n(\mathbf{z}_n^{(i)} - \tilde{\mathbf{H}}\mathbf{X}_n^{f,(i)})$$

for $i=1, \ldots, M$, where the Kalman gain matrix is computed as

$$\tilde{\mathbf{K}}_n = \tilde{\mathbf{C}}_n^f\tilde{\mathbf{H}}^{\mathrm{T}}(\tilde{\mathbf{H}}\tilde{\mathbf{C}}_n^f\tilde{\mathbf{H}}^{\mathrm{T}} + \tilde{\mathbf{R}})^{-1}.$$

Here the matrix $\tilde{\mathbf{C}}_n^f$ is the sample covariance of the forecast ensemble:

$$\tilde{\mathbf{C}}_n^f = \frac{1}{M-1}\sum_{i=1}^{M}(\mathbf{X}_n^{f,(i)} - \overline{\mathbf{X}_n^f})(\mathbf{X}_n^{f,(i)} - \overline{\mathbf{X}_n^f})^{\mathrm{T}},$$

where $\overline{\mathbf{X}_n^f} = 1/M\sum_{i=1}^{M}\mathbf{X}_n^{f,(i)}$. Then, we update the current L-step block by setting $\mathbf{x}_{n-L+1:n}^{a,(i)} = \mathbf{X}_n^{a,(i)}$.

When $L = 1$, the above algorithm is the same as the standard EnKF. When $L > 1$, it updates a block of the trajectory using the new observation. A natural choice of block length L is the length l of the memory in the forecast operator $f_n(\mathbf{x}_{n-l:n-1})$. This is the choice we make in this paper, and we leave it as future work to discuss of the optimal choice of L as well as of other issues such as covariance inflation and localization for this block update algorithm and its variants in applications to non-Markovian models.

3. Methods for accounting for model error

Let a forecast model at our disposal be represented as

$$\mathbf{x}_n = \mathbf{f}_0(\mathbf{x}_{n-1}), \tag{5}$$

where $\mathbf{x}_n$ is a vector in $\mathbb{R}^{d_x}$ representing the resolved scales at time t_n, and $\mathbf{f}_0$ is a forecast operator independent of time. This is a reduced model of a more complicated full model of the form

$$\begin{cases} \hat{\mathbf{x}}_n = \hat{\mathbf{F}}(\hat{\mathbf{x}}_{n-1}, \hat{\mathbf{y}}_{n-1}), \\ \hat{\mathbf{y}}_n = \hat{\mathbf{G}}(\hat{\mathbf{x}}_{n-1}, \hat{\mathbf{y}}_{n-1}), \end{cases} \tag{6}$$

where $\hat{\mathbf{x}}_n \in \mathbb{R}^{d_x}$ and $\hat{\mathbf{y}}_n \in \mathbb{R}^{d_y}$ are the resolved and unresolved scales at time t_n, respectively, with $d_x \ll d_y \leq \infty$, and where the functions $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$ map the states from time t_{n-1} to t_n. In general, this full model is a discrete representation of a system of differential equations. The reduced model is used when the full system is too difficult to solve or possibly not fully understood, and it is often obtained by truncating the full system. The difference between the solutions of the reduced model [(5)] and the full model [(6)] is the *model error due to unresolved scales*.

a. Covariance inflation and localization

1) Covariance localization

Covariance localization was originally designed to remove poorly estimated long-range spatial correlations due to insufficient ensemble size (Houtekamer and Mitchell 1998; Gaspari and Cohn 1999; Furrer and Bengtsson 2007; Anderson 2007b). The standard implementation of localization is through the Schur product (entry-wise product, also known as the Hadamard product) of the forecast covariance $\mathbf{C}_n^f$ by a localization matrix $\mathbf{C}_{\mathrm{loc}}$, which is a symmetric positive definite matrix with entries obtained from a predefined correlation-length function, known as a taper function. In this study, we employ the widely used Gaspari–Cohn taper function (Gaspari and Cohn 1999), which has compact support given by

$$g(s) = \begin{cases} 1 - \frac{5}{3}s^2 + \frac{5}{8}s^3 + \frac{1}{2}s^4 - \frac{1}{4}s^5, & \text{if} \quad 0 \leq s \leq 1; \\ -\frac{2}{3s} + 4 - 5s + \frac{5}{3}s^2 + \frac{5}{8}s^3 - \frac{1}{2}s^4 + \frac{1}{12}s^5, & \text{if} \quad 1 \leq s \leq 2; \\ 0, & \text{if} \quad s \geq 2. \end{cases} \tag{7}$$

The corresponding localization matrix is

$$\mathbf{C}_{r_{\mathrm{loc}}}(i, j) = g(|i-j|/r_{\mathrm{loc}}), \tag{8}$$

where r_{loc} is the localization radius. We refer to Furrer and Bengtsson (2007), Anderson (2007b), and Sakov and Bertino (2011) for analysis and comparison between different localization methods, and refer to Bishop and Hodyss (2007) and Anderson (2012), and the references therein, for recent developments in adaptive localization methods.

2) Covariance inflation

Covariance inflation algorithms account for the underestimation in the covariance of the forecast ensemble. There are two main types of covariance inflation: additive and multiplicative inflation. In additive inflation algorithms (Hamill and Whitaker 2005; Tong et al. 2016), the forecast covariance $\mathbf{C}_n^f$ in the EnKF is replaced by

$$\hat{\mathbf{C}}_n^f = \mathbf{C}_n^f + \lambda\mathbf{I},$$

for some $\lambda > 0$. In multiplicative inflation algorithms (Anderson and Anderson 1999; Hamill et al. 2001), the spread of the forecast ensemble is inflated by replacing $\mathbf{x}_n^{f,(i)}$ with $\mathbf{x}_n^{f,(i)} + \sqrt{1+\lambda}(\mathbf{x}_n^{f,(i)} - \bar{\mathbf{x}}_n^{f,(i)})$ for some $\lambda > 0$ in the analysis step, which is equivalent to replacing the covariance $\mathbf{C}_n^f$ by

$$\hat{\mathbf{C}}_n^f = (1+\lambda)\mathbf{C}_n^f.$$

This increases the covariance of the forecast ensemble, so as to account for the underestimation of covariance. Inflation has the effect of weighting the observations more than the deficient forecast model and pulling the filter back toward the observations so as to avoid filter divergence. Optimization of the inflation parameter λ is usually done by numerical tuning. To avoid ad hoc tuning and to account for the dynamical changes in the model error, adaptive inflation algorithms have been recently developed by Anderson (2007a, 2009) for multiplicative inflation and by Kelly et al. (2014) and Tong et al. (2015, 2016) for additive inflation.

In the numerical experiments that follow, we use inflation and localization simultaneously, test both additive and multiplicative inflation, and select the best combinations. The main cost is the generation of training data to tune the inflation parameter and the localization radius.

b. Discrete-time stochastic parameterization

The model error in the forecast model [(5)] is $\hat{\mathbf{F}}(\hat{\mathbf{x}}_{n-1}, \hat{\mathbf{y}}_{n-1}) - \mathbf{f}_0(\hat{\mathbf{x}}_{n-1})$, which can be seen by rewriting the first equation in the full system [(6)]:

$$\hat{\mathbf{x}}_n = \mathbf{f}_0(\hat{\mathbf{x}}_{n-1}) + [\hat{\mathbf{F}}(\hat{\mathbf{x}}_{n-1}, \hat{\mathbf{y}}_{n-1}) - \mathbf{f}_0(\hat{\mathbf{x}}_{n-1})].$$

The discrete-time stochastic parameterization method quantifies the model error and produces an improved forecast model. It constructs a non-Markovian nonlinear autoregression moving average (NARMA) forecast model of the form

$$\mathbf{x}_n = \mathbf{f}_0(\mathbf{x}_{n-1}) + \Phi(\mathbf{x}_{n-p:n-1}, \boldsymbol{\xi}_{n-q:n-1}) + \boldsymbol{\xi}_n, \tag{9}$$

where the $\{\boldsymbol{\xi}_n\}$ are independent Gaussian random variables with mean zero and covariance $\mathrm{diag}(\sigma_\xi^2)$. The function $\mathbf{f}_0$ comes from the original forecast model [(5)], and $\Phi(\mathbf{x}_{n-p:n-1}, \boldsymbol{\xi}_{n-q:n-1})$ is a parametric function of the form

$$\begin{aligned}\Phi_n &:= \Phi(\mathbf{x}_{n-p:n-1}, \boldsymbol{\xi}_{n-q:n-1}) \\ &= \sum_{j=1}^{p} a_j \mathbf{x}_{n-j} + \sum_{i=0}^{r}\sum_{j=1}^{p} b_{i,j}\mathbf{f}_i(\mathbf{x}_{n-j}) + \sum_{j=1}^{q} c_j \boldsymbol{\xi}_{n-j},\end{aligned} \tag{10}$$

where $\{a_j, b_{i,j}, c_j, \sigma_\xi^2\}$ are parameters to be estimated and $\{\mathbf{f}_i, i = 1, \ldots, r\}$ are functions to be provided by modelers. The appearance of $\mathbf{f}_0$ in Φ has the effect of modifying the coefficient of $\mathbf{f}_0(\mathbf{x}_{n-1})$ from what it was in the original forecast model.

The NARMA model can capture key statistical and dynamical features of the resolved scales and generate high quality forecast ensembles that have the correct mean and covariance if the ensemble size is sufficient. We emphasize that this is different from simply correcting the ensemble, because the forecast model is improved, and this treats the root of the model error problem.

The main difficulty in this construction is deriving and selecting the ansatz (i.e., the functions $\{\mathbf{f}_i\}$ and the orders $\{p, r, q\}$) of the NARMA model. The ansatz may be derived from the physical properties of the full system, and it may depend on the numerical scheme used in the original reduced model. We refer to Crommelin and Vanden-Eijnden (2008) Majda and Harlim (2013), Kondrashov et al. (2015), Harlim (2016), and Lu et al. (2017, 2016) for further discussion.

Once $\{\mathbf{f}_i\}$ and the orders $\{p, r, q\}$ are fixed, the parameters $\theta = \{a_j, b_{i,j}, c_j, \sigma_\xi^2\}$ are estimated by conditional likelihood methods. We first solve the full system [(6)] offline to generate a time series $\{\hat{\mathbf{x}}_n\}_{n=1}^N$ for a large N. Then, the parameters are estimated as follows. Conditional on $\xi_1, \ldots, \xi_m$, the negative log likelihood of $\{\mathbf{x}_n = \hat{\mathbf{x}}_n\}_{n=m+1}^N$ is

$$\begin{aligned}L(\theta \,|\, \xi_1, \ldots, \xi_m) = &\sum_{k=1}^{d_x}\sum_{n=m+1}^{N} \frac{(\mathbf{x}_{n,k} - \Phi_{n,k})^2}{2\sigma_{\xi,k}^2} \\ &+ \frac{N-q}{2}\log\sigma_{\xi,k}^2,\end{aligned} \tag{11}$$

where $m = \max\{p, q\}$ and $\theta = (a_j^k, b_{i,j}^k, c_j^k, \sigma_{\xi,k}^2)$. For a given value of θ, if $q = 0$, the values of $\{\Phi_n\}_{n=m+1}^N$ can be computed directly from data $\{\mathbf{x}_n\}_{n=1}^N$. If $q > 0$, the values of $\{\Phi_n\}_{n=m+1}^N$ and $\{\boldsymbol{\xi}_n\}_{n=m+1}^N$ can be computed recursively, conditional on $\xi_1 = \cdots = \xi_m = 0$. That is, one computes Φ_{m+1} from $\xi_{m-q+1:m}$ using (10), and computes ξ_{m+1} from Φ_{m+1} using (9); and repeats this process for the rest of the times $n \geq m+1$. The maximum likelihood estimator (MLE) of the parameter is the minimizer of the negative log-likelihood:

$$\hat{\theta}_N = \arg\min_\theta L_N(\theta \,|\, \xi_1, \ldots, \xi_m).$$

If $q = 0$, the minimization reduces to least squares regression. If $q > 0$, the minimization can be done by an iterative least squares approach (Ding and Chen 2005) or other optimization methods.

As in covariance inflation and localization algorithms, the main cost of the discrete-time stochastic parameterization method is the generation of the training

dataset. This requires solving the full system offline for a time interval long enough so that the maximum likelihood estimator, which converges at the rate $1/\sqrt{N}$, is close to its limit. The cost of parameter estimation depends on the NARMA model. It is negligible if the model does not have a moving average term (i.e., if $q = 0$). In this case the maximum likelihood estimator is equivalent to the least squares estimator. The cost varies when $q \neq 0$ since the minimization may need many iterations.

4. Numerical experiments on the Lorenz-96 system

In this section we carry out numerical experiments on the two-layer Lorenz-96 system (Lorenz 1996), which consists of K-resolved variables x_k coupled to $J \times K$ unresolved variables $y_{j,k}$:

$$\frac{d}{dt}x_k = x_{k-1}(x_{k+1} - x_{k-2}) - x_k + F + z_k \quad \text{and}$$
$$\frac{d}{dt}y_{j,k} = \frac{1}{\varepsilon}[y_{j+1,k}(y_{j-1,k} - y_{j+2,k}) - y_{j,k} + h_y x_k],$$

where $z_k = (h_x/J)\sum_j y_{j,k}$, $k = 1, \ldots, K$, and $j = 1, \ldots, J$. The indices are cyclic: $x_k = x_{k+K}$, $y_{j,k} = y_{j,k+K}$ and $y_{j+J,k} = y_{j,k+1}$. The system is invariant under spatial translations, and the statistical properties are identical for all x_k. The formulation here is equivalent to the original formulation by Lorenz (see, e.g., Fatkullin and Vanden-Eijnden 2004; Crommelin and Vanden-Eijnden 2008; Kwasniok 2012). The parameter ε measures the scale separation between the resolved variables x_k and the unresolved variables $y_{j,k}$. We set $\varepsilon = 0.5$, so that there is no significant scale separation between the resolved and unresolved processes, as is both more realistic and more difficult to handle for parameterizations [see Fatkullin and Vanden-Eijnden (2004) and references therein]. We take $K = 18$, $J = 20$, $F = 10$, $h_x = -1$ and $h_y = 1$. Here, one model time unit is approximately equal to five atmospheric days, deduced by comparing the error-doubling time of the model to that observed in the atmosphere (Lorenz 1996; Arnold et al. 2013; Mitchell and Carrassi 2015).

In the experiments, we take a trajectory of the resolved variables x in the full system to be the truth. We solve the full system by a fourth-order Runge–Kutta method with a time step $dt = 0.001$, and make recordings every 50 steps, that is, with observation spacing $h = 0.05$, approximately six atmospheric hours. To eliminate transients, we begin to make observations after running the full model for 100 time units. To create noisy observations, we add to the recorded trajectory independent Gaussian random vectors with mean zero and covariance $\mathbf{R} = \sigma_\varepsilon^2\mathbf{I}$.

In the data assimilation, we assume that we cannot afford to solve the full Lorenz-96 system for ensemble forecasts and use a reduced system obtained by discarding the y variables:

$$\frac{d}{dt}x_k = x_{k-1}(x_{k+1} - x_{k-2}) - x_k + F,$$

for $k = 1, \ldots, K$. After discretization by a fourth-order Runge–Kutta method with time step h (i.e., the observation spacing), one obtains a system of difference equations:

$$x_{k,n} = x_{k,n-1} + f_k^h(x_{\cdot,n-1}), \tag{12}$$

for $k = 1, \ldots, K$, where $x_{k,n}$ is the value of the component x_k at time n and $x_{\cdot,n-1}$ denotes the vector of the K-resolved variables at time $n - 1$. Hereafter, we refer to this reduced discretized model as the L96x model, and we refer to the discrete representation of the full L96 system as the full model.

In the following, we first implement the two methods reviewed in section 3 to account for such model error in sections 4a and 4b, and we then compare their filtering and forecasting performance in sections 4c and 4d.

a. *Accounting for model error by discrete-time stochastic parameterization*

Discrete-time stochastic parameterization quantifies the model error of the L96x model and produces an improved forecast model, which we call the NARMA model, as introduced in section 3b. Specifically, this is done by using the conditional likelihood method to fit a NARMA model to a set of training data, which is generated by solving the full model over a long time. The initial conditions in the simulation that generates training data can be arbitrary, because the estimated parameters of the NARMA model will converge as the length of the training data increases, because of the ergodicity of the full system (Chorin and Lu 2015). According to the results in Chorin and Lu (2015), we use a NARMA(2,0) model:

$$x_{k,n} = \sum_{j=1}^{2}[a_j x_{k,n-j} + b_j f_k^h(x_{\cdot,n-j})] + c_0 + c_1 x_{k,n-1}^2 + c_2 x_{k,n-1}^3 + \boldsymbol{\xi}_{k,n}, \tag{13}$$

where $f_k^h(x)$ comes from the right-hand side of (12) and $\{\boldsymbol{\xi}_{k,n}\}$ is a sequence of independent Gaussian random vectors with mean zero and covariance $\sigma_\xi^2\mathbf{I}$. The parameters in the different components are the same because of the symmetry in the equations.

The main cost in deriving the NARMA representation is the generation of training data. The cost of the NARMA parameter estimation is negligible compared

TABLE 1. Values of the parameters in the NARMA model.

a_1	a_2	b_1	b_2
1.8992	−0.9022	0.9946	−0.9058
c_0	c_1 ($\times 10^{-5}$)	c_2 ($\times 10^{-5}$)	σ_ξ
0.0024	−0.3903	0.9396	0.0084

with cost of generating the training data, because the model does not have moving average terms and the optimization reduces to linear least squares. In our tests, the training data were generated by solving the full system with step size $dt = 0.001$ and recording data every 50 steps (i.e., with observation spacing $h = 0.05$). Table 1 shows the values of the parameters (a_j, b_j, c_j, σ_ξ) estimated from a training dataset of length $N = 10^5$ (i.e., 5000 time units, approximately equal to 69 atmospheric years). Further tests showed that a data length of $N = 10^4$ could also lead to models with good statistical properties. The minimum data length necessary to identify a NARMA model is problem dependent, and a general criterion is beyond the scope of the current study.

NARMA AS AN IMPROVED FORECAST MODEL FOR THE L96X MODEL

Figure 1 shows the empirical probability density function (PDF) and the autocorrelation function (ACF) of the full model, the L96x model, and the NARMA model, computed from time averaging of a long trajectory of each model. The NARMA model reproduces the PDF and the ACF faithfully, while the L96x model misses the shape of the PDF and the oscillation of the ACF. The PDF approximates the invariant measure of the large-scale variables, and the ACF approximates the dynamical transition. Hence, the NARMA model captures the statistical and dynamical features of the large-scale variables much better than the L96x model.

By accounting for the model error, the NARMA model significantly improves state estimation of the filters over the L96x model. Table 2 shows the mean and standard deviation of the relative errors of state estimation from 100 simulations, in which the variance of the observation noise is $\sigma_\varepsilon = 0.2$ and the ensemble size is $M = 1000$. Here, we judge the quality of the state estimates by the relative error in the ensemble means, that is, the relative difference between the ensemble means and the truth:

$$e_{\text{rel}} = \left(\sum_{n=N_0+1}^{N} \sum_{k=1}^{K} |\overline{x_{k,n}} - x_{k,n}|^2 \Big/ \sum_{n=N_0+1}^{N} \sum_{k=1}^{K} |x_{k,n}|^2 \right)^{1/2}, \tag{14}$$

where $(\overline{x_{k,n}}, n = N_0 + 1, \ldots, N)$ are the ensemble means and $(x_{k,n}, n = N_0 + 1, \ldots, N)$ are the true state values.

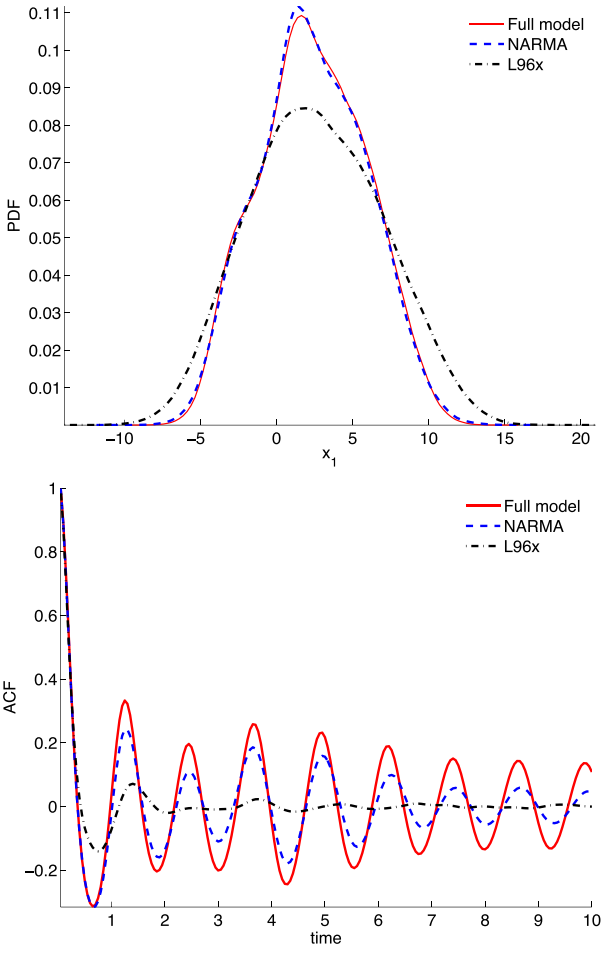

FIG. 1. Empirical (top) PDF and (bottom) ACF of the full model, the L96x model, and the NARMA model.

We skipped the first N_0 steps so as to eliminate the transients in assimilation. In the tests, it took only a few steps for the filters to reach a stationary state, so we took $N = 400$ and $N_0 = 200$. We implemented both the standard EnKF and the EnKF with the block update using block length $L = 2$. In both cases, the NARMA model successfully reduced the relative error in the state estimation to below 2.10%, which is the relative uncertainty induced by the observation noise; the filter with the L96x

TABLE 2. The mean and standard deviation of the relative errors of state estimation on 100 simulations, in which the ensemble size is $M = 1000$ and the variance of the observation noise is $\sigma_\varepsilon = 0.2$. Both the standard EnKF and the EnKF with the block update algorithm are implemented, with the L96x and NARMA models as the forecast model.

	Standard EnKF	EnKF with block update
L96x	0.7884 ± 0.0774	0.8022 ± 0.0818
NARMA	0.0182 ± 0.0016	0.0156 ± 0.0011

model performs very poorly, as a result of the model error.

We also tested a parameterization using a Markovian model in the form of NARMA (1, 0) similar to (13). The Markovian model reproduced the empirical PDF and ACF well, but is slightly inferior to the NARMA model (data not shown here). The Markovian model successfully reduced the relative error to 0.0210 ± 0.0022 in the above 100 simulations, which is slightly larger than those of the NARMA model (0.0156 ± 0.0019). Also, the NARMA model yielded better forecast performance than the Markovian model. Therefore, it is important to choose a good model for the model error, and we consider only the NARMA model in this study.

These results show that discrete-time stochastic parameterization can effectively account for model error and, therefore, improve the performance of ensemble filter.

Note also that the block update algorithm reduces the error of the state estimation for the NARMA model, but it does not improve the performance of the filter with the L96x model. Hence, in the following tests, we use the block update algorithm for the NARMA model and the standard EnKF for the L96x model.

b. *Accounting for model error by tuning inflation and localization*

Covariance inflation and localization can account for both model error and sampling error, but the parameterization can only reduce the model error. To compare their effectiveness in accounting for model error, we consider two situations: one with an ensemble sufficiently large for sampling error to be negligible and one with a practical small ensemble. In the first situation, we compare the filter performance of the NARMA model using no IL, with the performance of the L96x model using the best-tuned IL. This highlights the impact of the two methods on accounting for model error. In the second situation, we apply IL to both the L96x and the NARMA models; in the L96x model, IL accounts for both sampling error and model error; in the NARMA model, IL accounts mainly for sampling error.

We also test the standard EnKF using the full model, which has no model error, as the forecast model, so as to provide a useful yardstick for assessing the results.

We carry out the covariance localization with the localization matrix $\mathbf{C}_{r_{\rm loc}}$ defined in (8), using the Gaspari–Cohn taper function [(7)], where $r_{\rm loc}$ is the localization radius. We also tested a Toeplitz circulant matrix with exponential spectrum decay, but there is no clear improvement in filter performance over the Gaspari–Cohn matrix (data not shown here). In the EnKF with the block update, the localization matrix is an array containing L copies of $\mathbf{C}_{r_{\rm loc}}$ in the row and column dimensions.

We tune the localization and inflation by trying different values of $r_{\rm loc}$ and λ for filtering a single set of observations with noise variance $\sigma_\varepsilon = 0.2$. Both additive and multiplicative inflation were tested, and additive inflation led to slightly better filter performance for both the full and the L96x models (data are not shown here). Hence, in the following we only consider additive inflation.

1) TUNING IN THE CASE OF SUFFICIENT ENSEMBLE SIZE

We first discuss tuning in the case where the sample size is sufficiently large for the sampling error to be negligible. Here, a large ensemble with $M = 1000$ members is found to be sufficient. For the computational cost to be similar to that of the L96x and NARMA models, the full model uses an ensemble of size $M = 10$, with IL to account for the sampling error because of insufficient size. Tests showed that IL was able to effectively account for the sampling error, yielding state estimations almost as accurate as the full model with an ensemble size $M = 1000$.

Figure 2 shows the relative errors in scaled colors (the darker the color, the smaller the relative error in state estimation) for different $r_{\rm loc}$ and additive inflation λ. Here, a localization radius $r_{\rm loc} = 0$ means no localization, and an additive inflation value $\lambda = 0$ means no inflation. To demonstrate the need of tuning for different models, common values of $r_{\rm loc}$ and λ are plotted. The best-tuned values shown here may not be optimal, but they are close to the optimal values in finer tuning.

The left plot in Fig. 2 shows the relative errors of the L96x model with IL. Because of the model error, the L96x model performs poorly without IL ($r_{\rm loc} = 0, \lambda = 0$). As the additive inflation parameter λ increases, the relative error in state estimation first sharply decreases and then slightly increases; a similar pattern can be observed as the localization radius $r_{\rm loc}$ increases. To select the best values for $r_{\rm loc}$ and λ, we do not choose the pair $(r_{\rm loc}, \lambda)$ that produces the smallest relative error in the array, but rather the pair at the intersection of the column and the row that have the smallest sum of relative errors among the columns and rows, respectively. This is because tests show that the pair that yields the smallest error is sensitive to various factors, such as the number of observations and the initial conditions used to generate the training data, while the pairs at the intersection are much more robust to these factors. For the L96x model, this strategy yields $(r_{\rm loc}, \lambda) = (2, 0.1)$.

In the center panel in Fig. 2 we show the parameter values for tuning IL for the NARMA model. IL brings negligible improvements for the NARMA model: the relative error decreases only from 0.016 to 0.014 in this

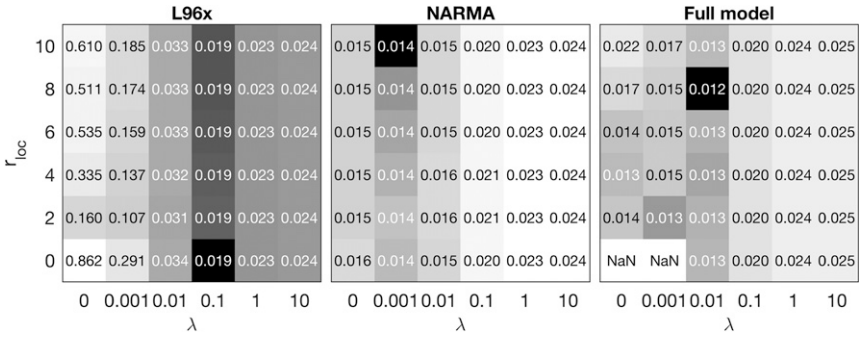

FIG. 2. Relative error of the ensemble fitter for different covariance localizations and additive inflations, with ensemble size $M = 10$ for the full model and $M = 1000$ for the reduced models. The letters NaN indicate that the filter diverged. Here, a localization radius $r_{loc} = 0$ means no localization. An additive inflation $\lambda = 0$ means no inflation.

simulation (results are similar for other simulations). This suggests that the NARMA model has accounted for the model error so well that IL cannot offer much help.

The right plot in Fig. 2 shows the relative errors in filtering with the full model. Because of the sampling error caused by the small ensemble size, the EnKF with the full model diverges if no localization or inflation is used. IL accounts for the sampling error, stabilizes the filter, and leads to accurate state estimation with relative error 0.013, while the relative error of the full model with $M = 1000$ is 0.011. The best values for the IL parameters in this setup are $r_{loc} = 2$ and $\lambda = 0.01$.

In summary, when comparing filter performance in the case of large ensemble size in section 4c, we use $r_{loc} = 2$ and $\lambda = 0.1$ for the L96x model and $r_{loc} = 2$ and $\lambda = 0.01$ for the full model. For the NARMA model, we use a block updating algorithm without any localization or inflation. We found that even when IL is tuned, the filter with the full model may diverge (with a frequency of about 2 out of 100 simulations). Since the full model only serves as a reference, we drop the simulation when the filter diverges.

2) TUNING IN THE CASE OF SMALL ENSEMBLE SIZE

We use the same tuning strategy as above for different small ensemble sizes, ranging from 10 to 100. Table 3 shows the best pair of the localization radius r_{loc} and the additive inflation parameter λ in the case of ensemble size $M = 10$. The best pair (r_{loc}, λ) for the L96x model did not change much when the ensemble size changed. The best pairs of (r_{loc}, λ) for the NARMA model and the full model were sensitive to changes in ensemble size, with λ varying between 0.001 and 0.01 and r_{loc} varying between 2 and 10. But the relative errors corresponding to these pairs in the array were very close to each other (data not shown here, but this can be readily seen from the center and right plots in Fig. 2). Therefore, we accept these suboptimal pairs and use them in Table 3 for other ensemble sizes when comparing filter performance in section 4d.

c. *Filter performance comparison: The case of sufficient ensemble size*

We consider first the case of a large ensemble size $M = 1000$. This setup aims to answer the main question of this paper: whether the parameterization can be as effective as IL in accounting for the model error due to unresolved scales. With this M, the sampling error in the ensemble covariance is negligible compared to the model error; therefore, the filter performance depends on how well the two methods can account for the model error.

Their performance is measured by the resulting state estimates and ensemble forecasts. We first compare them in a single simulation, and then we consider the statistics of the errors over 100 simulations. Results from the full model, with ensemble size $M = 10$ and IL with $(r_{loc}, \lambda) = (2, 0.01)$ are included to provide a sense of the best possible results at a comparable computational cost.

1) STATE ESTIMATION

The trajectories in a short single simulation with observation noise $\sigma_\varepsilon = 0.2$ are shown in Fig. 3. The filtered trajectories (the magenta lines) are in the time interval $[0, 6]$. The ensemble and its mean (the black dash dot

TABLE 3. The best-tuned values of localization radius r_{loc} and additive inflation parameter λ for the three models using ensemble size $M = 10$.

	L96x	NARMA	Full model
Localization radius r_{loc}	2	2	2
Additive inflation λ	0.1	0.01	0.01

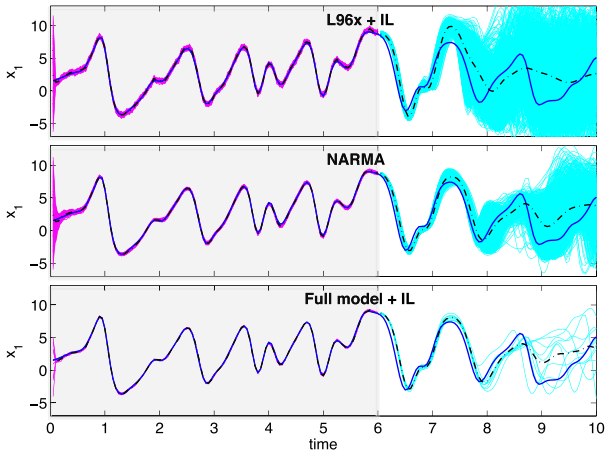

FIG. 3. Ensembles of trajectories in filtering and forecasting. Each plot contains a true trajectory (blue line), an ensemble of filtering trajectories (magenta lines) in the time interval [0, 6] (in gray shading) and forecasting trajectories (cyan lines) in the time interval (6, 10], and the ensemble mean (black dash–dot line). Covariance IL accounts for the model error of the L96x model in the filter, and the parameterization reduces the model error through the NARMA model, in the case of a large ensemble size ($M = 1000$). The full model provides a yardstick for performance at a comparable cost by using tuned IL with ensemble size $M = 10$.

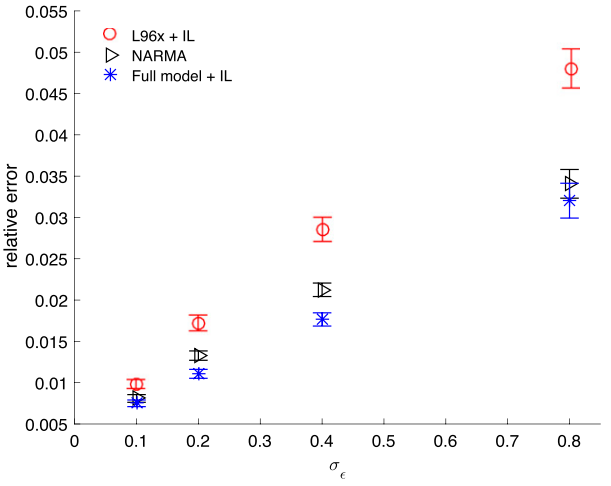

FIG. 4. Mean and standard deviation of the relative error of state estimation calculated with 100 simulations for different variances of observation noise. The ensemble size is 1000 for the L96x and NARMA model, and 10 for the full model. The L96x model and the full model use tuned IL, and NARMA uses neither.

lines) follow the true trajectory (the blue line) relatively well for all the three forecast models. The relative errors of state estimation are 1.91%, 1.59%, and 1.31% for L96x, NARMA, and the full model, respectively.

The difference in state estimation is clear in the statistics of the relative error in 100 simulations, as shown in Fig. 4. To test the robustness of the filter, we consider different variances of observation noise, with σ_ε taking the values {0.1, 0.2, 0.4, 0.8}, for which the relative error of the observation noise ranges from 1.05% to 8.40%. The NARMA model, using no covariance localization or inflation, has smaller errors than the L96x model using tuned IL. For example, in the case $\sigma_\varepsilon = 0.2$, the average relative errors are 1.73%, 1.33%, and 1.11% for L96x, NARMA and the full model, respectively. The relative error of the L96 model is about 1.3 times the relative error of the NARMA model. This shows that the stochastic parameterization is more effective than IL in dealing with model error. On the other hand, with the help of IL, the full model with a small ensemble size has slightly smaller errors than the NARMA model. This indicates that 1) tuned covariance IL is effective in dealing with sampling error and 2) there is still model error in the NARMA model.

2) FORECASTING

The goal of state estimation is to provide the initial conditions for the forecast model to use in forecasting the future evolution of the resolved scales. After assimilating the last observation, the filtering ensemble provides the desired initial conditions, and by running the forecast model, we obtain a forecasting ensemble.

The difference in the ensemble forecasts of these models is clear in the single simulation shown in Fig. 3. The forecasting trajectories (the cyan lines) are in the time interval (6, 10]. It is desirable that the ensemble of forecasting trajectories follows the true trajectory as long as possible before spreading out. The ensemble of the full model follows the true trajectory for about 2.5 time units (from $t = 6$ to 8.5), and the L96x and NARMA models for about 1 and 1.8 time units, respectively. The ensemble means keep following the true trajectory slightly longer. This shows that the NARMA model has better prediction skills than the L96x model in this simulation.

The improved forecast of the NARMA model over the L96x + IL combination can be attributed to two factors: a better forecast model and more accurate two-step initial distributions. To disentangle these two factors, we tested a Markovian model in the form of NARMA(1, 0) simultaneously with the above non-Markovian NARMA and L96x model. Results showed that the Markovian NARMA (1, 0) model made forecasts that were slightly inferior to the NARMA model but much better than L96x + IL, while it has relative errors in state estimation similar to L96x+ IL. This suggests that the improvement of the forecast of the NARMA model over L96x + IL comes mainly from the better forecast model.

We further compare the forecast performance over 100 simulations by studying the root-mean-square

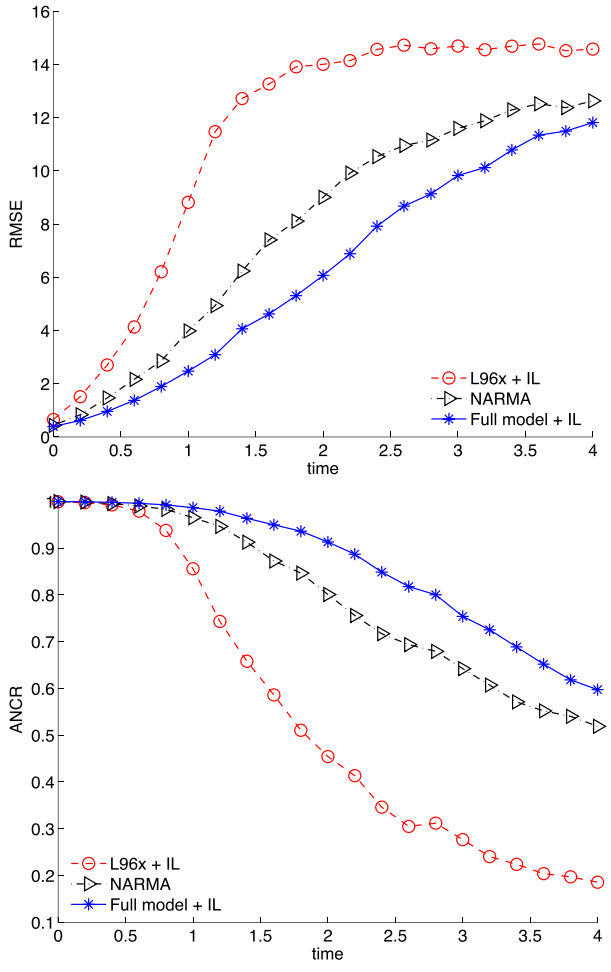

FIG. 5. (top) RMSE and (bottom) ANCR of ensemble forecasting on 100 simulations, with $\sigma_\varepsilon = 0.2$. The ensemble size is 1000 for the L96x and NARMA models, and the ensemble size is 10 for the full model. The L96x model and the full model use tuned IL, and NARMA uses neither.

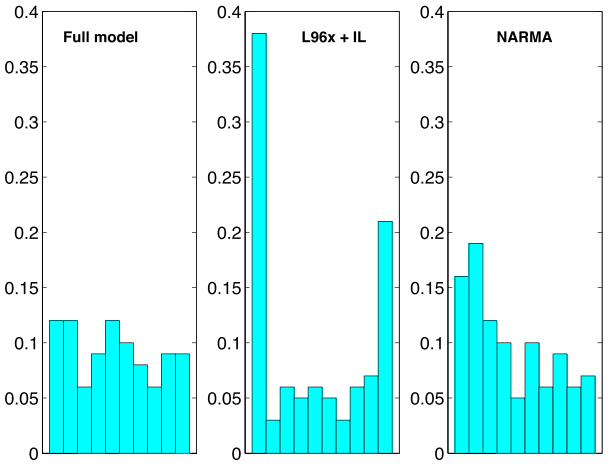

FIG. 6. Rank histograms of the 100 simulations at lead time $\tau = 1.6$. An ideal rank histogram should be flat. The full model has the flattest rank histogram and NARMA's rank histogram is close to flat, but L96x + IL has a U-shaped rank histogram.

error (RMSE) and the anomaly correlation (ANCR) between the mean trajectories of the forecast ensembles and the true trajectories. The RMSE measures the average difference between trajectories whereas the ANCR measures the average correlation between them (Crommelin and Vanden-Eijnden 2008). Figure 5 shows the RMSE and the ANCR results of the forecast ensemble in the 100 simulations, when $\sigma_\varepsilon = 0.2$. A small and slowly increasing RMSE, combined with a large and slowly decreasing ANCR, indicates a good level of forecast performance. The NARMA model shows a significant improvement over the L96x model and is close to the full model, which has the smallest RMSE and the largest ANCR. With the threshold of RMSE less than 9 and the ANCR larger than 0.8, the forecast time of the NARMA model is about two time units (approximately 10 atmospheric days), which is double that of the L96x model's one time unit (approximately 5 atmospheric days), and is slightly less than the full model's 2.5 time units (approximately 13 atmospheric days). We also computed the rank histogram (Crommelin and Vanden-Eijnden 2008) at lead time 1.6 (see Fig. 6). The rank histogram of the full model is almost flat, as desired, and NARMA has a rank histogram close to flat, but L96x + IL has a U-shaped rank histogram. We also compared the RMSE with ensemble spread (i.e., the trace of the ensemble covariance) in Fig. 7. The ensemble spread of the full model matches the RMSE well, and the ensemble spread of NARMA is close to the RMSE, but there is a sizeable mismatch between the ensemble spread and the RMSE for L96x + IL. The results for the other values of σ_ε are similar (data not shown here).

In short, the NARMA model delivers significantly better state estimation and prediction performance than the L96x model, and its performance is close to that of the full model. Recall that the NARMA accounts for the model error by discrete-time stochastic parameterization of the unresolved scales, while the L96x model accounts for the model error by covariance inflation and localization. This suggests that the discrete-time stochastic parameterization is more effective in dealing with model error than covariance inflation and localization.

d. Filter performance comparison: The case of small ensemble size

Because of limited computational resources, in many applications one can afford only a small ensemble, and significant sampling error may be present. In this case,

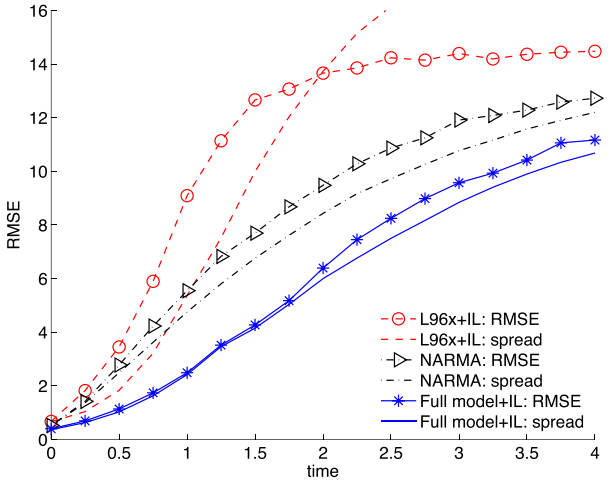

FIG. 7. Comparison of RMSE with ensemble spread (i.e., the trace of the ensemble covariance). The ensemble spread of the full model matches the RMSE well, and the ensemble spread of NARMA is close to the RMSE, but there is a sizeable mismatch between the ensemble spread and the RMSE for L96x + IL.

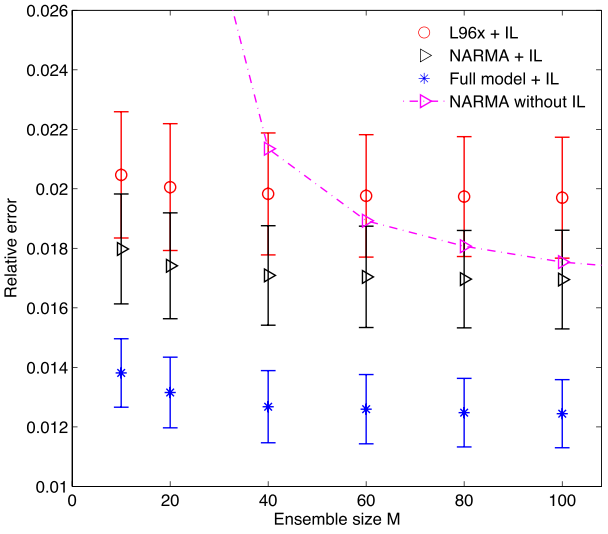

FIG. 8. Mean and standard deviation of the relative error of the state estimation in the EnKF with different ensemble sizes in 100 simulations. All three of the models use IL. The mean of the NARMA model without IL is also plotted to indicate that sampling error decreases as ensemble size increases.

localization and/or inflation are needed to account for the sampling error.

In this section, we compare the filter performance for several small ensemble sizes, ranging from $M = 10$ to 100, with all the models using tuned IL with parameters given in Table 3. In all of the cases, the variance of the observation noise is $\sigma_\varepsilon = 0.2$.

1) STATE ESTIMATION

Figure 8 shows the means and standard deviations of the relative errors in state estimation on 100 simulations, with several small ensemble sizes. With tuned IL, the NARMA model (black triangle) has smaller errors than the L96x model (red circle) for all sizes. Recall that IL accounts for both sampling error and model error for the L96x model, while in the filter with the NARMA model, the combination mainly accounts for the sampling error while the stochastic parameterization accounts for model error. This shows that the parameterization treats the model error more effectively than IL and improves the filter performance.

We also tested the NARMA model without using inflation or localization (cyan triangle with dash–dot line). Its error decreases much faster than those using inflation and localization as the ensemble size increases. In particular, NARMA has smaller errors than L96x with tuned IL when the ensemble size is larger than 60. Also, its performance becomes close to that of NARMA with IL when the ensemble size is 100. This indicates that 1) the NARMA model has effectively reduced the model error of the L96x model and 2) the sampling error becomes small when the ensemble size reaches $M = 100$. (It also verifies that the size $M = 1000$ used in section 4c is sufficiently large to make the sampling error negligible.)

2) FORECASTING

Figure 9 shows the RMSE and the ANCR of the forecast ensemble in 100 simulations, with all models using ensemble size $M = 10$ and tuned IL. The NARMA model is a clear improvement over the L96x model in forecasting: its RMSE increases much slower and its ANCR decreases much slower. But the gap between the NARMA model and the full model is slightly larger than the gap in Fig. 5, where a large ensemble size was used. Here, the forecast time of the NARMA model is about 1.5 time units (approximately 8 atmospheric days), which is 50% more than the L96x model's one time unit (approximately 5 atmospheric days), and it is less than the full model's 2.5 time units (approximately 13 atmospheric days).

In short, in cases with insufficient ensemble size, the NARMA model offers better state estimation and prediction properties than the L96x model, when both use tuned IL. Covariance inflation and localization account for both sampling error and model error for the L96x model; they mainly account for the sampling error for the NARMA model, which has quantified the model error by parameterization. Hence, the discrete-time stochastic parameterization can be combined with covariance inflation and localization to improve filter performance.

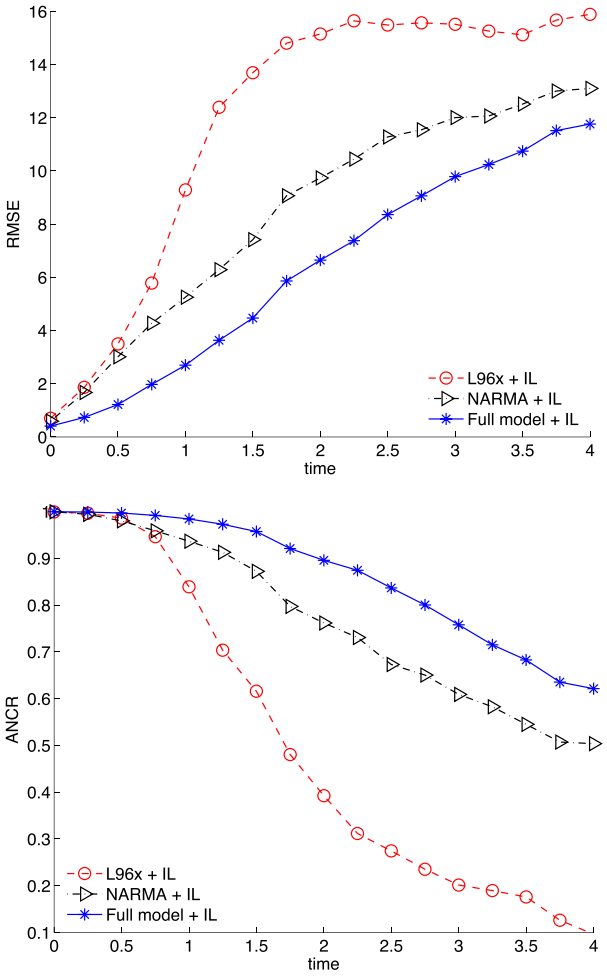

FIG. 9. (top) RMSE and (bottom) ANCR of ensemble forecasting on 100 simulations, with $\sigma_\varepsilon = 0.2$. All the models use ensemble size $M = 10$ and tuned IL.

5. Summary and discussion

We have examined discrete-time stochastic parameterization as a way of accounting for the model error due to unresolved scales in the version of EnKF with perturbed observations and compared it with covariance inflation and localization algorithms.

We carried out numerical experiments on the two-layer Lorenz-96 system, with the goal of predicting the future evolution of the observed variables on the basis of noisy observations of these variables. We assumed that a forecast model in the filter was a truncated system in which the unobserved variables were unresolved. The model error comes from this underresolution. We analyzed how the two methods accounted for this error. The stochastic parameterization method directly quantified the model error and led to an improved forecast model, while covariance inflation and localization corrected the ensemble covariance in the analysis step in the filter. When the ensemble size was sufficiently large for the sampling error to be negligible, the improved forecast model, without any inflation or localization, achieved significantly better performance in state estimation and prediction than the unmodified truncated forecast model with tuned inflation and localization. When the ensemble size was small, covariance inflation and localization were needed to account for the sampling error, but the improved forecast model provided further improvement in filter performance. These results show that the discrete-time stochastic parameterization approach was more effective than the inflation and localization approach in dealing with model error from unresolved scales.

As a consequence of this study, we advocate the direct approach, which works on the root of the problem: the deficiency of the model. The direct approach improves the forecast model and, therefore, improves the overall quality of the forecast ensemble as well as the filtering and prediction performance (Harlim 2016; Chorin et al. 2016). This is fundamentally different from the covariance inflation and localization approach, which corrects the sample covariance to improve ensemble quality but permits the model deficiency to remain. However, the parameterization can only account for model error, but covariance inflation and localization can account for both sampling and model error. When there are both model error and sampling error because of small ensemble size, these two methods can work together to achieve better performance than inflation/localization used alone.

This study has been carried out in a setting where the full model can be solved offline, and its solution used to tune inflation and localization or to infer parameters in stochastic parameterization. A more challenging and realistic setting would be one where the full model is unknown, and one has to use noisy observations to infer a parameterization (Li et al. 2009; Berry and Harlim 2014; Harlim 2016). This is the challenging topic of parameter estimation for hidden Markov and non-Markov models (Kantas et al. 2009). We leave it to future work.

Acknowledgments. The authors thank the anonymous reviewers for helpful comments, and thank Prof. Kevin Lin, Prof. Matthias Morzfeld, Dr. Robert Saye, and Mr. Michael Lindsey for helpful discussions. AJC and FL are supported in part by the director, Office of Science, Computational and Technology Research, U.S. Department of Energy, under Contract DE-AC02-05CH11231, and by the National Science Foundation under Grant DMS-1419044. XT is supported in part by

the National Science Foundation under Grant DMS-1419069.

REFERENCES

Anderson, J. L., 2007a: An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus*, **59A**, 210–224, doi:10.1111/j.1600-0870.2006.00216.x.

——, 2007b: Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D*, **230**, 99–111, doi:10.1016/j.physd.2006.02.011.

——, 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus*, **61A**, 72–83, doi:10.1111/j.1600-0870.2008.00361.x.

——, 2012: Localization and sampling error correction in ensemble Kalman filter data assimilation. *Mon. Wea. Rev.*, **140**, 2359–2371, doi:10.1175/MWR-D-11-00013.1.

——, and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758, doi:10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2.

Arnold, H. M., I. M. Moroz, and T. N. Palmer, 2013: Stochastic parametrization and model uncertainty in the Lorenz'96 system. *Philos. Trans. Roy. Soc. London*, **371A**, doi:10.1098/rsta.2011.0479.

Berry, T., and J. Harlim, 2014: Linear theory for filtering nonlinear multiscale systems with model error. *Proc. Roy. Soc. London*, **470A**, doi:10.1098/rspa.2014.0168.

Bishop, C. H., and D. Hodyss, 2007: Flow-adaptive moderation of spurious ensemble correlations and its use in ensemble-based data assimilation. *Quart. J. Roy. Meteor. Soc.*, **133**, 2029–2044, doi:10.1002/qj.169.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, doi:10.1002/qj.49712556006.

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, **126**, 1719–1724, doi:10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

Chekroun, M. D., D. Kondrashov, and M. Ghil, 2011: Predicting stochastic systems by noise sampling, and application to the El Niño-Southern Oscillation. *Proc. Natl. Acad. Sci. USA*, **108**, 11 766–11 771, doi:10.1073/pnas.1015753108.

Chorin, A. J., and O. H. Hald, 2013: *Stochastic Tools in Mathematics and Science.* 3rd ed. Texts in Applied Mathematics, Vol. 58, Springer-Verlag, 200 pp.

——, and F. Lu, 2015: Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. *Proc. Natl. Acad. Sci. USA*, **112**, 9804–9809, doi:10.1073/pnas.1512080112.

——, O. H. Hald, and R. Kupferman, 2000: Optimal prediction and the Mori–Zwanzig representation of irreversible processes. *Proc. Natl. Acad. Sci. USA*, **97**, 2968–2973, doi:10.1073/pnas.97.7.2968.

——, ——, and ——, 2002: Optimal prediction with memory. *Physica D*, **166**, 239–257, doi:10.1016/S0167-2789(02)00446-3.

——, F. Lu, R. M. Miller, M. Morzfeld, and X. Tu, 2016: Sampling, feasibility, and priors in data assimilation. *Discrete Contin. Dyn. Syst.*, **36A**, 4227–4246, doi:10.3934/dcds.2016.36.8i.

Crommelin, D., and E. Vanden-Eijnden, 2008: Subgrid-scale parameterization with conditional Markov chains. *J. Atmos. Sci.*, **65**, 2661–2675, doi:10.1175/2008JAS2566.1.

Danforth, C. M., and E. Kalnay, 2008: Using singular value decomposition to parameterize state-dependent model errors. *J. Atmos. Sci.*, **65**, 1467–1478, doi:10.1175/2007JAS2419.1.

Dee, D. P., and A. M. Da Silva, 1998: Data assimilation in the presence of forecast bias. *Quart. J. Roy. Meteor. Soc.*, **124**, 269–295, doi:10.1002/qj.49712454512.

Ding, F., and T. Chen, 2005: Identification of Hammerstein nonlinear ARMAX systems. *Automatica*, **41**, 1479–1489, doi:10.1016/j.automatica.2005.03.026.

Doucet, A., M. Briers, and S. Sénécal, 2006: Efficient block sampling strategies for sequential Monte Carlo methods. *J. Comput. Graph. Stat.*, **15**, 693–711, doi:10.1198/106186006X142744.

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, doi:10.1029/94JC00572.

——, 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, doi:10.1175/1520-0493(1996)124<0085:AOGADF>2.0.CO;2.

——, and P. Van Leeuwen, 2000: An ensemble Kalman smoother for nonlinear dynamics. *Mon. Wea. Rev.*, **128**, 1852–1867, doi:10.1175/1520-0493(2000)128<1852:AEKSFN>2.0.CO;2.

Fatkullin, I., and E. Vanden-Eijnden, 2004: A computational strategy for multi-scale systems with applications to Lorenz'96 model. *J. Comput. Phys.*, **200**, 605–638, doi:10.1016/j.jcp.2004.04.013.

Furrer, R., and T. Bengtsson, 2007: Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.*, **98**, 227–255, doi:10.1016/j.jmva.2006.08.003.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, doi:10.1002/qj.49712555417.

Gottwald, G. A., and J. Harlim, 2013: The role of additive and multiplicative noise in filtering complex dynamical systems. *Proc. Roy. Soc. London*, **469A**, doi:10.1098/rspa.2013.0096.

——, D. Crommelin, and C. Franzke, 2015: Stochastic climate theory. *Nonlinear and Stochastic Climate Dynamics*, C. L. E. Franzke and T. J. O'Kane, Eds., Cambridge University Press, 209–240.

Hamill, T. M., and J. S. Whitaker, 2005: Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Mon. Wea. Rev.*, **133**, 3132–3147, doi:10.1175/MWR3020.1.

——, ——, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790, doi:10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.

Harlim, J., 2016: Model error in data assimilation. *Nonlinear and Stochastic Climate Dynamics*, C. L. E. Franzke, and T. J. O'Kane, Eds., Cambridge University Press, 276–317.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811, doi:10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

——, ——, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Mon. Wea. Rev.*, **137**, 2126–2143, doi:10.1175/2008MWR2737.1.

Kantas, N., A. Doucet, S. S. Singh, and J. M. Maciejowski, 2009: An overview of sequential Monte Carlo methods for parameter

estimation in general state-space models. *Proc. Symp. on System Identification*, Saint-Malo, France, International Federation of Automatic Control. [Available online at http://wwwf.imperial.ac.uk/~nkantas/sysid09_final_normal_format.pdf.]

Kelly, D., K. Law, and A. Stuart, 2014: Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time. *Nonlinearity*, **27**, 2579, doi:10.1088/0951-7715/27/10/2579.

Khare, S., J. Anderson, T. Hoar, and D. Nychka, 2008: An investigation into the application of an ensemble Kalman smoother to high-dimensional geophysical systems. *Tellus*, **60A**, 97–112, doi:10.1111/j.1600-0870.2007.00281.x.

Kondrashov, D., M. D. Chekroun, and M. Ghil, 2015: Data-driven non-Markovian closure models. *Physica D*, **297**, 33–55, doi:10.1016/j.physd.2014.12.005.

Kwasniok, F., 2012: Data-based stochastic subgrid-scale parametrization: An approach using cluster-weighted modeling. *Philos. Trans. Roy. Soc. London*, **370A**, 1061–1086, doi:10.1098/rsta.2011.0384.

Lei, J., P. Bickel, and C. Snyder, 2010: Comparison of ensemble Kalman filters under non-Gaussianity. *Mon. Wea. Rev.*, **138**, 1293–1306, doi:10.1175/2009MWR3133.1.

Li, H., E. Kalnay, T. Miyoshi, and C. M. Danforth, 2009: Accounting for model errors in ensemble data assimilation. *Mon. Wea. Rev.*, **137**, 3407–3419, doi:10.1175/2009MWR2766.1.

Lorenz, E. N., 1996: Predictability: A problem partly solved. *Proc. Seminar on Predictability*, Reading, United Kingdom, ECMWF, 1–18.

Lu, F., K. K. Lin, and A. J. Chorin, 2016: Comparison of continuous and discrete-time data-based modeling for hypoelliptic systems. *Commun. Appl. Math. Comput. Sci.*, **11**, 187–216, doi:10.2140/camcos.2016.11.187.

——, ——, and ——, 2017: Data-based stochastic model reduction for the Kuramoto–Sivashinsky equation. *Physica D*, **340**, 46–57, doi:10.1016/j.physd.2016.09.007.

Majda, A. J., and J. Harlim, 2013: Physics constrained nonlinear regression models for time series. *Nonlinearity*, **26**, 201–217, doi:10.1088/0951-7715/26/1/201.

Meng, Z., and F. Zhang, 2007: Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part II: Imperfect model experiments. *Mon. Wea. Rev.*, **135**, 1403–1423, doi:10.1175/MWR3352.1.

Mitchell, H. L., and P. L. Houtekamer, 2000: An adaptive ensemble Kalman filter. *Mon. Wea. Rev.*, **128**, 416–433, doi:10.1175/1520-0493(2000)128<0416:AAEKF>2.0.CO;2.

Mitchell, L., and G. A. Gottwald, 2012: Data assimilation in slow–fast systems using homogenized climate models. *J. Atmos. Sci.*, **69**, 1359–1377, doi:10.1175/JAS-D-11-0145.1.

——, and A. Carrassi, 2015: Accounting for model error due to unresolved scales within ensemble Kalman filtering. *Quart. J. Roy. Meteor. Soc.*, **141**, 1417–1428, doi:10.1002/qj.2451.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quart. J. Roy. Meteor. Soc.*, **127**, 279–304, doi:10.1002/qj.49712757202.

Pavliotis, G. A., and A. Stuart, 2008: *Multiscale Methods: Averaging and Homogenization.* Texts in Applied Mathematics, Vol. 53, Springer-Verlag, 310 pp.

Sakov, P., and L. Bertino, 2011: Relation between two common localisation methods for the EnKF. *Comput. Geosci.*, **15**, 225–237, doi:10.1007/s10596-010-9202-6.

Tong, X. T., A. J. Majda, and D. B. Kelly, 2015: Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation. arXiv.org, 34 pp. [Available online at https://arxiv.org/abs/1507.08319.]

——, ——, and ——, 2016: Nonlinear stability and ergodicity of ensemble based Kalman filters. *Nonlinearity*, **29**, 657, doi:10.1088/0951-7715/29/2/657.

Whitaker, J., and G. Compo, 2002: An ensemble Kalman smoother for reanalysis. *Symp. on Observations, Data Assimilation, and Probabilistic Prediction*, Orlando, FL, Amer. Meteor. Soc., 6.2. [Available online at https://ams.confex.com/ams/pdfpapers/28864.pdf.]

Wilks, D. S., 2005: Effects of stochastic parameterizations in the Lorenz'96 system. *Quart. J. Roy. Meteor. Soc.*, **131**, 389–407, doi:10.1256/qj.04.03.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev.*, **132**, 1238–1253, doi:10.1175/1520-0493(2004)132<1238:IOIEAO>2.0.CO;2.

Zwanzig, R., 1973: Nonlinear generalized Langevin equations. *J. Stat. Phys.*, **9**, 215–220, doi:10.1007/BF01008729.

——, 2001: *Nonequilibrium Statistical Mechanics.* Oxford University Press, 240 pp.