A Bayesian Adaptive Ensemble Kalman Filter for Sequential State and Parameter Estimation

JONATHAN R. STROUD

McDonough School of Business, Georgetown University, Washington, D.C.

MATTHIAS KATZFUSS

Department of Statistics, Texas A&M University, College Station, Texas

CHRISTOPHER K. WIKLE

Department of Statistics, University of Missouri, Columbia, Missouri

(Manuscript received 11 November 2016, in final form 22 October 2017)

ABSTRACT

This paper proposes new methodology for sequential state and parameter estimation within the ensemble Kalman filter. The method is fully Bayesian and propagates the joint posterior distribution of states and parameters over time. To implement the method, the authors consider three representations of the marginal posterior distribution of the parameters: a grid-based approach, a Gaussian approximation, and a sequential importance sampling (SIR) approach with kernel resampling. In contrast to existing online parameter estimation algorithms, the new method explicitly accounts for parameter uncertainty and provides a formal way to combine information about the parameters from data at different time periods. The method is illustrated and compared to existing approaches using simulated and real data.

1. Introduction

Data assimilation refers to sequential inference on the state of a system by combining observations with a numerical model describing the evolution of the system over time. This is a ubiquitous task in many fields, including atmospheric science, where the system is typically high dimensional and consists of one or more spatial fields evaluated on a fine grid. From a statistical perspective, data assimilation is equivalent to filtering inference in a state-space model. In many applications, the evolution model and other parts of the state-space model are not fully known and are, instead, functions of parameters. Data assimilation then requires combined inference on the (temporally varying) system state and on (temporally static) model parameters. This setting is the focus of this article.

Sequential Monte Carlo methods, also known as particle filters (Gordon et al. 1993; Pitt and Shephard 1999; Doucet et al. 2001), are widely used for sequential estimation in general state-space models. Although there is an enormous literature on pure state estimation, there are fewer papers that consider sequential estimation of both states and parameters. The existing references include Kitagawa (1998), who proposed augmenting the state vector to include the static parameter and then estimating the augmented state using the particle filter. Liu and West (2001) proposed another state-augmentation approach that uses kernel density estimation of the parameter distribution within an auxiliary particle filter (Pitt and Shephard 1999) framework. Storvik (2002) suggested analytical updating of sufficient statistics, but this approach only applies to parameters with conjugate priors. Andrieu et al. (2005) proposed recursive and batch maximum-likelihood estimation (MLE) methods. These methods have all been shown to work well in nonlinear and non-Gaussian models, when the state dimension is fairly small, say, fewer than 10 dimensions. However, these particle filters rely on reweighting or resampling of particles, which results in filter collapse when the state dimension is high (e.g., Snyder et al. 2008). Hence, these

Corresponding author: Jonathan R. Stroud, jrs390@georgetown.edu

DOI: 10.1175/MWR-D-16-0427.1

particle-filter-based methods are not suited for the high-dimensional systems of interest here.

The ensemble Kalman filter (EnKF; Evensen 1994) is a sequential Monte Carlo algorithm designed for combining high-dimensional space–time models with observations. Several reviews of and tutorials on the EnKF are available (e.g., Wikle and Berliner 2007; Evensen 2009; Katzfuss et al. 2016; Houtekamer and Zhang 2016). While the EnKF is closely related to the Kalman filter (KF; Kalman 1960), it handles nonlinearities in a more flexible manner than analytic linearization schemes such as the extended Kalman filter (e.g., Grewal and Andrews 1993, chapter 5). Although much work has been done to improve EnKF estimation of state variables, little work has focused on estimation of model parameters. Anderson (2001) proposed adding the unknown parameters to the state vector and updating the augmented state using a standard EnKF scheme. Other state-augmentation schemes include the dual EnKF (Moradkhani et al. 2005) and iterative EnKF (Gu and Oliver 2007) and the related Kalman ensemble generator (Nowak 2009). But these state-augmentation approaches do not work well for parameters that exhibit small (linear) correlation with the state vector. For example, Stroud and Bengtsson (2007) and DelSole and Yang (2010) show that the augmentation method fails for variance parameters.

Here, we consider parameter inference in the EnKF based on the likelihood, which is the distribution or density of the observed data conditional on the parameters, viewed as a function of the parameters. Offline maximum likelihood (ML) estimation in the EnKF framework has been considered using Newton–Raphson (DelSole and Yang 2010; Stroud et al. 2010) as well as grid-based (Ueno et al. 2010) and expectation-maximization (Tandeo et al. 2015; Dreano et al. 2017) optimization techniques. ML estimation of parameters from an online perspective was considered by Mitchell and Houtekamer (2000), whose method combines ML estimates at each time point in a statistically inconsistent way [see section 3b(2) later]; by Ueno and Nakamura (2014), who estimate parameters in the noise covariance matrix via online expectation-maximization algorithm; and by De (2014), who proposed a method for sequentially updating the unknown parameters at each time point to find a stochastic approximation to the ML estimator in stationary systems.

The likelihood can also be used to conduct Bayesian inference on the parameters. Stroud and Bengtsson (2007) provide a Bayesian method for parameter inference within the EnKF, but their approach is limited to a scalar variance parameter describing the magnitude of additive evolution-model error. Frei and Künsch (2012) propose Bayesian inference on parameters by combining an EnKF for state inference with a particle filter to approximate the parameter distribution, but their focus is on temporally varying observation error covariance parameters. Vrugt et al. (2005) propose an offline Bayesian approach for model parameter estimation. Brankart et al. (2010) find the maximum a posteriori (MAP) estimators of temporally varying parameters, while Ueno and Nakamura (2016) focus on MAP estimation for parameters in the noise covariance matrix with temporal smoothing via online expectation maximization (EM).

Here, we propose a fully Bayesian method for sequential (i.e., online) inference on states and parameters within the EnKF framework. Our algorithms are designed to be applicable to temporally static parameters in nonlinear, high-dimensional systems. Unlike some of the other approaches (e.g., Mitchell and Houtekamer 2000), our method combines information about the parameters from data at different time points in a formal way using Bayesian updating. In contrast to the ML and MAP approaches discussed above, we quantify uncertainty in the parameters through analytic propagation of the entire filtering distribution of the parameters. Further, our approach is suitable for static parameters in various parts of the state-space model, including in the evolution-error and noise covariance matrices. To implement our algorithm, we propose three approximate methods: one based on a parameter grid, another is based on a normal (Laplace) approximation, and another is based on a particle approximation to the parameter distribution.

Note that there is also an extensive literature on estimation of specific tuning parameters in the EnKF, such as inflation and localization parameters (e.g., Wang and Bishop 2003; Anderson 2007a,b; Šmídl and Hofman 2011). Here, we focus instead on inference on general parameters that explicitly appear in the statistical model, and we regard the tuning parameters as known.

The remainder of the paper is outlined as follows. In section 2, we introduce the state-space model under consideration. In section 3, we motivate and introduce our proposed methodology. In section 4, we present numerical comparisons of our methods to existing approaches using simulated and real data. We conclude in section 5.

2. Additive Gaussian state-space models

Let $\mathbf{y}_t$ denote the $m_t \times 1$ observation vector and $\mathbf{x}_t$ the $n \times 1$ state vector. We consider the following class of additive Gaussian state-space models:

$$\text{Observation: } \mathbf{y}_t = \mathbf{H}_t(\theta)\,\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}[\mathbf{0}, \mathbf{R}_t(\theta)], \qquad (1)$$

$$\text{Evolution: } \mathbf{x}_t = \mathcal{M}_t(\mathbf{x}_{t-1}; \gamma) + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}[\mathbf{0}, \mathbf{Q}_t(\theta)], \quad (2)$$

for $t = 1, 2, \ldots,$ where the observation matrix $\mathbf{H}_t$ and the covariance matrices $\mathbf{R}_t$ and $\mathbf{Q}_t$ may depend on a vector of unknown parameters θ, and the possibly nonlinear evolution operator $\mathcal{M}_t(\cdot)$ may depend on a separate set of parameters γ. For now, we will focus on inference for θ and consider γ to be fixed and known [and so we simply write $\mathcal{M}_t(\mathbf{x}_{t-1})$], but we will describe in section 3d how our algorithm for inference on θ can be combined with a state-augmentation approach to perform inference on γ. The model is completed with a prior on the initial state $p(\mathbf{x}_0|\mathbf{Y}_0) = \mathcal{N}(\mathbf{a}_0, \mathbf{P}_0)$ and a prior distribution of the parameters $p(\theta|\mathbf{Y}_0)$, where $\mathbf{Y}_0$ denotes the initial information. In applications where the relationship between the observations and the state is nonlinear, we take $\mathbf{H}_t$ in (1) to be the linearization of the nonlinear mapping.

3. Sequential Bayesian inference on state and parameters

The Bayesian filtering problem requires computing the joint posterior distribution $p(\mathbf{x}_t, \theta \mid \mathbf{Y}_t)$ of the current state and the parameters at each time $t = 1, \ldots, T$, where $\mathbf{Y}_t = \{\mathbf{Y}_0, \ldots, \mathbf{y}_1, \ldots, \mathbf{y}_t\}$ denotes the information available at time t, and $\mathbf{Y}_0$ is the initial information. This joint posterior encodes all available information about the states and parameters contained in the data, and it is typically summarized through marginal distributions, posterior means, standard deviations, or credible intervals. As we will see, Bayesian inference has two advantages over frequentist or more ad hoc methods: it allows accounting for parameter uncertainty, and information about the parameters can be naturally combined over time following a consistent probabilistic framework.

Except in very special cases (see Stroud and Bengtsson 2007), the joint posterior distribution is unavailable in closed form, so Monte Carlo methods must be used to approximate the distribution. In what follows, we propose a method for combined state and parameter estimation that scales to high-dimensional states.

Our approach relies on the decomposition of the joint posterior distribution of the state and parameters into two terms: the conditional posterior distribution for the states given the parameters and the marginal posterior distribution for the parameters:

$$p(\mathbf{x}_t, \theta|\mathbf{Y}_t) = p(\mathbf{x}_t|\theta, \mathbf{Y}_t)\, p(\theta|\mathbf{Y}_t). \quad (3)$$

In the following subsections, we describe how $p(\mathbf{x}_t, |\theta, \mathbf{Y}_t)$ can be obtained via the EnKF (section 3a), we examine the marginal parameter posterior $p(\theta, |\mathbf{Y}_t)$ (section 3b), we propose three approximation methods for $p(\theta, |\mathbf{Y}_t)$ (section 3c), and, finally, we describe the full algorithm that combines these ideas and results (section 3d).

a. EnKF for state inference

The first term on the right-hand side of (3) is the filtering distribution of the state, given the parameters. Because our algorithm must be implemented sequentially, it is useful to write this distribution in recursive form:

$$p(\mathbf{x}_t|\theta, \mathbf{Y}_t) \propto p(\mathbf{y}_t|\mathbf{x}_t, \theta) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta) p(\mathbf{x}_{t-1}|\theta, \mathbf{Y}_{t-1})\, d\mathbf{x}_{t-1}, \quad (4)$$

that is, the observation density times the state forecast density $p(\mathbf{x}_t, |\theta, \mathbf{Y}_{t-1})$, which is defined by the integral. In a linear Gaussian model, these recursions can be computed analytically using the Kalman filter (provided the dimension of the state is not excessively large).

Here, we are interested in high-dimensional systems with possibly nonlinear evolution (see section 2), for which we instead employ an EnKF. Throughout the paper, the superscripts p and f refer to the predictive and forecast distributions, the superscript i is the ensemble index, and N is the ensemble size.

Assume we have an ensemble of states $\{\mathbf{x}_{t-1}^i\}_{i=1}^N$ representing the filtering distribution at time $t - 1$. The EnKF then propagates each state vector forward, $\mathbf{x}_t^{pi} = \mathcal{M}(\mathbf{x}_{t-1}^i), i = 1, \ldots, N$ and estimates the covariance matrix from the prior ensemble. In most applications, we have $n \gg N$, and some form of regularization of this covariance matrix is necessary. Denoting the prior ensemble mean as $\hat{\mathbf{a}}_t^p = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_t^{pi}$, we assume here that

$$\hat{\mathbf{P}}_t^p = \mathbf{C} \circ \left[\frac{1}{N-1}\sum_{i=1}^N (\mathbf{x}_t^{pi} - \hat{\mathbf{a}}_t^p)(\mathbf{x}_t^{pi} - \hat{\mathbf{a}}_t^p)'\right] \quad (5)$$

is given by an elementwise product of the empirical covariance matrix with a sparse tapering correlation matrix $\mathbf{C}$ (e.g., Houtekamer and Mitchell 1998; Anderson 2007b; Furrer and Bengtsson 2007). The estimated Kalman gain is a function of θ:

$$\hat{\mathbf{K}}_t(\theta) = \mathbf{H}_t(\theta)\, \hat{\mathbf{P}}_t^f(\theta)\, \mathbf{H}_t(\theta)'\, \hat{\mathbf{S}}_t(\theta)^{-1}, \quad (6)$$

where $\hat{\mathbf{P}}_t^f(\theta) = \hat{\mathbf{P}}_t^p + \mathbf{Q}_t(\theta)$, and

$$\hat{\mathbf{S}}_t(\theta) = \mathbf{H}_t(\theta)\hat{\mathbf{P}}_t^f(\theta)\mathbf{H}_t(\theta)' + \mathbf{R}_t(\theta) \quad (7)$$

is the innovation covariance matrix. Then, after generating the forecast ensemble by setting $\mathbf{x}_t^{fi} = \mathbf{x}_t^{pi} + \mathbf{w}_t^i$, where $\mathbf{w}_t^i \sim \mathcal{N}[\mathbf{0}, \mathbf{Q}(\theta)], i = 1, \ldots, N$, and simulating observation errors as $\mathbf{v}_t^i \sim \mathcal{N}[0, \mathbf{R}_t(\theta)], i = 1, \ldots, N$, we can obtain the posterior ensemble at time t based on

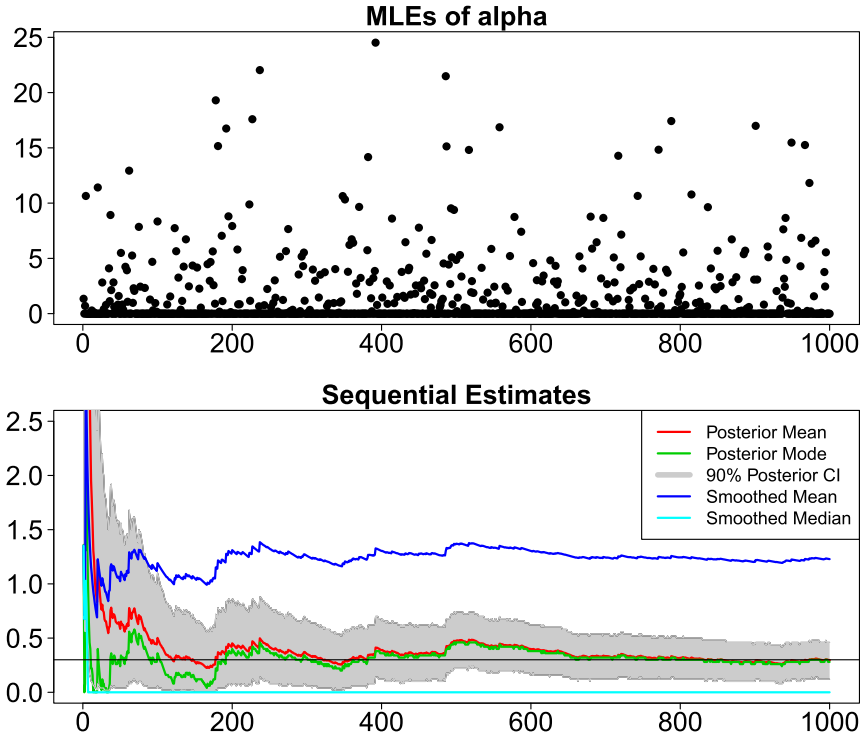

FIG. 1. For the static model in section 3a, (top) individual maximum likelihood estimates and (bottom) cumulative inference on the parameter α over time. The true parameter value is $\alpha_* = 0.3$. In the bottom panel, the blue (cyan) line represent the cumulative mean (median) of the individual estimates as proposed by Mitchell and Houtekamer (2000), while the red and green lines and the gray band represent summaries of the posterior distribution of α.

parameter value θ using the analysis scheme of Burgers et al. (1998):

$$\mathbf{x}_t^i = \mathbf{x}_t^{fi} + \hat{\mathbf{K}}_t(\theta)[\mathbf{y}_t + \mathbf{v}_t^i - \mathbf{H}_t(\theta)\,\mathbf{x}_t^{fi}].$$

b. *The marginal posterior of parameters*

The second term on the right-hand side of (3) is the marginal posterior for the parameters. It can be written recursively via Bayes's theorem as

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

The above formula is crucial, as it defines a recursion for the parameter distribution over time. Note that the first term on the right side is the posterior distribution for θ at the previous time $t - 1$ (which serves as the prior with respect to $\mathbf{y}_t$). The second term is the likelihood at time t. Recall that $\mathbf{Y}_t = \{\mathbf{Y}_{t-1}, \mathbf{y}_t\}$. Then, if we ignore $\mathbf{Y}_{t-1}$ in the conditioning sets, Eq. (8) can be written as $p(\theta|\mathbf{y}_t) \propto p(\theta)\,p(\mathbf{y}_t|\theta)$, which can easily be recognized as Bayes's theorem.

Note that under a flat initial prior for the parameters (i.e., $p(\theta|\mathbf{Y}_0) \propto 1$), the marginal posterior for θ in (8) is exactly proportional to the cumulative likelihood $L_t(\theta) = \prod_{j=1}^{t} p(\mathbf{y}_j|\theta, \mathbf{Y}_{j-1})$ that is often considered in frequentist (i.e., non-Bayesian) inference.

1) ACCUMULATION OF EVIDENCE OVER TIME

The recursion in (8) provides a natural way to propagate information about the parameter over time, as opposed to the ad hoc methods used by Dee (1995) and Mitchell and Houtekamer (2000). To illustrate this, we replicated a static model example presented in Mitchell and Houtekamer (2000). The model assumes that (scalar) observations $y_t \sim \mathcal{N}(0,\ 2 + \alpha)$ are generated independently from a Gaussian distribution with mean zero and variance $(2 + \alpha)$, where α is an unknown variance parameter, and the goal is to estimate α sequentially as new data arrive. Mitchell and Houtekamer (2000) considered the cumulative mean and median of the single-stage ML estimates $\hat{\alpha}_t = \text{argmax}_\alpha p(y_t|\alpha)$, where $p(y_t|\alpha)$ is the likelihood considering only the data at time t. Because α is a variance parameter and, thus, must be nonnegative, the estimates are given by $\hat{\alpha}_t = \max(0, y_t^2 - 2)$.

The results of a simulation using a true parameter value $\alpha_* = 0.3$ are shown in Fig. 1. The majority of the individual estimates $\hat{\alpha}_t$ are equal to zero, and the distribution of the estimates is heavily right-skewed.

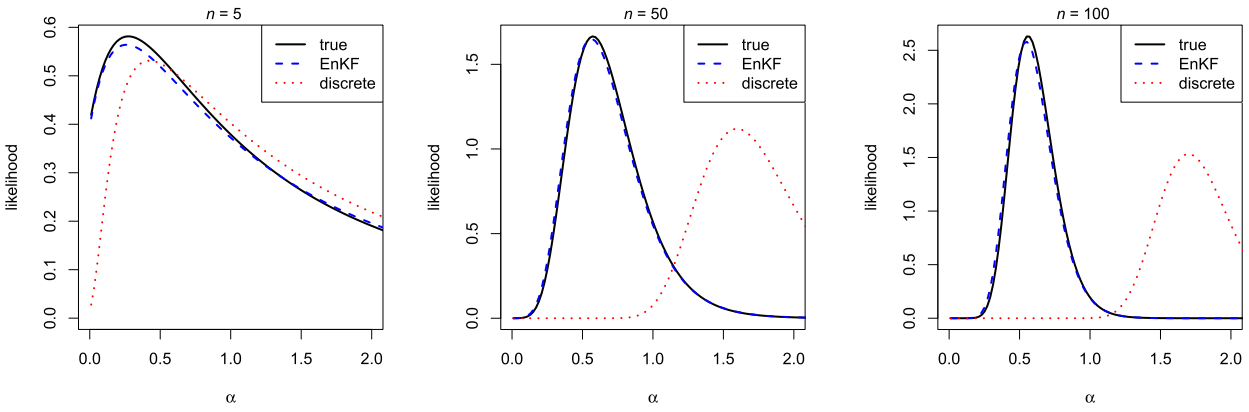

FIG. 2. Comparison of the likelihoods discussed in section 3b in a simple simulated example. Data are simulated at a single time point from the prior covariance matrix $\mathbf{P}_t^p$, with $\mathbf{Q} = \alpha\mathbf{I}_n$, $\mathbf{H} = \mathbf{I}_n$, $\mathbf{R} = 0.1\mathbf{I}_n$. The true (solid black) α is 0.5. The observations are on a one-dimensional spatial domain at locations $1, 2, \ldots, n$ and $\mathbf{P}_t^p$ is based on an exponential covariance function with range parameter 3. We use the same $N = 50$ draws from the forecast distribution for the EnKF (dashed blue) and discrete (dotted red) approximations and a Wendland taper with range 12 for the EnKF. The likelihoods are normalized to integrate to 1 and can, hence, be viewed as posterior distributions for α under a uniform (flat) prior.

Hence, as in Mitchell and Houtekamer (2000), we find an upward bias in the cumulative mean with respect to the true value α_*, while the cumulative median is zero. Moreover, after 10 000 observations, the estimates did not converge to α_*, indicating statistical inconsistency of the estimators. In contrast, the posterior distribution of α from (8) can be seen to become more concentrated over time and to converge to the true value of α_*. Corresponding point estimates, such as the posterior mode and posterior mean, do converge to the true value.

Thus, it is desirable to use the recursive expression for the posterior distribution of the parameters from (8) for rigorously combining information about the parameters from data at different time points.

2) FEASIBLE LIKELIHOOD APPROXIMATION FOR HIGH-DIMENSIONAL STATES

For high-dimensional models, evaluation of the likelihood $p(\mathbf{y}_t|\theta, \mathbf{Y}_{t-1}) \propto$ in (8) is computationally infeasible. Given that we use an ensemble representation for the state distributions in section 3a above, it is natural to use the same ensemble representation in order to approximate the likelihood. Specifically, given that the filtering distribution at time $t-1$ is a discrete distribution with equal weights at the filtering ensemble $\{\mathbf{x}_{t-1}^i\}_{i=1}^N$, an approximation of the likelihood (called the "discrete" likelihood approximation here) is given by

$$p_{\text{disc}}(\mathbf{y}_t|\theta,\mathbf{Y}_{t-1}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{N}[\mathbf{y}_t|\,\mathbf{H}_t(\theta)\mathbf{x}_t^{pi}, \mathbf{H}_t(\theta)\mathbf{Q}_t(\theta)\mathbf{H_t}(\theta)' + \mathbf{R}_t(\theta)].$$

However, as illustrated in Fig. 2, this approximation can break down when the data are informative (i.e., when m_t and the signal-to-noise ratio are large relative to N). Instead, we employ a likelihood approximation based on the EnKF, which approximates the forecast distribution by a multivariate Gaussian distribution (Mitchell and Houtekamer 2000). This EnKF likelihood approximation is given by

$$p_{\text{enkf}}(\mathbf{y}_t|\theta,\mathbf{Y}_{t-1}) \propto |\hat{\mathbf{S}}_t(\theta)|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\hat{\mathbf{e}}_t(\theta)'\hat{\mathbf{S}}_t(\theta)^{-1}\hat{\mathbf{e}}_t(\theta)\right], \tag{9}$$

where $\hat{\mathbf{e}}_t(\theta) = \mathbf{y}_t - \mathbf{H}_t(\theta)\hat{\mathbf{a}}_t^p$ is the innovation mean vector, and $\hat{\mathbf{S}}_t(\theta)$ is the innovation covariance matrix defined in (7).

c. *Approximations of the parameter posterior*

Note that the likelihood $p_{\text{enkf}}(\mathbf{y}_t|\theta, \mathbf{Y}_{t-1})$ in (9) is a complicated nonlinear function of θ, which arises in both the determinant and inverse of the $m_t \times m_t$ innovation covariance matrix $\hat{\mathbf{S}}_t(\theta)$. Thus, generally, no conjugate prior is available for θ, so the marginal posterior distribution $p(\theta|\mathbf{Y}_t)$ in (8) is typically unavailable in closed form.

To implement parameter learning in the Bayesian framework, we need a representation of the parameter distribution that allows for recursive updating. In what follows, we consider three representations of the parameter distribution: a discrete (grid based) distribution, a Gaussian approximation, and a particle approximation with kernel resampling. The first provides an exact recursive updating method on a discretized parameter space, while the second and third provide an approximate method over the full parameter space.

1) GRID-BASED REPRESENTATION OF $p(\theta\,|\,\mathbf{Y}_T)$

This approach treats the parameter space as discrete. The parameter distribution is specified by a set of points $\{\theta_1^*, \ldots, \theta_K^*\}$ and associated probability weights $\{\pi_{t,1}, \ldots, \pi_{t,K}\}$ with normalization constraint $\sum_{k=1}^{K} \pi_{t,k} = 1$. The discrete parameter distribution at time t is defined by

$$p_{\text{grid}}(\theta|\mathbf{Y}_t) = \sum_{k=1}^{K} \delta_{\theta_k^*}(\theta)\,\pi_{t,k}, \tag{10}$$

where $\delta(\cdot)$ is the Dirac delta function. The initial probability weights at time $t = 0$ are obtained by defining $\pi_{0,k} \propto p(\theta_k^*|\mathbf{Y}_0)$ for $k = 1, \ldots, K$, with weights normalized to sum to 1. Samples from this distribution can be easily generated at any time t by selecting θ_i with replacement from the discrete set $\{\theta_1^*, \ldots, \theta_K^*\}$ with corresponding weights $\{\pi_{t,1}, \ldots, \pi_{t,K}\}$. With this representation for the prior, the posterior is given as the product of the prior and the likelihood, that is,

$$p_{\text{grid}}(\theta\,|\,\mathbf{Y}_t) \propto p_{\text{grid}}(\theta\,|\,\mathbf{Y}_{t-1})\,p_{\text{enkf}}(\mathbf{y}_t|\,\theta, \mathbf{Y}_{t-1}), \tag{11}$$

where the EnKF likelihood function (9) is used instead of the exact likelihood. The updating formula in (8) reduces to a recursion on the weights: $\pi_{t,k} \propto \pi_{t-1,k} \times p_{\text{enkf}}(\mathbf{y}_t|\theta_k^*, \mathbf{Y}_{t-1})$ for $k = 1, \ldots, K$, where the weights $\pi_{t,k}$ are normalized to sum to 1. Therefore, the computational cost of the update is K likelihood evaluations, one for each grid point θ_k^*.

To simplify implementation, we assume the parameter grid is fixed over time. For example, the grid points could be based on the initial prior distribution, with equally spaced points between, say, the 1st and 99th percentiles of the distribution. Or we could choose unequally spaced points based on evenly spaced percentiles of the initial prior.

While the discrete approach is conceptually appealing, it has some limitations. First, the method does not extend beyond a few parameters because the computational cost of the update step grows exponentially in the dimension of the parameter space. Second, the grid of parameter values is specified a priori and is not adaptive over time. Thus, as the posterior becomes more concentrated over time, the posterior distribution eventually concentrates on a single grid point. This implies falsely that there is no posterior uncertainty about θ. Finally, the initial grid may be poorly specified and may not cover the high probability region of the posterior at later time points. To alleviate these problems, we next consider an adaptive method based on a Gaussian approximation.

2) NORMAL APPROXIMATION TO $p(\theta\,|\,\mathbf{Y}_T)$

Here, the parameter distribution at each time t is approximated by a normal distribution with mean $\mathbf{m}_t$ and covariance matrix $\mathbf{C}_t$. The posterior density is then given by

$$p_{\text{norm}}(\theta|\mathbf{Y}_t) \propto \exp\left[-\frac{1}{2}(\theta - \mathbf{m}_t)'\mathbf{C}_t^{-1}(\theta - \mathbf{m}_t)\right]. \tag{12}$$

The updating recursions for the posterior moments are then derived as follows. Assume the parameter distribution at time $t - 1$ is normal with mean $\mathbf{m}_{t-1}$ and covariance $\mathbf{C}_{t-1}$. The posterior is proportional to the product of the prior and likelihood, that is,

$$\exp[\ell(\theta)] = p_{\text{norm}}(\theta\,|\,\mathbf{Y}_{t-1})p_{\text{enkf}}(\mathbf{y}_t\,|\,\theta, \mathbf{Y}_{t-1}), \tag{13}$$

where we use the EnKF likelihood function $p_{\text{enk},f}(\mathbf{y}_t|\theta, \mathbf{Y}_{t-1})$ in place of the exact likelihood. The function $\ell(\theta)$ represents the log posterior distribution, that is, $\ell(\theta) = \log(\text{prior} \times \text{likelihood})$. Because this posterior is not of a recognizable form, a normal (or Laplace) approximation is used. We define the normal approximation $p_{\text{norm}}(\theta\,|\,\mathbf{Y}_t)$ based on a 2nd-order expansion of $\ell(\theta)$ at the mode. The posterior mean and covariance matrix $\mathbf{m}_t$ and $\mathbf{C}_t$ are defined by

$$\mathbf{m}_t = \underset{\theta}{\text{argmax}}\ \ell(\theta) \quad \text{and} \quad \mathbf{C}_t = -\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\,\partial\theta'}\right]_{\theta=\mathbf{m}_t}^{-1}. \tag{14}$$

The posterior mean (mode) $\mathbf{m}_t$ is obtained using a numerical optimization scheme. In our applications, we use a modified version of the Subplex algorithm (Rowan 1990), a derivative-free method, implemented using the SBPLX function in the NLopt package (Johnson 2011). The starting value for the optimization is the prior mean $\mathbf{m}_{t-1}$. Because $\ell(\theta)$ may not be concave, a global optimum is not guaranteed. However, for our examples, we find the optimization to be quite reliable, with convergence to a mode at nearly every time step. After obtaining the posterior mode $\mathbf{m}_t$, the Hessian matrix $\mathbf{C}_t^{-1}$ is computed numerically using finite differences. In nearly all of the simulations, we found that $\mathbf{C}_t^{-1}$ was invertible. When it was noninvertible, a small value was added to the diagonal elements to make it invertible.

3) PARTICLE APPROXIMATION TO $p(\theta\,|\,\mathbf{Y}_T)$

The third method, suggested by one of the referees, uses a particle filter with kernel resampling to approximate the parameter distribution. Under a sampling importance resampling (SIR) approach without kernel resampling, the posterior distribution is approximated by a discrete distribution:

$$p_{\text{sir}}(\theta \mid \mathbf{Y}_t) = \sum_{i=1}^{N} \delta_{\theta_t^i}(\theta)\, \omega_t^i, \tag{15}$$

where θ_t^i are the particles at time t, and ω_t^i are the corresponding weights normalized such that $\sum_{i=1}^N \omega_t^i = 1$. The algorithm is initialized by sampling particles from the initial prior $\theta_0^i \sim p(\theta \mid \mathbf{Y}_0)$ and setting the initial weights to $\omega_0^i = 1/N$ for each $i = 1, \ldots, N$. For each subsequent observation $\mathbf{y}^t$, the posterior is updated by first updating the weights for each particle $i = 1, \ldots, N$ via

$$\omega_t^i \propto \omega_{t-1}^i \times p_{\text{enkf}}(\mathbf{y}_t \mid \theta_{t-1}^i, \mathbf{Y}_{t-1}),$$

where the weights are normalized to $\sum_{i=1}^N \omega_t^i = 1$. We then sample each θ_t^i, $i = 1, \ldots, N$ from the discrete set $\{\theta_{t-1}^i\}$ (with replacement) with weights equal to $\{\omega_t^i\}$. After resampling N particles, the particle weights are reset to $\omega_t^i = 1/N$ for $i = 1, \ldots, N$. Unfortunately, this SIR scheme suffers from the same degeneracy problem as the grid-based method, that is, the support points are fixed, so as more data are collected, the posterior eventually concentrates on a single particle.

To mitigate this degeneracy problem, a kernel resampling scheme, as in Liu and West (2001), can be used to generate new parameter values at each time t. Here, the posterior distribution is represented by a kernel density estimate (mixture of normals) of the form

$$p_{\text{kern}}(\theta \mid \mathbf{Y}_t) = \sum_{i=1}^{N} \mathcal{N}[\theta \mid a\theta_t^i + (1-a)\overline{\theta}_t, (1-a^2)\mathbf{V}_t]\, \omega_t^i, \tag{16}$$

where $\{\theta_t^i\}$ are the particles at time t, $\{\omega_t^i\}$ are the particle weights normalized so that $\sum_{i=1}^N \omega_t^i = 1$, $\overline{\theta}_t = \sum_{i=1}^N \omega_t^i \theta_t^i$ is the posterior mean and $\mathbf{V}_t = \sum_{i=1}^N \omega_t^i (\theta_t^i - \overline{\theta}_t)(\theta_t^i - \overline{\theta}_t)'$ is the posterior variance-covariance matrix, and $a \in [0, 1]$ is a smoothing parameter. Note that $a \to 1$ implies a discrete distribution, as in (15), while $a = 0$ implies a Gaussian approximation similar to (12). Liu and West (2001) recommend choosing a between 0.975 and 0.995.

Estimation proceeds in a similar manner as the SIR approach above. However, instead of resampling θ_t^i from the discrete set $\{\theta_{t-1}^i\}$, we use a kernel resampling approach. Here, we first resample $\tilde{\theta}_t^i$ from the set $\{\theta_{t-1}^i\}$ with weights $\{\omega_t^i\}$ and then generate posterior samples as

$$\theta_t^i \sim \mathcal{N}[a\tilde{\theta}_t^i + (1-a)\overline{\theta}_{t-1}, (1-a^2)\mathbf{V}_{t-1}].$$

Thus, even if all of the resampled particles $\tilde{\theta}_t^i$ are identical, the posterior draws θ_t^i will be unique because they are independently drawn from a normal distribution.

d. *Combined state and parameter learning in the EnKF*

Given the developments above, an ensemble-based algorithm is proposed to generate a sample from the joint posterior distribution of the state and parameters in (3) at each time point. At each t, we have an analytical (discrete, normal, or particle) representation of the parameter distribution $\hat{p}(\theta \mid \mathbf{Y}_t)$, along with an ensemble of states and parameters $(\mathbf{x}_t^i, \theta^i)_{i=1}^N$ from $p(x_t, \theta \mid \mathbf{Y}_t)$.

To implement our algorithm, we make the assumption of forecast independence between the states and parameters, that is, $\mathcal{M}(\mathbf{x}_{t-1})$, and θ are independent conditional on $\mathbf{Y}_{t-1}$. This implies that the joint forecast distribution can be written as

$$p[\mathcal{M}(\mathbf{x}_{t-1}), \theta \mid \mathbf{Y}_{t-1}] = p[\mathcal{M}(\mathbf{x}_{t-1}) \mid \mathbf{Y}_{t-1}] p(\theta \mid \mathbf{Y}_{t-1}).$$

The advantage of this assumption is that it allows us to use a single ensemble for both states and parameters, that is, we do not need a separate state ensemble for each member of the parameter ensemble. This provides enormous computational savings, and we find that the assumption is quite accurate in our examples. We note that Frei and Künsch (2012) also made the assumption of forecast independence and justified it based on asymptotic independence arguments. However, they considered only unknown parameters in the observation error covariance matrix $\mathbf{R}(\theta)$, and they assumed forecast independence of $\mathbf{x}_t$ and θ.

Our approach is closely related to that of Mitchell and Houtekamer (2000), but it includes steps to update and simulate from the parameter distribution rather than obtaining θ through maximum likelihood. Our approach naturally quantifies uncertainty in the parameters and takes this uncertainty into account when obtaining the filtering ensemble of the state.

Algorithm 1: The algorithm is initialized by drawing from the initial prior: $\theta^i \sim p(\theta \mid \mathbf{Y}_0)$ and $\mathbf{x}^i \sim \mathcal{N}(\mathbf{a}_0, \mathbf{P}_0)$ for $i = 1, \ldots, N$. Each assimilation cycle $t = 1, 2, \ldots$ proceeds as follows:

1) Propagate each state vector forward: $\mathbf{x}_t^{pi} = \mathcal{M}(\mathbf{x}_{t-1}^i)$, $i = 1, \ldots, N$.

2) Approximate the likelihood function using the prior ensemble as in (9) by $\hat{p}_{\text{enkf}}(\mathbf{y}_t \mid \theta, \mathbf{Y}_{t-1}) \propto |\hat{\mathbf{S}}_t(\theta)|^{-1/2} \exp[-(1/2)\, \hat{\mathbf{e}}_t(\theta)' \hat{\mathbf{S}}_t(\theta)^{-1} \hat{\mathbf{e}}_t(\theta)]$.

3) Update the analytical parameter distribution using the grid, normal, or particle approximation [section 3c(1), 3c(2), or 3c(3), respectively]: $\hat{p}(\theta \mid \mathbf{Y}_t) \propto \hat{p}(\theta \mid \mathbf{Y}_{t-1}) \hat{p}_{\text{enkf}}(\mathbf{y}_t \mid \theta, \mathbf{Y}_{t-1})$.

4) Draw parameters from the updated posterior distribution: $\theta^i \sim \hat{p}(\theta|\mathbf{Y}_t)$, $i = 1, \ldots, N$.

5) Generate the forecast ensemble by setting $\mathbf{x}_t^{fi} = \mathbf{x}_t^{pi} + \mathbf{w}_t^i$, where $\mathbf{w}_t^i \sim \mathcal{N}[0, \mathbf{Q}(\theta^i)]$, $i = 1, \ldots, N$.

6) Draw a posterior ensemble using the analysis scheme of Burgers et al. (1998): $\mathbf{x}_t^i = \mathbf{x}_t^{fi} + \hat{\mathbf{K}}_t(\theta^i)[\mathbf{y}_t + \mathbf{v}_t^i - \mathbf{H}_t(\theta^i)\mathbf{x}_t^{fi}]$, where $\mathbf{v}_t^i \sim \mathcal{N}[0, \mathbf{R}_t(\theta^i)]$, $i = 1, \ldots, N$, and the estimated Kalman gain $\hat{\mathbf{K}}_t(\theta^i)$ is given in (6).

Depending on which approximation method is used in step 3 of this algorithm, we refer to it as EnKF-Grid, EnKF-Normal, or EnKF-Particle.

MERGING OUR METHOD WITH EXISTING APPROACHES

Note that our algorithm works best when the number of unknown parameters in θ is small. Hence, we recommend combining our algorithm with other approaches as much as possible.

For example, it can be combined with state augmentation (Anderson 2001), which works well for parameters (γ, say) that have a strong correlation with the state [e.g., parameters in $\mathcal{M}$ as introduced in (2)]. In the algorithm above, this means replacing the state $\mathbf{x}_t$ with the augmented state $(\mathbf{x}_t', \gamma')'$ and $\mathbf{H}_t$ with the matrix $(\mathbf{H}_t, 0)$. The transition of the parameters γ is typically assumed to be constant, although it is also possible to treat γ as a time-varying parameter γ_t with small artificial evolution noise (e.g., Kitagawa 1998; Liu and West 2001). Equivalently, we could use covariance inflation for the parameters, which has a similar effect.

The EnKF-Grid method can also be combined with the approach of Stroud and Bengtsson (2007) to make inference on a scalar multiplicative parameter that appears in both $\mathbf{Q}_t$ and $\mathbf{R}_t$. If this parameter has an inverse-gamma prior distribution, its marginal posterior distribution is also inverse gamma and available in closed form. Stroud and Bengtsson (2007) provide an EnKF algorithm to update the hyperparameters of the inverse-gamma distribution at each time point and sample from the joint filtering distribution of the state vector and the scalar parameter.

4. Numerical comparison and applications

a. Linear evolution

We first consider a linear dynamic spatiotemporal model from Xu and Wikle (2007). The model is a vector autoregression plus noise, where the state vector $\mathbf{x}_t = (x_{t_1}, \ldots, x_{t_n})'$ corresponds to n equally spaced locations $\{1, 2, 3, \ldots, n\}$ along a spatial transect. Following the notation in (1) and (2), the evolution mean function is linear, $\mathcal{M}(\mathbf{x}_{t-1}) = \mathbf{M}\mathbf{x}_{t-1}$ where the propagator matrix is tridiagonal with parameters $\gamma = \gamma_1, \gamma_2, \gamma_3$:

$$\mathbf{M}(\gamma) = \begin{pmatrix} \gamma_1 & \gamma_2 & & 0 \\ \gamma_3 & \gamma_1 & \ddots & \\ & \ddots & \ddots & \gamma_2 \\ 0 & & \gamma_3 & \gamma_1 \end{pmatrix}.$$

The evolution errors are spatially correlated with covariance $\mathbf{Q}(\theta) = \sigma_\eta^2\mathbf{C}(\tau)$, where $\mathbf{C}(\tau)$ is defined by the exponential correlation function $c(d;\tau) = \exp(-\tau d)$, and d is the distance between locations. The initial state distribution is given by $p(\mathbf{x}_0|\theta) = \mathcal{N}(0, \sigma_\varepsilon^2\mathbf{I})$. For the data model, we assume that observations are taken at each location $\mathbf{y}_t = (y_{t1}, \ldots, y_{tn})'$ and the observation matrix and error covariance matrix are given by $\mathbf{H} = \mathbf{I}$ and $\mathbf{R} = \sigma_\varepsilon^2\mathbf{I}$. The signal-to-noise ratio is denoted by $\beta = \sigma_\eta^2/\sigma_\varepsilon^2$.

We consider two relatively low-dimensional examples here, which (along with the assumption of linear evolution) allows us to compute the true posterior distribution of θ at each time using the Markov chain Monte Carlo (MCMC) procedure of Carter and Kohn (1994) and to compare ours and other approaches to the true posterior distribution.

1) SIMULATION

First, we simulated observations from the true model with dimensions $n = m = 20$ for $T = 100$ time points. The true parameters were taken to be $\gamma = (0.3, 0.6, 0.1)$, $\beta = 5$, $\tau = 1$, and $\sigma_\varepsilon^2 = 1$. For this simulation, we assumed that γ and σ_ε^2 were known, so that $\theta = (\beta, \tau)'$ were the unknown parameters with independent prior distributions $\beta \sim \mathcal{N}^+(5, 10)$ and $\tau \sim \mathcal{N}^+(2, 0.16)$, where $\mathcal{N}^+$ denotes a truncated normal distribution on the positive real line.

We obtained the posterior distribution of the parameters θ for each time $t = 1, \ldots, T$ using algorithm 1 with $N = 100$ ensemble members and no tapering or covariance inflation. The results are shown in Fig. 3. As we can see, the posterior distributions, as approximated by our EnKF-Norm and EnKF-Grid procedures from algorithm 1, are very close to the true posterior distribution obtained via MCMC, and they seem to converge to the true values of $\sigma_\eta^2(=\beta\,\sigma_\varepsilon^2) = 5$ and $\tau = 1$. This is in contrast to the approximation of the posterior obtained by state augmentation. The flat bands for both parameters indicate that the augmentation approach does not work in this case, likely because the relationship

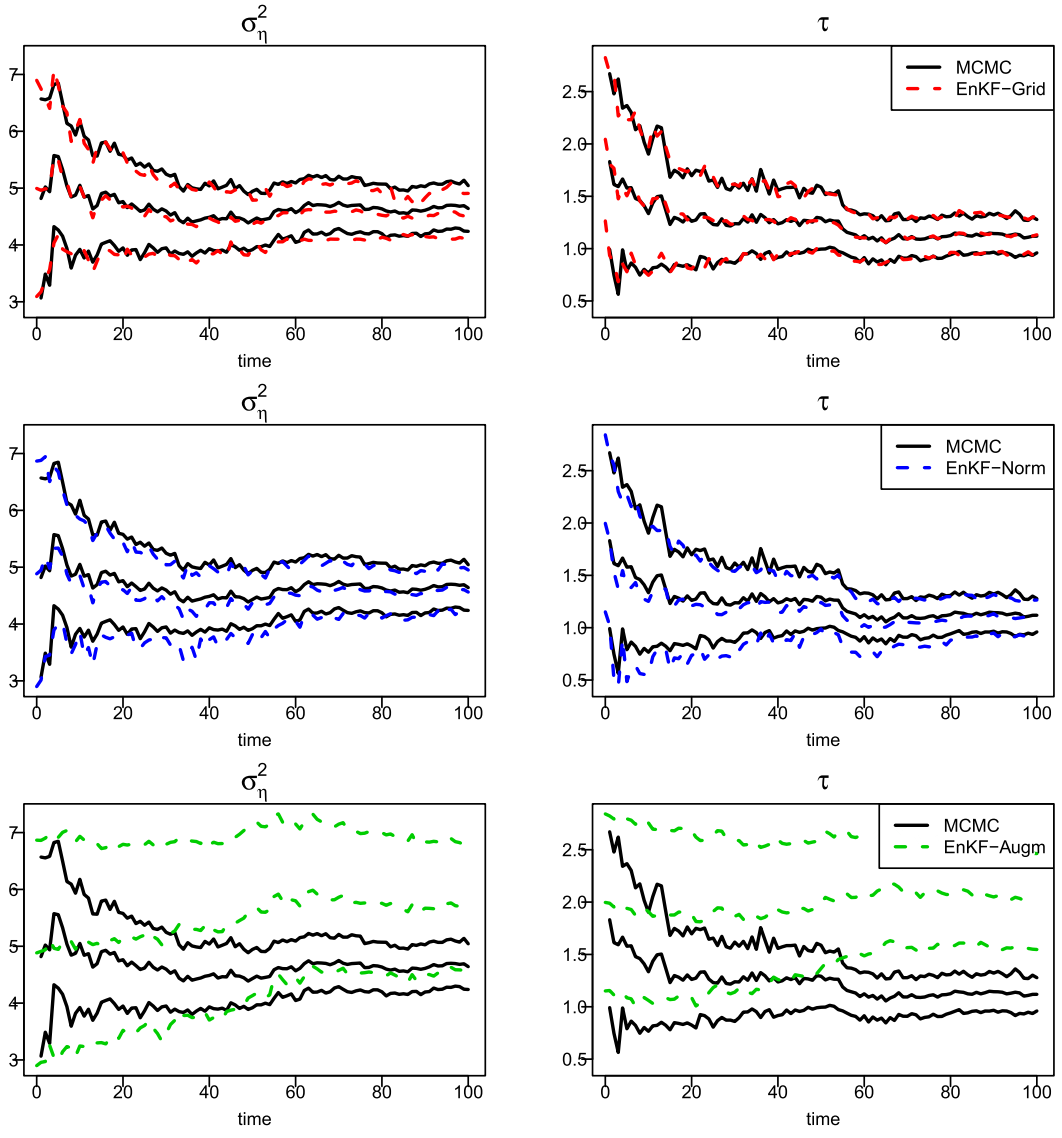

FIG. 3. For the simulated data using a linear state-space model described in section 4a(1), true posterior distributions (filtered mean and 95% bands) over time of the parameters (left) $\sigma_\eta^2$and (right) τ (in black)and the corresponding approximations (colored lines) using the (top) EnKF-Grid , (middle) EnKF-Norm, and (bottom) EnKF with state augmentation.

between the covariance parameters and the observations is not linear.

2) CLOUD DATA

Next, we apply the proposed methods to the cloud motion data of Wikle (2002). The data are cloud intensities at $n = 60$ equally spaced locations along a transect at $T = 80$ time periods. Wikle (2002) used non-Gaussian spatiotemporal kernel models to analyze the data. Here, because the original data z_{ti} are counts with a large number of zeros, we work with the transformed observations y_{ti} 4 = log(1 + z_{ti}). Using again the model of Xu and Wikle (2007) described above, we now treat all parameters $\theta = (\gamma', \beta, \tau, \sigma_\varepsilon^2)'$ as unknown with the following prior distributions: $\gamma|\sigma_\varepsilon^2 \sim \mathcal{N}[(0.3, 0.3, 0.3)', 0.01\sigma_\varepsilon^2\mathbf{I}]$, $\beta \sim \mathcal{N}^+(1.0, 0.01)$, $\tau \sim \mathcal{N}^+(0.1, 0.0004)$, and $\sigma_\varepsilon^2 \sim \mathcal{IG}(25, 2)$.

We applied our algorithm 1 to the data using $N = 100$ ensemble members and no tapering or covariance inflation. While the model includes six unknown parameters in total, the autoregressive parameters γ were included in the state and handled with state

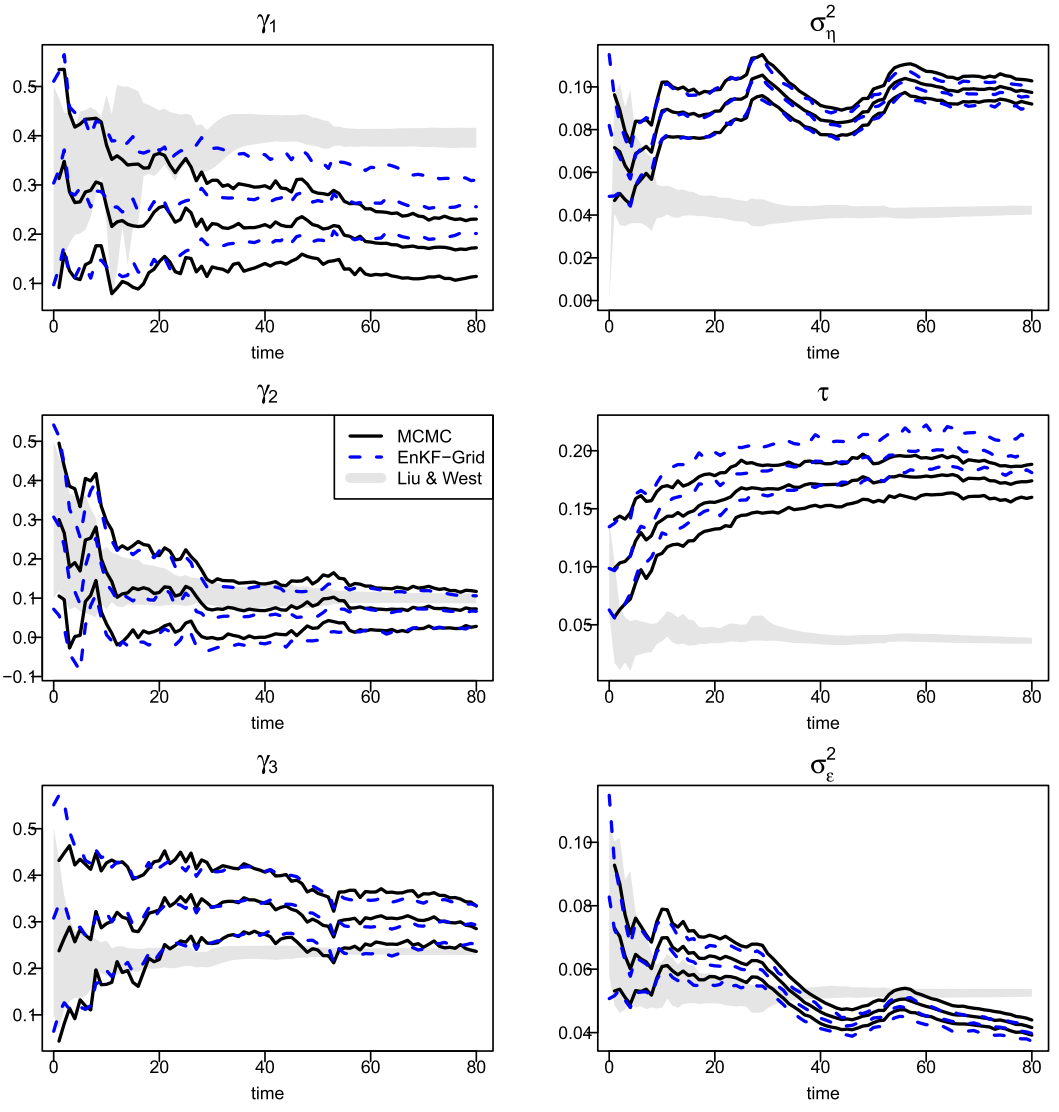

FIG. 4. As in Fig. 3, but for the cloud data in section 4a(2), using the EnKF-Grid (blue dashed lines) and the particle filter (gray bands) of Liu and West (2001).

augmentation, and the technique of Stroud and Bengtsson (2007) was used for inference on σ_ε^2 (see section 3d). This leaves us with the parameter vector $\theta = (\tau, \beta)'$, to which we apply our method.

In Fig. 4, the results are compared to the true posterior distribution, as obtained by an MCMC procedure, and to Liu and West (2001)'s particle filter with state augmentation using $N = 10\,000$ particles and a tuning parameter $\delta = 0.98$ (see appendix). As we can see, the EnKF-Grid posterior does well and is close to the true posterior. The results for EnKF-Normal (not shown) are similar. Despite the very large ensemble size, the APF does not perform well, in that the means are way off, and the posterior uncertainty appears to be strongly underestimated.

b. *The Lorenz-96 model*

We now consider the 40-variable system of Lorenz (1996), commonly referred to as the Lorenz-96 model, which mimics advection at equally spaced locations along a latitude circle. The differential equations defining the time evolution of the system are given by

$$\dot{x}_{t,k} = (x_{t,k+1} - x_{t,k-2})\, x_{t,k-1} - x_{t,k} + F,$$

for $k = 1, \ldots, n = 40$, with periodic boundary conditions. We note that the system equations contain quadratic nonlinearities that define a nonlinear transition function $\mathcal{M}(\cdot)$ and also that $\mathbf{Q} = 0$ [cf. Eq. (2)]. Here, we set the

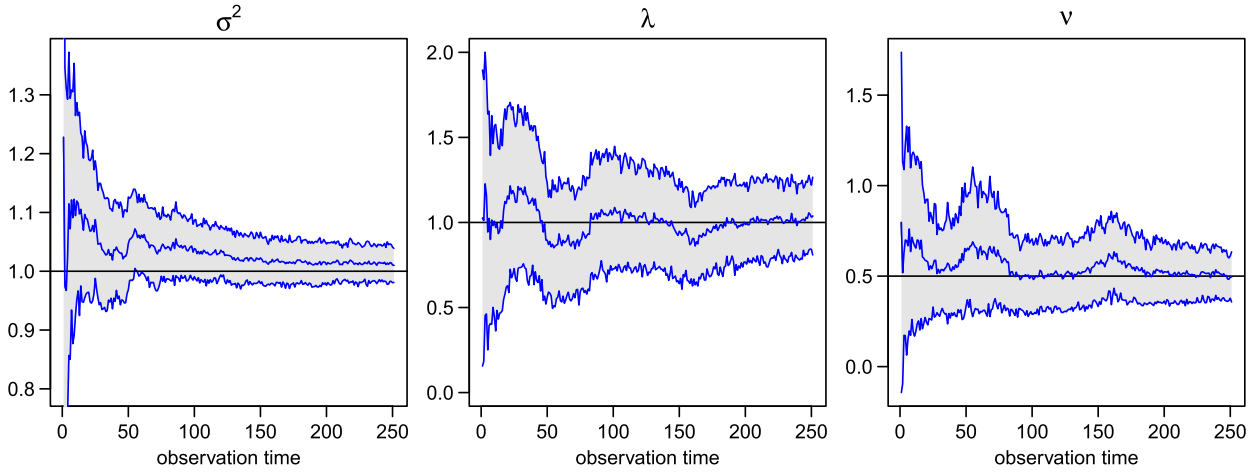

FIG. 5. For data simulated from the Lorenz-96 model (see section 4b), the marginal posterior distributions (filtered mean and 95% bands) of the parameters over time: (left to right) σ^2, λ, and ν.

forcing parameter $F = 8$ and the time step $\delta = 0.25$, resulting in a forward map with significant nonlinearities yielding distinctly non-Gaussian forecast distributions (see Fig. 2 in Bengtsson et al. 2003). A numerical solver is used to propagate the system over time.

We simulate the true value $\mathbf{x}_0^*$ of the initial state from a long run of the Lorenz-96 model. At each time δt, $t = 1, 2, \ldots, 250$ we take $m = n$ noisy observations according to (1) with $\mathbf{H} = \mathbf{I}$. We assume spatially correlated observation errors, with $\mathbf{R} = \mathbf{R}(\theta)$ defined by the Matérn covariance model

$$K(d;\theta) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\frac{d}{\lambda}\right)^{\nu}\mathcal{K}_{\nu}\left(\frac{d}{\lambda}\right),$$

where σ^2 is the sill parameter, λ is the spatial range parameter, ν is the smoothness parameter, and the distance d between $x_{t,i}$ and $x_{t,j}$ is defined as $\min\{|i-j|, 40-|i-j|\}$. Data are simulated using the parameter values $(\sigma^2, \lambda, \nu) = (1, 1, 0.5)$. We take the initial state distribution to be $\mathbf{x}_0 \sim (\mathbf{x}_0^*, 0.25\mathbf{I})$ and assume the following independent prior distributions for the parameters: $\sigma^2 \sim \mathcal{IG}(5, 5)$, $\lambda \sim \mathcal{N}^+(1.0, 0{,}64)$, and $\nu \sim \mathcal{N}^+(0.25, 0.25)$.

Using the method of Stroud and Bengtsson (2007) to handle σ^2, we applied our EnKF-Grid algorithm for inference on λ and ν to the resulting simulated data with 20 grid points per parameter, $N = 100$ ensemble members, and a Gaspari and Cohn (1999) taper with range 12. The marginal posterior distributions of the three parameters over time are shown in Fig. 5. As we can see, the posterior distributions again seem to be converging to the true values. The EnKF-Grid also produces estimates of the joint posterior distribution of the parameters. Figure 6 shows the strong posterior dependence between λ and ν at several time points.

5. Conclusions

We have presented new algorithms for sequential state and parameter estimation that combine information about the parameters from data at different time points in a consistent, probabilistic framework. The algorithms obtain the marginal posterior distribution of the parameters at each time point using a grid, normal, or particle approximation, while the distribution of the states, given the parameters, is obtained by the EnKF. The methods can also be combined with existing approaches for parameter estimation in the EnKF, such as state augmentation. We have shown in several numerical examples that the posterior distribution of the parameters, as approximated by our methods, is close to the true posterior, converges to the true parameter value, and strongly outperforms popular existing approaches.

While the current software implementation of our approach is not suitable for applications with truly high-dimensional states, we expect our methods to work in high dimensions as well, as long as the embedded EnKF is well suited and well tuned to the application if the parameters are known.

A separate question is how our method will scale to high-dimensional parameter vectors (i.e., a large number of unknown parameters). The computational cost of the EnKF-Grid approach is exponential in the number of unknown parameters, and this approach is, hence, most suitable when the number of parameters (minus the parameters that can be handled by state augmentation and other methods) is in the single digits. The computational cost of the EnKF-Normal approach is cubic in the number of parameters and should, thus, scale to moderately high parameter dimensions,

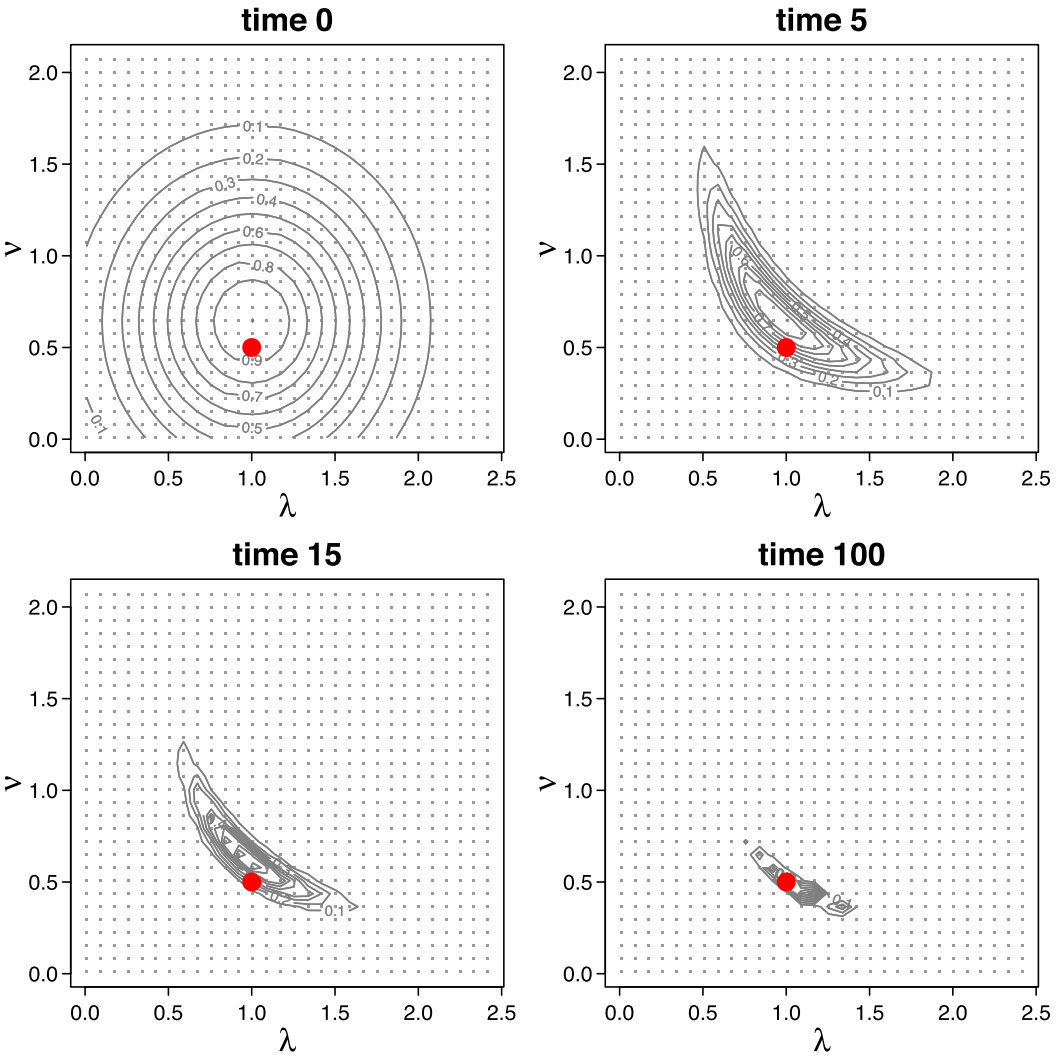

FIG. 6. For data simulated from the Lorenz-96 model (see section 4b), contours of the joint posterior distribution of the parameters λ and ν at time points: (top left) 0, (top right) 5, (bottom left) 15, and (bottom right) 100. The contour values are normalized to yield a maximum posterior density of 1 at each time point. The true value of $(\lambda, \nu) = (1, 0.5)$ is indicated by the red dot.

although the cost of the optimization procedure to find the posterior mode at each time point might become prohibitive. The EnKF-Particle approach is also cubic in the number of parameters due to the mixture of normals approximation.

Note that our methods ''break'' the dependence between the parameter approximations at successive time points and are, in their present form, unable to approximate the joint dependence structure in the posterior distribution of the parameters at different time points. This could potentially be remedied using a shift-based update to the parameters. In the case of the normal approximation, this shift would be similar to the one used for the state update in the EnKF, while in the grid-based approximation, a shift based on a piecewise linear approximation (cf. Anderson 2010) to the parameter density might be possible.

Acknowledgments. Stroud acknowledges the support of the Stallkamp Teaching Skills Program Fund at Georgetown University. Katzfuss' research was partially supported by U.S. National Science Foundation (NSF) Grant DMS-1521676. Wikle acknowledges the support of the NSF and the U.S. Census Bureau under NSF Grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program.

APPENDIX

Liu and West's Particle Filter

We describe the particle filter algorithm of Liu and West (2001) for sequential state and parameter estimation for a state-space model with transition density $p(\mathbf{x}_{t+1}|\mathbf{x}_t, \theta) = \mathcal{N}[\mathcal{M}[(\mathbf{x}_t), \mathbf{Q}(\theta)]$ and measurement density $p(\mathbf{y}_t|\mathbf{x}_t, \theta) = \mathcal{N}[\mathbf{H}(\boldsymbol{\theta})\mathbf{x}_t, \mathbf{R}(\theta)]$. The algorithm uses state augmentation and generates samples $[\mathbf{x}_t^{(i)}, \theta_t^{(i)}] \sim p(\mathbf{x}_t, \theta|\mathbf{Y}_t)$ for each time $t = 0, 1, \ldots, T$. The main idea is to use kernel density estimation to approximate the parameter distribution

$$p(\theta|\mathbf{Y}_t) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{N}[\theta|a\theta_t^{(i)} + (1-a)\overline{\theta}_t, (1-a^2)\mathbf{V}_t],$$

where $\theta_t^{(i)}$ are the parameter samples (particles) at time t, $\overline{\theta}_t$ is the posterior mean, $\mathbf{V}_t$ is the posterior variance-covariance matrix for the particles at time t, and a is a tuning parameter between 0 and 1. The joint posterior distribution for the states and parameters is defined recursively as

$$p(\mathbf{x}_{t+1}, \theta|\mathbf{Y}_{t+1}) \propto p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \theta)\int p(\mathbf{x}_t, \theta|\mathbf{Y}_t)p(\mathbf{x}_{t+1}|\mathbf{x}_t, \theta)d\mathbf{x}_t.$$

The algorithm is as follows:

1) Start with samples $[\mathbf{x}_0^{(i)}, \theta_0^{(i)}]$, $i = 1, \ldots, N$ from $p(\mathbf{x}_0, \theta \mid \mathbf{Y}_0)$.
2) For each observation, time $t + 1 = 1, \ldots, T$:
   (i) Generate state and parameter forecasts: $\mu_{t+1}^{(i)} = \mathcal{M}[\mathbf{x}_t^{(i)}]$ and $\mathbf{m}_t^{(i)} = a\theta_t^{(i)} + (1-a)\overline{\theta}_t$.
   (ii) Compute first-stage weights: $p[\mathbf{y}_{t+1}|\mu_{t+1}^{(i)}, \mathbf{m}_t^{(i)}]$.
   (iii) Sample the index k_i from $\{1, \ldots, N\}$ with probabilities proportional to $\{\lambda_{t+1}^{(i)}\}$.
   (iv) Draw $\tilde{\theta}_{t+1}^{(i)}$ from a normal density $\tilde{\theta}_{t+1}^{(i)} \sim \mathcal{N}[\mathbf{m}_t^{(k_i)}, (1-a^2)\mathbf{V}_t]$.
   (v) Draw $\tilde{x}_{t+1}^{(i)}$ from the transition density $\tilde{x}_{t+1}^{(i)} \sim \mathcal{N}[\mu_{t+1}^{(k_i)}, \mathbf{Q}(\tilde{\theta}_{t+1}^{(i)})]$.
   (vi) Compute the second-stage weights: $\omega_{t+1}^{(i)} \propto p[\mathbf{y}_{t+1}|\tilde{\mathbf{x}}_{t+1}^{(i)}, \tilde{\theta}_{t+1}^{(i)}]/p[\mathbf{y}_{t+1} \mid \mu_{t+1}^{(i)}, \mathbf{m}_t^{(i)}]$.
   (vii) Resample $[\mathbf{x}_{t+1}^{(i)}, \theta_{t+1}^{(i)}]$ from $\{[\tilde{\mathbf{x}}_{t+1}^{(i)}, \tilde{\theta}_{t+1}^{(i)}]\}$ with weights proportional to $[\omega_{t+1}^{(i)}]$.

Step 2(vii) provides samples from the joint posterior distribution $p(\mathbf{x}_{t+1}, \theta \mid \mathbf{Y}_{t+1})$, as desired. The algorithm requires the choice of a discount factor $\delta \in (0, 1)$, which determines the smoothing parameters as $a = (3\delta - 1)/2\delta$. The discount factor controls the degree of smoothing in the parameter distribution $p(\mathbf{x}, \theta|\mathbf{Y}_t)$, with large values corresponding to less smoothing and small values to more smoothing. Typical values of δ are between 0.95 and 1. For all of the examples in section 4, we use a value of $\delta = 0.98$.

REFERENCES

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

——, 2007a: An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus*, **59A**, 210–224, https://doi.org/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, https://doi.org/10.1016/j.physd.2006.02.011.

——, 2010: A non-Gaussian ensemble filter update for data assimilation. *Mon. Wea. Rev.*, **138**, 4186–4198, https://doi.org/10.1175/2010MWR3253.1.

Andrieu, C., A. Doucet, and V. Tadić, 2005: On-line parameter estimation in general state-space models. *Proc. 44th IEEE Conf. on Decision and Control/2005 European Control Conf.*, Seville, Spain, IEEE, 332–337, https://doi.org/10.1109/CDC.2005.1582177.

Bengtsson, T., C. Snyder, and D. Nychka, 2003: Toward a nonlinear ensemble filter for high-dimensional systems. *J. Geophys. Res.*, **108**, 8775, https://doi.org/10.1029/2002JD002900.

Brankart, J.-M., E. Cosme, C.-E. Testut, P. Brasseur, and J. Verron, 2010: Efficient adaptive error parameterizations for square root or ensemble Kalman filters: Application to the control of ocean mesoscale signals. *Mon. Wea. Rev.*, **138**, 932–950, https://doi.org/10.1175/2009MWR3085.1.

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

Carter, C. K., and R. Kohn, 1994: On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553, https://doi.org/10.1093/biomet/81.3.541.

De, D., 2014: Essays on Bayesian time series and variable selection. Ph.D. thesis, Texas A&M University, 83 pp., http://oaktrust.library.tamu.edu/bitstream/handle/1969.1/152793/DE-DISSERTATION-2014.pdf?sequence=1.

Dee, D. P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.*, **123**, 1128–1145, https://doi.org/10.1175/1520-0493(1995)123<1128:OLEOEC>2.0.CO;2.

DelSole, T., and X. Yang, 2010: State and parameter estimation in stochastic dynamical models. *Physica D*, **239**, 1781–1788, https://doi.org/10.1016/j.physd.2010.06.001.

Doucet, A., N. de Freitas, and N. Gordon, Eds., 2001: *Sequential Monte Carlo Methods in Practice*. Springer, 582 pp.

Dreano, D., P. Tandeo, M. Pulido, B. Ait-El-Fquih, T. Chonavel, and I. Hoteit, 2017: Estimating model-error covariances in nonlinear state-space models using Kalman smoothing and the expectation-maximization algorithm. *Quart. J. Roy. Meteor. Soc.*, **143**, 1877–1885, https://doi.org/10.1002/qj.3048.

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, https://doi.org/10.1029/94JC00572.

——, 2009: *Data Assimilation: The Ensemble Kalman Filter*. 2nd ed. Springer, 307 pp.

Frei, M., and H. Künsch, 2012: Sequential state and observation noise covariance estimation using combined ensemble Kalman and particle filters. *Mon. Wea. Rev.*, **140**, 1476–1495, https://doi.org/10.1175/MWR-D-10-05088.1.

Furrer, R., and T. Bengtsson, 2007: Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter

variants. *J. Multivar. Anal.*, **98**, 227–255, https://doi.org/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, https://doi.org/10.1002/qj.49712555417.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith, 1993: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc.*, **140**, 107–113, https://doi.org/10.1049/ip-f-2.1993.0015.

Grewal, M. S., and A. P. Andrews, 1993: *Kalman Filtering: Theory and Practice.* Prentice Hall, 368 pp.

Gu, Y., and D. S. Oliver, 2007: An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. *SPE J.*, **12**, 438–446, https://doi.org/10.2118/108438-PA.

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

——, and F. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **144**, 4489–4532, https://doi.org/10.1175/MWR-D-15-0440.1.

Johnson, S. G., 2011: The NLopt nonlinear-optimization package. Accessed 10 January 2018, https://nlopt.readthedocs.io/en/latest/.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **82**, 34–45.

Katzfuss, M., J. R. Stroud, and C. K. Wikle, 2016: Understanding the ensemble Kalman filter. *Amer. Stat.*, **70**, 350–357, https://doi.org/10.1080/00031305.2016.1141709.

Kitagawa, G., 1998: A self-organizing state-space model. *J. Amer. Stat. Assoc.*, **93**, 1203–1212.

Liu, J., and M. West, 2001: Combined parameter and state estimation in simulation-based filtering. *Sequential Monte Carlo Methods in Practice*, A. Doucet, J. de Freitas, and N. Gordon, Eds., Statistics for Engineering and Information Science Series, Springer, 197–223, https://doi.org/10.1007/978-1-4757-3437-9_10.

Lorenz, E., 1996: Predictability: A problem partially solved. *Proc. Seminar on Predictability*, Reading, United Kingdom, ECMWF, 1–18, https://www.ecmwf.int/sites/default/files/elibrary/1995/10829-predictability-problem-partly-solved.pdf.

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

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Houser, 2005: Dual state–parameter estimation of hydrological models using ensemble Kalman filter. *Adv. Water Resour.*, **28**, 135–147, https://doi.org/10.1016/j.advwatres.2004.09.002.

Nowak, W., 2009: Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator. *Water Resour. Res.*, **45**, W04431, https://doi.org/10.1029/2008WR007328.

Pitt, M. K., and N. Shephard, 1999: Filtering via simulation: Auxiliary particle filters. *J. Amer. Stat. Assoc.*, **94**, 590–599, https://doi.org/10.1080/01621459.1999.10474153.

Rowan, T., 1990: Functional stability analysis of numerical algorithms. Ph.D. dissertation, University of Texas at Austin, 218 pp.

Šmídl, V., and R. Hofman, 2011: Marginalized particle filtering framework for tuning of ensemble filters. *Mon. Wea. Rev.*, **139**, 3589–3599, https://doi.org/10.1175/2011MWR3586.1.

Snyder, C., T. Bengtsson, P. Bickel, and J. L. Anderson, 2008: Obstacles to high-dimensional particle filtering. *Mon. Wea. Rev.*, **136**, 4629–4640, https://doi.org/10.1175/2008MWR2529.1.

Storvik, G., 2002: Particle filters for state-space models with the presence of unknown static parameters. *IEEE Trans. Signal Process.*, **50**, 281–289, https://doi.org/10.1109/78.978383.

Stroud, J. R., and T. Bengtsson, 2007: Sequential state and variance estimation within the ensemble Kalman filter. *Mon. Wea. Rev.*, **135**, 3194–3208, https://doi.org/10.1175/MWR3460.1.

——, M. L. Stein, B. M. Lesht, D. J. Schwab, and D. Beletsky, 2010: An ensemble Kalman filter and smoother for satellite data assimilation. *J. Amer. Stat. Assoc.*, **105**, 978–990, https://doi.org/10.1198/jasa.2010.ap07636.

Tandeo, P., M. Pulido, and F. Lott, 2015: Offline parameter estimation using EnKF and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Quart. J. Roy. Meteor. Soc.*, **141**, 383–395, https://doi.org/10.1002/qj.2357.

Ueno, G., and N. Nakamura, 2014: Iterative algorithm for maximum-likelihood estimation of the observation-error covariance matrix for ensemble-based filters. *Quart. J. Roy. Meteor. Soc.*, **140**, 295–315, https://doi.org/10.1002/qj.2134.

——, and ——, 2016: Bayesian estimation of observation-error covariance matrix in ensemble-based filters. *Quart. J. Roy. Meteor. Soc.*, **142**, 2055–2080, https://doi.org/10.1002/qj.2803.

——, T. Higuchi, T. Kagimoto, and N. Hirose, 2010: Maximum likelihood estimation of error covariances in ensemble-based filters and its application to a coupled atmosphere–ocean model. *Quart. J. Roy. Meteor. Soc.*, **136**, 1316–1343, https://doi.org/10.1002/qj.654.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten, 2005: Improved treatment of uncertainty in hydrological modeling: Combining the strengths of global optimization and data assimilation. *Water Resour. Res.*, **41**, W01017, http://doi.org/10.1029/2004WR003059.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, https://doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2.

Wikle, C. K., 2002: A kernel-based spectral model for non-Gaussian spatio-temporal processes. *Stat. Modell.*, **2**, 299–314, https://doi.org/10.1191/1471082x02st036oa.

——, and L. M. Berliner, 2007: A Bayesian tutorial for data assimilation. *Physica D*, **230**, 1–16, https://doi.org/10.1016/j.physd.2006.09.017.

Xu, K., and C. K. Wikle, 2007: Estimation of parameterized spatio-temporal dynamic models. *J. Stat. Plan. Inference*, **137**, 567–588, https://doi.org/10.1016/j.jspi.2005.12.005.