⁶Ensemble Kalman Filter Updates Based on Regularized Sparse Inverse Cholesky Factors

WILL BOYLES^a AND MATTHIAS KATZFUSS^a

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

(Manuscript received 11 September 2020, in final form 13 April 2021)

ABSTRACT: The ensemble Kalman filter (EnKF) is a popular technique for data assimilation in high-dimensional nonlinear state-space models. The EnKF represents distributions of interest by an ensemble, which is a form of dimension reduction that enables straightforward forecasting even for complicated and expensive evolution operators. However, the EnKF update step involves estimation of the forecast covariance matrix based on the (often small) ensemble, which requires regularization. Many existing regularization techniques rely on spatial localization, which may ignore long-range dependence. Instead, our proposed approach assumes a sparse Cholesky factor of the inverse covariance matrix, and the nonzero Cholesky entries are further regularized. The resulting method is highly flexible and computationally scalable. In our numerical experiments, our approach was more accurate and less sensitive to misspecification of tuning parameters than tapering-based localization.

KEYWORDS: Bayesian methods; Filtering techniques; Kalman filters; Statistical techniques; Ensembles; Data assimilation

1. Introduction

In spatiotemporal data assimilation, the goal is to sequentially infer the state of a spatial field by combining noisy and incomplete observations with an evolution operator, which describes how the field evolves over time. This is a ubiquitous task across many scientific areas, with numerical weather prediction as a prominent example. From a statistical perspective, data assimilation can be viewed as filtering inference in a spatiotemporal state-space model, with the aim of obtaining the conditional distribution of the state given all data observed up to the current time point. Because the spatial field is often discretized at a high resolution, the state dimension may be very high, making approximations to the inference problem necessary.

A highly successful approximate filtering technique is the ensemble Kalman filter (EnKF; e.g., Evensen 1994; Burgers et al. 1998; Houtekamer and Mitchell 1998; Anderson 2001; Evensen 2007; Katzfuss et al. 2016; Houtekamer and Zhang 2016), a sequential Monte Carlo algorithm that represents distributions of interest by an ensemble. However, the EnKF update step involves estimation of the large forecast covariance matrix based on an often small ensemble, which requires regularization. See Ueno and Tsuchiya (2009) for a brief review of regularization in spectral or wavelet space.

Many existing EnKF regularization techniques rely on spatial localization via covariance tapering or local updates (e.g., Houtekamer and Mitchell 2001; Hamill et al. 2001; Ott et al. 2004; Furrer and Bengtsson 2007; Hunt et al. 2007; Bishop and Hodyss 2009; Anderson 2012; Bishop et al. 2017). Covariance tapering results in a sparse forecast covariance matrix, with zero entries corresponding to state variables that are more than a small number of grid points apart, which means that longer-range correlation in the forecast is ignored.

An alternative EnKF-regularization approach is to assume sparsity in the forecast precision (i.e., inverse covariance) matrix (Ueno and Tsuchiya 2009). A zero entry in the precision matrix means that the two corresponding state variables are conditionally uncorrelated given all other state variables. Importantly, this is often a weaker assumption than sparsity in the covariance matrix, in that two variables can be (approximately) conditionally uncorrelated even if they are strongly marginally (i.e., unconditionally) correlated. Thus, even a highly sparse precision matrix can capture nonzero long-range correlations, and hence may be preferable to sparsity in the covariance matrix. However, it can be challenging to ensure that an estimated precision matrix is valid (e.g., positive definite). In addition, the resulting EnKF update requires Cholesky factorization (or other decompositions) of the precision matrix, which introduces additional nonzero entries and thus increased computational cost.

Hence, it is advantageous to work directly with a sparse Cholesky factor of the precision matrix, sometimes called an inverse Cholesky factor. Positive diagonal entries are sufficient to ensure a valid inverse Cholesky factor. In addition, a zero inverse-Cholesky entry implies that the corresponding variables are conditionally uncorrelated given subsequently ordered variables (according to the row/column ordering of the Cholesky factor); thus, highly sparse inverse-Cholesky factors can often capture nonnegligible correlations at all spatial scales. This sparse-inverse-Cholesky idea has been very successful in general covariance-estimation problems (e.g., Smith and Kohn 2002; Huang et al. 2006). Related approaches have recently also been proposed in the EnKF context by Yang (2017, chapter 3) and

DOI: 10.1175/MWR-D-20-0299.1

^o Denotes content that is immediately available upon publication as open access.

[©] Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-20-0299.s1.

Corresponding author: Matthias Katzfuss, katzfuss@gmail.com

^{© 2021} American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Nino-Ruiz et al. (2018). These approaches require a suitable ordering of the state variables, which is typically done according to their corresponding spatial coordinates. Without further approximations or assumptions, these existing approaches may not efficiently scale to high dimensions and large numbers of observations, or they may be inaccurate for small ensemble sizes.

Here, we extend a recently proposed method for nonparametric inference on spatial covariance matrices (Kidd and Katzfuss 2021) for use in stochastic EnKF updates, estimating the forecast covariance matrix based on a sparse inverse Cholesky factor. In contrast to the coordinate ordering of spatial locations used in existing Cholesky approaches, our method uses a maximum-minimum-distance ordering, which can lead to more accurate approximations for a given sparsity level (e.g., Guinness 2018; Katzfuss and Guinness 2021). Our approach provides additional regularization motivated by recent results on the exponential decay of inverse Cholesky factors under this ordering. Our method is highly flexible and computationally scalable to highdimensional state and observation vectors. We demonstrate that, in contrast to localization-based updates, our approach is easier to tune and can accommodate various dependence scales.

The remainder of this document is organized as follows. In section 2, we review existing results and describe our proposed EnKF algorithm. In section 3, we provide numerical comparisons of our method to existing approaches. Section 4 concludes. A separate supplemental material document contains additional details and plots.

2. Methodology

a. State-space model

At time points t = 1, 2, ..., let \mathbf{x}_t denote the *n*-dimensional latent spatial field of interest and \mathbf{y}_t a corresponding vector of n_t noisy observations. Given an initial state distribution, $\mathbf{x}_0 \sim p(\mathbf{x}_0)$, we assume a state-space model that, at each time t = 1, 2, ..., consists of a linear Gaussian observation model with diagonal noise covariance \mathbf{R}_t :

$$\mathbf{y}_t | \mathbf{x}_t \sim \mathcal{N}_n \left(\mathbf{H}_t \mathbf{x}_t, \mathbf{R}_t \right), \tag{1}$$

and an evolution model with Markov structure:

$$\mathbf{x}_t | \mathbf{x}_{1:t-1} \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}), \tag{2}$$

which is often a computationally expensive, black-box model determined by a system of differential equations. Data assimilation essentially means sequential filtering inference on the distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})p(\mathbf{x}_t|\mathbf{y}_{1:t})$ of the current state \mathbf{x}_t given all data $\mathbf{y}_{1:t}$ observed so far, for t = 1, 2, ... We are interested in the setting where *n* and the n_t are large. For simplicity, we assume that the \mathbf{H}_t , \mathbf{R}_t , and the evolution model are all known, and that the observation distributions in (1) are Gaussian, although extensions for unknown model parameters and non-Gaussian observations are possible (e.g., Katzfuss et al. 2020).

b. Review of the ensemble Kalman filter (EnKF)

A popular and successful technique for data assimilation is the EnKF (Evensen 1994), which approximates the state distribution by an ensemble. At time *t*, assuming a previous filtering ensemble $\mathbf{x}_{t-1|t-1}^{(1:N)} = {\mathbf{x}_{t-1|t-1}^{(1)}, \dots, \mathbf{x}_{t-1|t-1}^{(N)}}$ with $\mathbf{x}_{t-1|t-1}^{(j)} \sim p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$, we can draw $\mathbf{x}_{t|t-1}^{(j)} \sim p(\mathbf{x}_t|\mathbf{x}_{t-1} = \mathbf{x}_{t-1}^{(j)})$ for $j = 1, \dots, N$ using the evolution model (2), to obtain the forecast ensemble $\mathbf{x}_{t|t-1}^{(1:N)}$ with $\mathbf{x}_{t|t-1}^{(j)} \sim p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$. This forecast ensemble, which can be thought of as a sample from a prior distribution, must then be updated based on the new data \mathbf{y}_t , to obtain the filtering ensemble $\mathbf{x}_{t|t}^{(1:N)}$ as a sample from the posterior. The stochastic EnKF update can be shown (e.g., Hunt et al. 2007) to be equivalent to

$$\mathbf{x}_{t|t}^{(j)} = \mathbf{\Sigma}_{t|t} (\mathbf{\Sigma}_{t|t-1}^{-1} \mathbf{x}_{t|t-1}^{(j)} + \mathbf{H}_{t}^{\mathsf{T}} \mathbf{R}_{t}^{-1} \mathbf{y}_{t}^{(j)}), \quad j = 1, \dots, N,$$
(3)

which requires perturbed observations $\mathbf{y}_{t}^{(j)} \sim \mathcal{N}(\mathbf{y}_{t}, \mathbf{R}_{t})$, the forecast covariance matrix $\boldsymbol{\Sigma}_{t|t-1}$, and the posterior precision $\boldsymbol{\Sigma}_{|t|t-1}^{-1} = \boldsymbol{\Sigma}_{|t|t-1}^{-1} + \mathbf{H}_{t}^{T} \mathbf{R}_{t}^{-1} \mathbf{H}_{t}$.

In practice, the forecast covariance matrix $\Sigma_{t|t-1}$ is unknown and must be estimated from the forecast ensemble $\mathbf{x}_{t|t-1}^{(1:N)}$. Estimating a large $n \times n$ matrix $\Sigma_{t|t-1}$ from a sample of small to moderate size N requires regularization. Existing EnKF regularization approaches often rely on spatial localization; discussions and comparisons will be presented in section 3.

c. Review of spatial covariance estimation based on sparse inverse Cholesky factors

Kidd and Katzfuss (2021) proposed a Bayesian nonparametric estimation method for spatial covariance matrices, which we briefly review here. Instead of estimating $O(n^2)$ entries in the covariance matrix, the basic idea is to infer a nearlinear number of nonzero entries in a sparse Cholesky factor of the inverse covariance matrix, whose nonzero entries are further regularized via prior distributions.

Temporarily drop time subscripts and assume an ensemble $\mathbf{x}^{(1:N)}$ from a distribution with covariance matrix $\boldsymbol{\Sigma}$. Order the variables $x_1^{(j)}, \ldots, x_n^{(j)}$ in $\mathbf{x}^{(j)}$ according to a maximum-minimumdistance (maximin) ordering (Guinness 2018; Schäfer et al. 2021a), which sequentially selects each variable in the ordering to maximize the minimum distance to all previously ordered variables. The ordering is assumed to be the same in each ensemble member $\mathbf{x}^{(j)}$ and implies a corresponding ordering of the rows and columns of Σ . For i = 2, 3, ..., n, let $g_m(i) \subset$ $(1, \ldots, i-1)$ be an index vector consisting of the indices of the $\min(m, i-1)$ nearest neighbors (ordered by increasing distance) among those ordered previously, for some positive integer $m \ll$ n. Thus, the distance to the neighbors decreases with the index i in the maximin ordering. If each state variable is associated with a geospatial location, the ordering and conditioning can be carried out using the physical distance between the corresponding spatial locations. (Other distance measures can also be used; see section 4.) The ordering and neighbor-selection scheme is illustrated in Fig. 1.

Consider a modified Cholesky decomposition of the precision matrix:

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\mathsf{U}}\boldsymbol{\mathsf{D}}^{-1}\boldsymbol{\mathsf{U}}^{\mathrm{T}},\tag{4}$$

where **D** = diag (d_1, \ldots, d_n) is a diagonal matrix with positive entries $d_i > 0$, and **U** is an upper triangular matrix with unit



FIG. 1. Illustration of maximin ordering and nearest-neighbor selection for $n = 10\,000$ locations on a 100×100 grid (small gray dots). The panels show, for three different indices *i*, the *i*th-ordered location (blue) and the previous i - 1 locations (black diamonds), including the nearest m = 6 previously ordered neighbors with indices in $g_m(i)$ (orange crosses).

diagonal, $\mathbf{U}_{ii} = 1$. Assume that \mathbf{U} is sparse, with at most *m* nonzero off-diagonal elements per column, and define $\mathbf{u}_i = \mathbf{U}_{g_m(i),i}$ as the nonzero off-diagonal entries in the *i*th column.

Kidd and Katzfuss (2021) assumed independent normalinverse-gamma (NIG) prior distributions for \mathbf{u}_i , d_i , such that $\mathbf{u}_i|d_i \sim \mathcal{N}(0, d_i\mathbf{V}_i)$ and $d_i \sim \mathcal{IG}(\alpha_i, \beta_i)$. The parameters of the NIG prior (and *m*, which determines the length of the \mathbf{u}_i) are determined by a vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^{\mathrm{T}}$ of tuning parameters as follows:

$$\alpha_i = 6, \quad \beta_i = 5 \,\theta_1 (1 - e^{-\theta_2/\sqrt{i}}),$$
 (5)

$$\mathbf{V}_{i} = \text{diag}(v_{i1}, \dots, v_{im}), \quad v_{ik} = e^{-\theta_{3}k} 5/\beta_{i}, \\
m = \max\{k: e^{-\theta_{3}k} > 10^{-2}\},$$
(6)

for i = 1, ..., n. To give some intuition, the parameterization of the v_{ik} is inspired by results for Matérn and similar covariances, which essentially imply an exponential decay in the entries of the inverse Cholesky factor as a function of the neighbor number k (Schäfer et al. 2021a, section 6.2), as assumed in (6); Cholesky entries for large k > m are assumed to be exactly zero, where m is a function of θ_3 . Further, β_i is proportional to the prior mean of the residual variance d_i for predicting $x_i^{(j)}$ based on its neighbors $\mathbf{x}_{g_m(i)}^{(j)}$; this residual variance decreases with the index i in the maximin ordering, because the neighbors $\mathbf{x}_{g_m(i)}^{(j)}$ are closer to $x_i^{(j)}$ for large i than for small i (e.g., compare the left and right panels in Fig. 1).

Assuming that the $\mathbf{x}^{(j)}$ follow independent *n*-variate Gaussian distributions with mean zero and covariance matrix $\boldsymbol{\Sigma}$, the NIG priors for \mathbf{u}_i , d_i are conjugate, resulting in closed-form NIG posteriors. Here, instead of considering the full posterior distributions, we obtain point estimates given by the posterior means:

$$\hat{\mathbf{u}}_{i} = \mathbf{G}_{i}^{-1} \mathbf{X}_{i}^{\mathrm{T}} \mathbf{x}_{i}, \quad \hat{d}_{i} = \tilde{\boldsymbol{\beta}}_{i} / (\tilde{\alpha}_{i} - 1),$$
(7)

where $\mathbf{G}_i = \mathbf{X}_i^T \mathbf{X}_i + \mathbf{V}_i^{-1}, \tilde{\alpha}_i = \alpha_i + N/2, \tilde{\beta}_i = \beta_i + (\mathbf{x}_i^T \mathbf{y}_i - \hat{\mathbf{u}}_i^T \mathbf{G}_i \hat{\mathbf{u}}_i^T)/2,$ $\mathbf{x}_i = (x_i^{(1)}, \dots, x_i^{(N)})^T$ consists of the *N* ensemble values at the *i*th grid location, and \mathbf{X}_i is an $N \times m$ matrix with *j*th row $-\mathbf{x}_{g_i}^{(j)T}$ consisting of the ensemble values at the *m* neighbor locations. Hence, given $\boldsymbol{\theta}$, the entries of **U** and **D** can be easily calculated based on the Cholesky factors of the $m \times m$ matrices \mathbf{G}_i , $i = 1, \ldots, n$.

The tuning parameters θ can be determined by maximizing (the log of) the integrated likelihood, which is the distribution of the ensemble $\mathbf{x}^{(1:N)}$ under the Gaussian assumption, with Σ (i.e., the \mathbf{u}_i and d_i) integrated out:

$$p(\mathbf{x}^{(1:N)}|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left\{ (|\mathbf{G}_{i}|^{-1}/|\mathbf{V}_{i}|)^{1/2} \times (\boldsymbol{\beta}_{i}^{\alpha_{i}}/\tilde{\boldsymbol{\beta}}_{i}^{\tilde{\alpha}_{i}}) \times [\Gamma(\tilde{\alpha}_{i})/\Gamma(\alpha_{i})] \right\},$$
(8)

where Γ denotes the gamma function, and (8) depends on θ via (5)–(6). We assume that θ_1 , θ_2 , θ_3 are all positive, and so the optimization of (8) is performed on the log scale.

This concludes the review of Kidd and Katzfuss (2021).

d. EnKF update using regularized sparse inverse Cholesky

Let us return to the EnKF setting in section 2b. At time *t*, given the forecast ensemble $\mathbf{x}_{t|t-1}^{(1:N)}$ (and tuning parameters $\boldsymbol{\theta}$), the goal is to carry out the stochastic EnKF update in (3) based on a regularized estimate of the forecast covariance matrix $\boldsymbol{\Sigma}_{t|t-1}$.

We propose to estimate $\Sigma_{t|t-1}$ using the method from section 2c, defining $\mathbf{x}^{(1:N)}$ as the centered forecast ensemble $\mathbf{x}_{i|t-1}^{(1:N)}$ (i.e., after subtracting the ensemble mean at each location). Using the point estimates of \mathbf{u}_i and d_i , i = 1, ..., n, as in (7), we can form the sparse triangular matrix \mathbf{U} and the diagonal matrix \mathbf{D} in (4). From this, it is straightforward to compute a prior Cholesky factor $\mathbf{L}_{t|t-1} = \mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}$, the prior precision $\Sigma_{t|t-1}^{-1} = \mathbf{L}_{t|t-1}^{\mathrm{T}} \mathbf{L}_{t|t-1}$, the posterior precision $\Sigma_{t|t}^{-1} = \mathbf{L}_{t|t-1}^{\mathrm{T}} \mathbf{H}_{t}$, and finally the posterior Cholesky factor $\mathbf{L}_{t|t}$ with $\Sigma_{t|t}^{-1} = \mathbf{L}_{t|t}^{\mathrm{T}} \mathbf{L}_{t|t-1}$.

$$\mathbf{x}_{t|t}^{(j)} = \mathbf{L}_{t|t}^{-1} \mathbf{L}_{t|t}^{-T} (\mathbf{L}_{t|t-1}^{T} \mathbf{L}_{t|t-1} \mathbf{x}_{t|t-1}^{(j)} + \mathbf{H}_{t}^{T} \mathbf{R}_{t}^{-1} \mathbf{y}_{t}^{(j)}), \quad j = 1, \dots, N.$$
(9)

Note that we only compute and work with the sparse precision matrices $\Sigma_{t|t-1}^{-1}$ and $\Sigma_{t|t}^{-1}$ and their respective Cholesky

factors $L_{t|t-1}$ and $L_{t|t}$; the dense covariance matrices $\Sigma_{t|t-1}$ and $\Sigma_{t|t}$ are never explicitly computed.

ALGORITHM 1: RSIC-ENKF

- 1) Using state grid locations, compute maximin ordering and nearest-neighbor indices $g_{m^*}(i), i = 1, ..., n$, for large m^* $(e.g., m^* = 50)$
- 2) Initialize ensemble: $\mathbf{x}_{0|0}^{(j) \text{ ind}} \sim p(\mathbf{x}_0)$ for $j = 1, \dots, N$
- 3) for t = 1, 2..., do
- Compute forecast ensemble using (2): $\mathbf{x}_{t|t-1}^{(j)} \sim$ 4) $p(\mathbf{x}_t | \mathbf{x}_{t-1} = \mathbf{x}_{t-1}^{(j)})$ for j = 1, ..., N
- Optimize $\boldsymbol{\theta} = \arg \max_{\boldsymbol{\theta}} \log p(\mathbf{x}^{(1:N)}|\boldsymbol{\theta})$ [see (8)], with the centered $\mathbf{x}_{l|t-1}^{(1:N)}$ as $\mathbf{x}^{(1:N)}$ 5)
- 6) Update $m, g_m(i) \subset g_{m^*(i)}, \alpha_i, \beta_i, \mathbf{V}_i, i = 1, \dots, n$, as in (5) and (6) based on θ
- Compute $\hat{\mathbf{u}}_i$, \hat{d}_i , i = 1, ..., n as in (7); form sparse 7) matrices **U** and **D** [see (4)]
- Compute Cholesky factors $\mathbf{L}_{t|t-1} = \mathbf{D}^{-1/2} \mathbf{U}^{\mathrm{T}}$ and $\mathbf{L}_{t|t} =$ 8) $\left(\mathbf{L}_{t|t-1}^{\mathrm{T}}\mathbf{L}_{t|t-1} + \mathbf{H}_{t}^{\mathrm{T}}\mathbf{R}_{t}^{-1}\mathbf{H}_{t}\right)^{1/2}$
- Update ensemble: $\mathbf{x}_{l|t}^{(j)} = \mathbf{L}_{l|t}^{-T} \mathbf{L}_{t|t}^{-T} (\mathbf{L}_{t|t-1}^{T} \mathbf{L}_{t|t-1} \mathbf{x}_{t|t-1}^{(j)} + \mathbf{H}_{t}^{T} \mathbf{R}_{t}^{-1} \mathbf{y}_{t}^{(j)})$ with $\mathbf{y}_{t}^{(j)} \sim \mathcal{N}(\mathbf{y}_{t}, \mathbf{R}_{t})$, for $j = 1, \dots, N$ 9)
- 10) end for

e. Summary and computational complexity

Our regularized sparse inverse Cholesky (RSIC) EnKF procedure is summarized in algorithm 1.

The computational complexity of our RSIC update is essentially linear in n for fixed m and N. Specifically, in algorithm 1, lines 6, 7, and 9, and evaluation of $p(\mathbf{x}^{(1:N)}|\boldsymbol{\theta})$ in line 5 require $\mathcal{O}[n(m^2N+m^3)]$ time, where $m \le 20$ (and often even m < 10) in our numerical experiments. However, while Σ_{tlt}^{-1} is very sparse, its Cholesky factor $L_{t|t}$ can contain additional nonzeroes, resulting in a complexity for line 8 that may not be linear in n anymore; in this case, we can maintain the linear complexity by computing $L_{t|t}$ using an incomplete Cholesky algorithm with zero fill-in as described in Schäfer et al. (2021b, section 4a). This incomplete-Cholesky approach also allows us to deal with nondiagonal \mathbf{R}_t , provided that $\mathbf{R}_t^{-1} \mathbf{y}_t^{(j)}$ and entries of $\mathbf{H}_t^{\mathrm{T}} \mathbf{R}_t^{-1} \mathbf{H}_t$ can be computed relatively cheaply. The ordering and nearest neighbors for a large m^* in line 1 can also be computed in near-linear time in n(Schäfer et al. 2021a,b). For any $m \le m^*$ implied by a specific θ , $g_m(i)$ can then simply be selected as the first m entries of $g_{m^*}(i)$; if a $m > m^*$ does happen to occur, then the $g_m(i)$ must be recomputed. While the optimization in line 5 requires multiple evaluations of the integrated likelihood, section S2 in the supplemental material shows that similar values of θ may be obtained even when reducing computational cost via warm starts (i.e., initializing the optimization at the previous θ) or skipping lines 5 and 6 altogether at most time points.

In addition, many of the most expensive computations, such as evaluating the integrated likelihood in (8) and computing the estimates in (7), are perfectly parallel for i = 1, ..., n, while the ensemble updates in line 9 can be computed in parallel over j = 1, ..., N. Hence, for systems with expensive evolution operators, we expect the cost of the RSIC update to be



simple one-dimensional setting: posterior/filtering means $\mathbb{E}(\mathbf{x}|y)$ given a single observation y = 1 at location 0.5 with noise variance $\mathbf{R} = 0.01$, assuming a Gaussian prior/forecast distribution for x with mean zero and exponential covariance matrix with unit variance and range 0.4, on a regular grid of size n = 500 on the unit interval. We show the exact posterior mean, the RSIC-EnKF update (with m = 2), a local update based on observations within a distance of 0.1, and a tapered update with radius 0.1, all based on the same prior ensemble of size N = 1000.

negligible relative to the cost of the forecast step in line 4 (see Figs. S7 and S8 in the online supplemental material).

3. Numerical comparisons

a. Qualitative comparison to localization in a toy example

EnKF updates typically require regularization of the forecast covariance matrix (see section 2b), often via spatial localization. We conducted a qualitative comparison of our updating procedure to two popular localization methods in a simple toy example with long-range dependence shown in Fig. 2. Our RSIC-EnKF update (lines 5–9 in algorithm 1) produced an ensemble mean that was close to the exact posterior mean, despite RSIC being based on a very sparse inverse Cholesky factor $\mathbf{L}_{t|t-1}$ with at most m = 2 nonzero entries per row. Local updating computed the exact update for each state element x_i , but only based on data within a distance of 0.1 (i.e., only based on observations at the nearest 100 grid points); by definition, the posterior mean of state elements located outside of the interval [0.4, 0.6] ignored the observation at location 0.5 and was thus equal to the prior ensemble mean, which was roughly zero.

For localization via tapering, the EnKF update in (3) was based on an estimate of the prior/forecast covariance matrix $\Sigma_{t|t-1}$ given by the sample covariance of the prior ensemble multiplied element-wise by a Wendland correlation matrix with a radius of 0.1. Thus, the 500 \times 500 forecast covariance matrix had roughly 100 nonzero entries per row, and state elements more than 50 grid points apart were assumed to be independent a priori. Despite this fairly dense covariance matrix, the taper update was highly inaccurate. This illustrates a severe limitation of





FIG. 3. Comparison of energy score (relative to exact update) for posterior ensemble based on different updates at a single time point for a Gaussian forecast distribution with exponential covariance with unit variance and range λ on a grid of size $n = 35 \times 35 = 1225$ on the unit square, with $\mathbf{H} = \mathbf{R} = \mathbf{I}_n$. Exact denotes update using the exact forecast covariance matrix; Taper Exact and Taper Sample multiplied the exact and sample covariance, respectively, by a Wendland taper; and the taper radii 0.1 and 0.5 indicated in the legends result in covariance matrices with roughly 38 and 960 nonzero entries per row, respectively. RSIC performed almost as well as the exact update across all considered spatial ranges in the forecast distribution, even when *m* was misspecified and fixed at 5 and 10; in contrast, tapering with a small (0.1) or large (0.5) radius performed poorly for large and small ranges, respectively.

tapering-based localization for long-range forecast dependence on high-resolution grids: A small tapering radius can make the updates inaccurate, while a larger tapering radius will result in a fairly dense forecast covariance matrix and hence high computational cost. In addition, decomposing or inverting a large covariance matrix is often expensive due to fill-in even if the matrix is very sparse (e.g., Lipton et al. 1979).

b. Update at a single time point based on a Gaussian forecast

Figure 3 shows a comparison of different EnKF updates at a single time point for a Gaussian forecast distribution on a grid of size $n = 35 \times 35 = 1225$ on the unit square. We randomly sampled a true state vector **x** from the forecast distribution, generated a corresponding observation vector **y** from (1), and quantified the accuracy of the methods by computing the energy score of the resulting posterior or filtering ensembles relative to the true **x**, averaged over 20 repetitions. The energy score (Gneiting et al. 2008) is a proper scoring rule (e.g., Gneiting and Katzfuss 2014) that simultaneously quantifies the calibration and sharpness of the joint posterior distribution, as characterized by the updated ensemble. (We also conducted comparisons based on the mean square error of the posterior ensemble mean, and the relative results were very similar).

In Fig. 3a, RSIC-EnKF updates were almost as accurate as updates using the exact forecast covariance for moderate to large N, despite using highly sparse Cholesky factors with m < 10 in all settings. Updates using the sample covariance (without regularization) performed poorly for small N, as expected. Updates using tapering-based localization were also relatively inaccurate, even when based on the true forecast covariance

matrix, which (for fixed radius) provides a lower bound on the error no matter the ensemble size.

In Fig. 3b, we examined the robustness of various updates to varying dependence ranges in the forecast distribution. Tapering updates with a small radius of 0.1 performed increasingly poorly as the range parameter increased; tapering updates with a large tapering radius of 0.5 (which resulted in almost completely dense estimates of the forecast covariance matrix) exhibited the opposite behavior. RSIC-EnKF was more accurate than all tapering approaches for all levels of the range parameter; interestingly, when we arbitrarily fixed m = 5 or m = 10 (instead of using the *m* implied by the estimated θ as in line 6 of algorithm 1), the results were almost indistinguishable.

In summary, tapering-based localization was highly sensitive to the chosen tapering radius, and it generally performed relatively poorly despite relatively dense covariance estimates and hence high computational cost. In contrast, RSIC-EnKF was robust with respect to a wide range of dependence scales and to "misspecification" of *m*. Section S1 in the supplemental material contains additional simulation results for various signal-to-noise ratios, observation densities, and tapering radii; the results were consistent with Fig. 3, with RSIC-EnKF performing better than even the optimal tapering updates in virtually all considered settings.

c. Nonlinear Lorenz model

We considered a nonlinear evolution model (Lorenz 2005, section 4) that replicates features of atmospheric variables along a latitudinal band. This model (called Model III in Lorenz 2005) produces trajectories with both long-range and short-range spatial dependence; thus, we consider it a more realistic data-assimilation testbed than the popular model of



FIG. 4. For Lorenz05 with n = 1920 at a single time point [section 3c(1)], comparison of average energy score as a function of ensemble size N for different EnKF update methods. RSIC's Cholesky factors were highly sparse with $m \le 20$, while the taper radii 0.025, 0.1, and 0.3 (on a circular domain with unit circumference) indicated in the legend result in covariance matrices with roughly 15, 61, and 185 nonzero entries per row, respectively.

Lorenz (1996), for which spatial dependence is negligible for state elements more than one grid point apart. We considered Model III (henceforth called Lorenz05) with n = 1920 state grid points, and parameters K = 64, F = 15, b = 9, c = 4, and I = 10 (see Lorenz 2005, for a detailed description of the model and its parameters). We solved Lorenz05 on a circle with unit circumference using a fourth-order Runge–Kutta scheme, using code from Jurek and Katzfuss (2020).

1) LORENZ SIMULATION AT A SINGLE TIME POINT

First, we compared different updating schemes at a single time point, with the forecast ensemble and true state drawn from a very long run of Lorenz05. For each of 30 simulations and true sampled states, we generated $n_t = 96$ observations based on the true state as in (1), using a noise variance of $\mathbf{R}_t = \tau^2 \mathbf{I}_{n_t}$ with $\tau = 4$ and with \mathbf{H}_t comprised of randomly sampled rows from the identity matrix \mathbf{I}_n .

Figure 4 shows the averaged energy scores for different EnKF updates and different ensemble sizes. The fast decay of the entries of the inverse Cholesky factor exploited by RSIC in (6) seems to hold for the Lorenz05 covariance matrix (see Fig. S5 in the supplemental material), and so RSIC again performed very well for all ensemble sizes despite highly sparse Cholesky factors with $m \le 20$ in all settings. Tapering-based updates were only competitive for small ensemble size and a large tapering radius of 0.3, which implies a fairly dense covariance matrix and hence high computational cost.

2) SEQUENTIAL LORENZ SIMULATION

We then considered data assimilation using Lorenz05. Starting with an initial draw and ensemble as in section 3c(1) at time t = 0, we used Lorenz05 as the evolution model in (2), assimilating $n_t = 96$ observations at each time t that were simulated based on the true state \mathbf{x}_t as in section 3c(1).

As shown in Fig. 5, RSIC-EnKF was able to track the true state very well. Figure 6 shows a comparison of RSIC to several tapering approaches. RSIC was more accurate in all settings, despite relying on much sparser matrices. We also tried to



FIG. 5. For time t = 21 in the sequential Lorenz05 setting of section 3c(2), the observations \mathbf{y}_t , the true state \mathbf{x}_t , and the RSIC filtering ensemble with N = 50 are plotted against the n = 1920 spatial grid indices.

include tapering with a smaller radius of 0.03, but the resulting filtering ensembles were wildly oscillatory and resulted in numerical errors (see Fig. S7 in the supplemental material).

Additional figures in section S2 in the supplemental material show that most of the computational time for RSIC-EnKF was spent on the forecast steps, with negligible time spent on the update steps. Also, the estimated tuning parameters θ were relatively stable over time after a short burn-in period, indicating that θ optimization may not be necessary at most time points. Figure S9 provides further comparison results in this setting on other metrics such as spread and MSE, showing that RSIC had both lower spread and lower MSE than the tapering updates. All scores were stable and did not change much after t = 20 when running the simulations for more time points. Further exploratory comparisons of prior (i.e., forecast) scores and using variance inflation (with inflation factors between 1.05 and 1.2) led to similar conclusions, with RSIC performing better than tapering updates.

4. Conclusions and future work

We have proposed a new EnKF updating procedure that relies on a regularized sparse Cholesky factor of the inverse forecast covariance matrix. The proposed method is scalable to high dimensions, is easy to tune, can accommodate various dependence scales, and was considerably more accurate than tapering-based localization in our numerical experiments.

It is straightforward to extend our approach to non-Gaussian observations or unknown parameters in the evolution or observation models by combining it with the hierarchical EnKF methods in (Katzfuss et al. 2020). Our approach can also be extended to multivariate or strongly nonstationary spatial processes, by carrying out the maximin ordering and nearestneighbor selection using a correlation distance based on an offline or preliminary estimate of the correlation matrix (M. Kang and M. Katzfuss 2021, unpublished manuscript; Kidd and Katzfuss 2020). If the state vector consists of multiple different geophysical variables, we recommend standardizing them to have comparable marginal variances, to ensure a sensible





FIG. 6. Comparison of average energy scores for the filtering distributions in the sequential Lorenz05 setting. RSIC's Cholesky factors were highly sparse with $m \le 15$, while the taper radii 0.1 and 0.3 indicated in the legend result in covariance matrices with roughly 61 and 185 nonzero entries per row, respectively.

regularization of the inverse-Cholesky entries. Our numerical comparisons focused on the accuracy of the EnKF update step under the assumption that the observation and evolution models in (1) and (2) were correctly specified; for real-data applications, an investigation of the sensitivity of our approach to model misspecification is warranted. Further, it would be of interest to extend our method to deterministic updates and to handle nonlinear observation operators. Finally, while we only used point estimates of the entries of the inverse Cholesky factor given by their posterior means, it would be useful to account for uncertainty in their full posterior distribution.

Acknowledgments. Katzfuss's research was partially supported by National Science Foundation (NSF) Grants DMS-1654083, DMS-1953005, and CCF-1934904. We thank Brian Kidd, Marcin Jurek, Florian Schäfer, Jonathan R. Stroud, Christopher K. Wikle, and Joseph Guinness for helpful comments and discussions.

Data availability statement. R code implementing our method (based on code from Kidd and Katzfuss 2021) and numerical comparisons are provided with this article.

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.
- —, 2012: Localization and sampling error correction in ensemble Kalman filter data assimilation. *Mon. Wea. Rev.*, **140**, 2359– 2371, https://doi.org/10.1175/MWR-D-11-00013.1.
- Bishop, C. H., and D. Hodyss, 2009: Ensemble covariances adaptively localized with ECO-RAP. Part 2: A strategy for the atmosphere. *Tellus*, 61A, 97–111, https://doi.org/10.1111/ j.1600-0870.2008.00372.x.
- —, J. S. Whitaker, and L. Lei, 2017: Gain form of the ensemble transform Kalman filter and its relevance to satellite data assimilation with model space ensemble covariance localization.

Mon. Wea. Rev., 145, 4575–4592, https://doi.org/10.1175/MWR-D-17-0102.1.

- Burgers, G., P. Jan 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.
- Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res., 99, 10143–10162, https://doi.org/10.1029/94JC00572.
- —, 2007: Data Assimilation: The Ensemble Kalman Filter. Springer, 272 pp., https://doi.org/10.1007/978-3-642-03711-5.
- 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.
- Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. Annu. Rev. Stat. Appl., 1, 125–151, https://doi.org/10.1146/ annurev-statistics-062713-085831.
- —, L. Stanberry, E. Grimit, L. Held, and N. A. Johnson, 2008: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, **17**, 211, https://doi.org/10.1007/s11749-008-0114-x.
- Guinness, J., 2018: Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics*, 60, 415–429, https://doi.org/10.1080/00401706.2018.1437476.
- Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790, https://doi.org/ 10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO:2.
- 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 —, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137, https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF> 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.

- Huang, J. Z., N. Liu, M. Pourahmadi, and L. Liu, 2006: Covariance matrix selection and estimation via penalized normal likelihood. *Biometrika*, **93**, 85–98, https://doi.org/10.1093/biomet/ 93.1.85.
- Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, 230, 112–126, https:// doi.org/10.1016/j.physd.2006.11.008.
- Jurek, M., and M. Katzfuss, 2020: Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatiotemporal filtering. http://arxiv.org/abs/2006.16901.
- Katzfuss, M., and J. Guinness, 2021: A general framework for Vecchia approximations of Gaussian processes. *Stat. Sci.*, 36, 124–141, https://doi.org/10.1214/19-STS755.
- —, 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.
- —, —, and —, 2020: Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models. J. Amer. Stat. Assoc., 115, 866–885, https://doi.org/ 10.1080/01621459.2019.1592753.
- Kidd, B., and M. Katzfuss, 2021: Bayesian nonstationary and nonparametric covariance estimation for large spatial data. https://arxiv.org/abs/2012.05967.
- Lipton, R. J., D. J. Rose, and R. E. Tarjan, 1979: Generalized nested dissection. *SIAM J. Numer. Anal.*, 16, 346–358, https:// doi.org/10.1137/0716027.

- Lorenz, E. N., 1996: Predictability—A problem partly solved. Seminar on Predictability, Shinfield Park, Reading, ECMWF, 18 pp., https://doi.org/10.1017/CBO9780511617652.004.
 —, 2005: Designing chaotic models. J. Atmos. Sci., 62, 1574–1587,
- https://doi.org/10.1175/JAS3430.1. Nino-Ruiz, E. D., A. Sandu, and X. Deng, 2018: An ensemble Kalman
- filter implementation based on modified Cholesky decomposition for inverse covariance matrix estimation. *SIAM J. Sci. Comput.*, 40, A867–A886, https://doi.org/10.1137/16M1097031.
- Ott, E., and Coauthors, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, 56A, 415–428, https:// doi.org/10.3402/tellusa.v56i5.14462.
- Schäfer, F., T. J. Sullivan, and H. Owhadi, 2021a: Compression, inversion, and approximate PCA of dense kernel matrices at nearlinear computational complexity. *Multiscale Model. Simul.*, 19, 688–730, https://doi.org/10.1137/19M129526X.
- —, M. Katzfuss, and H. Owhadi, 2021b: Sparse Cholesky factorization by Kullback–Leibler minimization. https://arxiv.org/ abs/2004.14455.
- Smith, M., and R. Kohn, 2002: Parsimonious covariance matrix estimation for longitudinal data. J. Amer. Stat. Assoc., 97, 1141–1153, https://doi.org/10.1198/016214502388618942.
- Ueno, G., and T. Tsuchiya, 2009: Covariance regularization in inverse space. *Quart. J. Roy. Meteor. Soc.*, **135**, 1133–1156, https://doi.org/10.1002/qj.445.
- Yang, B., 2017: Particle and ensemble methods for state space models. Ph.D. thesis, George Washington University, 144 pp.