Statistica Sinica **29** (2019), 2205-2227
doi:https://doi.org/10.5705/ss.202017.0217

A MATRIX-FREE METHOD FOR SPATIAL-TEMPORAL GAUSSIAN STATE-SPACE MODELS

Debashis Mondal and Chunxiao Wang

Oregon State University

Abstract: This study develops a scalable matrix-free h-likelihood method for spatial-temporal Gaussian state-space models. The state vectors are constructed in such a way that they follow spatial-temporal Gaussian autoregressions that are consistent with the conditional formulation of auto-normal spatial fields. The proposed h-likelihood method provides the same inferences as those obtained from the Kalman filter and residual maximum likelihood analyses. However, for data from a large number of spatial sites, our method has significant computational advantages. Furthermore, we describe inferences in small time steps and indicate how the proposed method can be adapted to other complex spatial-temporal dynamical models based on stochastic partial differential equations. The proposed method applies to data with regularly or irregularly sampled spatial locations. Lastly, we illustrate our method by means of a simulation study and a data example on atmospheric concentrations of total nitrate across eastern North America.

Key words and phrases: Advection-diffusion, conditional autoregression, discrete cosine transform, Gaussian Markov random field, H-likelihood, incomplete Cholesky, Kalman filter, Lanczos algorithm, Residual likelihood, stochastic partial differential equation, trust region.

1. Introduction

This study develops statistical inferences for spatial-temporal Gaussian state-space models. We focus on dynamical models that are consistent with the conditional formulation of lattice-based Gaussian random fields and that can be used as building blocks to develop subsequent and more complex spatial-temporal inferences. Following Besag (1974, 1986), Künsch (1987), and Cressie (1993), among others, an extensive body of literature has been developed on the conditional modeling of spatial variables. In a spatial setting, conditional modeling gives rise to sparse dependence structures and swift statistical computations. In contrast, the aforementioned extensions to spatial-temporal settings require the exponential of the negative of the spatial precision matrix to define temporal dependence, which destroys sparse structures and presents computational challenges. For example, traditional time-series methods used for the estimation and conditional simulation of state variables, such as Kalman filtering (Kalman (1960); Kalman and Bucy (1961)), work well for data with small to moderately large spatial locations. However, for data from a large number of spatial locations, implementing Kalman filtering is not possible without reducing the number of dimensions or replacing certain spatial covariance matrices with sample versions constructed from ensembles of stochastic simulations; see Mardia et al. (1998), Wikle and Cressie (1999) and Evensen (1994, 2009). Whereas dimension reductions and ensembles of stochastic simulations can result in a loss of information, parameter estimations, conditional simulations, and log-likelihood computations introduce further challenges, because they require evaluating the square root or the determinant of large covariance matrices.

As an alternative, we draw upon the works of Lee and Nelder (1996, 2001) and Dutta and Mondal (2015, 2016) to develop a scalable matrix-free h-likelihood method that yields the same inferences obtained from Kalman filtering and residual maximum likelihood (REML) analyses. The h-likelihood method is faster than the iteratively nested Laplacian (and related) approximations of Rue, Martino and Chopin (2009) and Lindgren, Rue and Lindström (2011), as shown explicitly and numerically for the spatial case in Dutta and Mondal (2015, 2016). Furthermore, the h-likelihood method explains why Kalman filtering suffers from computational challenges and resolves the inferential challenges discussed in Sigrist, Künsch and Stahel (2015). The novel elements of the method include how we represent a spatial-temporal state-space model as a linear regression model and how the estimations employ matrix-free computations. These computations include the following: (i) an adaptation of the two-dimensional discrete cosine transformation that arises in the spectral decomposition of spatial-temporal autoregressions and that allows fast matrix-free matrix-vector multiplications; (ii) a preconditioned matrix-free scalable Lanczos algorithm that solves nonsparse matrix equations; (iii) a matrix-free Hutchinson trace estimator that stochastically approximates the trace of a matrix; (iv) a robust matrix-free trust region method that solves the REML score equations; and (v) a stochastic approximation of the differences in log-REML functions.

The past two decades have witnessed significant advances in general theory and applications of continuum spatial-temporal dynamical processes. Several noteworthy works on dynamical models provide different perspectives on the statistical analyses, including those of Brown et al. (2000), Brix and Diggle (2001), Cressie and Wikle (2011) and Sigrist, Künsch and Stahel (2015). Thus, as a second development, we show that the spatial-temporal Gaussian autoregressions and matrix-free method for inferences presented here can be adapted to and embodied in the above-mentioned research. To this end, we also describe inferences at small time steps, discuss how the proposed method applies to spatial-temporal models based on stochastic partial differential equations (SPDEs), such as advection diffusions, and outline an extension to lattice-based dynamical models that are consistent with fractional and Matérn spatial fields. We illustrate our method by means of a simulation study and analyze two data sets. The first contains monthly data on soil moisture content across North America (see Supplementary Material), and the second contains data on atmospheric concentrations of total nitrate across eastern North America (Section 6.2).

2. Spatial-temporal Models

2.1. A class of spatial-temporal dynamical models

Let $\psi(t)$ be a stationary spatial-temporal Gaussian process on the two-dimensional integer lattice $\mathcal{Z}^2$ at time $t = 0, 1, \ldots$ that evolves from the continuous-time dynamical model

$$d\psi(t) + B\psi(t)dt = dz(t). \tag{2.1}$$

In equation (2.1), B is a suitable infinite-dimensional normal matrix, and $z(t)$ represents an infinite-dimensional vector of independent Brownian motions with mean zero and variance τ^2. The matrix B determines how the instantaneous change in a spatial location depends on the current values at that location and the surrounding spatial locations. The solution ψ_t, for discrete time $t = 1, 2, \ldots,$ is

$$\psi_t = \exp\left\{\frac{-(B + B^{\scriptscriptstyle T})}{2}\right\}\psi_{t-1} + \nu_t, \tag{2.2}$$

where ν_t is an infinite-dimensional Gaussian vector with mean zero and covariance matrix $\tau^2(B + B^{\scriptscriptstyle T})^{-1}[I - \exp\{-(B + B^{\scriptscriptstyle T})\}]$. Furthermore, it follows that components of ψ_t are Gaussian with mean zero and have the infinite-dimensional covariance matrix $\tau^2(B + B^{\scriptscriptstyle T})^{-1}$. Let

$$C = \tau^{-2}(B + B^{\scriptscriptstyle T}), \quad V = C^{-1}\{I - \exp(-\lambda_0 C)\}, \quad \lambda_0 = \tau^2.$$

We focus on spatial-temporal lattice processes of the form (2.1)–(2.2), following Besag (1977). Besag (1977) noted that there is no loss in the distributional properties of ψ_t as a result of replacing B in equation (2.1) with $(B + B^{\scriptscriptstyle T})/2$. Furthermore, there is a one-to-one correspondence between the inverse covariance matrix (or precision matrix) C and the conditional probability structures of ψ_t; see, for example Besag (1974) and Besag and Kooperberg (1995). Thus, choices of C made by specifying different sets of spatial conditional probability structures yield different spatial-temporal Gaussian lattice processes ψ_t. For example, consider the Gaussian conditional autoregression $\psi_{t,x}$, for $x \in \mathcal{Z}^2$, with a conditional mean and variance structure of

$$E\{\psi_{t,x} \mid \psi_{t,x'}, x' \neq x\} = \sum \beta_{x'} \psi_{t,x-x'} \quad \text{and} \quad \text{var}\{\psi_{t,x} \mid \psi_{t,x'}, x' \neq x\} = \sigma^2$$

respectively, where the real coefficients β_x are nonzero only on a finite set $\mathcal{N}$, such that $\beta_0 = 0$, $\beta_x = \beta_{-x}$, and $\sum_x \beta_x \cos(\omega^{\scriptscriptstyle T} x) < 1$, for $\omega \in (-\pi, \pi]^2$. The set $\mathcal{N}$ defines the neighbors of the lattice point 0, and two lattice points x and x' are said to be neighbors (written as $x \sim x'$) if $x - x'$ belongs to $\mathcal{N}$. The matrix C is then specified by

$$C_{x,x} = \sigma^{-2}\left(1 - \sum_x \beta_x\right), \quad C_{x,x'} = -\sigma^{-2}\beta_{x-x'}.$$

Equation (2.2) together with this choice of C provides a spatial-temporal Gaussian lattice model parametrized by σ^2 and β_x, $x \in \mathcal{N}$, that is consistent with the lattice-based conditional formulation of spatial fields.

The interpretation of the above discrete-time dynamical model and its parameters is straightforward. The parameters β_x, $x \in \mathcal{N}$ control the spatial dependence. In fact, $\beta_{x-x'}$ is the partial correlation coefficient between $\psi_{t,x}$ and $\psi_{t,x'}$ for $x \neq x'$ and t. The conditional mean and variance structure suggest that, for any time point t, the conditional dependence of $\psi_{t,x}$ is linear on its surrounding values, and that the fluctuation around the conditional mean is relatively constant. The parameter λ_0 controls the strength of the temporal dependence between ψ_t and ψ_{t+1}. Furthermore, ψ_{t+1} depends linearly on its own previous value ψ_t through the matrix $\exp(-\lambda_0 C/2)$. This dependence generalizes the notion of the correlation parameter of an autoregressive time series of order one. In particular, all eigenvalues of $\exp(-\lambda_0 C/2)$ lie in $(0, 1]$, suggesting that we are focusing on the positive dependence between ψ_t and ψ_{t+1}.

To see how different choices of $\mathcal{N}$ lead to different spatial-temporal dynamical models, we consider two examples. In the first example, we take $\mathcal{N} = \{(0, \pm 1), (\pm 1, 0)\}$. This choice yields the first neighborhood-order spatial-temporal autoregression with

$$E\{\psi_{t,k,l} \mid \psi_{t,k',l'}, (k', l') \neq (k, l)\} = \beta_{1,0}(\psi_{t,k-1,l} + \psi_{t,k+1,l}) + \beta_{0,1}(\psi_{t,k,l-1} + \psi_{t,k,l+1})$$

and

$$\mathrm{var}\,\{\psi_{t,k,l} \mid \psi_{t,k',l'}, (k',l') \neq (k,l)\} = \sigma^2.$$

We also need $|\beta_{1,0}| + |\beta_{0,1}| < 1/2$ for the nonnegative definiteness of the matrix C. In the second example, we consider $\mathcal{N} = \{(0,\pm 1), (\pm 1, 0), (\pm 1, \pm 1)\}$. This choice results in the second neighborhood-order spatial-temporal autoregression with $E\{\psi_{t,k,l} \mid \psi_{t,k',l'}, (k',l') \neq (k,l)\}$ equal to

$$\begin{aligned}&\beta_{1,0}(\psi_{t,k-1,l} + \psi_{t,k+1,l}) + \beta_{0,1}(\psi_{t,k,l-1} + \psi_{t,k,l+1})\\ &+\beta_{1,-1}(\psi_{t,k-1,l+1} + \psi_{t,k+1,l-1}) + \beta_{1,1}(\psi_{t,k-1,l-1} + \psi_{t,k+1,l+1})\end{aligned}$$

and

$$\mathrm{var}\,\{\psi_{t,k,l} \mid \psi_{t,k',l'}, (k',l') \neq (k,l)\} = \sigma^2.$$

The parameter values must ensure that C is nonnegative definite, requiring a stringent sufficient condition such as $|\beta_{0,1}| + |\beta_{1,0}| + |\beta_{1,-1}| + |\beta_{1,1}| < 1/2$. Higher neighborhood-order versions, involving more neighbors, are constructed in a similar fashion.

In what follows, we work with the matrix C that arises in the first neighborhood-order symmetric spatial-temporal Gaussian autoregressions. Specifically, for $k, l \in \mathcal{Z}$, we assume that

$$E\{\psi_{t,k,l} \mid \psi_{t,k',l'}, (k',l') \neq (k,l)\} = \beta(\psi_{t,k-1,l} + \psi_{t,k+1,l} + \psi_{t,k,l-1} + \psi_{t,k,l+1}),$$

where $0 < \beta < 1/4$ and

$$\mathrm{var}\,\{\psi_{t,k,l} \mid \psi_{t,k',l'}, (k',l') \neq (k,l)\} = \sigma^2.$$

Whereas λ_0 controls the strength of the temporal dependence, the parameter β takes into account the spatial dependence. When β is close to 1/4, the spatial correlation decays very slowly and roughly logarithmically, as shown in Besag (1981). This is in contrast to the geometric rate of decay of the temporal autoregression.

Inferences for higher neighborhood-order spatial-temporal autoregressions and lattice-based fractional and Matérn dynamical systems are discussed in the Supplemental Material.

2.2. Restrictions to finite rectangular arrays

Practical applications of spatial-temporal Gaussian autoregressions often involve random variables on a finite regular lattice. Examples include brain-imaging, where sites represent pixels (or voxels), and satellite imaging, where sites represent approximate rectangular areas. Furthermore, when sites are irregularly distributed, it is often possible to embed the spatial locations in a fine-scale rectangular lattice, treating unobserved lattice cells as missing data. Thus, in what follows, we consider a finite restriction of ψ_t on a two-dimensional $r \times c$ regular lattice, with $n = rc$. However, to ensure nonnegative definiteness and the sparsity of the precision matrix, a finite restriction of ψ_t requires suitable boundary approximations. Here, we follow the boundary conditions suggested in Besag and Higdon (1999), Dutta and Mondal (2015), and Mondal (2018). Let W_m denote an $m \times m$ tridiagonal matrix with nonzero off-diagonal elements $W_{m,i,i\pm 1} = -1$ and $W_{m,i,i} = -\sum_{j \neq i} W_{m,i,j}$. Then, the restriction of C on a two-dimensional $r \times c$ regular lattice takes the form

$$C = \lambda_1 \frac{(I_c \otimes W_r + W_c \otimes I_r)}{2} + \lambda_2 I_n, \quad \lambda_1 > 0, \quad \lambda_0 > 0, \tag{2.3}$$

with precision parameters λ_1 and λ_2, where

$$\beta = \frac{\lambda_1}{(4\lambda_1 + 2\lambda_2)}, \quad \sigma^2 = \frac{1}{(4\lambda_1 + 2\lambda_2)}.$$

Stationary lattice processes are studied using elegant spectral representations, and the finite-dimensional matrix C in (2.3) also has an elegant and analytically known spectral decomposition. Let P_m denote the $m \times m$ matrix corresponding to the discrete cosine transformation with entries

$$p_{m,1,j} = m^{-1/2}, \quad p_{m,i,j} = \left(\frac{2}{m}\right)^{1/2} \cos\left\{\frac{\pi(i-1)(j-1/2)}{m}\right\},$$

for $i = 2, \ldots, m$ and $j = 1, \ldots, m$. Suppose that D_k is a diagonal matrix, where the ith diagonal entry is

$$d_{m,i} = 2\left[1 - \cos\left\{\frac{\pi(i-1)}{m}\right\}\right].$$

It follows that $P = P_c \otimes P_r$ diagonalizes C and V. Specifically,

$$PCP^{T} = \frac{\lambda_1(I_c \otimes D_r + D_c \otimes I_r)}{2} + \lambda_2 I = \Lambda, \quad PVP^{T} = \Lambda^{-1}(I - e^{-\lambda_0 \Lambda}) = \Lambda_1,$$

where Λ and Λ_1 are both $n \times n$ diagonal matrices. The matrices P and P^{T} correspond to the two-dimensional discrete cosine transformation and its inverse transformation, respectively. For any vector θ, matrix-vector multiplications of $P\theta$ and $P^{T}\theta$ require no storage of the matrices P and P^{T}, respectively, and only $O(n \log n)$ computations; see, for example Rao and Yip (2014), Frigo and Johnson (2005), and the discussion in Dutta and Mondal (2016).

2.3. A state-space model

Let the response variable y_t be observed at n_t sampling locations and

$$y_t = F_t \psi_t + \epsilon_t, \quad t = 1, 2, \ldots, s. \tag{2.4}$$

Assume that the latent state vector ψ_t obeys spatial-temporal autoregressions on a fine $r \times c$ regular array on which the sampling locations are embedded. The incidence (or averaging) matrix F_t is known, and indicates whether an observation corresponds to a particular array cell. The vector $F_t\psi_t$ returns the latent spatial-temporal variable values for the observed y_t. The vector ϵ_t represents the residual terms left unexplained by the variations in $F_t\psi_t$, and its entries are assumed to be independent and identically distributed (i.i.d.) Gaussian random variables with mean zero and precision λ_3. Furthermore, suppose that ψ_t evolves as in (2.2) and (2.3), that is,

$$\psi_t = G\psi_{t-1} + \nu_t, \quad G = \exp\left(\frac{-\lambda_0 C}{2}\right), \tag{2.5}$$

where $\nu_t \sim N(0, V)$ with $V = C^{-1}\{I - \exp(-\lambda_0 C)\}$ and C as in (2.3).

3. H-likelihood Estimation

3.1. Estimation of state vectors

Let $y^T = (y_1^T, \ldots, y_s^T)$, $\psi^T = (\psi_1^T, \ldots, \psi_s^T)$, and $\lambda = (\lambda_0, \ldots, \lambda_3)^T$. Denote by n_+ the total number of observations $n_1 + \cdots + n_s$. For a fixed precision parameter λ, the objective is to compute the best linear unbiased predictor of ψ by maximizing the joint distribution of y and ψ. It follows from equation (2.5) and the spectral decomposition of C that ψ is normally distributed with mean zero and a precision matrix Γ, the spectral decomposition of which takes the form $\Gamma = R^T M R$. The $ns \times ns$ matrix R is a block diagonal with all $n \times n$ diagonal blocks equal to P. The matrix M is a block tridiagonal matrix with blocks $M_{(i,i)}$ (diagonal), $M_{(i,i+1)}$ (upper diagonal), and $M_{(i-1,i)}$ (lower diagonal), such that

$$M_{(1,1)} = M_{(s,s)} = \Lambda_1^{-1}, \quad M_{(i,i)} = \Lambda_1^{-1}(I + e^{-\lambda_0 \Lambda}), \quad i = 2, \ldots, s-1,$$
$$M_{(j,j+1)} = M_{(k-1,k)} = -e^{-\lambda_0 \Lambda/2}\Lambda_1^{-1}, \quad j = 1, \ldots, s-1, \quad k = 2, \ldots, s.$$

Therefore, all blocks in M are diagonal matrices. We rewrite equations (2.3)–(2.5) as

$$y_t = F_t\psi_t + \epsilon_t, \quad 0 = P\psi_t + \eta_t, \quad t = 1, 2, \ldots, s,$$

where ϵ_ts and η_ts are independent random error vectors. Let $\epsilon^T = (\epsilon_1^T, \ldots, \epsilon_s^T)$ and $\eta^T = (\eta_1^T, \ldots, \eta_s^T)$. It follows immediately that $\epsilon \sim N(0, \lambda_3^{-1} I_{n_+})$ and $\eta \sim N(0, M^{-1})$. Next, let F be the $n_+ \times ns$ rectangular block diagonal matrix with diagonal blocks $F_1, \ldots, F_s$, and assume that

$$u = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad X = \begin{pmatrix} F \\ R \end{pmatrix}, \quad \zeta = \begin{pmatrix} \epsilon \\ \eta \end{pmatrix}, \quad Q = \begin{pmatrix} \lambda_3 I_{n+} & 0 \\ 0 & M \end{pmatrix}.$$

The state-space model in equations (2.3)–(2.5) then becomes the linear regression model

$$u = X\psi + \zeta, \tag{3.1}$$

where $\zeta \sim N(0, Q^{-1})$. Following Lee and Nelder (1996, 2001), we can thus obtain the best linear unbiased prediction of ψ by solving the generalized least squares estimating equation

$$X^T Q X \hat{\psi} = X^T Q u, \tag{3.2}$$

which is equivalent to solving $A\psi = b$ with $A = X^T Q X$ and $b = X^T Q u$.

The matrix A is block tridiagonal. When A is tridiagonal, the solution to the linear equation $A\theta = b$ can be obtained using the forward-backward Thomas algorithm, which is a simplified version of the Gaussian elimination method. Thus, when A is block tridiagonal, the traditional backward-forward Kalman filtering algorithm for solving $A\theta = b$ can be viewed as an extension of the Gaussian elimination method with blocks. However, unless $n_t = n$ and $F_t = I_n$ for all t, Kalman filtering requires computations of order $O(n^3 s)$ and the storage of $O(n^2 s)$ variables. Thus, as n becomes large, Kalman filtering becomes computationally challenging.

3.2. Estimation of the precision parameters

Let $\hat{\psi}$ be the estimate of ψ from equation (3.2). The results of Harville (1977) and Lee and Nelder (1996) imply that the log-residual likelihood function of λ obtained from (3.1) is

$$2l(\lambda) = \log|Q| - \log|X^T Q X| - (u - X\hat{\psi})^T Q (u - X\hat{\psi}).$$

Let $Q_i = \partial Q / \partial \lambda_i$ and $Q_{ij} = \partial^2 Q / (\partial \lambda_i \partial \lambda_j)$, for $i, j = 0, \ldots, 3$. Denote by $H = X(X^T Q X)^{-1} X^T Q$ the hat matrix of the regression in (3.1). Treating $\hat{\psi}$ as fixed, the score equations become

$$\left(\frac{1}{2}\right)\{tr(Q^{-1}Q_i) - tr(HQ^{-1}Q_i) - (u - X\hat{\psi})^T Q_i (u - X\hat{\psi})\} = 0, \tag{3.3}$$

for $i = 0, \ldots, 3$. Furthermore, the second derivatives provide the information matrix $\mathscr{I}$ with an (i, j)th entry, for $i, j = 0, \ldots, 3$, equal to

$$\mathscr{I}(i, j) = \left(\frac{1}{2}\right) tr(Q^{-1}Q_i Q^{-1}Q_j - HQ^{-1}Q_i HQ^{-1}Q_j + HQ^{-1}Q_{ij}).$$

The estimation follows an iterative algorithm. It starts with an initial value of λ. It then computes $\hat{\psi}$ in (3.2) and updates the estimate of λ by solving the score equations in (3.3). The algorithm continues until successive estimates of λ become sufficiently close.

4. Matrix-free Computation

We open with a brief discussion on the importance and necessity of matrix-free methods. There is no denying that matrices and matrix notation are useful for succinctly representing arrays of numbers and studying algebraic representations, including the properties of variances, covariances, and linear mixed models. However, efficient numerical computations involving matrices (e.g., the multiplication of two matrices) often require refined approaches and sophisticated algorithms that minimize the complexity of the computations (e.g., the number of elementary operations, such as additions and multiplications, of scalars necessary to multiply two matrices) and the use of the computer memory. In Section 6, in the simulation study, we take the size of the matrix A to be 163,840 by 163,840, whereas in our applications, the size of the matrix A is 153,600 by 153,600 (for the analysis of soil moisture data) and 333,312 by 333,312 (for the analysis of atmospheric concentrations of total nitrate). The standard R or Matlab packages cannot store matrices of these sizes. Thus, direct computations with such matrices are not available. In contrast, the matrix-free algorithm presented in this section only stores minimal relevant information, such as the analytically derived nonzero diagonal entries of each block of the block tridiagonal matrix M, the data vector y, the estimates for ψ and λ, and a few additional vectors of size ns. The matrix-free algorithm reduces the memory requirement significantly from $O(n^2s)$ to $O(ns)$. All computations are performed using iterative methods and matrix-vector multiplications (e.g., A times x) in a matrix-free way, that is, by storing a few vectors. The complexity of these matrix-vector computations is $O(ns\log n)$, whereas a direct matrix-vector multiplication has complexity $O(n^2s)$. We now present a set of matrix-free methods for statistical inferences of the state-space model defined in (2.3)–(2.5).

4.1. Lanczos algorithm for estimating state vectors

To solve $A\psi = b$ in (3.2), we follow Paige and Saunders (1975, 1982) and Dutta and Mondal (2015, 2016) and adapt a matrix-free scalable Lanczos algorithm. Starting from $u_1 = b/\|b\|$, the algorithm sequentially computes a set of orthonormal vectors $u_1, u_2, \ldots$ from the span of $b, Ab, A^2b, \ldots$ At the ith iteration, the Lanczos orthonormal vectors $u_1, \ldots, u_i$ reduce A to a partial tridiagonal form $AU_i \approx U_iT_i$, where $U_i = (u_1, \ldots, u_i)$ and T_i is a suitable $i \times i$ positive-definite tridiagonal matrix. As the algorithm progresses, the solution ψ is sequentially updated by computing

$$T_i\tilde{\psi}_i = \|b\|e_1, \quad \psi_i = U_i\tilde{\psi}_i, \quad e_1 = (1, 0, \ldots, 0)^T,$$

which requires only a linear solution with the lower bidiagonal Cholesky decomposition of T_i. The algorithm stops when the solution converges with sufficient numerical accuracy. The multiplication by A is the only large-scale linear operation, which we compute using the discrete cosine transformation and its inverse transformation, and requires only $O(ns \log n)$ operations. Greenbaum and Strakos (1992) proved that, for a well-conditioned matrix A (i.e., one with a bounded conditioned number), the Lanczos algorithm converges geometrically. Thus, the number of iterations required by the algorithm remains nearly constant or does not grow more than $O(\log(ns))$. When $\lambda > 0$, the algorithm requires only $O(ns \log(ns) \log n)$ operations to solve (3.2).

4.2. Stochastic trace approximation

To solve (3.3), we need to compute the trace terms $tr(Q^{-1}Q_i)$ and $tr((X^TQ X)^{-1}X^TQ_iX)$, which can be difficult. Here we follow the stochastic approximations derived in Hutchinson (1990). Let z_k, for $k = 1, 2, \ldots, K$, be i.i.d.. Rademacher random variables, where K is an integer of order at most $\log(ns)$. Assume that

$$g_i(\lambda) = (2K)^{-1} \sum_{k=1}^{K} z_k^T (Q^{-1}Q_i - HQ^{-1}Q_i) z_k - 2^{-1}(u - X\hat{\psi})^T Q_i (u - X\hat{\psi}),$$

for $i = 0, \ldots, 3$. We approximate (3.3) using the unbiased estimating equations

$$g_i(\lambda) = 0, \quad i = 0, \ldots, 3. \tag{4.1}$$

Each term in (4.1) can be computed in a matrix-free scalable way in conjunction with the discrete cosine transformation and the Lanczos algorithm.

4.3. Trust region method for precision parameter estimation

To solve (3.3), we follow Powell (1984) and Nocedal and Wright (2006) and develop an iterative matrix-free trust region algorithm. The trust region algorithm considers the objective function $(1/2)\|\nabla g(\lambda)\|^2$, where $\nabla g(\lambda)$ is the gradient of $g(\lambda) = (g_0(\lambda), g_1(\lambda), g_2(\lambda), g_3(\lambda))$ in equation (4.1). The objective function is minimized at $\lambda = \phi$ if and only if $\nabla g(\phi) = 0$, provided that $\nabla^2 g(\lambda)$ is positive-definite. Furthermore, we use the following quadratic function to approximate $(1/2)\|\nabla g(\lambda)\|^2$ around a point ϕ_k:

$$\Omega_k(\alpha) = \left(\frac{1}{2}\right)\|\nabla g(\phi_k)\|^2 + \alpha^T\{\nabla^2 g(\phi_k)\}^T \nabla g(\phi_k) + \left(\frac{1}{2}\right)\alpha^T\{\nabla^2 \nabla g(\phi_k)\}^2\alpha.$$

Evidently, each term in the above quadratic function can also be computed in a scalable matrix-free way in conjunction with the discrete cosine transformation, the Lanczos algorithm, and the stochastic trace approximation. The trust region algorithm is given as follows. First, the algorithm computes the step size α_{k+1} by minimizing $\Omega_k(\alpha)$. Second, it updates the trust region radius and decides whether α_{k+1} should be accepted. If α_{k+1} is accepted, it computes $\phi_{k+1} = \phi_k + \alpha_{k+1}$, otherwise, it proceeds with $\phi_{k+1} = \phi_k$. The iteration stops when there is no significant change in ϕ_k and the value of the objective function is sufficiently close to zero. As a byproduct, the algorithm also computes the Hessian matrix of (3.3), from which we can derive the standard errors of $\hat{\lambda}$. Refer to Nocedal and Wright (2006) and Dutta and Mondal (2016) for further details.

The trust region algorithms are dual to global line search methods, and are known to have excellent convergence and scalability properties. Global convergence for the trust region method holds when the effective step size is zero, as proved by Powell (2004). Nocedal and Wright (2006) provide convergence results when the effective threshold lies in $(0, 1/4)$, assuming that the objective function is Lipschitz and continuously differentiable and that the corresponding Hessian is bounded.

4.4. Preconditioning

To solve $A\psi = b$, a preconditioning matrix L facilitate convergence of the Lanczos algorithm if the condition number (i.e., the ratio of the largest to the smallest eigenvalues) of LAL^T is small compared to that of A, or if its eigenvalues are clustered around few values. The solution also requires that the scalable matrix-free matrix vector multiplication of Lx and $L^T x$ is possible. Therefore, a judicious choice of L arises if we have L or L^{-1} sparse and $LL^T \approx A^{-1}$. In our case, $A = \lambda_3 F^T F + R^T M R$, and our objective is to find L that solves equation (3.2) in three steps:

$$L(\lambda_3 R F^T F R^T + M)L^T \tilde{x} = LRb, \quad x = L\tilde{x}, \quad R\psi = x.$$

To this end, we consider $L = (\lambda_3 I + M)^{-1/2}$. This choice of L ensures that the eigenvalues of $(\lambda_3 I + M)^{-1/2}(\lambda_3 R F^T F R^T + M)(\lambda_3 I + M)^{-1/2}$ are bounded below by one and that the minimum eigenvalue is bounded above by $\lambda_0/(1+\lambda_0)$. In many practical applications, F^TF is a binary diagonal matrix and, in such instances, a fraction of eigenvalues of $(\lambda_3 I + M)^{-1/2}(\lambda_3 RF^TFR^T + M)(\lambda_2 I + M)^{-1/2}$ also clusters around one (see the Supplementary Material). Furthermore, when F^TF is the identity matrix, that is, the regular grid has no missing values, all eigenvalues of $(\lambda_3 I + M)^{-1/2}(\lambda_3 RF^TFR^T + M)(\lambda_3 I + M)^{-1/2}$ are equal to one. Then, $L = (\lambda_3 I + M)^{-1/2}$ is the best choice. In practice, we do not actually work with $L = (\lambda_3 I + M)^{-1/2}$, but instead use the inverse of the sparse incomplete Cholesky factorization of the sparse matrix $\lambda_3 I + M$. Unlike the sparse Cholesky factorization, which incurs a cost of at least $O((ns)^{3/2})$ in terms of computation and storage, an incomplete Cholesky factorization requires $O(ns)$ computations and storage. In addition, the incomplete Cholesky factorization ensures that $L^T(\lambda_3 I + M)^{-1}L$ is still a good approximation of the identity matrix. The using this factorization facilitates faster convergence of the matrix-free Lanczos algorithm. In our applications, we only considered a no-fill incomplete Cholesky decomposition of the sparse matrix $\lambda_2 I + M$. Such a decomposition of a matrix S contains nonzero elements in the same positions that S contains nonzeros. Thus the storage requirement is $O(ns)$ and the decomposition can be computed in a matrix-free way in $O(ns)$ steps. An algorithm for this operation is provided by the "*ichol*" function in Matlab. Various modifications of an incomplete Cholesky decomposition are also possible. However, these require some fill-in (see, e.g., Lin and Moré (1999)) with a memory requirement $O(ns)$ to improve the approximation and reduce the number of steps required for the convergence of the Lanczos algorithm in order to solve $Ax = b$. Because the storage requirement does not increase by more than a constant factor, these modifications of a sparse incomplete Cholesky decomposition can also be applied in a matrix-free way. Refer to Kershaw (1978) and Benzi (2002) for further details on incomplete Cholesky factorizations.

4.5. Conditional simulation and log-likelihood calculation

An advantage of the scalable matrix-free h-likelihood method is that it can be used to generate samples from the conditional distribution of ψ, given observations y. This is done as follows. First, vectors v_i, for $i = 1, \ldots, s$, are generated from a Gaussian distribution with mean zero and covariance matrix Λ_1^{-1}. Then, we compute

$$\tilde{\psi}_1 = P^{T} v_1, \quad \tilde{\psi}_s = P^{T} v_s, \quad \tilde{\psi}_i = P^{T} v_{i-1} + G^{T} P^{T} v_i, \quad i = 2, \ldots, s-1.$$

Next, let $\tilde{\psi}^{\scriptscriptstyle T} = (\tilde{\psi}_1^{\scriptscriptstyle T}, \ldots, \tilde{\psi}_s^{\scriptscriptstyle T})$. Then,

$$\psi = \hat{\psi} + (R^T M R)^{-1} \tilde{\psi}$$

provides a realization from the conditional distribution of $\psi_1, \ldots, \psi_s$, given observations $y_1, \ldots, y_s$. Here, the last term $(R^{\scriptscriptstyle T} M R)^{-1} \tilde{\psi}$, particularly the multiplication of M^{-1} with $R\tilde{\psi}$, can be computed using the incomplete Cholesky preconditioned Lanczos algorithm. Thus, a matrix-free conditional simulation of one ψ requires only $O(ns \log(ns) \log n)$ operations.

The computation of the log-residual likelihood function presents further challenges, but can be performed using the results of Dutta and Mondal (2016).

5. Extension to Other Spatial-temporal State-space Models

5.1. Inferences for small time steps

If observations $y_1, \ldots y_s$ are made at small time steps $\Delta, 2\Delta, \ldots, s\Delta$, the state-space model in (2.3)–(2.5) can be approximated using finite differencing. In particular, the dynamical model in (2.1) becomes

$$(\psi_{t+\Delta} - \psi_t) + \left(\frac{\lambda_0 C}{2} \right) \Delta \psi_t = \nu_t, \quad \nu_t \sim N(0, \lambda_0 \Delta I_n).$$

Taking $\varphi_i = \psi_{\Delta i}$ for $i = 1, \ldots, s$, we can rewrite the state model in (2.5) as

$$\varphi_i = G_\Delta \varphi_{i-1} + \nu_{i\Delta}, \quad \nu_{i\Delta} \sim N(0, \lambda_0 \Delta I_n), \quad G_\Delta = I - \frac{\lambda_0 \Delta C}{2}. \tag{5.1}$$

Next, let $\varphi^{\scriptscriptstyle T} = (\varphi_1^{\scriptscriptstyle T}, \ldots, \varphi_s^{\scriptscriptstyle T})$. Then, φ follows a multivariate normal distribution with mean zero and a sparse precision matrix Γ_Δ, with a spectral decomposition $\Gamma_\Delta = R^{\scriptscriptstyle T} M_\Delta R$. Here, M_Δ is a block tridiagonal matrix with diagonal blocks $M_{\Delta(i,i)}, M_{\Delta(i,i+1)}$, and $M_{(\Delta i-1,i)}$, such that

$$M_{\Delta(1,1)} = M_{\Delta(s,s)} = (\lambda_0 \Delta)^{-1} I_n, \quad M_{\Delta(i,i)} = 2(\lambda_0 \Delta)^{-1} I_n - \Lambda + 4^{-1} \lambda_0 \Delta \Lambda^2,$$

for $i = 2, \ldots, s-1$, and

$$M_{\Delta(j,j+1)} = M_{\Delta(j-1,j)} = -(\lambda_0 \Delta)^{-1} I_n + 2^{-1} \Lambda, \quad j = 1, \ldots, s-1.$$

In the h-likelihood formulation, we obtain

$$y_i = F_i \varphi_i + \epsilon_i, \quad 0 = P \varphi_i + \eta_i, \quad i = 1, \ldots, s, \quad \eta \sim N(0, M_\Delta^{-1}).$$

Hence, equation (3.1) simplifies to

$$u = X\varphi + \zeta, \quad \zeta \sim N(0, Q_\Delta^{-1}),$$

where Q_Δ is derived from Q by replacing M with M_Δ. Consequently, the estimates of the state vectors φ are obtained by solving

$$X^T Q_\Delta X \hat{\varphi} = X^T Q_\Delta u.$$

The matrix $A_\Delta = X^T Q_\Delta X$ is sparse. This sparsity allows us to simplify the steps in the matrix-free statistical calculations. In particular, the sparsity allows faster matrix-vector computations and the derivation of an efficient preconditioning matrix for the Lanczos algorithm by using the incomplete Cholesky decomposition.

5.2. Approximation to stochastic advection-diffusions

The small time step approximations in (5.1) may look naive, but they have wider relevance in deriving scalable matrix-free statistical computations for spatial-temporal models based on other complex SPDEs. For example, consider a stochastic advection-diffusion equation of the form

$$\frac{\partial\psi(t,x)}{\partial t} = -\left(\frac{1}{2}\right)\{\mathscr{A}\psi_t\} + z(t,x), \tag{5.2}$$

where $z(t,x)$ is a temporally uncorrelated Gaussian process and A is a linear operator given by

$$\mathscr{A}\psi(t,x) = \frac{2\mu^T\partial\psi(t,x)}{\partial x} - tr\left\{\frac{\partial^2\psi(t,x)}{(\partial x\partial x^T)}\right\}\Sigma + 2\tau\psi(t,x). \tag{5.3}$$

Here, the first term $\mu^T\partial\psi(t,x)/\partial x$ models the transport effect, with velocity or drifting rate μ, the second term $\mathrm{tr}\{\partial^2\psi(t,x)/(\partial x\partial x^T)\}\Sigma$ models the diffusion, with Σ controlling the rate and the anisotropy of the diffusion or blurring, and the third term $\tau\psi(t,x)$ controls the damping or decay, with rate τ. For discussions of stochastic advection-diffusion equations, see Whittle (1963), Brown et al. (2000), Cressie and Wikle (2011), and Sigrist, Künsch and Stahel (2015). Unlike (2.1), the model in (5.2)–(5.3) has no simple and explicit analytic solution. Thus, approximations are necessary before we can pursue a statistical analysis. Here, we focus on approximations in small time steps using finite differencing. Specifically, we consider an approximation of the continuum process $\psi(t,x)$ at time points $t = \Delta,\ldots,s\Delta$ and regular spatial lattice points $x = (x_1,x_2)$ at spacing Δ_0. For brevity, we consider $\mu = 0$, $\Sigma = \gamma_1 I$, and $\tau = -\gamma_2$. Then, the advection-diffusion equation takes the form

$$\frac{\partial\psi(t,x)}{\partial t} = \gamma_1\frac{\{\partial^2\psi(t,x)/\partial x_1^2 + \partial^2\psi(t,x)/\partial x_1^2\}}{2} - \gamma_2\psi(t,x) + z(t,x). \tag{5.4}$$

By replacing the first-order derivative with a forward difference and the second-order derivative with a centered difference and discretizing $\varphi_{i,x} = \psi(i\Delta, x\Delta_0)$, equation (5.4) is now approximated as

$$\begin{aligned}&\Delta^{-1}(\varphi_{i+1,x}-\psi_{i,x})\\&=\gamma_1\Delta_0^{-2}\frac{(\varphi_{i,x_1+1,x_2}+\varphi_{i,x_1-1,x_2}+\varphi_{i,x_1,x_2+1}+\varphi_{i,x_1,x_2-1}-4\varphi_{i,x})}{2}-\gamma_2\varphi_{i,x}+z_{i,x}.\end{aligned}$$

On a finite $r\times c$ spatial array, with boundary approximations that use a forward difference to replace the second-order derivative, (5.4) reduces to the state model

$$\varphi_{i+1}=G^{\dagger}\varphi_i+z_i, \tag{5.5}$$

where

$$G^{\dagger}=(1-\gamma_2\Delta)I_n-\gamma_1\Delta{\Delta_0}^{-2}\frac{(I_c\otimes W_r+W_c\otimes I_r)}{2}.$$

Thus, the state equation (5.5) has a form very similar to that of (5.1). The matrix $G^{\dagger}$ has the spectral representation $PG^{\dagger}P^{\mathrm{T}}=\Lambda^{\dagger}=(1-\gamma_2\Delta)I-\gamma_1\Delta{\Delta_0}^{-2}(I_c\otimes D_r+D_c\otimes I_r)/2$, with the diagonal matrix $\Lambda^{\dagger}$ providing the eigenvalues.

Assuming z_t is normally distributed, with mean zero and precision matrix $\gamma_3 I_n$, the state-space model takes the form

$$y_t=F_t\varphi_t+\epsilon_t,\quad \varphi_t=G^{\dagger}\varphi_{t-1}+z_t,\quad i=1,\ldots,s,$$

where ϵ_t and z_t are independent and $\epsilon_t\sim N(0,\gamma_4^{-1}I)$. Here, φ is normally distributed with mean zero and precision matrix $\Gamma^{\dagger}$. Furthermore, $\Gamma^{\dagger}=R^{\mathrm{T}}M^{\dagger}R$, where $M^{\dagger}$ is a block tridiagonal matrix, with diagonal blocks $M^{\dagger}_{(i,i)}$, $M^{\dagger}_{(i,i+1)}$, and $M^{\dagger}_{(i-1,i)}$, such that

$$\begin{aligned}M^{\dagger}_{(1,1)}&=M^{\dagger}_{(s,s)}=\gamma_3 I,\quad M^{\dagger}_{(i,i)}=\gamma_3(I+(\Lambda^{\dagger})^2),\quad i=2,\ldots,s-1,\\M^{\dagger}_{(j,j+1)}&=M^{\dagger}_{(k-1,k)}=-\gamma_3\Lambda^{\dagger},\quad j=1,\ldots,s-1,\quad k=2,\ldots,s.\end{aligned}$$

In the h-likelihood formulation, the discretized stochastic advection-diffusion equation has the regression form

$$u=X\varphi+\zeta,\quad \zeta\sim N(0,(Q^{\dagger})^{-1}),$$

which is very similar to that in (3.1). The state vectors are estimated as

$$X^{\mathrm{T}}Q^{\dagger}X\hat{\varphi}=X^{\mathrm{T}}Q^{\dagger}u.$$

As in Section 5.1, the matrix $A^{\dagger}=X^{\mathrm{T}}Q^{\dagger}X$ is sparse, which allows us to use no-fill incomplete Cholesky preconditioning. This enables fast and efficient matrix-free computations and resolves the inferential challenges discussed in Sigrist, Künsch and Stahel (2015).

Table 1. Summary of the simulation study with spatial data generated in an 128 × 128 array with $s = 10$. The standard errors are given in parentheses.

Parameters (true value)	λ_0 (1.0)	λ_1 (2.0)	λ_2 (0.01)	λ_3 (1.0)
Initial estimates	1.0 –	1.949 (0.127)	0.007 (0.003)	1.009 (0.025)
Final REML estimates	0.999 (0.002)	2.015 (0.006)	0.007 (0.001)	0.979 (0.014)

6. A Simulation Experiment and Data Examples

6.1. A simulation study

To illustrate how the method works, we sample $y_1, y_2, \ldots, y_{10}$ from the state-space model (2.3)–(2.5) on a 128 × 128 array at time $t = 1, \ldots, 10$, with $\lambda_0 = 1$, $\lambda_1 = 2$, $\lambda_2 = 0.01$, and $\lambda_3 = 1$. The sampling is done in three steps. First, using the spectral representation, we generate $\theta_1, \ldots, \theta_{10}$ on a 128×128 array, with $\lambda_0 = 1$, $\lambda_1 = 2$, and $\lambda_2 = 0.01$. Second, we generate Gaussian white noise $\epsilon_1, \ldots, \epsilon_{10}$, with mean zero and precision $\lambda_3 = 1$. Third, we compute $\epsilon_t + \theta_t$ and, for each t, discard 20% of the entries at random to obtain $y_1, \ldots, y_{10}$. Thus, $r = c = 128$, $s = 10$, and $F_t \neq I_n$. The total sample size is $n_+ = 131{,}072$ and $rcs = 163{,}840$. The method requires working with matrices of size 163,840 × 163,840, which is quite daunting.

Next, we apply the matrix-free computation in Section 4 to estimate the precision parameters. The initial estimate $\hat{\lambda}^{(0)}$ of λ is derived by fitting the corresponding spatial model on y_1 and setting $\hat{\lambda}_0^{(0)} = 1$. To compute the trace terms in (4.1), we use $p = 50$ Rademacher vectors. The trust region iteration stops when sufficient numerical accuracy is achieved, yielding the overall REML estimates $\hat{\lambda}$ for the precision parameters.

Table 1 summarizes these results, along with the standard error values. Note that the initial estimates based on y_1 are fairly accurate. The final estimates based on $y_1, \ldots, y_{10}$ are consistent with the true values and have smaller standard errors.

6.2. Analysis of atmospheric concentrations of total nitrate

The Environmental Protection Agency provides output from its numerical model called Models-3, which includes the atmospheric total nitrogen concentra-

Table 2. REML estimates of the precision parameters for total nitrogen concentration data under no splitting (Scenario 1) and 2×2 splitting (Scenario 2) of the original array. The standard errors are given in parentheses.

Parameters	λ_0	λ_1	λ_3
Scenario 1	7.536 (0.062)	7.894 (0.088)	14.571 (0.062)
Scenario 2	28.726 (0.118)	2.285 (0.028)	13.931 (0.119)

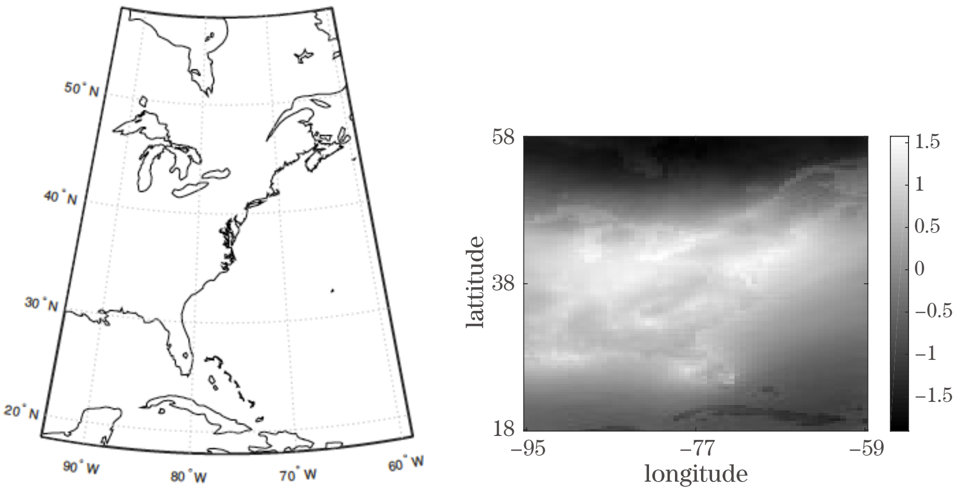

Figure 1. The left panel shows a map of the study region. The right panel displays the average total nitrogen concentration over 12 months.

tion level as one component. This concentration level is defined as the gas-phase nitric acid plus particle-phase nitrate, and is vital for air quality assessment. Models-3 estimates the average concentration levels over regions of size 36×36 km^2 and a period of 28 days by combining pollution emissions data with numerical models of regional weather, the emission process, and land use. For further inferences using Model-3 output and their statistical analyses, see Fuentes and Raftery (2005) and Ghosh et al. (2010). In particular, Fuentes and Raftery (2005) do not consider any spatial-temporal modeling. Instead, they combine Model-3 data with observations from certain monitoring stations. On the other hand, Ghosh et al. (2010) use an elaborate Bayesian dynamical model for atmospheric total nitrate.

Here, we consider the Models-3 output (provided by Montse Fuentes) for the total nitrogen concentrations for the first 12 time periods of 2001. The left panel of Figure 1 provides a map of the region. The latitudes are between 18°N and 58°N and the longitudes are between 59°W and 94°W. The total nitrate data are embedded in a spatial array of size 62 × 112, and we take a log-transformation of the data to reduce skewness and to improve the normality assumption. The right panel of Figure 1 displays an image plot of the mean log-total nitrogen concentrations across the region. We see that there is strong spatial dependence across the entire region. The nitrogen concentration is high across continental North America, where the population is dense and urbanism and industrialization are compact. Furthermore, the nitrogen concentrations thin out over the North Atlantic Ocean, the Gulf of Mexico, and northern Canada.

Our objective is to investigate how the spatial-temporal state-space models in (2.3)–(2.5) perform in terms of explaining the variation in the Models-3 output. To this end, let $y_1, \ldots, y_{12}$ be the standardized monthly log-averaged total nitrogen concentrations from the numerical models. It is not unreasonable to consider that large-scale atmospheric pollution is driven by forms of advection-diffusion equations. However, we lack certain information, such as pollution sources, the effects of wind speed and wind direction, and the rates of atmospheric deposits of pollution on land and into the ocean. Instead, we fit the parsimonious stochastic advection-diffusion equation given in Section 5.2, with $\mu = 0$, $\vartriangle = 0.01$, $\vartriangle_0 = 1$,

$$G^{\dagger} = (1 - 0.01\gamma_2)I_n - \frac{0.01\gamma_1(I_c \otimes W_r + W_c \otimes I_r)}{2},$$

and

$$\Lambda^{\dagger} = (1 - 0.01\gamma_2)I_n - \frac{0.01\gamma_1(I_c \otimes D_r + D_c \otimes I_r)}{2}.$$

We apply two scenarios. In Scenario 1, we assume that the underlying state vectors ψ_t, for $t = 1, \ldots, 12$, follow spatial-temporal autoregressions, as in equation (5.5), at the original spatial resolution of 36×36 km^2. In Scenario 2, we assume that the state vectors follow spatial-temporal autoregressions at a finer spatial resolution of 18×18 km^2. Here, we split each region into 2×2 subregions so that ψ_t, for $t = 1, \ldots, 12$, lie on a 124×224 spatial array and $rcs = 333{,}312$. Accordingly, we construct the averaging matrix F_t, of order $27{,}776 \times 6{,}944$, such that $F_t\psi_t$ provides a vector of average state values at the original 36×36 km^2 spatial resolution. Scenario 2 is particularly useful for obtaining a spatial interpolation at a finer resolution. Furthermore, the finer resolution allows us to achieve an approximate inference from the limiting continuum geostatistical model. See Besag and Mondal (2005) and Dutta and Mondal (2015, 2016) for examples of such inferences in spatial statistics. At the 18×18 km^2 spatial res-

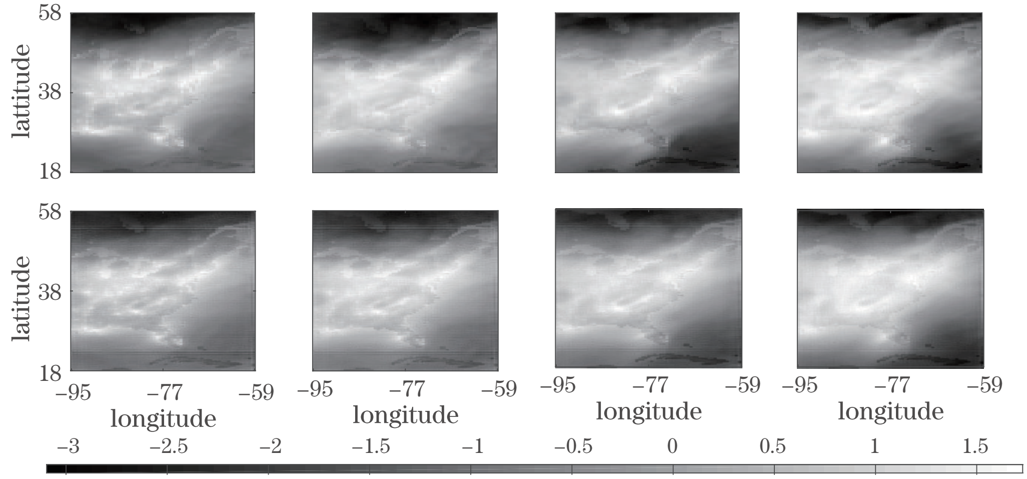

Figure 2. The top panel shows image plots of $y_9, \ldots, y_{12}$. The bottom panel displays $\hat{\psi}_9, \ldots, \hat{\psi}_{12}$ at the 18×18 km^2 spatial resolution.

olution, computations are particularly challenging because we need to deal with a 333,321 × 333,312 sparse block triangular matrix A, where each block of A is of order 27,776 × 27,776, and F_t is not an identify matrix. Thus, the traditional Kalman filtering algorithm does not work in this case.

For this model, the REML estimation encounters a boundary problem. Specifically, we found that the estimate of γ_2 tends to the boundary point 0. Therefore, we set $\gamma_2 = 0$, which then results in a simpler model with

$$G^{\dagger} = I_n - \frac{0.01\gamma_1(I_c \otimes W_r + W_c \otimes I_r)}{2}, \quad \Lambda^{\dagger} = I_n - \frac{0.01\gamma_1(I_c \otimes D_r + D_c \otimes I_r)}{2}.$$

This is the same as the small-time-step state-space model (5.1) with $\Delta = 0.01$ and

$$C = \frac{\lambda_1(I_c \otimes W_r + W_c \otimes I_r)}{2}, \quad \Lambda = \frac{\lambda_1(I_c \otimes D_r + D_c \otimes I_r)}{2}.$$

The relationship between λ and γ is given by

$$\gamma_1 = \frac{\lambda_0\lambda_1}{2}, \quad \gamma_3 = \frac{1}{(\lambda_0\Delta)}, \quad \text{and} \quad \gamma_4 = \lambda_3.$$

Furthermore, for this model, $\beta = \lambda_1/(4\lambda_1 + 2\lambda_2) = 1/4$ and $\sigma^2 = 1/((4\lambda_1)$. Thus, we are considering an intrinsic spatial-temporal autoregression model, rather than a second-order stationary version.

Table 2 summarizes the REML estimates for the standardized log-transformed numerical values for total nitrogen concentrations using the methods in Section 5.1. As before, we used $p = 50$ Rademacher vectors to approximate the trace terms. For verification, we also compute the precision parameters using the methods in Section 5.2. The estimate of $(\gamma_1, \gamma_3, \gamma_4)$ is found to be (29.745, 13.271, 14.569) in Scenario 1 and (32.851, 3.483, 13.926) in Scenario 2. Therefore, the relationship defined in (5.1) between $(\lambda_0, \lambda_1, \lambda_3)$ and $(\gamma_1, \gamma_3, \gamma_4)$ is satisfied. Finally, Figure 2 displays the actual observations and the predictions for the latent variables at a finer spatial resolution of $18 \times 18\ km^2$ for the final four months. Here, we find that a parsimonious advection-diffusion model is quite effective in downscaling and in explaining a large fraction (about 96%) of the total variation in the data.

7. Discussion

Using circulant embedding (see the Supplementary Material), the score equation (3.3) can be shown to be equivalent to a gamma nonlinear regression model. This contrasts with the results of Lee and Nelder (1996) and Dutta and Mondal (2015), where the estimation of the precision parameters is equivalent to fitting a gamma generalized linear model. Thus, the optimization in (4.1) is nonconvex and can suffer from multiple modes, boundary problems, and long flat ridges. While finding the global maxima can become challenging if multiple modes are present, long flat ridges around the maxima make the information matrix nearly singular and the standard error computation difficult. We encountered some of these issues in this study. Thus, the starting value of λ can play an important role, and in certain applications, we may need to run the algorithm several times with different initial values for the parameters. For spatial models, multi-modality has been studied by Mardia and Watkins (1989) and Dietrich (1991), among others, and is often related to range parameter estimation. Similarly, long flat ridges in the likelihood function of spatial models have been studied by Zhang (2004). However, further work is needed to reveal the nature of multi-modality and long flat ridges in a spatial-temporal setting.

Notwithstanding these expected issues, this study advances the computations and methods in spatial-temporal settings and resolves the inferential challenges discussed in Sigrist, Künsch and Stahel (2015). To achieve this, we needed to impose assumptions in addition to the modeling assumptions in equations (2.1) and (5.3). First, some edge correction is necessary when we restrict an infinite lattice spatial process to a finite rectangular lattice. The edge corrected matrix C used in equation (2.3) is introduced by Besag and Kooperberg (1995) and Besag and Higdon (1999), and is linked to reflective boundary conditions; see Mondal

(2018) for further details. Second, in Section 5.2, we considered a discrete time approximation of the stochastic advection-diffusions, as discussed in Cressie and Wikle (2011). Third, we assumed that spatial sampling locations are embeddable on a (fine) rectangular lattice of size $r \times c$ and that observations are sampled at regular time intervals. These are the only assumptions imposed in this study.

Supplementary Material

The Supplementary Material includes the following: (1) an example of the effect of preconditioning; (2) a derivation of the connection between the REML estimation and a nonlinear gamma regression; (3) an analysis of monthly soil moistures across North America; and (4) possible extensions to the proposed method.

Acknowledgments

We thank an anonymous reviewer, an associate editor, and the editor for their constructive suggestions. The work is supported by the NSF Career Award DMS 1519890.

References

Benzi, M. (2002). Preconditioning techniques for large linear systems: a survey. *Journal of Computational Physics* **182**, 418–477.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **36**, 192–236.

Besag, J. (1977). On spatial-temporal models and markov fields. In: *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians*, 47-55. Springer.

Besag, J. (1981). On a system of two-dimensional recurrence equations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **43**, 302–309.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **48**, 259–302.

Besag, J. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 691–746.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregression. *Biometrika* **82**, 733–746.

Besag, J. and Mondal, D. (2005). First-order intrinsic autoregressions and the de Wijs process. *Biometrika* **92**, 909–920.

Brix, A. and Diggle, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**, 823–841.

Brown, P. E., Roberts, G. O., Kåresen, K. F. and Tonellato, S. (2000). Blur-generated nonseparable space–time models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 847–860.

Cressie, N. (1993). *Statistics for Spatial Data.* New York: John Wiley and Sons.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data.* John Wiley & Sons, New York.

Dietrich, C. (1991). Modality of the restricted likelihood for spatial Gaussian random fields. *Biometrika* **78**, 833–839.

Dutta, S. and Mondal, D. (2015). An h-likelihood method for spatial mixed linear models based on intrinsic auto-regressions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**, 699–726.

Dutta, S. and Mondal, D. (2016). REML estimation with intrinsic Matérn dependence in the spatial linear mixed model. *Electronic Journal of Statistics* **10**, 2856–2893.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans* **99**, 10143–10162.

Evensen, G. (2009). *Data Assimilation: the Ensemble Kalman Filter.* Springer Science & Business Media, Berlin.

Frigo, M. and Johnson, S. G. (2005). The design and implementation of FFTW3. *Proceedings of the IEEE* **93**, 216–231.

Fuentes, M. and Raftery, A. E. (2005). Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics* **61**, 36–45.

Ghosh, S. K., Bhave, P. V., Davis, J. M. and Lee, H. (2010). Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. *Journal of the American Statistical Association* **105**, 538–551.

Greenbaum, A. and Strakos, Z. (1992). Predicting the behavior of finite precision Lanczos and conjugate gradient computations. *SIAM Journal on Matrix Analysis and Applications* **13**, 121–137.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338.

Hutchinson, M. F. (1990). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation* **19**, 433–450.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**, 35–45.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering* **83**, 95–108.

Kershaw, D. S. (1978). The incomplete cholesky conjugate gradient method for the iterative solution of systems of linear equations. *Journal of Computational Physics* **26**, 43–65.

Künsch, H. R. (1987). Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika* **74**, 517–524.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **58**, 619–678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88**, 987–1006.

Lin, C.-J. and Moré, J. J. (1999). Newton's method for large bound-constrained optimization problems. *SIAM Journal on Optimization* **9**, 1100–1127.

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 423–498.

Mardia, K. and Watkins, A. (1989). On multimodality of the likelihood in the spatial linear model. *Biometrika* **76**, 289–295.

Mardia, K. V., Goodall, C., Redfern, E. J. and Alonso, F. J. (1998). The kriged Kalman filter. *Test* **7**, 217–282.

Mondal, D. (2018). On edge correction of conditional and intrinsic autoregressions. *Biometrika* **105**, 447–454.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization.* Springer Science & Business Media, Berlin.

Paige, C. C. and Saunders, M. A. (1975). Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis* **12**, 617–629.

Paige, C. C. and Saunders, M. A. (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)* **8**, 43–71.

Powell, M. (1984). On the global convergence of trust region algorithms for unconstrained minimization. *Mathematical Programming* **29**, 297–303.

Powell, M. (2004). On the use of quadratic models in unconstrained minimization without derivatives. *Optimization Methods and Software* **19**, 399–411.

Rao, K. R. and Yip, P. (2014). *Discrete cosine transform: algorithms, advantages, applications.* Academic press, Boston.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 319–392.

Sigrist, F., Künsch, H. R. and Stahel, W. A. (2015). Stochastic partial differential equation based modelling of large space–time data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**, 3–33.

Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute* **40**, 974–994.

Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika* **86**, 815–829.

Zhang, H. (2004). nconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99**, 250–261.

Department of Statistics, Oregon State University, Corvallis, Oregon, 97331, U.S.A.

E-mail: debashis@stat.oregonstate.edu

Department of Statistics, Oregon State University, Corvallis, Oregon, 97331, U.S.A.

E-mail: wangc@stat.oregonstate.edu

(Received May 2017; accepted February 2018)