Statistica Sinica

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

Debashis Mondal and Chunxiao Wang

Oregon State University

Abstract: This paper develops a scalable and 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 h-likelihood method proposed here provides the same inference as that obtained from the Kalman filter and residual maximum likelihood analysis. However, for data from a large number of spatial sites, the proposed method is shown to have significant computational advantages. Furthermore, the paper details inference in small time steps and indicates how the proposed method can be adapted to other complex spatial-temporal dynamical models based on stochastic partial differential equations. The method applies to data with both regularly and irregularly sampled spatial locations, and is illustrated through a simulation study and two data examples, one with monthly soil moistures across North America and the other with atmospheric concentrations of total nitrate across Eastern North America.

Key words and phrases: Advection-diffusion, Conditional autoregression, Dis-

crete 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 paper is concerned with developing statistical inference for spatial-temporal Gaussian state-space model. The focus here is 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 inference. Following Besag (1974, 1986), Künsch (1987), Cressie (1993) and others, an extensive literature has been developed on conditional modeling of spatial variables. However, unlike in spatial setting where conditional modeling gives rise to sparse dependence structures and swift statistical computations, their extensions in 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, the traditional time series method of estimation and conditional simulation of state variables such as Kalman filtering, arising out of the seminal contributions by Kalman (1960) and Kalman and Bucy (1961), works for data with small to moderately large spatial locations. However,

for data from a large number of spatial locations, it becomes impossible to implement Kalman filtering without dimension reduction or replacement of certain spatial covariance matrices by their sample versions contrusted from ensembles of stochastic simulations. For references on these issues of Kalman filtering, see Mardia et al. (1998), Houtekamer and Mitchell (1998), Wikle and Cressie (1999) and Evensen (1994, 2009). While dimension reduction and ensembles of stochastic simulations can incur a loss of information, the parameter estimation, conditional simulations, and log-likelihood computations introduce further challenges because they require evaluations of 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), and develop a scalable matrix-free h-likelihood method that yields the same inference that one would obtain from Kalman filtering and residual maximum likelihood (REML) analysis. The h-likelihood method is faster than iteratively nested Laplacian and related approximations of Rue et al. (2009) and Lindgren et al. (2011), as shown explicitly and numerically in the spatial case in Dutta and Mondal (2015, 2016). Furthermore, the h-likelihood method explains why Kalman filtering is stymied by computational challenges and resolves the inferential challenges discussed in Sigrist et al. (2015). The novel elements of

the method include how we represent a spatial-temporal state-space model as a linear regression model and develop estimation by matrix-free computations. These computations include: (i) an adaptation of the two-dimensional discrete cosine transformation that arises in the spectral decomposition of spatial-temopral autoregressions and that allows fast matrix-free matrix-vector multiplications; (ii) a preconditioned matrix-free scalable Lanczos algorithm that solves non-sparse matrix equations; (iii) a matrix-free Hutchinson's trace estimator that stochastically approximates the trace of a matrix; (iv) a robust matrix-free trust region method that finds the solution of the REML score equations; and (v) stochastic approximation of differences in log-REML functions.

Finally, the last two decades have also witnessed significant advances in general theory and application of continuum spatial-temporal dynamical processes with several noteworthy works on dynamical models providing different perspectives on the statistical analysis. These works include Brown et al. (2000), Brix and Diggle (2001), Stroud et al. (2010), Cressie and Wikle (2011), Sigrist et al. (2015) and many subsequent references. Thus, as a second development, we show that the construction of a spatial-temporal Gaussian autoregressions and the matrix-free method of inference presented here can largely be adapted to and embodied in the above-mentioned re-

search. To this end, we also detail inference 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 the fractional and Matérn spatial fields. We illustrate our method through a simulation study and analyze two data sets; one on monthly soil moistures across North America and the other on atmospheric concentrations of total nitrate across Eastern North America.

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 0 and variance τ^2. The matrix B determines how the instantaneous change in any particular spatial location depends on the current values at that particular location and the surrounding spatial locations.

The solution ψ_t, for discrete time $t = 1, 2, \ldots$, is

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

where ν_t is an infinite-dimensional Gaussian vector with mean 0 and covariance matrix $\tau^2(B + B^T)^{-1}[I - \exp\{-(B + B^T)\}]$. Furthermore, it follows that components of ψ_t are Gaussian with mean 0 and have the infinite-dimensional covariance matrix $\tau^2(B + B^T)^{-1}$. Let

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

Spatial-temporal lattice processes of the form (2.1)–(2.2) are the focus of this paper and have their origins in Besag (1977). Besag (1977) noted that there is no loss in distributional properties of ψ_t if we replace B in equation (2.1) with $(B + B^T)/2$. 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, e.g., Besag (1974) and Besag and Kooperberg (1995). Thus, different choices of C made by specifying different sets of spatial conditional probability structures give rise to different spatial-temporal Gaussian lattice processes ψ_t. For example, consider Gaussian conditional autoregressions $\psi_{t,x}$, $x \in \mathcal{Z}^2$ with conditional mean and variance structure

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

where the real coefficients β_x are non-zero only on a finite set $\mathcal{N}$ such that $\beta_0 = 0$, $\beta_x = \beta_{-x}$, and $\sum_x \beta_x \cos(\omega^T x) < 1$, $\omega \in (-\pi, \pi]^2$. 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 $x \sim x'$) of each other if $x - x'$ belongs to $\mathcal{N}$. The matrix C is then specified by

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

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

The interpretation of the above discrete-time dynamical model and its parameters are straightforward. The parameters β_x, $x \in \mathcal{N}$ controls 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 the fluctuation around the conditional mean is just constant. The parameter λ_0 control the strength of 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 $\exp(-\lambda_0 C/2)$. The dependence generalizes the notion of correlation parameter of an autoregressive time series of order 1. In particular, all eigenvalues of $\exp(-\lambda_0 C/2)$

lie in $(0, 1]$, and this suggests that we are focusing on positive dependence between ψ_t and ψ_{t+1}.

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

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

We also need $|\beta_{1,0}| + |\beta_{0,1}| < \frac{1}{2}$ for 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

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

$$+\beta_{1,-1}(\psi_{t,k-1,l+1} + \psi_{t,k+1,l-1}) + \beta_{1,1}(\psi_{t,k-1,l-1} + \psi_{t,k+1,l+1})$$

and

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

The parameters values must ensure that C is nonnegative definite, for which a stringent sufficient condition is that $|\beta_{0,1}| + |\beta_{1,0}| + |\beta_{1,-1}| + |\beta_{1,1}| < \frac{1}{2}$.

Higher neighborhood-order versions, involving more neighbors, are constructed in a similar fashion.

In what follows, we shall primarily work with C that arises in the first neighborhood-order symmetric spatial-temporal Gaussian autoregressions. Specifically, for $k, l \in \mathcal{Z}$, we shall 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 $0 < \beta < \frac{1}{4}$, and

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

While λ_0 controls the strength of temporal dependence, the parameter β take into account of spatial dependence. When β is close to $\frac{1}{4}$, the spatial correlation decays very slowly and roughly logarithmically, as shown in Besag (1981). This is in contrast with the geometric rate of decay of the temporal autoregression.

Inference and computations for higher neighborhood-order spatial-temporal autoregressions and lattice-based fractional and Matérn dynamical system will be discussed in Section 7.

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 shall consider finite restriction of ψ_t on a two-dimensional $r \times c$ regular lattice with $n = rc$. However, to ensure non-negative definiteness and sparsity of the precision matrix, a finite restriction of ψ_t requires suitable boundary approximations. Here, we follow boundary conditions suggested in Besag and Higdon (1999), Dutta and Mondal (2015) and Mondal (2016). Let W_m denote a $m \times m$ tridiagonal matrix whose non-zero off-diagonal elements are $W_{m,i,i\pm1} = -1$ 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

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

with precision parameters λ_1 and λ_2, where

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

Stationary lattice processes are studied through elegant spectral representations, and the finite-dimensional matrix C in (2.3) also has an elegant and analytically known spectral decomposition. Specifically, let P_m denote

the $m \times m$ matrix corresponding to the discrete cosine transformation with entries

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

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

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

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

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

where both Λ and Λ_1 are $n \times n$ 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 and only $O(n \log n)$ computations; see, e.g., 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 it indicates whether an observation corresponds to a particular array cell. The vector $F_t\psi_t$ gives back the latent spatial-temporal variable values for the observed y_t. The vector ϵ_t represents the residual terms that are left unexplained by the variations in $F_t\psi_t$, and its entires are assumed to be independent and identically distributed Gaussian random variables with mean 0 and precision λ_3. Furthermore, suppose that ψ_t evolves as in (2.2) and (2.3), i.e.,

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

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 parameters λ, the objective here 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 0 and a precision matrix Γ, whose spectral

decomposition takes the form of $\Gamma = R^T M R$. The $ns \times ns$ matrix R is a block diagonal with all $n \times n$ diagonal blocks equal to P. The matrix M is 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 is immediate 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 of a 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 to solve $A\theta = b$ can be seen 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 would require computations of order $O(n^3 s)$ and a storage of $O(n^2 s)$ variables. Thus, as n becomes large, Kalman filtering becomes computationally challenging.

3.2 Estimation for precision parameters

Let $\hat{\psi}$ be the estimate of ψ from equation (3.2). The results of Patterson and Thompson (1971), 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 $Q_i = \partial Q/\partial\lambda_i$, $Q_{ij} = \partial^2 Q/(\partial\lambda_i\partial\lambda_j)$, $i, j = 0, \ldots, 3$. Denote by $H = X(X^TQX)^{-1}X^TQ$ the hat matrix of the regression (3.1). Treating $\hat{\psi}$ as fixed, the score equations become

$$(1/2)\{tr(Q^{-1}Q_i) - tr(HQ^{-1}Q_i) - (u - X\hat{\psi})^TQ_i(u - X\hat{\psi})\} = 0, \quad (3.3)$$

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

$$\mathscr{I}(i, j) = (1/2)tr(Q^{-1}Q_iQ^{-1}Q_j - HQ^{-1}Q_iHQ^{-1}Q_j + HQ^{-1}Q_{ij}).$$

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

4. Matrix-free computation

4.1 Lanczos algorithm to estimate state vectors

To solve $A\psi = b$ in (3.2), we follow the works of 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, Lanczos orthonormal vectors $u_1, \ldots, u_i$ reduce A to a partial tridiagonal form $AU_i \approx U_iT_i$ with $U_i = (u_1, \ldots, u_i)$

and T_i a suitable $i \times i$ positive definite tridiagonal matrix. As the algorithm progresses, the solution ψ is sequentially updated by computing

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

which only requires linear solving 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. In our case it is computed 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 that a Lanczos algorithm takes remains nearly constant or does not grow more than $O(\log(ns))$. When $\lambda > 0$, the algorithm only takes $O(ns \log(ns) \log n)$ 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 $tr((X^TQX)^{-1}X^TQ_iX)$ which can be difficult. As a solution, we follow stochastic approximations derived in Hutchinson (1990). Let z_k, $k = 1, 2 \ldots K$ be independent and identically distributed Rademacher random variables,

where K is an integer which is of order $\log(ns)$ at most. 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 then approximate (3.3) with unbiased estimating equations

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

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

4.3 Trust region method for precision parameter estimation

To solve (3.3), we follow the works of 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,

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

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 iterates as follows. First, the algorithm computes the step size α_{k+1} by minimizing $\Omega_k(\alpha)$. Second, it updates 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 values 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 $\hat{\lambda}$. We 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 0 and is proved by Powell (2004). Nocedal and Wright (2006) provide convergence results when the effective threshold is in $(0, 1/4)$ under the assumptions that the objective function is Lipschitz, continuously differentiable, and that the corresponding Hessian is bounded.

4.4 Preconditioning

To solve $A\psi = b$, a preconditioning matrix L can 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 the eigenvalues are clustered into few values. It also requires that 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 to solve 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 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 1 and the minimum eigenvalue is bounded above by $\lambda_0/(1 + \lambda_0)$. In many practical applications, $F^T F$ is a binary diagonal matrix and, in such instances, a fraction of eigenvalues of $(\lambda_3 I + M)^{-1/2}(\lambda_3 R F^T F R^T + M)(\lambda_2 I + M)^{-1/2}$ also clumps at 1. Furthermore, when $F^T F$ is the identity matrix, all eigenvalues of $(\lambda_3 I + M)^{-1/2}(\lambda_3 R F^T F R^T + M)(\lambda_3 I + M)^{-1/2}$ are equal to 1. $F^T F$ is the identity matrix when the regular grid has no missing values, then $L = (\lambda_3 I + M)^{-1/2}$ is the best choice. In practice, we don't 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 costs at least $O((ns)^{3/2})$ computations 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 to the identity matrix. The use of the incomplete Cholesky factorization facilitates faster convergence of the matrix-free Lanczos algorithm. We refer to Benzi (2002) and seminal work of Kershaw (1978) on incomplete Cholesky factorizations.

4.5 Conditional Simulation and log*-likelihood calculation*

One 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, $i = 1, \ldots, s$ are generated from a Gaussian distribution with mean 0 and covariance matrix Λ_1^{-1}. Then, we compute

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

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

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

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

Computation of log-residual likelihood function presents further challenges, and can be pursued in a matrix-free way as outlined in Dutta and Mondal (2017).

5. Extension to other spatial-temporal state-space models

5.1 Inference 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

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

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

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

Next, let $\varphi^T = (\varphi_1^T, \ldots, \varphi_s^T)$. Then φ follows a multivariate normal distribution with mean 0 and a sparse precision matrix Γ_Δ with a spectral decomposition $\Gamma_\Delta = R^T M_\Delta R$, where 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 get

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

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, 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 derivation of an efficient preconditioning matrix for the Lanczos algorithm through the use of the incomplete Cholesky decomposition.

5.2 Approximation to stochastic advection-diffusions

The small time steps 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. As a specific example, we consider a stochastic advection-diffusion equation that has the form

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

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

$$\mathscr{A}\psi(t,x) = 2\mu^T \partial\psi(t,x)/\partial x - tr\{\partial^2\psi(t,x)/(\partial x \partial x^T)\}\Sigma + 2\tau\psi(t,x), \quad (5.3)$$

In the above, 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 $\tau\psi(t,x)$ control the damping or decay with rate τ. For references on stochastic advection diffusion equations, see Whittle (1963), Brown et al. (2000), Cressie and Wikle (2011) and Sigrist et al. (2015). Unlike (2.1), the model in (5.2)–(5.3) has no simple and explicit analytic solution, and approximations are necessary to pursue 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 $\mu = 0$, $\Sigma = \gamma_1 I$, $\tau = -\gamma_2$. The advection-diffusion equation then takes the form

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

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

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

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 a state model

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

where

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

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

Assuming z_t is normally distributed with mean 0 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 0 and precision matrix $\Gamma^\dagger$. Furthermore,

$\Gamma^\dagger = R^T M^\dagger R$, 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

$$M^\dagger_{(1,1)} = M^\dagger_{(s,s)} = \gamma_3 I, \quad M^\dagger_{(i,i)} = \gamma_3(I + (\Lambda^\dagger)^2), \quad i = 2, \ldots, s-1,$$

$$M^\dagger_{(j,j+1)} = M^\dagger_{(k-1,k)} = -\gamma_3 \Lambda^\dagger, \quad j = 1, \ldots, s-1, \quad k = 2, \ldots, s.$$

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

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

which is very similar to (3.1) and the state vectors are estimated as

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

As was the case with small time-step approximations in Section 5.1, the matrix $A^\dagger = X^T Q^\dagger X$ is sparse which allows us to use incomplete Cholesky preconditioning. This enables faster and more efficient matrix-free computations and resolves the inferential challenges discussed in Sigrist et al. (2015).

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 spectral representation, we generate $\theta_1, \ldots, \theta_{10}$ on a 128×128 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 0 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 $n_+ = 131072$, and $rcs = 163840$. The method requires working with matrices of size 163840×163840, 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, and we obtain the overall REML estimates $\hat{\lambda}$ of the precision parameters.

Table 1 summarizes these results and along with standard error values. We note that 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 they have smaller standard errors.

To illustrate the effect of preconditioning in Section 4, we next focus on the eigenvalues for $\lambda_2 RF^T FR^T + M$ and $(\lambda_2 I + M)^{-1/2}(\lambda_2 RF^T FR^T +$

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

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

$M)(\lambda_2 I + M)^{-1/2}$. However, most available softwares such as Matlab, Python and R failed to accurately compute the eigenvalues decomposition of the 163840×163840 matrices as used in the above simulation study. Instead, we choose $r = 64$, $c = 64$, and $s = 3$ for which the dimension of matrices $\lambda_2 RF^T FR^T + M$ and $(\lambda_2 I + M)^{-1/2}(\lambda_2 RF^T FR^T + M)(\lambda_2 I + M)^{-1/2}$ reduces to 12288×12288. Figure 1 provides the eigenvalues plot for these matrices. For the latter, we see that a large proportion of eigenvalues are clusters at 1 and others are below 1 but strictly about from 0, which results in a much smaller conditioning number compared with

$$\lambda_2 RF^T FR^T + M$$

and the original A matrix. As a consequence of this preconditioning, we get a speed up in the convergence of the Lanczos algorithm.

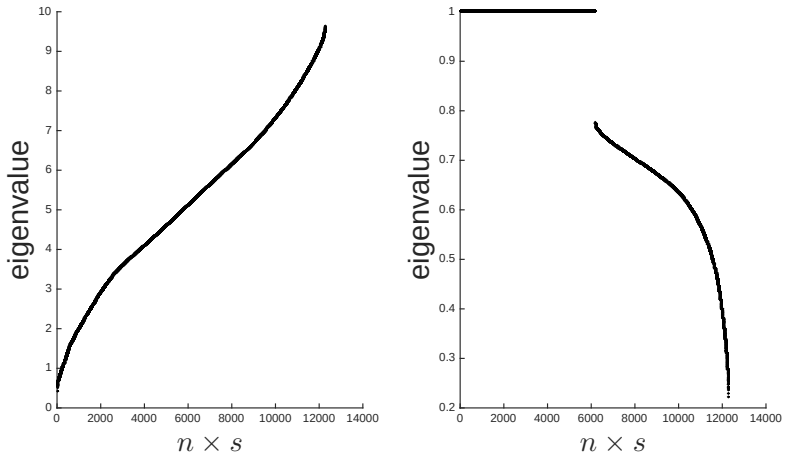

Figure 1: The left plot shows the eigenvalues of $\lambda_2 RF^T FR^T + M$. The right plot shows eigenvalues of $(\lambda_2 I + M)^{-1/2}(\lambda_2 RF^T FR^T + M)(\lambda_2 I + M)^{-1/2}$

6.2 Analysis of soil moisture data

The Climate Prediction Center of the National Weather Service provides monthly mean soil moistures at $0.5^o \times 0.5^o$ spatial resolution for the time period 1948 to 2014. For further references, see Robock et al. (2000) and Fan and van den Dool (2004). Here, we consider a subset of the data with latitudes between 40^oN and 50^oN, longitudes between 95^oW and 75^oW, for the time period from January, 2009 to December, 2009. The subset constitutes $s = 12$ months of data, and is spatially embedded into

Table 2: REML estimates of precision parameters for soil moisture data under no splitting (Scenario 1) and 4×4 splitting (Scenario 2) of the original array. The standard errors are in parentheses.

Parameters	λ_0	λ_1	λ_2	λ_3
Scenario 1	0.325	220.289	473.386	0.977
	(0.005)	(22.772)	(258.081)	(0.205)
Scenario 2	0.193	589.850	327.251	0.976
	(0.003)	(89.233)	(35.156)	(0.002)

a 40×20 array. Due to presence of lakes and water bodies, there are 36 array cells with no observations. Let $y_1, \ldots y_{12}$ be the observed monthly mean soil moisture in this subset. Next, we consider the state-space model in (2.4)–(2.5). We consider two different scenarios. First, the underlying state vectors ψ_t, $t = 1, \ldots, 12$, follow spatial-temporal autoregressions as in equation (2.4) at the original spatial resolution $0.5^o \times 0.5^o$. Second, the state vectors follow spatial-temporal autoregressions at a finer spatial resolution $0.125^o \times 0.125^o$.

In Scenario 2, we split each array cell into 4×4 sub-cells so that ψ_t, $t = 1, \ldots, 12$ lie on a160×80 spatial array, and $rcs = 153600$. Accordingly,

we construct the averaging matrix F_t, each of order 12800×764, such that $F_t\psi_t$ provides the vector of average of state values at the original $0.5^o \times 0.5^o$ spatial resolution. This scenario is particularly useful if we want to obtain spatial interpolation at a finer resolution. Furthermore, finer resolution allows us to achieve approximate inference from the limiting continuum geostatistical model; see, e.g., Besag and Mondal (2005), and Dutta and Mondal (2015, 2016) for examples of such inferences in spatial statistics. At spatial resolution $0.125^o \times 0.125^o$, computations are particularly challenging as we have a 153600×153600 non-sparse block triangular matrix A, where each block of A is of order 12800×12800.

Table 2 summarizes the REML estimates for standardized monthly soil moisture values. In Scenario 2, the estimate of λ_0 decreases, and the estimate of $\beta = \lambda_1/(4\lambda_1 + 2\lambda_2)$ increases. This explains that ψ_ts are spatially and temporally more dependent at spatial resolution $0.125^o \times 0.125^o$. Finally, Figure 1 displays the actual observations, and the prediction for latent variables for the last four months. We see that the model performs reasonably well in predicting soil moisture at spatial resolution $0.125^o \times 0.125^o$. In particular, in Scenario 2, about 76% percentage of total variation is explained by spatial-temporal state-space model. The algorithms detailed in Section 4 made these computations possible without dimension reduction

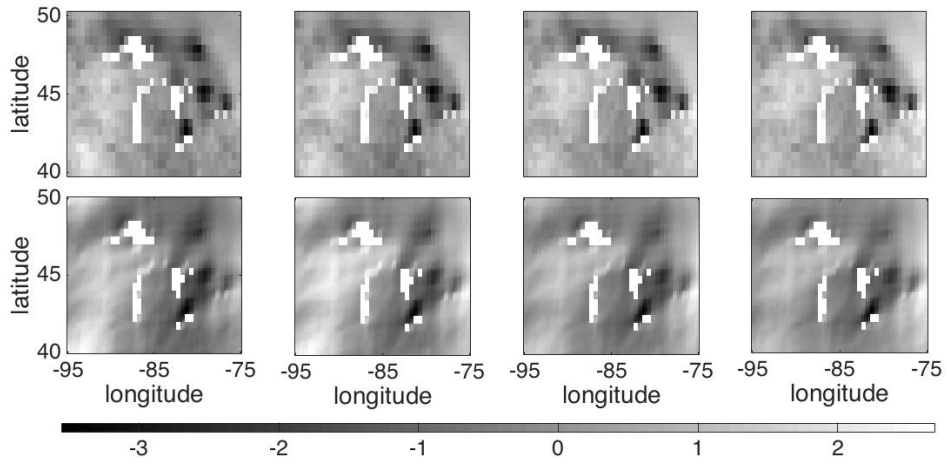

Figure 2: The top panel shows image plots of $y_9, \ldots, y_{12}$. The bottom panel displays $\hat{\psi}_9, \ldots, \hat{\psi}_{12}$ at $0.125^o \times 0.125^o$ spatial resolution. The horizontal bar gives the gray scale.

or ensembles of stochastic simulations.

6.3 Analysis of atmospheric concentrations of total nitrate

The Environmental Protection Agency provides output from the numerical model called Models-3 that contains atmospheric total Nitrogen concentration level as one component. The atmospheric total Nitrogen concentration level is defined as 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

Table 3: REML estimates of precision parameters for total Nitrogen concentrations 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	7.894	14.571
	(0.062)	(0.088)	(0.062)
Scenario 2	28.726	2.285	13.931
	(0.118)	(0.028)	(0.119)

of regional weather, the emission process and land use and cover information. For further inference on Model-3 output and their statistical analyses, we refer to Fuentes and Raftery (2005) and Ghosh et al. (2010). In particular, Fuentes and Raftery (2005) did not consider any spatial-temporal modeling, but focused on combining Model-3 data with observations from certain monitoring stations. On the other hand Ghosh et al. (2010) used an elaborate Bayesian dynamical model to atmospheric total nitrate.

Here we consider Models-3 output (provided by Montse Fuentes) for total Nitrogen concentrations for the year 2001 for first 12 time periods. The left panel of Figure 3 provides the map of the region. The latitudes

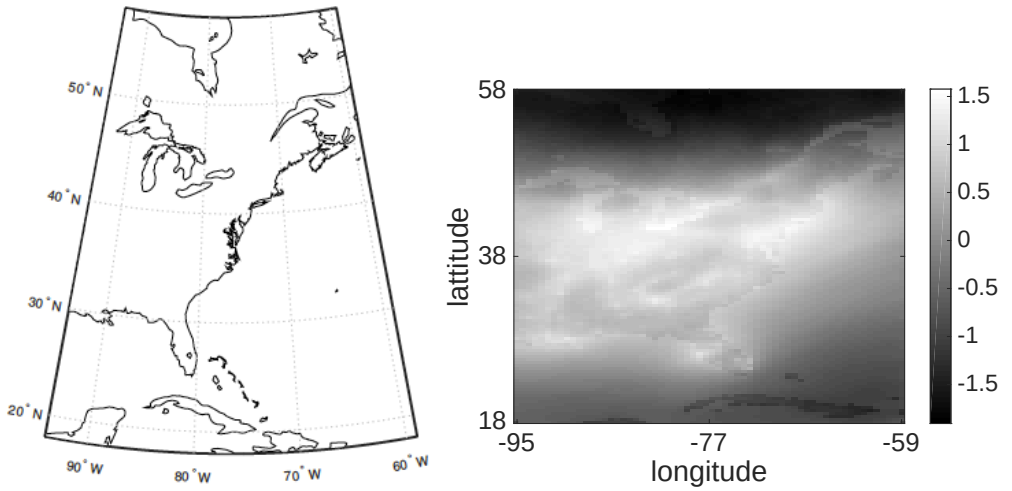

Figure 3: The left panel shows image plot of average total Nitrogen concentration over 12 months. The right panel displays map information of the sample regions.

are between 18^oN and 58^oN and longitudes are between 59^oW and 94^oW. The total nitrate data are embedded in a spatial array of size 62×112. and we take log-transformation of the data to reduce skewness and to improve the normality assumption. The right panel of Figure 3 displays the image plot of the mean log-total Nitrogen concentrations across the region. We see that there is very strong spatial dependence across the entire region. The Nitrogen concentrations is hight across the continental North America where population density is high, human settlement is dense and there are compact urbanism and industrialization. Furthermore, the mean total

Nitrogen concentrations thin out over the North Atlantic Ocean, the Gulf of Mexico, the Hudson Bay and the northern Canada.

Our objective here is to investigate how spatial-temporal state-space models in (2.1)–(2.5) perform in explaining variations in the Models-3 output. To this end, let $y_1, \ldots, y_{12}$ be the standardized monthly log-averaged total Nitrogen concentrations from numerical models. It is not unreasonable to think that a large-scale atmospheric pollution are driven by some form of advection-diffusion equations. However, we lack other information such as pollution sources, effect of wind speed and wind direction, and rate of atmospheric depositions of pollutions into land and ocean. Instead, we focus on fitting a parsimonious stochastic advection-diffusion equation in Section 5.2 with $\mu = 0$, $\Delta = 0.01$, $\Delta_0 = 1$,

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

and

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

We apply two scenarios. In Scenario 1, we assume that the underlying state vectors ψ_t, $t = 1, \ldots, 12$, 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, $t = 1, \ldots, 12$ lie on a124×224 spatial array, and $rcs = 333312$. Scenario 2 is particularly useful for downscaling and for approximating continuum geo-statistical processes. As in soil moisture data analysis, we next construct the averaging matrix F_t, each of order 27776×6944. Here, we need to deal with a 333321×333312 sparse block triangular matrix A, where each block of A is of order 27776×27776 and F_ts are not identify matrix. Thus, statistical computations are quite daunting and traditional Kalman filtering algorithm does not work in this case.

For this model, REML estimation, however, ran into boundary problem. Specifically, we found that the estimate of the decay rate γ_2 to be tending to the boundary point 0. Therefore, we set $\gamma_2 = 0$, which then results in a simpler model with

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

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

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

The relationship between λ and γ are

$$\gamma_1 = \lambda_0\lambda_1/2, \quad \gamma_3 = 1/(\lambda_0\Delta), \quad \gamma_4 = \lambda_3.$$

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

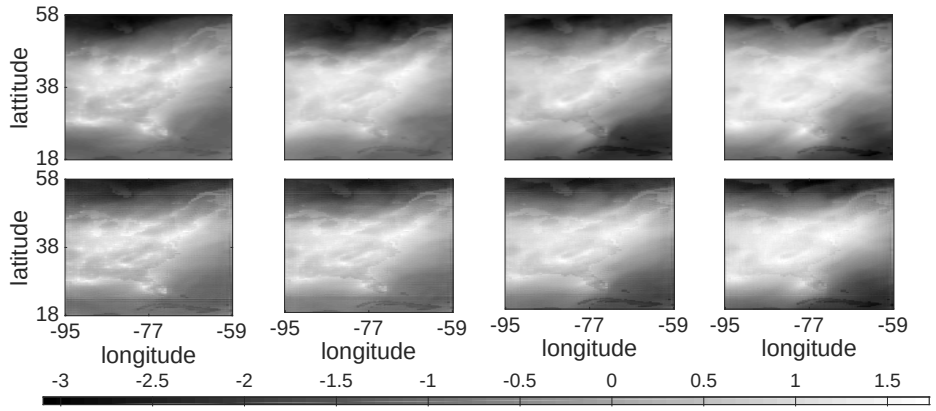

Figure 4: The top panel shows image plots of $y_9, \ldots, y_{12}$. The bottom panel displays $\hat{\psi}_9, \ldots, \hat{\psi}_{12}$ at 18×18 km^2 spatial resolution.

Table 3 provides the summary of REML estimates for standardized log-transformed numerical values for total Nitrogen concentrations using methods in Section 5.1. As before, we used $p = 50$ Rademacher vectors to approximate trace terms. For verification, we also compute the precision parameters using 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 respectively. Therefore, the relationship (5.1) between $(\lambda_0, \lambda_1, \lambda_3)$ and $(\gamma_1, \gamma_3, \gamma_4)$ is satisfied. Finally, Figure 2 displays the actual observa-

tions and the prediction for latent variables at a finer spatial resolution of 18×18 km^2 for the last four months and we see that a parsimonious advection-diffusion model is quite effective in downscaling and in explaining a large fraction (about 96%) of the total variations in the data.

7. Discussion

Using circulant embedding (see Appendix), the score equation (3.3) can be shown to be equivalent to a gamma non-linear regression model, in contrast to the results in Lee and Nelder (1996) and Dutta and Mondal (2015) where estimation of precision parameters is equivalent to fitting a gamma generalized linear model. Thus, the optimization in (4.1) is a non-convex one, and it 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 maxima make the information matrix nearly singular and the standard error computation difficult. In our limited experience with numerical computation, we have encountered some of these issues. 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 was studied by Mardia and Watkins (1989), Dietrich (1991) and others, and is often related to range parameter estimation. Similarly,

long flat ridges in the likelihood function of spatial models were studied in Zhang (2004). However, further work is needed to uncover the nature of multi-modality and long flat ridges in a spatial-temporal setting.

Notwithstanding these expected issues the paper advances computations and methods in spatial-temporal settings and resolves inferential challenges discussed in Sigrist et al. (2015). In fact, there are further possible extensions of the spate-space dynamical model in (2.4) and (2.5). For example, we can consider higher neighborhood-order conditional autoregressions, or fractional spatial random fields. In the former, we replace the precision matrix C in equation (2.5) with $C = J(W)$ where J is a suitable positive polynomial; see Mondal (2016) for details. In the latter, C in equation (2.5) is replaced with C^κ, $\kappa > 0$, which corresponds to the fractional Laplacian differenced random fields and approximate spatial Matérn processes. For both the higher neighborhood-order conditional autoregressions and the fractional Laplacian differenced random fields, M still provides the eigenvectors of the precision matrix of the state variables. We can thus calculate fast matrix-vector products without storing any matrices and pursue computations as presented in Section 4. If needed, we can also include covariate information and consider mixed effect models. Furthermore, various complex spatial-temporal autoregressions arise from small time-step

discretization of a wide variety of SDPEs. Sections 5 and 6 demonstrate how we can implement these elaborate models. The computations proposed here are better than those presented in Rue et al. (2009) and Lindgren et al. (2011). The method presented here will have further applications in data assimilation and computer simulations involving discrete linearizations of complex non-linear stochastic particle differential equations. These applications typically involve a small or moderate value of s and a large or very large n and are ideal for out matrix-free computations.

If in certain applications, both s and n are very large, one can also adopt various strategies including parallel and distributive computing. The discrete cosine transform, the matrix-vector product calculations, the Lanczos algorithm, the stochastic trace approximations are in fact all parallelizable. For references on this topic, see e.g., Frigo and Johnson (2005) and Kim and Chronopoulos (1991).

In this paper, we focused on spatial-temporal model where the time dynamics follow an autoregressive structure of order 1 (i.e., ψ_{t+1} given $\psi_1, \ldots, \psi_t$ depends only on ψ_t through the dependence matrix $\exp(-\lambda_0 C/2)$). Furthermore, equations (A.1)–(A.2) explicitly show how both the temporal autoregressive structure and the spatial dependence structure enter into the spectral factorization of the spatial-temporal models. This was pos-

sible because the two-dimensional discrete cosine transform of ψ_t breaks the spatial dependence structure and convert ψ_t into independent components. This can be extended further to construct spatial-temporal models where the temporal dynamics have an autoregressive structure of order p or have a fractional dependence structure. The construction of general spatial-temporal lattice systems will be pursued in a future work.

Finally, spatial-temporal non-Gaussian state-space models and Bayesian computations are of interest. Examples include binomial or Poisson models and for application see Brix and Diggle (2001). They often arise in generalized linear model when data y_t is a response to a linear predictor χ_t that can be represented as

$$\chi_t = T_t\delta + F_t\psi_t + \epsilon_t. \tag{7.1}$$

In the above δ is covariate effect, T_t provides covariate information at time t, ψ_t is underlying (often unobserved) spatial-temporal effect, F_t is sparse incident matrix or sparse averaging operator, and ϵ_t is residual effect. The model (7.1) generalizes the Gaussian state-space model in (2.3)–(2.5). REML analysis does not extend to the non-Gaussian model in (7.1) and statistical inference typically requires (Bayesian) Markov chain Monte Carlo (MCMC) and other simulation-based computations. The best linear unbiased prediction calculations and the method of conditional simu-

lations presented in this paper is however relevant in this context. These can computations can be used to develop various matrix-free scalable block Gibbs-Metropolis-Hastings and Hamiltonian MCMC computations.

Acknowledgements

The work is supported by a NSF Career award DMS 1519890.

Appendix: Non-convexity in REML estimation

To characterize the non-convexity in the REML estimation, we apply circulant embedding and rewrite spatial-temporal autoregressions (2.5) as

$$\psi_1 = G\psi_{2s} + \nu_1, \psi_i = G\psi_{i-1} + \nu_i, \quad i = 2, \ldots, 2s.$$

Let $\psi_E^T = (\psi_1^T, \ldots, \psi_{2s}^T)$, ψ_E is normally distributed with mean 0 and a precision matrix Γ_E, whose spectral decomposition takes the form of $\Gamma_E = R_E^T M_E R_E$. The $2ns \times 2ns$ matrix R_E is a block diagonal matrix with all $n \times n$ diagonal blocks equal to P. The matrix M_E is block circulant matrix with non-zero blocks $M_{E,(i,i)}$ (diagonal), $M_{E,(i,i+1)}$ (upper diagonal), $M_{E,(i-1,i)}$ (lower diagonal), $M_{E,(1,2s)}$(last block in first row) and $M_{E,(2s,1)}$(first block in last row) are such that

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

$$M_{E,(1,2s)} = M_{E,(2s,1)} = -e^{-\lambda_0 \Lambda/2}\Lambda_1^{-1},$$

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

Furthermore, symmetric permutation of rows and columns on M_E results in a block diagonal matrix with circulant blocks. Let $\mathscr{P}$ be the permutation matrix with $\mathscr{P}^T \mathscr{P} = I$, then $M_E = \mathscr{P}^T \mathscr{K} \mathscr{P}$. The matrix $\mathscr{K}$ is a block diagonal matrix with $2s \times 2s$ diagonal blocks $\mathscr{K}_i$, $i = 1, \ldots, n$. Each $\mathscr{K}_i$ is circulant matrix with spectral decomposition denoted as $\mathscr{K}_i = \Phi^T \mathscr{T}_i \Phi$, where Φ corresponds to the discrete Fourier transform and $\mathscr{T}_i$ is a diagonal matrix with jth element

$$\mathscr{T}_{i,j} = \rho_i(1+e^{-\lambda_0 \rho_i})/(1-e^{-\lambda_0 \rho_i}) - 2\{\rho_i e^{-\lambda_0 \rho_i/2}/(1-e^{-\lambda_0 \rho_i})\} \cos\{\pi(j-1)/(2s)\} \tag{A.1}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, 2s$, where ρ_i is the ith diagonal element of Λ. Let Φ_E denote a $2ns \times 2ns$ block diagonal matrix with all $2s \times 2s$ diagonal blocks equal to Φ. Then, $\mathscr{K} = \Phi_E^T \mathscr{T} \Phi$, where $\mathscr{T}$ is a diagonal matrix with diagonal blocks $\mathscr{T}_1, \ldots, \mathscr{T}_n$. Therefore, the spectral decomposition of the original precision matrix Γ_E is

$$\Gamma_E = \Phi_E^T \mathscr{P}^T R_E^T \mathscr{T} R_E \mathscr{P} \Phi_E. \tag{A.2}$$

Now we follow the settings in Section 3.1. Let $\eta_E^T = (\eta_1^T, \ldots, \eta_{2s}^T)$. It is immediate that $\eta_E \sim N(0, M_E^{-1})$. Next, let $n_+ \times 2ns$ matrix F_E be the

column binding of F and a $n_+ \times 2ns$ 0 matrix, and assume that

$$u_E = \begin{pmatrix} y \\ 0 \end{pmatrix}, X_E = \begin{pmatrix} F_E \\ R_E \mathscr{P} \Phi_E \end{pmatrix}, \zeta_E = \begin{pmatrix} \epsilon \\ \eta_E \end{pmatrix}, Q_E = \begin{pmatrix} \lambda_3 I_{n+} & 0 \\ 0 & \mathscr{T} \end{pmatrix},$$

where u_E is a $(n_+ + 2ns)$ column vector and Q_E is a $(n_+ + 2ns) \times (n_+ + 2ns)$ diagonal matrix. The state-space model with circulant embedding takes the regression form

$$u_E = X_E \psi_E + \zeta_E,$$

where $\zeta_E \sim N(0, Q_E^{-1})$, which is very similar to (3.1) and the state vectors can be reestimated as

$$X_E^T Q_E X_E \hat{\psi}_E = X_E^T Q_E u_E.$$

The log-residual likelihood function in Section 3.2 now can be rewritten as

$$2l_E(\lambda) = \log |Q_E| - \log |X_E^T Q_E X_E| - (u_E - X_E \hat{\psi}_E)^T Q_E (u_E - X_E \hat{\psi}_E).$$

Next, denote by $H_E = X_E (X_E^T Q_E X_E)^{-1} X_E^T Q_E$ the hat matrix. Let δ_i be the square of the ith element of the residual vector $u_E - X_E \hat{\psi}_E$ and let $q_{E,i}$ be the ith diagonal element of Q_E. Take $\delta_i^* = \delta_i / (1 - h_{E,i})$, where $h_{E,i}$ the ith diagonal element of the hat matrix H_E. The score equations in (8) are then same as

$$\partial l_E / \partial \lambda_3 = (1/2) \sum_{i=1}^{2ns} (1 - h_{E,i})(\delta_i^* - \lambda_3) / \lambda_3^2,$$

$$\partial l_E/\partial \lambda_i = (1/2) \sum_{i=n_++1}^{n_++2ns} (\partial q_{E,i}/\partial \lambda_j)(1-h_{E,i})(\delta_i^* - q_{E,i})/q_{E,i}^2, \quad j = 0, 1, 2.$$

These score equations coincide with estimating equations of a gamma regression where the response variables are adjusted squared residuals δ_i^* and follow independent gamma distribution. Furthermore, we have inverse link, nonlinear predictors $q_{E,i}$ as a non-linear function of λ_0, ..., λ_3 and prior weights $(1 - h_{E,i})$. The non-linearly in the gamma regression specifies the exact nature of the non-convexity in the REML estimation.

References

Benzi, M. (2002). Preconditioning techniques for large linear systems: a survey. *Journal of Computational Physics 182*(2), 418–477.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological) 36*(2), 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*, pp. 47–55. Springer.

Besag, J. (1981). On a system of two-dimensional recurrence equations.

Journal of the Royal Statistical Society. Series B (Methodological) 43, 302–309.

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

Besag, J. and D. Higdon (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(4), 691–746.

Besag, J. and C. Kooperberg (1995). On conditional and intrinsic autoregression. *Biometrika 82*(4), 733–746.

Besag, J. and D. Mondal (2005). First-order intrinsic autoregressions and the de Wijs process. *Biometrika 92*(4), 909–920.

Brix, A. and P. J. Diggle (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(4), 823–841.

Brown, P. E., G. O. Roberts, K. F. Kåresen, and S. Tonellato (2000). Blur-generated non-separable space–time models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 847–860.

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

Cressie, N. and C. K. Wikle (2011). *Statistics for spatio-temporal data*. New York: John Wiley & Sons.

Dietrich, C. (1991). Modality of the restricted likelihood for spatial Gaussian random fields. *Biometrika 78*(4), 833–839.

Dutta, S. and D. Mondal (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*(3), 699–726.

Dutta, S. and D. Mondal (2016). Reml estimation with intrinsic Matérn dependence in the spatial linear mixed model. *Electronic Journal of Statistics 10*(2), 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*(C5), 10143–10162.

Evensen, G. (2009). *Data assimilation: the ensemble Kalman filter*. Berlin: Springer Science & Business Media.

Fan, Y. and H. van den Dool (2004). Climate prediction center global monthly soil moisture data set at 0.5 resolution for 1948 to present. *Journal of Geophysical Research: Atmospheres 109*, D10102.

Frigo, M. and S. G. Johnson (2005). The design and implementation of FFTW3. *Proceedings of the IEEE 93*(2), 216–231.

Fuentes, M. and A. E. Raftery (2005). Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics 61*(1), 36–45.

Ghosh, S. K., P. V. Bhave, J. M. Davis, and H. Lee (2010). Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. *Journal of the American Statistical Association 105*(490), 538–551.

Greenbaum, A. and Z. Strakos (1992). Predicting the behavior of finite precision Lanczos and conjugate gradient computations. *SIAM Journal on Matrix Analysis and Applications 13*(1), 121–137.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*(358), 320–338.

Houtekamer, P. L. and H. L. Mitchell (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review 126*(3), 796–811.

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*(2), 433–450.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*(1), 35–45.

Kalman, R. E. and R. S. Bucy (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering 83*(1), 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*(1), 43–65.

Kim, S. K. and A. T. Chronopoulos (1991). A class of lanczos-like algorithms implemented on parallel computers. *Parallel Computing 17*, 763–778.

Künsch, H. R. (1987). Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika 74*(3), 517–524.

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

Lee, Y. and J. A. Nelder (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika 88*(4), 987–1006.

Lindgren, F., H. Rue, and J. Lindström (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*(4), 423–498.

Mardia, K. and A. Watkins (1989). On multimodality of the likelihood in the spatial linear model. *Biometrika 76*(2), 289–295.

Mardia, K. V., C. Goodall, E. J. Redfern, and F. J. Alonso (1998). The kriged Kalman filter. *Test 7*(2), 217–282.

Mondal, D. (2016). On edge correction of conditional and intrinsic autoregressions. *Biometrika*, Under revision.

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

Paige, C. C. and M. A. Saunders (1975). Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis 12*(4), 617–629.

Paige, C. C. and M. A. Saunders (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS) 8*(1), 43–71.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*, 545–554.

Powell, M. (1984). On the global convergence of trust region algorithms for unconstrained minimization. *Mathematical Programming 29*(3), 297–303.

Powell, M. (2004). On the use of quadratic models in unconstrained minimization without derivatives. *Optimization Methods and Software 19*(3-4), 399–411.

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

Robock, A., K. Y. Vinnikov, G. Srinivasan, J. K. Entin, S. E. Hollinger, N. A. Speranskaya, S. Liu, and A. Namkhai (2000). The global soil moisture data bank. *Bulletin of the American Meteorological Society 81*(6), 1281–1299.

Rue, H., S. Martino, and N. Chopin (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*(2), 319–392.

Sigrist, F., H. R. Künsch, and W. A. Stahel (2015). Stochastic partial differential equation based modelling of large space–time data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(1), 3–33.

Stroud, J. R., M. L. Stein, B. M. Lesht, D. J. Schwab, and D. Beletsky (2010). An ensemble Kalman filter and smoother for satellite data assimilation. *Journal of the American Statistical Association 105*(491),

978–990.

Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute 40*(2), 974–994.

Wikle, C. K. and N. Cressie (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika 86*(4), 815–829.

Zhang, H. (2004). nconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association 99*(465), 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)