Statistica Sinica **25** (2015), 61-79
doi:http://dx.doi.org/10.5705/ss.2013.085w

A BAYESIAN SPATIO-TEMPORAL GEOSTATISTICAL MODEL WITH AN AUXILIARY LATTICE FOR LARGE DATASETS

Ganggang Xu[1], Faming Liang[1] and Marc G. Genton[2]

[1] *Texas A&M University and* [2] *King Abdullah University of Science and Technology*

Abstract: When spatio-temporal datasets are large, the computational burden can lead to failures in the implementation of traditional geostatistical tools. In this paper, we propose a computationally efficient Bayesian hierarchical spatio-temporal model in which the spatial dependence is approximated by a Gaussian Markov random field (GMRF) while the temporal correlation is described using a vector autoregressive model. By introducing an auxiliary lattice on the spatial region of interest, the proposed method is not only able to handle irregularly spaced observations in the spatial domain, but it is also able to bypass the missing data problem in a spatio-temporal process. Because the computational complexity of the proposed Markov chain Monte Carlo algorithm is of the order $O(n)$ with n the total number of observations in space and time, our method can be used to handle very large spatio-temporal datasets with reasonable CPU times. The performance of the proposed model is illustrated using simulation studies and a dataset of precipitation data from the coterminous United States.

Key words and phrases: Auxiliary Lattice, Bayesian hierarchical spatio-temporal model, Gaussian Markov random field, large datasets, spatio-temporal kriging.

1. Introduction

The analysis of spatio-temporal data is a current research topic in such areas as geophysical and environmental sciences. Due to technological advances in data collection, large amounts of observations can be obtained from many spatial locations over time. These datasets impose computational challenges to the implementation of traditional spatial statistical tools, such as maximum likelihood estimation and kriging. For a spatial process, various approaches have been proposed to facilitate the computation of large datasets; examples include covariance tapering (Furrer, Genton, and Nychka (2006)), Gaussian predictive processes (Banerjee et al. (2008)), fixed rank kriging (Cressie and Johannesson (2008)) and Gaussian Markov random fields (GMRF, Rue and Held (2005)). See Sun, Li, and Genton (2012) for a review. Much less work has been done on spatio-temporal modelling of geostatistical processes with large amounts of observations and we addresses this issue.

Let $Y(\boldsymbol{s},t)$ denote a real-valued spatio-temporal Gaussian process observed on $\mathbb{R}^d \times \mathbb{Z}$. We consider the model

$$Y(\boldsymbol{s},t) = \mu(\boldsymbol{s},t) + Z(\boldsymbol{s},t) + e(\boldsymbol{s},t), \tag{1.1}$$

where $Z(\boldsymbol{s},t)$ is an unobserved spatio-temporal process and $e(\boldsymbol{s},t)$ is a temporally and spatially uncorrelated Gaussian measurement error observed at location $\boldsymbol{s}$ and time t with mean 0 and variance σ_e^2. We suppose that

$$\mu(\boldsymbol{s},t) = \xi_0 + \xi_1 c_1(\boldsymbol{s},t) + \cdots + \xi_p c_p(\boldsymbol{s},t), \tag{1.2}$$

where $c_1,\ldots,c_p$ are some observed covariates at location $\boldsymbol{s}$ and time t, and $\boldsymbol{\xi} = (\xi_0,\ldots,\xi_p)^{\mathrm{T}}$ is a vector of parameters. In this model, $\mu(\boldsymbol{s},t)$, $Z(\boldsymbol{s},t)$, and $e(\boldsymbol{s},t)$ represent the large-scale, small-scale, and fine-scale variability, respectively.

We focus on spatio-temporal datasets with a large number of spatial locations, assuming that the same locations are monitored over time but that observations from some locations may be missing at some time points, which is common in geostatistical practice. Denote by $(\boldsymbol{s}_{i,t},t)$ the ith location at time t for $i = 1,\ldots,n_t$ and $t = 1,\ldots,T$. The n_t's may be different and the total number of observations is $n = \sum_{t=1}^{\mathrm{T}} n_t$. A motivating example is the annual total precipitation data for the coterminous United States, available at (`http://www.image.ucar.edu/Data/US.monthly.met/`). The original dataset consists of monthly precipitation observations from 11,918 stations throughout the United States and we manually converted them into annual total precipitation observations. In our dataset, for each year, observations from about 6,000–7,000 of these 11,918 stations are recorded. Due to missing observations, we have 6,905, 5,744, 6,595, 6,438, and 6,159 stations that have complete yearly records from 1980 to 1984. Since the stations with recorded observations are different from year to year, it is challenging to analyze the data using a spatio-temporal model. Our primary goal is to conduct computationally efficient spatio-temporal kriging for any given location and time point, using the complete dataset. We revisit this dataset in Section 4.

There are two main approaches to modelling $Z(\boldsymbol{s},t)$. The first treats time as an additional dimension and uses a $(d+1)$-dimensional covariance function to model the correlation among different locations and time points; see, e.g., Cressie and Huang (1999) and Allcroft and Glasbey (2003). This approach has drawbacks. To define a valid spatio-temporal covariance function, it is critical to define a meaningful distance that involves both space and time coordinates and this is not easy when spatial distance and temporal distance have different units and physical interpretations. The spatial domain is usually fixed while the time domain usually keeps increasing and asymptotically, infill asymptotics

suit the spatial domain process while increasing domain asymptotics are more appropriate for temporal processes; see Stein (1999). Then too, computation can be prohibitive for a random field where the number of spatial locations is large.

A second approach to modelling $Z(\boldsymbol{s}, t)$ is to use dynamical probabilistic models (Stroud, Müller, and Sansó (2001); Cressie and Wikle (2011)). There are some recent developments in this direction. Cressie, Shi, and Kang (2010) proposed modelling the spatial correlation function using a low rank basis approximation and the temporal dependence with a vector autoregressive process. Katzfuss and Cressie (2012) further proposed a Bayesian hierarchical spatio-temporal random effects model that uses the Markov chain Monte Carlo (MCMC) method to efficiently generate samples from posterior distributions. Finley, Banerjee, and Gelfand (2012) developed a space-time version of a Gaussian predictive process to conduct Bayesian dynamic modeling for large spatio-temporal datasets. These works focus on approximating the Gaussian random field by a lower dimensional spatial process using smoothing techniques such as basis function approximations. Computational cost can be reduced to a certain extent, but can still be high if the number of knots used is large, which is desirable if the primary goal is to make accurate predictions. As pointed out in Banerjee et al. (2008) and Sang and Huang (2011), a predictive process with a small number of knots provides a poor approximation of the dependence structure between the pairs of observations obtained at locations very close to each other. Yet the nearest observations have the largest impact on the prediction at a particular location (Stein (1999)). Another approach is to use a Gaussian Markov random field (GMRF, Rue and Tjelmeland (2002)) model that can approximate a Gaussian random field well with small neighborhoods (Rue and Held (2005); Lindgren, Rue, and Lindstrom (2011)). For example, Lemos and Sansó (2009) model an irregularly spaced spatio-temporal process as a kernel convolution of a latent GMRF, where the choice of the kernel is largely subjective and can have a substantial impact.

We propose a Bayesian hierarchical model using a latent GMRF defined on an auxiliary lattice to approximate the spatial correlation at a given time point, and a vector autoregressive model for the temporal transition of latent states. The Bayesian model is appealing in that it can provide a better quantification of uncertainty in estimating parameters and making predictions. By using precision and transition matrices of particular forms, the proposed model can be used to model more complex spatial and temporal correlations. At the same time, one can avoid matrix inversion by taking advantage of analytical properties of block circulant matrices and thus reduce the computational cost for MCMC iterations to $O(n)$, where n is the total number of observations in space and time.

The paper is organized as follows. Section 2 gives the details of the proposed method and its implementation. In Section 3, simulations are assessed for the

predictive performance of the proposed method. The example of precipitation data is used in Section 4 to illustrate the use of the proposed model when the data size is large. Some discussions are given in Section 5. The proof of Proposition 1 and a detailed MCMC algorithm are given in the supplementary document.

2. A Bayesian Spatio-Temporal Geostatistical Model

2.1. Auxiliary Gaussian Markov random fields

Suppose that we have a spatio-temporal dataset $Y(\boldsymbol{s}_{it}, t)$ observed at location $\boldsymbol{s}_{it}$ at time t, where $i = 1, \ldots, n_t$ and $t = 1, \ldots, T$. Let $\mathbf{Y}_t = \{Y(\boldsymbol{s}_{it}, t), i = 1, \ldots, n_t\}$. To model the data, we introduce an $m_1 \times m_2$ auxiliary lattice, $W = \{(k, l) : k = 1, \ldots, m_1, l = 1, \ldots, m_2\}$, to cover the spatial region of interest and use the same grid points for all time points. We define a series of latent GMRFs, $\mathbf{U}_1, \ldots, \mathbf{U}_T$, on the auxiliary lattice where $\mathbf{U}_t = \{U_{kl,t}, (k, l) \in W\}$. Denote $\mathbf{Y}_{1:t} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$, $\mathbf{U}_{1:t} = \{\mathbf{U}_1, \ldots, \mathbf{U}_t\}$ and let $\boldsymbol{\theta}$ be the vector of all parameters. We assume that the structure on $(\mathbf{Y}_t, \mathbf{U}_t|\boldsymbol{\theta})$ is Markovian,

$$\mathbf{Y}_t|\mathbf{U}_{1:t}, \mathbf{Y}_{1:(t-1)}, \boldsymbol{\theta} \sim f(\cdot|\mathbf{U}_t, \boldsymbol{\theta}), \quad \mathbf{U}_t|\mathbf{U}_{1:(t-1)}, \mathbf{Y}_{1:(t-1)}, \boldsymbol{\theta} \sim g(\cdot|\mathbf{U}_{t-1}, \boldsymbol{\theta}),$$

where $f(\cdot)$ and $g(\cdot)$ are some density functions. One can use a GMRF to model the spatial dependence within each $\mathbf{U}_t$ and can model the temporal dependence between $\mathbf{U}_t$'s using a vector autoregressive model. Let $\vec{U}_t$ be the prolonged vector of $\mathbf{U}_t$ arranged as

$$\vec{U}_t = (U_{11,t}, U_{12,t}, \ldots, U_{1m_2,t}, U_{21,t}, \ldots, U_{2m_2,t}, \ldots, U_{m_1m_2,t})^{\mathrm{T}}. \tag{2.1}$$

Then $g(\cdot)$ can be modelled using the space-time autoregressive (STAR) model

$$\vec{U}_t = \boldsymbol{\Phi}(\boldsymbol{\beta}_t)\vec{U}_{t-1} + \vec{\boldsymbol{\zeta}}_t, \quad \vec{\boldsymbol{\zeta}}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\beta}_t)), \quad t = 1, \ldots, T, \tag{2.2}$$

where $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T$ is a sequence of vectors of parameters and, for each time point t, $\boldsymbol{\Phi}(\boldsymbol{\beta}_t)$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}_t)$ are $m_1m_2 \times m_1m_2$ matrices depending on $\boldsymbol{\beta}_t$. Here $\vec{\boldsymbol{\zeta}}_t$ is a spatial GMRF defined on the lattice W, and $\vec{\boldsymbol{\zeta}}_t$ and $\vec{\boldsymbol{\zeta}}_{t-1}$ are assumed to be independent. Although not required, we take $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_T = \boldsymbol{\beta}$ (Cressie, Shi, and Kang (2010); Katzfuss and Cressie (2012)) to ease the model identifiability. The complexity of the $\boldsymbol{\Phi}(\boldsymbol{\beta}_t)$'s and $\boldsymbol{\Sigma}(\boldsymbol{\beta}_t)$'s determines the complexity of the spatio-temporal dependence structure that the STAR model can handle and the computational cost of the MCMC algorithm. A simple model can assume that $\boldsymbol{\Phi}(\boldsymbol{\beta}_t) = \beta\mathbf{I}_{m_1m_2}$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}_t) = \sigma_q^2\mathbf{I}_{m_1m_2}$ for all t's, where β and σ_q^2 are some scalars and $\mathbf{I}_{m_1m_2}$ is the $m_1m_2 \times m_1m_2$ identity matrix. In fact, a much simpler spatio-temporal AR(1) model has been used in Finley, Banerjee, and Gelfand (2012). However, such an overly simple model may not be able to capture both the

spatial and temporal dependence, which is crucial for accurate predictions (Wikle, Berliner, and Cressie (1998)). On the other hand, non-diagonal choices of the $\boldsymbol{\Phi}(\boldsymbol{\beta}_t)$'s and $\boldsymbol{\Sigma}(\boldsymbol{\beta}_t)$'s can quickly lead to a significant increase of the computational expense. Issue with (2.2) is the trade-off between the richness of the $\boldsymbol{\Phi}(\boldsymbol{\beta}_t)$'s and $\boldsymbol{\Sigma}(\boldsymbol{\beta}_t)$'s and computational feasibility.

2.2. GMRF with block circulant precision matrices

For ease of presentation, we take $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_T = \boldsymbol{\beta}$. We model the spatial dependence of the elements within the $\vec{\boldsymbol{\zeta}}_t$'s by a first-order GMRF with a block circulant precision matrix, $\sigma_q^{-2}\boldsymbol{\Lambda}(\boldsymbol{\beta})$, where $\sigma_q > 0$ is a scaling factor such that all diagonal elements of $\boldsymbol{\Lambda}(\boldsymbol{\beta})$ are 1. Thus, $\boldsymbol{\Sigma}(\boldsymbol{\beta}_1) = \cdots = \boldsymbol{\Sigma}(\boldsymbol{\beta}_T) = \sigma_q^2\boldsymbol{\Lambda}(\boldsymbol{\beta})^{-1}$ and $\vec{\boldsymbol{\zeta}}_t \sim N(\mathbf{0}, \sigma_q^2\boldsymbol{\Lambda}(\boldsymbol{\beta})^{-1})$ for $t = 1, \ldots, T$. Suppose that (k_0, l_0, t_0) is a reference point and that the element in $\boldsymbol{\Lambda}(\boldsymbol{\beta})$ corresponding to the spatial interaction between locations (k_0, l_0) and (k, l) at the time point $t = t_0$ in the auxiliary lattice W is 0 unless $|k - k_0| \leq 1$ and $|l - l_0| \leq 1$, as illustrated in the right panel of (2.3). For example, β_{110} represents the strength of the interactions between $U_{k_0 l_0, t_0}$ and $\{U_{k_0-1 l_0-1, t_0}, U_{k_0-1 l_0+1, t_0}, U_{k_0+1 l_0-1, t_0}, U_{k_0+1 l_0+1, t_0}\}$. Temporal dependence is modelled through the transition matrices $\boldsymbol{\Phi}(\boldsymbol{\beta}_1) = \cdots = \boldsymbol{\Phi}(\boldsymbol{\beta}_T) = \boldsymbol{\Phi}(\boldsymbol{\beta})$ in (2.2), by assuming that, given all $\mathbf{U}_{1:(t_0-1)}$, $U_{k_0 l_0, t_0}$ depends only on the first-order neighboring grid points, U_{kl, t_0-1}, where $|k - k_0| \leq 1$ and $|l - l_0| \leq 1$, as illustrated in the left panel of (2.3). For example, β_{001} represents the temporal dependence strength between $U_{k_0 l_0, t_0}$ and $U_{k_0 l_0, t_0-1}$:

$$\underbrace{\left|\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_{111} & \beta_{101} & \beta_{111} & 0 \\ 0 & \beta_{011} & \beta_{001} & \beta_{011} & 0 \\ 0 & \beta_{111} & \beta_{101} & \beta_{111} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right|}_{t_0 \quad \text{v.s.} \quad t_0-1 \quad (\Phi(\boldsymbol{\beta}))}, \qquad \underbrace{\left|\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_{110} & \beta_{100} & \beta_{110} & 0 \\ 0 & \beta_{010} & 1 & \beta_{010} & 0 \\ 0 & \beta_{110} & \beta_{100} & \beta_{110} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right|}_{t_0 \quad \text{v.s.} \quad t_0 \quad (\Lambda(\boldsymbol{\beta}))}. \tag{2.3}$$

This parameterizations give $\boldsymbol{\beta} = (\beta_{100}, \beta_{010}, \beta_{110}, \beta_{001}, \beta_{011}, \beta_{011}, \beta_{111})^{\mathrm{T}}$, where the first three parameters control the spatial correlation within the GMRF $\mathbf{U}_{t_0}$ and the last four parameters model the temporal dependence between $\mathbf{U}_{t_0}$ and $\mathbf{U}_{t_0-1}$. For the $m \times m$ spatial adjacency matrix

$$\mathbf{S}_m = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & 1 & 0 & 1 & 0 \\ 0 & 0 & \vdots & \ddots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}_{m \times m},$$

it is straightforward to show that

$$\mathbf{\Phi}(\boldsymbol{\beta}) = \beta_{001}\mathbf{I}_{m_1m_2} + \beta_{011}\mathbf{I}_{m_1} \otimes \mathbf{S}_{m_2} + \beta_{101}\mathbf{S}_{m_1} \otimes \mathbf{I}_{m_2} + \beta_{111}\mathbf{S}_{m_1} \otimes \mathbf{S}_{m_2}, \quad (2.4)$$

$$\mathbf{\Lambda}(\boldsymbol{\beta}) = \mathbf{I}_{m_1m_2} + \beta_{010}\mathbf{I}_{m_1} \otimes \mathbf{S}_{m_2} + \beta_{100}\mathbf{S}_{m_1} \otimes \mathbf{I}_{m_2} + \beta_{110}\mathbf{S}_{m_1} \otimes \mathbf{S}_{m_2}, \quad (2.5)$$

where $\otimes$ is the Kronecker product. Since $\mathbf{\Phi}(\boldsymbol{\beta})$ and $\mathbf{\Lambda}(\boldsymbol{\beta})$ in (2.4) and (2.5) have more complicated structures, they can model more complex space-time dependence. It is straightforward to extend our results to a higher order GRMF, where $\Phi(\boldsymbol{\beta})$ and $\Lambda(\boldsymbol{\beta})$ can have more complicated structures. As long as they maintain circulant structures, the computational complexity is the same.

Remark 1. The form of $\mathbf{S}_m$, a spatial lattice wrapped on a torus, is adopted to reduce the computational cost because of the properties of circulant matrices. While somewhat artificial, this has been used in many geostatistical models (Rue and Tjelmeland (2002); Allcroft and Glasbey (2003); Rue and Held (2005)).

The joint distribution of $\mathbf{U}_1, \ldots, \mathbf{U}_T$ can be expressed in terms of the log-likelihood of $\vec{U} = (\vec{U}_1^{\mathrm{T}}, \ldots, \vec{U}_T^{\mathrm{T}})^{\mathrm{T}}$ as

$$\begin{aligned}
&\log f(\vec{U}|\boldsymbol{\beta}, \sigma_q^2) \\
&\quad = \sum_{t=2}^{\mathrm{T}} \log f(\vec{U}_t|\vec{U}_{t-1}, \boldsymbol{\beta}) + \log f(\vec{U}_1|\boldsymbol{\beta}) \\
&\quad \propto -\frac{1}{2\sigma_q^2} \sum_{t=2}^{\mathrm{T}} (\vec{U}_t - \mathbf{\Phi}(\boldsymbol{\beta})\vec{U}_{t-1})^{\mathrm{T}} \mathbf{\Lambda}(\boldsymbol{\beta})(\vec{U}_t - \mathbf{\Phi}(\boldsymbol{\beta})\vec{U}_{t-1}) - \frac{1}{2}\vec{U}_1^{\mathrm{T}}\mathbf{Q}(\boldsymbol{\beta})\vec{U}_1 \\
&\qquad -\frac{m_1m_2T}{2}\log\sigma_q^2 + \frac{T-1}{2}\log|\mathbf{\Lambda}(\boldsymbol{\beta})| + \frac{1}{2}\log|\mathbf{Q}(\boldsymbol{\beta})|, \qquad (2.6)
\end{aligned}$$

where $\vec{U}_1|\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_q^2\mathbf{Q}^{-1}(\boldsymbol{\beta}))$. Since we focus on the stationary model, we assume that (2.2) is a stationary vector autoregressive model, so all eigenvalues of $\mathbf{\Phi}(\boldsymbol{\beta})$ are between -1 and 1.

Proposition 1. *If the STAR model* (2.2) *is stationary and if* $\mathbf{\Phi}(\boldsymbol{\beta})$, $\mathbf{\Lambda}(\boldsymbol{\beta})$ *are block circulant matrices such that* $|\lambda_{kl}(\mathbf{\Phi}(\boldsymbol{\beta}))| < 1$ *and* $\lambda_{kl}(\mathbf{\Lambda}(\boldsymbol{\beta})) > 0$ *for all* $k = 1, \ldots, m_1, l = 1, \ldots, m_2$, *then the marginal distribution of* $\vec{U}_t$ *is* $N(\mathbf{0}, \sigma_q^2\mathbf{Q}^{-1}(\boldsymbol{\beta}))$ *with* $\mathbf{Q}(\boldsymbol{\beta}) = (\mathbf{I} - \mathbf{\Phi}^2(\boldsymbol{\beta}))\mathbf{\Lambda}(\boldsymbol{\beta})$.

By the properties of block circulant matrices, $\mathbf{Q}(\boldsymbol{\beta})$ is a block circulant matrix and, since both $\mathbf{\Phi}(\boldsymbol{\beta})$ and $\mathbf{\Lambda}(\boldsymbol{\beta})$ are sparse, $\mathbf{Q}(\boldsymbol{\beta})$ is sparse. Hence, the most computationally expensive parts in evaluating (2.6) are $\log|\mathbf{\Lambda}(\boldsymbol{\beta})|$ and $\log|\mathbf{Q}(\boldsymbol{\beta})|$, which usually require $O(m_1^3m_2^3)$ floating operations. For block circulant matrices $\mathbf{\Phi}(\boldsymbol{\beta})$ and $\mathbf{\Lambda}(\boldsymbol{\beta})$ as (2.4) and (2.5), using the fact that $\lambda_k(\mathbf{S}_m) =$

$2\cos\{2(k-1)\pi/m\}$ (Jain (1979)) and properties of the Kronecker product, we have

$$\lambda_{k,l}(\boldsymbol{\Phi}(\boldsymbol{\beta})) = \beta_{001} + 2\beta_{101}\cos\left\{\frac{2(k-1)\pi}{m_1}\right\} + 2\beta_{011}\cos\left\{\frac{2(l-1)\pi}{m_2}\right\} + 4\beta_{111}\cos\left\{\frac{2(k-1)\pi}{m_1}\right\}\cos\left\{\frac{2(l-1)\pi}{m_2}\right\}, \tag{2.7}$$

$$\lambda_{k,l}(\boldsymbol{\Lambda}(\boldsymbol{\beta})) = 1 + 2\beta_{100}\cos\left\{\frac{2(k-1)\pi}{m_1}\right\} + 2\beta_{010}\left\{\cos\frac{2(l-1)\pi}{m_2}\right\} + 4\beta_{110}\cos\left\{\frac{2(k-1)\pi}{m_1}\right\}\cos\left\{\frac{2(l-1)\pi}{m_2}\right\}, \tag{2.8}$$

for $k = 1, \ldots, m_1$, $l = 1, \ldots, m_2$. Plugging eigenvalues back into (2.6), we have

$$\begin{aligned}&\log f(\vec{U}|\boldsymbol{\beta}, \sigma_q^2)\\ &\propto -\frac{1}{2\sigma_q^2}\sum_{t=2}^{\mathrm{T}}(\vec{U}_t - \boldsymbol{\Phi}(\boldsymbol{\beta})\vec{U}_{t-1})^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{\beta})(\vec{U}_t - \boldsymbol{\Phi}(\boldsymbol{\beta})\vec{U}_{t-1}) - \frac{1}{2\sigma_q^2}\vec{U}_1^{\mathrm{T}}(I - \boldsymbol{\Phi}^2(\boldsymbol{\beta}))\boldsymbol{\Lambda}(\boldsymbol{\beta})\vec{U}_1\\ &\quad + \frac{T}{2}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2}\log\lambda_{k,l}(\boldsymbol{\Lambda}(\boldsymbol{\beta})) + \frac{1}{2}\log\left(1 - \lambda_{k,l}^2(\boldsymbol{\Phi}(\boldsymbol{\beta}))\right) - \frac{m_1 m_2 T}{2}\log\sigma_q^2,\end{aligned} \tag{2.9}$$

where, again, $\lambda_{k,l}(\boldsymbol{\Lambda}(\boldsymbol{\beta})) > 0$ and $|\lambda_{k,l}(\boldsymbol{\Phi}(\boldsymbol{\beta}))| < 1$.

2.3. Conditional distribution of $\boldsymbol{Z}_t$ given $\mathbf{U}_t$

Since for a time point t, $\{Z(\boldsymbol{s}_{it}, t)\}$ is irregularly spaced over the spatial domain, we need to make connections between $\{Z(\boldsymbol{s}_{it}, t)\}$ and the latent GMRF $\mathbf{U}_t$'s. To do so, we assume that the $Z(\boldsymbol{s}_{it}, t)$'s are mutually independent given the GMRF $\mathbf{U}_t$. We then have

$$f(Z(\boldsymbol{s}_{1t}, t), \ldots, Z(\boldsymbol{s}_{n_t t}, t)|\mathbf{U}_t) = \prod_{i=1}^{n_t} f(Z(\boldsymbol{s}_{it}, t)|\mathbf{U}_t).$$

The idea of inducing conditional independence in spatial processes by using a Markov field as a latent process is not new, see Hughes and Guttorp (1999) and Park and Liang (2012). Here, the conditional distribution, $f(Z(\boldsymbol{s}_{it}, t)|\mathbf{U}_t)$, is of particular importance. We model $f(Z(\boldsymbol{s}_{it}, t)|\mathbf{U}_t)$ by assuming that, for a given t, $\{Z(\boldsymbol{s}_{it}, t), i = 1, \ldots, n_t\}$ and $\mathbf{U}_t$ are generated from the same spatial process, utilizing the fact that a Gaussian random field can be approximated well by a GMRF with small neighborhoods (Rue and Held (2005); Lindgren, Rue, and Lindstrom (2011)). Since $\vec{U}_t \sim N(\mathbf{0}, \sigma_q^2\mathbf{Q}^{-1}(\boldsymbol{\beta}))$, we assume that $Z(\boldsymbol{s}_{it}, t)$ and $\vec{U}_t$ are jointly normally distributed with mean zero and covariance matrix

$$\boldsymbol{\Omega}_{it} = \sigma_q^2\begin{bmatrix} c(\boldsymbol{\beta}) & c(\boldsymbol{\beta})\boldsymbol{r}_{it}^{\mathrm{T}} \\ c(\boldsymbol{\beta})\boldsymbol{r}_{it} & \mathbf{Q}^{-1}(\boldsymbol{\beta}) \end{bmatrix}, \tag{2.10}$$

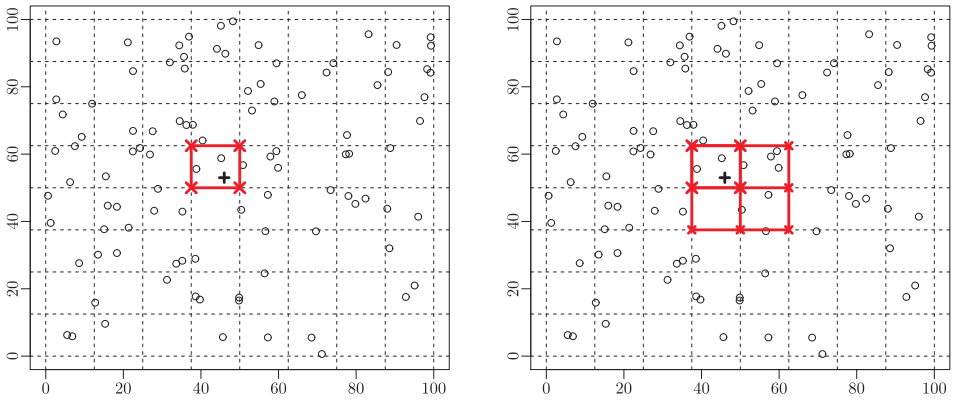

Figure 1. A 2×2 neighborhood (left) and a 3×3 neighborhood (right). The '+' is the target location and the set of '×'s are its neighboring points used in the auxiliary lattice.

where $c(\boldsymbol{\beta}) = (1/m_1 m_2) \sum_{kl} [\lambda_{kl}(\boldsymbol{\Lambda}(\boldsymbol{\beta}))\{1 - \lambda_{kl}^2(\boldsymbol{\Phi}(\boldsymbol{\beta}))\}]^{-1}$ and $\boldsymbol{r}_{it} = \mathrm{Corr}(Z(\boldsymbol{s}_{it}, t), \vec{U}_t)$ for $i = 1, \ldots, n_t$. Here, $c(\boldsymbol{\beta})$ is introduced to match the variance of $Z(\boldsymbol{s}_{it}, t)$ and that of the GMRF $\vec{U}_t$. Following the argument in Park and Liang (2012), we can show that if $1 - c(\boldsymbol{\beta}) \boldsymbol{r}_{it}^{\mathrm{T}} \mathbf{Q}(\boldsymbol{\beta}) \boldsymbol{r}_{it} > 0$, then $\boldsymbol{\Omega}_{it}$ is positive definite. By the property of the multivariate Gaussian distribution, we have

$$Z(\boldsymbol{s}_{it}, t) | \mathbf{U}_t \sim N \left[c(\boldsymbol{\beta}) \boldsymbol{r}_{it}^{\mathrm{T}} \mathbf{Q}(\boldsymbol{\beta}) \vec{U}_t, \sigma_q^2 c(\boldsymbol{\beta}) \{1 - c(\boldsymbol{\beta}) \boldsymbol{r}_{it}^{\mathrm{T}} \mathbf{Q}(\boldsymbol{\beta}) \boldsymbol{r}_{it}\} \right]. \tag{2.11}$$

The computational costs of evaluating $c(\boldsymbol{\beta}) \boldsymbol{r}_{it}^{\mathrm{T}} \mathbf{Q}(\boldsymbol{\beta}) \vec{U}_t$ is $O(m_1^2 m_2^2)$, which can be quite expensive if the grid size is large, and the constraint $1 - c(\boldsymbol{\beta}) \boldsymbol{r}_{it}^{\mathrm{T}} \mathbf{Q}(\boldsymbol{\beta}) \boldsymbol{r}_{it} > 0$ is restrictive, especially when the n_t's, m_1, and m_2 are large. The mean $c(\boldsymbol{\beta}) \boldsymbol{r}_{it}^{\mathrm{T}} \mathbf{Q}(\boldsymbol{\beta}) \vec{U}_t$ is the simple kriging prediction at $\boldsymbol{s}_{it}$ based on $\mathbf{U}_t$ and generally most of the kriging coefficients are close to 0, known as the "screen" effect (Stein (1999)). This motivates us to assume that, conditioned on $\mathbf{U}_t$, $Z(\boldsymbol{s}_{it}, t)$ depends only on a fixed subset of $\mathbf{U}_t$ in the neighborhoods of the location $\boldsymbol{s}_{it}$ (denoted by $\partial \boldsymbol{s}_{it}$), say $\mathbf{U}_{\partial \boldsymbol{s}_{it}, t}$. In this paper, we consider the 2×2 and 3×3 neighborhood structures, as illustrated in Figure 1.

As in (2.10), we assume that the joint distribution of $Z(\boldsymbol{s}_{it}, t)$ and $\mathbf{U}_{\partial \boldsymbol{s}_{it}, t}$ from a $m \times m$ neighborhood is Gaussian with mean zero and covariance matrix

$$\boldsymbol{\Omega}_{\partial \boldsymbol{s}_{it}, t} = \sigma_q^2 \begin{bmatrix} c(\boldsymbol{\beta}) & c(\boldsymbol{\beta}) \boldsymbol{r}_{\partial \boldsymbol{s}_{it}}^{\mathrm{T}} \\ c(\boldsymbol{\beta}) \boldsymbol{r}_{\partial \boldsymbol{s}_{it}} & \boldsymbol{\Sigma}_{\partial}(\boldsymbol{\beta}) \end{bmatrix},$$

where $c(\boldsymbol{\beta})$ is defined as in (2.10), $\boldsymbol{r}_{\partial \boldsymbol{s}_{it}} = \mathrm{Corr}(Z(\boldsymbol{s}_{it}, t), \vec{U}_{\partial \boldsymbol{s}_{it}, t})$, and $\boldsymbol{\Sigma}_{\partial}(\boldsymbol{\beta})$ is a $m^2 \times m^2$ submatrix of $\mathbf{Q}^{-1}(\boldsymbol{\beta})$ corresponding to the covariance matrix of $\mathbf{U}_{\partial \boldsymbol{s}_{it}, t}$. In general, we cannot obtain $\boldsymbol{\Sigma}_{\partial}(\boldsymbol{\beta})$ without actually inverting $\mathbf{Q}(\boldsymbol{\beta})$, but

because $\mathbf{\Phi}(\boldsymbol{\beta})$ and $\mathbf{\Lambda}(\boldsymbol{\beta})$ are block circulant matrices, $\mathbf{Q}(\boldsymbol{\beta})$ and $\mathbf{Q}^{-1}(\boldsymbol{\beta})$ are as well. This enables us to use the two-dimensional inverse Fourier transform of the eigenvalues of $\mathbf{Q}^{-1}(\boldsymbol{\beta})$ to obtain each element in $\mathbf{Q}^{-1}(\boldsymbol{\beta})$; see, e.g., Rue and Held (2005, Chap. 2.6). Suppose that $a_{i_1j_1,i_2j_2}$ is the entry in $\mathbf{Q}^{-1}(\boldsymbol{\beta})$ corresponding to the interaction between $U_{i_1j_1,t}$ and $U_{i_2j_2,t}$. Then, we have

$$a_{i_1j_1,i_2j_2} = \frac{1}{m_1m_2}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2}\frac{\cos\{2|i_1-i_2|(k-1)\pi/m_1 + (2|j_1-j_2|(l-1)\pi)/m_2\}}{\{1-\lambda_{kl}^2(\mathbf{\Phi}(\boldsymbol{\beta}))\}\lambda_{kl}(\mathbf{\Lambda}(\boldsymbol{\beta}))},$$

where $\lambda_{kl}(\mathbf{\Phi}(\boldsymbol{\beta}))$ and $\lambda_{kl}(\mathbf{\Lambda}(\boldsymbol{\beta}))$ are as defined in (2.7) and (2.8), $k, i_1, i_2 = 1,\ldots,m_1$, $l, j_1, j_2 = 1,\ldots,m_2$. For an $m \times m$ neighborhood, it suffices to calculate $a_{i_1j_1,i_2j_2}$ for $i_1, i_2, j_1, j_2 = 1,\ldots,m$ to obtain $\mathbf{\Sigma}_\partial(\boldsymbol{\beta})$, which requires only $O(m_1m_2)$ floating operations when m is fixed to a small number (2 or 3). Then, it follows that

$$Z(\boldsymbol{s}_{it},t)|\mathbf{U}_t \sim N\left[c(\boldsymbol{\beta})\boldsymbol{r}_{\partial\boldsymbol{s}_{it}}^{\mathrm{T}}\mathbf{\Sigma}_\partial^{-1}(\boldsymbol{\beta})\vec{U}_{\partial\boldsymbol{s}_{it},t}, \sigma_q^2 c(\boldsymbol{\beta})\{1-c(\boldsymbol{\beta})\boldsymbol{r}_{\partial\boldsymbol{s}_{it}}^{\mathrm{T}}\mathbf{\Sigma}_\partial^{-1}(\boldsymbol{\beta})\boldsymbol{r}_{\partial\boldsymbol{s}_{it}}\}\right], \tag{2.12}$$

subject to the constraint

$$1 - c(\boldsymbol{\beta})\boldsymbol{r}_{\partial\boldsymbol{s}_{it}}^{\mathrm{T}}\mathbf{\Sigma}_\partial^{-1}(\boldsymbol{\beta})\boldsymbol{r}_{\partial\boldsymbol{s}_{it}} > 0. \tag{2.13}$$

Following Park and Liang (2012), the spherical correlation function is used to model $\boldsymbol{r}_{\partial\boldsymbol{s}_{it}}$, with

$$r_{i,klt} = \mathrm{Corr}(Z(\boldsymbol{s}_{it},t), U_{klt}) = 1 - \frac{3}{2}\frac{h_{i(klt)}}{\phi} + \frac{1}{2}\left(\frac{h_{i(klt)}}{\phi}\right)^3, \tag{2.14}$$

if $0 \le h_{i(klt)} \le \phi$ and 0 otherwise, where $h_{i(klt)}$ is the spatial distance between the sites of $Z(\boldsymbol{s}_{it},t)$ and $U_{kl,t}$. Although the primary reason for this choice is because of its compact support, our empirical experience indicates that this choice provides good spatio-temporal predictions even when the underlying spatial correlation is not spherical (see Section 3.2.2).

2.4. Conditional distribution of $\mathbf{Y}_t$ given $\mathbf{U}_t$

Denote by $\boldsymbol{\theta}$ the parameter vector consisting of σ_e^2, the regression coefficients $\boldsymbol{\xi} = (\xi_0,\ldots,\xi_p)^{\mathrm{T}}$ as defined in (1.2), and the interaction parameters $\boldsymbol{\beta}$ of the auxiliary GMRF. The likelihood function is $f(\mathbf{Y}|\mathbf{U},\boldsymbol{\theta}) = \prod_{t=1}^{\mathrm{T}} f(\mathbf{Y}_t|\mathbf{U}_t,\boldsymbol{\theta})$ with

$$\begin{aligned} f(\mathbf{Y}_t|\mathbf{U}_t,\boldsymbol{\theta}) &= \int f(\mathbf{Y}_t|\mathbf{Z}_t,\boldsymbol{\theta}) f(\mathbf{Z}_t|\mathbf{U}_t,\boldsymbol{\theta})\, d\mathbf{Z}_t \\ &= \prod_{i=1}^{n_t}\int f(Y(\boldsymbol{s}_{it},t)|Z(\boldsymbol{s}_{it},t),\boldsymbol{\theta}) f(Z(\boldsymbol{s}_{it},t)|\mathbf{U}_t,\boldsymbol{\theta})\, dZ(\boldsymbol{s}_{it},t) \\ &= \prod_{i=1}^{n_t}\frac{1}{\sqrt{2\pi(\sigma_e^2+\sigma_{it}^2)}}\exp\left\{-\frac{(Y(\boldsymbol{s}_{it},t)-\mu(\boldsymbol{s}_{it},t)-\nu_{it})^2}{2(\sigma_e^2+\sigma_{it}^2)}\right\}, \end{aligned} \tag{2.15}$$

where $\nu_{it} = c(\boldsymbol{\beta})\boldsymbol{r}_{it}^{\mathrm{T}}\boldsymbol{\Sigma}_{\partial}^{-1}(\boldsymbol{\beta})\vec{U}_{\partial \boldsymbol{s}_{it},t}$, $\sigma_{it}^2 = \sigma_q^2 c(\boldsymbol{\beta})\{1 - c(\boldsymbol{\beta})\boldsymbol{r}_{\partial \boldsymbol{s}_{it}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\partial}^{-1}(\boldsymbol{\beta})\boldsymbol{r}_{\partial \boldsymbol{s}_{it}}\}$ and $\vec{U}_{\partial \boldsymbol{s}_{it},t}$ is a stacked version of $\mathbf{U}_{\partial \boldsymbol{s}_{it},t}$ arranged in the same way as in (2.1). Thus

$$Y(\boldsymbol{s}_i, t) = \mu(\boldsymbol{s}_i, t) + c(\boldsymbol{\beta})\boldsymbol{r}_{it}^{\mathrm{T}}\boldsymbol{\Sigma}_{\partial}^{-1}(\boldsymbol{\beta})\vec{U}_{\partial \boldsymbol{s}_{it},t} + \varepsilon_{it}, \tag{2.16}$$

where $\varepsilon_{it} \sim N(0, \sigma_e^2 + \sigma_{it}^2)$, $i = 1, \ldots, n_t, t = 1, \ldots, T$.

2.5. The Markov chain Monte Carlo algorithm

In this subsection, we describe a Markov chain Monte Carlo algorithm for the proposed STAR model (2.2). For the regression parameters $\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_p)^{\mathrm{T}}$, we use the non-informative prior $\pi(\boldsymbol{\xi}) \propto 1$. It is generally reasonable to believe that $\sigma_e^2 \le \sigma_q^2$ and $\sigma_q^2 + \sigma_e^2 \le \sigma_s^2$, where σ_s^2 is the sample variance of the observed data (Park and Liang (2012)). We use the priors for

$$\pi(\sigma_q^2) \propto \frac{1}{\sigma_q^2} I(L_1 \le \sigma_q^2 \le U_1), \quad \pi(\sigma_e^2|\sigma_q^2) \propto \frac{1}{\sigma_e^2} I(L_2 \le \sigma_e^2 \le \sigma_q^2),$$

where $L_1 = 0.01\sigma_s^2$, $U_1 = 2\sigma_s^2$, and $L_2 = 0.001\sigma_s^2$. The joint prior of $(\boldsymbol{\beta}, \phi)$ is

$$\begin{aligned}\pi(\boldsymbol{\beta}, \phi) \propto & \prod_{k=1}^{m_1}\prod_{l=1}^{m_2} I(\lambda_{kl}(\boldsymbol{\Lambda}(\boldsymbol{\beta})) > 0) I(|\lambda_{kl}(\boldsymbol{\Phi}(\boldsymbol{\beta}))| < 1) \\ & \times \prod_{t=1}^{\mathrm{T}}\prod_{i=1}^{n_t} I(1 - c(\boldsymbol{\beta})\boldsymbol{r}_{\partial \boldsymbol{s}_{it}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\partial}^{-1}(\boldsymbol{\beta})\boldsymbol{r}_{\partial \boldsymbol{s}_{it}} > 0) I(\phi > 0).\end{aligned}$$

Remark 2. Although in principle (2.13) is required for all observed locations, $\boldsymbol{s}_{it}$'s, this is too restrictive and tends to result in under-estimation of the range parameter ϕ. In our examples, we relaxed the constraint by forcing 95% of the observed locations to meet (2.13) while making sure that $\sigma_{it}^2 + \sigma_e^2 > 0$ for all $i = 1, \ldots, n$, where σ_{it}^2 is defined as (2.16). This does not cause numerical problems because $Z(\boldsymbol{s}, t)$ is integrated out in (2.16) in the MCMC algorithm. Numerical studies indicate that this is a useful strategy to achieve better prediction results.

With the above priors, the posterior of our model is

$$f(\sigma_q^2, \sigma_e^2, \phi, \boldsymbol{\beta}, \boldsymbol{\xi}, \mathbf{U}|\mathbf{Y}) \propto \pi(\boldsymbol{\xi})\pi(\sigma_q^2)\pi(\sigma_e^2|\sigma_q^2)\pi(\boldsymbol{\beta}, \phi)\prod_{t=1}^{\mathrm{T}} f(\mathbf{U}_t|\sigma_q^2, \boldsymbol{\beta}) f(\mathbf{Y}_t|\mathbf{U}_t, \boldsymbol{\theta}), \tag{2.17}$$

with $f(\mathbf{U}_t|\sigma_q^2, \boldsymbol{\beta})$ and $f(\mathbf{Y}_t|\mathbf{U}_t, \boldsymbol{\theta})$ given in (2.9) and (2.15). It is easy to show that the joint posterior (2.17) is proper. For a new location, $\boldsymbol{s}_{pt}$, at time t, $Y(\boldsymbol{s}_{pt}, t)$ can be predicted by

$$E\{Y(\boldsymbol{s}_{pt}, t)|\mathbf{Y}\} = \int\int E\{Y(\boldsymbol{s}_{pt}, t)|\mathbf{Y}, \boldsymbol{\theta}, \mathbf{U}_t\} f(\boldsymbol{\theta}, \mathbf{U}_t|\mathbf{Y})\, d\boldsymbol{\theta}\, d\mathbf{U}_t, \tag{2.18}$$

where $E\{Y(\boldsymbol{s}_{pt},t)|\mathbf{Y},\boldsymbol{\theta},\mathbf{U}_t\} = \mu(\boldsymbol{s}_{pt},t) + c(\boldsymbol{\beta})\boldsymbol{r}^{\mathrm{T}}_{\partial \boldsymbol{s}_{pt},t}\boldsymbol{\Sigma}^{-1}_{\partial \boldsymbol{s}_{pt},t}(\boldsymbol{\beta})\vec{U}_{\partial \boldsymbol{s}_{pt},t}$, with $\boldsymbol{r}_{\partial \boldsymbol{s}_{pt},t}$ $= \mathrm{Corr}(Z(\boldsymbol{s}_{pt},t),\vec{U}_{\partial \boldsymbol{s}_{pt},t})$. Hence, (2.18) can be numerically estimated using MCMC samples. A detailed MCMC algorithm is given in the supplementary document.

2.6. Choice of the grid size

Each iteration of the MCMC algorithm costs only $O(m_1m_2T)+O(n)$ floating operations with $n = \sum_{i=1}^{\mathrm{T}} n_t$, where the first term is due to the imputation of the GMRFs $\mathbf{U}_1,\ldots,\mathbf{U}_T$ and the second term is the cost of the likelihood evaluation when drawing samples of $\boldsymbol{\theta}$. Computational cost of the proposed algorithm is thus determined by the grid size and the sample size. Our simulation studies and data example indicate that the predictive performance of the proposed method generally improves as the grid size increases. However, after the grid size reaches a certain level, the benefits of using more grid points gradually vanish. In the purely spatial case, Park and Liang (2012) suggested that using $m_1m_2 = n$ is sufficient for most applications. Our limited numerical experience suggests that taking $m_1m_2 = (1/T)\sum_{i=1}^{\mathrm{T}} n_t$ yields sufficiently good prediction. With this choice of m_1 and m_2, the computational complexity of our method is of order $O(n)$.

As pointed out by a referee, when modelling irregularly spaced data the resolution of the auxiliary lattice is highly dependent on the spatial pattern of the data. In many cases, a much larger grid size than the sample size may be necessary to capture important spatial correlations. In addition, the $O(n)$ computational cost is for each MCMC step using the Gibbs sampler. When the dimension of the $\mathbf{U}_t$'s increases, the mixing property of the Gibbs sampler may be questionable and thus more steps may be needed to collect well-behaved posterior MCMC samples. Therefore, the aforementioned $O(n)$ complexity might be a little overly optimistic. Alternatively, the block Gibbs sampler can be used to achieve better mixing properties. How to balance the block size of the Gibbs sampler and the computation cost is an interesting research question.

3. Simulation Studies

3.1. Model estimation

To show that the MCMC algorithm in Subsection 2.5 can correctly estimate the parameters in the proposed model, we simulated data as follows. First, we set $\boldsymbol{\beta} = (\beta_{010},\beta_{100},\beta_{110},\beta_{101},\beta_{011},\beta_{101},\beta_{111}) = (-0.2,-0.2,0.0,0.4,0.1,0.1,0.0)$ and $(\sigma_q^2,\sigma_e^2) = (3,1)$. For ease of presentation, we put no $\boldsymbol{\xi}$ term in the model. With $T = 5$, we used (2.2) to simulate auxiliary GMRFs $\mathbf{U}_1,\ldots,\mathbf{U}_T$ on a 32×32 lattice that covers the region $[0,100] \times [0,100]$. For each time point

t, $n_t = 1,000$ observation sites, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{1000}$, were randomly drawn on the region $[0, 100] \times [0, 100]$ such that (2.13) held. We considered 2×2 and 3×3 neighborhoods. We generated $Z(\boldsymbol{s}_{it}, t)$ and $Y(\boldsymbol{s}_{it}, t)$ for $i = 1, \ldots,$ 1,000 and $t = 1, \ldots, 5$ according to (2.16) and (2.12). For each dataset, the Metropolis-within-Gibbs sampler of Subsection 2.5 was run for $30,000$ iterations with the first 10,000 iterations discarded for the burn-in process. Then, 1,000 samples were collected from the remaining 20,000 iterations at equally spaced time points. The estimation results are summarized in Table S1 in the supplementary document, where SE stands for the standard error. The numerical results indicate that the sampling scheme described in Subsection 2.5 can correctly estimate the parameters from the proposed model, with either a 2×2 or a 3×3 neighborhood structures.

3.2. Approximation to spatio-temporal Gaussian random fields

In this subsection, we simulated the data from a spatio-temporal Gaussian random field with mean 0 and covariance function

$$\gamma((\boldsymbol{s}_1, t_1), (\boldsymbol{s}_2, t_2)) = \sigma_0^2 \exp(-|t_1 - t_2|/\tau)\rho_{sp}(||\boldsymbol{s}_1 - \boldsymbol{s}_2||/\phi) + \sigma_e^2 I(t_1 = t_2, \boldsymbol{s}_1 = \boldsymbol{s}_2), \tag{3.1}$$

where τ determines the strength of the temporal dependence and ϕ is the range parameter that determines the length of the spatial correlation. In the simulation studies, we take $\tau = 2$ and $\tau = 5$, where the former represents moderate temporal dependence and the later represents strong temporal dependence. We took $\phi = 20$ and $\phi = 40$ to show the prediction accuracy of the proposed method for a spatio-temporal Gaussian random field with short and long spatial correlation lengths. Other parameters were $\sigma_0^2 = 7$ and $\sigma_e^2 = 1$. For each scenario, 20 independent datasets of size $1,000 \times 5$ using the spatio-temporal covariance function (3.1) were simulated. The function GaussRF() in the R package RandomFields (Schlather and Menck (2013)) was used to generate the data. For each dataset, 1,000 locations, $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{1000}\}$, were uniformly drawn from the region $[0, 100] \times [0, 100]$ and the spatial locations were the same for each time point. Then, at each time point t, $n_t = 800$ samples were randomly drawn from the 1,000 samples as the training dataset and the remaining 200 samples were used for the prediction. We therefore had 800×5 samples for the model estimation and 200×5 samples for the prediction. We designed this simulation scheme to mimic the situation when some observations are missing from a set of irregularly located stations over time. To study the effect of the grid size on the prediction performance, four grid sizes were used for comparisons.

3.2.1 Kriging with correctly specified correlation functions

As (2.14) relates an observation from any given location to the GMRF defined on the auxiliary lattice, we supposed that the underlying covariance function was

spherical. We simulated $1,000 \times 5$ observations from the model (3.1) with $\rho_{sp}(\cdot)$ as in (2.14). Because observations were missed randomly at different time points, we used two approaches to evaluate the prediction performance.

We treated the random field at different time points as if they were independent of each other. For a given time point t, we used the maximum likelihood approach to estimate $\boldsymbol{\theta}$ and then plugged in the estimated parameters to conduct simple kriging in a purely spatial manner. Predictions from different time points were then collected together to calculate the mean square prediction error (MSPE), denoted as MSPE(SP), where "SP" stands for "spatial kriging".

Our second approach was to plug the true values of $\boldsymbol{\theta}$ back into (3.1) and then use the estimated (3.1) to conduct spatio-temporal kriging for all prediction locations at different time points. The resulting MSPE is denoted as MSPE(SPT), where "SPT" stands for "spatio-temporal kriging", which can be viewed as a surrogate of the minimal possible prediction error. The function Kriging() from the R package RandomFields (Schlather and Menck (2013)) was used to conduct spatio-temporal kriging.

For the proposed STAR model, we used a lattice of size 32×32 to conduct spatio-temporal kriging and denoted the corresponding mean square prediction error as MSPE. For each dataset, we ran $30,000$ MCMC iterations and discarded the first $10,000$ iterations as the burn-in period. Then, 1,000 samples were collected from the remaining $20,000$ iterations at equally-spaced time points. The results are summarized in Table S2 in the supplementary document. It can be seen that, in this case, by taking into account the temporal dependence, our method always outperforms the purely spatial kriging method and the difference between MSPE and MSPE(SP) grows as the strength of the temporal dependence increases. We see that the MSPE and MSPE(SPT) are reasonably close, which implies that the proposed method provides good spatio-temporal predictions. And, as the grid size increases, the MSPE becomes smaller in most cases. The smallest grid size here, $24 \times 24 = 576$, is much smaller than the averaged sample size 800 and does not do a sufficiently good job.

3.2.2 Kriging with mis-specified correlation functions

We wanted to evaluate the predictive performance of the proposed STAR model when the underlying spatio-temporal Gaussian random field does not have a spherical spatial correlation function. We simulated $1,000 \times 5$ observations using model (3.1) with $\rho_{sp}(\cdot)$ from the Matérn family

$$\rho_{sp}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(4\frac{||\boldsymbol{s}_1 - \boldsymbol{s}_2||}{\phi}\right)^{\nu} K_{\nu}\left(4\frac{||\boldsymbol{s}_1 - \boldsymbol{s}_2||}{\phi}\right),$$

where $\nu = 1$, $\Gamma(\cdot)$ is the gamma function and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind.

For the proposed STAR model, we used a lattice of size 32×32 to conduct spatio-temporal kriging and denoted the corresponding mean square prediction error as MSPE. For each data set, we ran $30,000$ MCMC iterations and discarded the first $10,000$ iterations as the burn-in process. Then, 1,000 samples were collected from the remaining $20,000$ iterations at equally-spaced time points. The results are summarized in Table S3 in the supplementary document. MSPE(SP) was obtained by plugging in the maximum likelihood estimator based on the assumption that the underlying correlation function is spherical. MSPE(SPT) was obtained by spatio-temporal kriging using the true spatio-temporal covariance function. By comparing MSPE(SP) and MSPE, we see that our method yielded better prediction results, especially in the case of $\phi = 20$. This indicates that our approach is more robust to the choice of the correlation functions than is the likelihood approach, at least in terms of the predictive performance. Perhaps, since most of the spatial dependence has been taken into account by the auxiliary GMRF, the correlation function plays a less important role in our method. Again, MSPE(SPT) and MSPE are close, so even though the correlation function is mis-specified, our method can still yield good prediction results. As the grid size increases, the MSPE decreases in most cases, and the smallest grid sizes, $24 \times 24 = 576$, seems not sufficient.

3.3. Computational complexity analysis

The computational complexity of the proposed method is of the order $O(n)$. We conducted a small simulation study to numerically demonstrate this point. The data were generated in the same way as in Subsection 3.2.2 with sample sizes 400×5, 625×5, 900×5, 1225×5, 1600×5, and 2025×5. The corresponding lattice sizes were chosen to match the sample size at a specific time point. Thus, with $n = 400 \times 5$, the lattice size was 20×20. For each case, the CPU times of running $5,000$ MCMC iterations were recorded and the average CPU times over 10 repetitions was calculated. We fit a linear function to the CPU times used by a 2×2 neighborhood structure, $\text{CPU}(n) = 11.52 + 0.039n$ and by a 3×3 neighborhood structure, $\text{CPU}(n) = -23.29 + 0.062n$. The fitted functions are plotted in Figure 2, which shows a linear pattern between the CPU times and the sample sizes. As one would expect, for the 3×3 neighborhood structure, the computation cost increases at a faster rate as the sample size grows.

4. Precipitation Data

To illustrate the effectiveness of our method, we used annual total precipitation data. The longitudes of the stations ranged from -124.6 to -67.0 and the latitudes ranged from 24.55 to 49.00. Besides the longitude and latitude, there was also elevation (in meters) information available for each location. We made

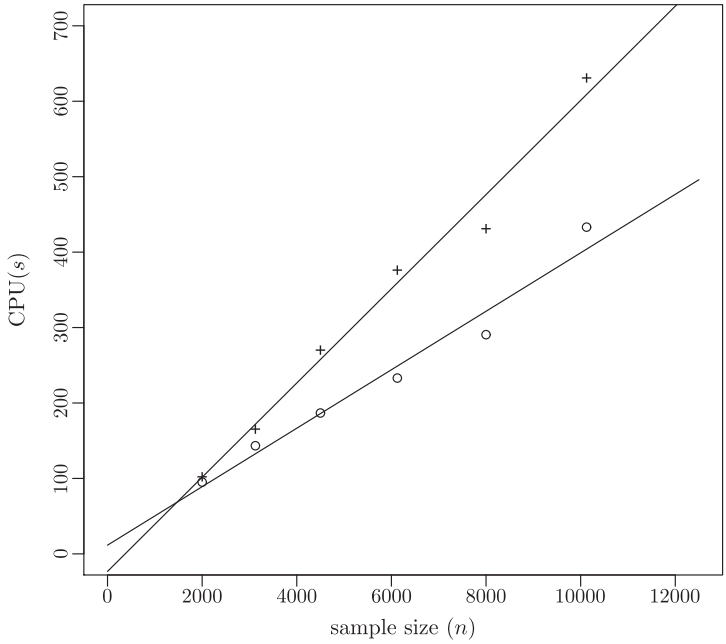

Figure 2. Computational complexity (CPU time measured in seconds on a 2.80Ghz Intel Xeon X5560 computer) of the proposed method with 2×2 neighborhood ("o") and a 3×3 neighborhood ("+") as a function of sample size, n.

a square root transform of the original data to make them more normal. Let $Y(\boldsymbol{s}, t)$ be the square root of the annual precipitation for location $\boldsymbol{s}$ in year t. We used model (1.1) to analyze the observed process, where an intercept term, ξ_0, and the elevation, $elev(\boldsymbol{s}, t)$ (calculated as the elevation (in meters) divided by 100), were included in the mean structure $\mu(\boldsymbol{s}, t)$. Since the region of interest can be roughly considered as flat, we used a grid of size 115×55 to cover the region $[-125, -65] \times [20, 50]$; for comparison, we took the grid sizes 100×50, 120×60 and 125×65. The ratio of the grid numbers in longitude and latitude roughly close to $2 : 1$ so that the grid points are evenly spaced in the spatial region. For prediction performance of the proposed method, we created 20 datasets by randomly taking 90% of the available data as the training dataset and the rest as prediction locations. For each dataset, we ran $23,000$ MCMC iterations and took 1,000 samples from the last $20,000$ iterations at equally-spaced time points to do the estimation and prediction. The estimation results are summarized in Table 1. As the maximum likelihood approach brings prohibitive computation cost, we compared the purely spatial prediction accuracy using our method (by assuming that $T = 1$ and $\boldsymbol{\Phi}(\boldsymbol{\beta}) = 0$ in (2.2)) and the spatio-temporal prediction accuracy using the proposed STAR model. The ability of using the auxiliary

GMRF to approximate the Gaussian random field in a purely spatial scenario was discussed and illustrated by Park and Liang (2012). The results are summarized in Table 1. The results indicate that, by taking into account the temporal dependence, the STAR model spatio-temporal kriging yields better prediction results with either a 2×2 or a 3×3 neighborhood structure. Increasing the grid size will give better prediction performance.

We also used the univariate Bayesian dynamic space-time regression models proposed in Finley, Banerjee, and Gelfand (2012) to do spatio-temporal kriging on the same dataset. The implementation was carried out using the function spDynLM() in the R package spBayes (Finley and Banerjee (2013)). The variable "elevation" was used as the only predictor in the regression model. Predictions were made based on $5,000$ MCMC samples with a burn in period of $5,000$ iterations. Knots were chosen as the grid points on 10×5, 14×7, and 20×10 lattices, which give the MSPE for the hold out data as 0.67, 0.53 and 0.42, respectively. As expected, the MSPE decreases as the number of knots increases. Using 200 knots gives roughly the prediction accuracy of our method, but its computation time for each iteration is almost 7 times as much as ours using a 115×55 grid size and a 2×2 neighborhood.

In Figure S1 in the supplementary document, we present the spatio-temporal prediction of the annual total precipitation in 1982 for all $11,918$ stations. There are 6,595 stations that have records in 1982. The locations of these stations are plotted as green dots in Figure S1(a). The red dots in Figure S1(a) represent stations that do not have records in 1982. An image of the observed precipitation data is shown in Figure S1(b). Figures S1(c) and (d) present images of the predicted precipitation for all $11,918$ stations in 1982 using the STAR model spatio-temporal kriging with 2×2 and 3×3 neighborhood structures, respectively. The observations from the $6,595$ stations with data from year 1982 are also used in the model estimation step. The similarity between Figure S1(b) and Figures S1(c) and (d) indicates that our method yields good prediction performance. Figures S1(e) and (f) give the prediction standard errors calculated using MCMC samples for all $11,918$ locations.

5. Discussion

We have proposed a computationally efficient Bayesian hierarchical model for large spatio-temporal data. Our approach completely avoids matrix inversion in MCMC sampling and its computational cost increases only linearly with the sample size. Numerical examples show that, by using the STAR model (2.2), the spatio-temporal prediction is more accurate than the purely spatial prediction. Our method can be extended to more general treatments of spatio-temporal processes in different ways. We can make the coefficients, $\boldsymbol{\xi}$, in (1.2) dependent

Table 1. The mean of the estimated parameters averaged over 20 datasets drawn from the precipitation data. The numbers in the parentheses are the standard errors of the estimates. The size of the auxiliary lattice is 115×55.

Neighborhoods	2×2		3×3	
	SPTKrig	SPKrig	SPTKrig	SPKrig
$\boldsymbol{\beta}_{010}$	-0.08(0.05)	-0.08(0.05)	-0.10(0.06)	-0.10(0.05)
$\boldsymbol{\beta}_{100}$	-0.12(0.04)	-0.12(0.04)	-0.15(0.05)	-0.14(0.06)
$\boldsymbol{\beta}_{110}$	-0.15(0.03)	-0.15(0.03)	-0.11(0.04)	-0.12(0.04)
$\boldsymbol{\beta}_{001}$	0.69(0.06)	0	0.55(0.18)	0
$\boldsymbol{\beta}_{011}$	0.05(0.03)	0	0.05(0.06)	0
$\boldsymbol{\beta}_{101}$	0.05(0.03)	0	0.07(0.04)	0
$\boldsymbol{\beta}_{111}$	0.01(0.02)	0	0.03(0.03)	0
ξ_0	7.21(0.34)	7.36(0.42)	6.95(0.20)	7.70(0.28)
$elev$	0.30(0.00)	0.29(0.01)	0.28(0.03)	0.18(0.04)
ϕ	9.90(2.22)	2.94(0.33)	4.43(1.87)	1.61(0.21)
σ_q^2	0.37(0.04)	0.97(0.15)	0.55(0.28)	1.35(0.21)
σ_e^2	0.19(0.04)	0.22(0.03)	0.15(0.06)	0.28(0.06)
MSPE(100×50)	0.44(0.03)	0.53(0.02)	0.47(0.05)	0.64(0.06)
MSPE(115×55)	0.43(0.02)	0.52(0.02)	0.46(0.07)	0.64(0.05)
MSPE(120×60)	0.40(0.01)	0.49(0.02)	0.43(0.05)	0.63(0.06)
MSPE(125×65)	0.40(0.01)	0.49(0.01)	0.41(0.02)	0.65(0.07)
CPU (m)	75.34	23.80	121.10	64.20

on time, and model them as in Katzfuss and Cressie (2012) as a Markovian Gaussian process evolving over time. We can also model the coefficients, $\boldsymbol{\beta}_t$, in the STAR model (2.2) as a Markovian Gaussian process evolving over time. This enables one to deal with a spatio-temporal process with a nonstationary temporal dependence structure. As well, we can use higher-order GMRF to model spatial correlations and a more complicated transition matrix, $\boldsymbol{\Phi}(\boldsymbol{\beta})$, than those in (2.2). If $\boldsymbol{\Phi}(\boldsymbol{\beta})$ and $\boldsymbol{\Lambda}(\boldsymbol{\beta})$ are circulant, the computational complexity can be maintained as $O(n)$.

There are some limitations of the proposed method. The use of the circulant matrices makes it difficult to generalize our method to deal with a spatially nonstationary process. Although the use of circulant matrices helps lower the computational expense, it can be difficult to verify the corresponding boundary conditions in a real data problem. Then too, the autoregressive structure used in (2.2) may not be flexible enough to handle a spatio-temporal process where the spatial and temporal interaction is very strong and complicated.

Acknowledgement

This research was support in part by Award No. KUS-C1-016-04 made by King Abdullah University of Science and Technology. Liang's research was partially supported by NSF grants DMS-1007457 and DMS-1106494. Genton's research was partially supported by NSF grants DMS-1007504 and DMS-1106494.

References

Allcroft, D.J. and Glasbey, C.A. (2003). A latent Gaussian Markov random field model for spatio-temporal rainfall disaggregation. *J. Roy. Statist. Soc. Ser. B* **52**, 487-498.

Banerjee, S., Gelfand, A.E., Finley, A.O. and Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *J. Roy. Statist. Soc. Ser. B* **70**, 825-848.

Cressie, N. and Huang, H.C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94**, 1330-1340.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial processes. *J. Roy. Statist. Soc. Ser. B* **70**, 209-226.

Cressie, N., Shi, T. and Kang, E.L. (2010). Fixed rank filtering for spatio-temporal data. *J. Comput. Graph. Statist.* **19**, 724-745.

Cressie, N. and Wikle, C. (2011). *Statistics for Spatial-Temporal Data.* Wiley, New York.

Finley, A.O. and Banerjee, S. (2013). Univariate and multivariate spatial-temporal modeling. *CRAN package repository.*

Finley, A.O., Banerjee, S. and Gelfand, A.E. (2012). Bayesian dynamic modeling for large sapce-time datasets using Gaussian predictive processes. *J. Geographical Systems* **14**, 29-47.

Furrer, R., Genton, M.G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15**, 502-523.

Hughes, J.P. and Guttorp, P. (1999). A non-homogeneous hidden Markov model for precipitation. *Appl. Statist.* **48**, 15-20.

Jain, A.K. (1979). A sinusoidal family of unitary transforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 356-365.

Katzfuss, M. and Cressie, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics* **23**, 94-107.

Lemos, R. and Sansó, B. (2009). A spatio-temporal model for mean, anomaly, and trend fields of north Atlantic sea surface temperature. *J. Amer. Statist. Assoc.* **104**, 5-25.

Lindgren, F., Rue, H. and Lindstrom, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: SPDE approach. *J. Roy. Statist. Soc. Ser. B* **73**, 423-498.

Park, J. and Liang, F. (2012). Bayesian analysis of geostatistical models with an auxiliary lattice. *J. Comput. Graph. Statist.* **21**, 453-475.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* Chapman & Hall/CRC.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.* **29**, 31-50.

Sang, H. and Huang, J.Z. (2011). A full scale approximation of covariance functions for large spatial data sets. *J. Roy. Statist. Soc. Ser. B* **74**, 111-132.

Schlather, M. and Menck, P. (2013). Simulation and analysis of random fields. *CRAN package repository.*

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory of Kriging.* Springer, New York.

Stroud,J.R., Müller, P. and Sansó, B. (2001). Dynamic models for spatiotemporal data. *J. Roy. Statist. Soc. Ser. B* **64**, 673-689.

Sun, Y., Li, B. and Genton, M.G. (2012). Geostatistics for large datasets. In *Space-Time Processes and Challenges Related to Environmental Problems,* (Porcu, E., Montero, J.M., Schlather, M. eds), 55-77, Springer.

Wikle, C., Berliner, L.M. and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics* **5**, 117-154.

Institute for Applied Mathematics and Computational Science (IAMCS), Texas A&M University, College Station, TX 77843, USA.

E-mail: gang@stat.tamu.edu

Department of Statistics, Texas A&M University, College Station, TX 77843, USA.

E-mail: fliang@stat.tamu.edu

CEMSE Division, King Abdullah University of Science and Technology, Thuwal,23955-6900, Saudi Arabia.

E-mail: marc.genton@kaust.edu.sa

(Received April 2013; accepted December 2013)