Supplementary materials for this article are available at https://doi.org/10.1007/s13253-022-00525-y.

Scalable Semiparametric Spatio-temporal Regression for Large Data Analysis

Ting Fung MA, Fangfang WANG, Jun ZHU, Anthony R. IVES, and Katarzyna E. LEWIŃSKA

With the rapid advances of data acquisition techniques, spatio-temporal data are becoming increasingly abundant in a diverse array of disciplines. Here, we develop spatio-temporal regression methodology for analyzing large amounts of spatially referenced data collected over time, motivated by environmental studies utilizing remotely sensed satellite data. In particular, we specify a semiparametric autoregressive model without the usual Gaussian assumption and devise a computationally scalable procedure that enables the regression analysis of large datasets. We estimate the model parameters by maximum pseudolikelihood and show that the computational complexity can be reduced from cubic to linear of the sample size. Asymptotic properties under suitable regularity conditions are further established that inform the computational procedure to be efficient and scalable. A simulation study is conducted to evaluate the finite-sample properties of the parameter estimation and statistical inference. We illustrate our methodology by a dataset with over 2.96 million observations of annual land surface temperature, and comparison with an existing state-of-the-art approach to spatio-temporal regression highlights the advantages of our method.

Supplementary materials accompanying this paper appear online.

Key Words: Environmental statistics; Remote sensing; Sparse matrix operations; Spatio-temporal autoregression.

Ting Fung Ma, Department of Statistics, University of South Carolina, Columbia, USA (E-mail: *tingfung@mailbox.sc.edu*).
Fangfang Wang (✉) Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, USA (E-mail: *fwang4@wpi.edu*).
Jun Zhu, Department of Statistics, University of Wisconsin-Madison, Madison, USA (E-mail: *jzhu@stat.wisc.edu*).
Anthony R. Ives, Department of Integrative Biology, University of Wisconsin-Madison, Madison, USA (E-mail: *arives@wisc.edu*).
Katarzyna E. Lewińska, Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, Madison, USA (E-mail: *lewinska@wisc.edu*).
Geography Department, Humboldt-Universität zu Berlin, Unter den Linden, 10099 Berlin, Germany.

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-022-00525-y

Published online: 28 November 2022

1. INTRODUCTION

With the rapid advances of data acquisition techniques, spatio-temporal data are becoming increasingly abundant in a diverse array of disciplines including the physical, biological, and social sciences (see, e.g., Cressie and Wikle 2011; Dutilleul 2011; Anselin 2013; Agirbas et al. 2017; Belgiu and Stein 2019; Wikle et al. 2019; Zhang and Cressie 2020; Guinness 2021). Here we consider developing novel spatio-temporal regression methods for analyzing large amounts of spatially referenced data collected over time, motivated by environmental studies utilizing remotely sensed satellite data.

For illustration, we consider an environmental study of the land surface temperature (LST), which quantifies thermal energy flow among land surface, atmosphere, and biosphere, and thus characterizes local environmental conditions. Changes in LST have multiple causes, but LST is strongly related to air temperature, which is currently increasing due to global warming (IPCC 2021; NOAA 2021). Monitoring LST and understanding drivers of the observed changes are critical for agriculture, biochemical processes, bioecology, economy, and health (Hanewinkel et al. 2013; Asseng et al. 2015; Hu et al. 2016; Gasparrini et al. 2017; Zhao et al. 2017; Thompson et al. 2018).

In the USA alone, increasing temperatures contribute to crop insurance losses [($27.0 billion for the 1991–2017 period (Diffenbaugh et al. 2021)], drive the spread of tree pests (Lesk et al. 2017), and elevate risks of wildfires (Westerling et al. 2006; Mueller et al. 2020). Notably, the increase in LST temperature is not spatially uniform because it is shaped by many factors including differences in received solar radiation and the dominant land-cover type (Chakraborty et al. 2020; Yan et al. 2020). For example, urban areas have experienced relatively rapid increases in LST over the past decades (Fu and Weng 2016; Oleson et al. 2018), particularly during nighttime (Sarangi et al. 2021), whereas shrub encroachment in the Southwest United States has had a cooling effect (Shen et al. 2022). Consequently, investigation of time trends of LST across the USA (and elsewhere) is essential to better understand ongoing changes and provide needed management and mitigation strategies.

To examine spatio-temporal trends in LST, we analyze the nighttime LST derived from 2001 to 2019 MOD11A2 version 6 data (Wan et al. 2014), which we resampled to annual averages at 8 km spatial resolution. We selected the nighttime LST over the daytime measurements to limit the direct impact of solar radiation; because nighttime LST is produced by the reradiation of thermal infrared radiation generated largely through heating during the day, it gives a synoptic measure of heating that affects vegetation water stress and other variables used in ecological modeling. Because LST is related to environmental conditions, we analyzed time trends in LST in relation to latitude and longitude, which regulate vegetation zonation and amount of incoming solar radiation energy. Furthermore, to capture regional variability in LST trends, we used Level III ecoregions (Fig. S.3) defined as areas with similar landform, soil, vegetation, land use, wildlife, and hydrology (Omernik and Griffith 2014).

We could cast this research on LST in a spatio-temporal regression framework, regressing LST on the predictor variables of time trend, ecoregion classes, and interactions between the time trend and ecoregions, as well as the environmental covariates of elevation and latitude. However, there are multiple challenges with using the existing spatio-temporal regression

methods. First, the sample size of the dataset is large. With $T = 19$ years and $N = 155{,}900$ image pixels per year, there are over 2.96 million LST observations in the dataset. The traditional spatio-temporal regression models with a regression mean and a spatio-temporal covariance function would be infeasible to implement, as the computations are on the order of $\mathcal{O}(N^3T^3)$ for evaluating the likelihood function and $\mathcal{O}(N^2T^2)$ for memory usage (see, e.g., Cressie 1993; Cressie and Wikle 2011). There is ample room for innovations to reduce the computational burden and to make spatio-temporal regression analysis feasible for practical applications.

Second, although spatio-temporal statistics have advanced greatly in the past two decades, most of the state-of-the-art methods focus on the spatio-temporal dependence structure and the prediction (i.e., kriging) of the underlying spatio-temporal processes. Even when the mean function is considered, the computation of the regression coefficients is of secondary interest and may not be scalable for large data, calling for further research on the statistical inference of the mean function (see, e.g., Wikle et al. 2019).

Third, the distribution of the data is not necessarily Gaussian as is assumed by the traditional spatio-temporal models. Indeed, the histograms depicted in Fig. S.5 in Supplementary Materials suggest a possible departure of the LST distribution from Gaussian.

There has been much research on the development of statistical methodology for analyzing spatio-temporal data (see, e.g., Huang and Cressie 1996; Zhang et al. 2003; Johannesson et al. 2007; Lu et al. 2009; Cressie et al. 2010; Lee and Yu 2015; Zhang et al. 2015; Chu et al. 2019). Cressie and Wikle (2011) and Wikle et al. (2019) give excellent reviews. For Gaussian errors, Cressie et al. (2010) proposed a fixed-rank filtering method for spatio-temporal data focusing on fast computation by dimension reduction spatially and fast smoothing, filtering, or forecasting over time, which in principle can be adapted to perform regression analysis but in practice is not quite feasible yet for the scale of our LST data. Guinness (2021) developed a Gaussian process (GpGp) method that scales up more readily and can be adapted to spatio-temporal regression analysis. GpGp type of methodology approximates the full likelihood of a Gaussian process by a product of conditional likelihoods on subsets, where the subsets are formed by reordering and grouping the data (see, e.g., Vecchia 1988; Guinness 2018; Katzfuss and Guinness 2021). For non-Gaussian errors, Chu et al. (2019) and Lee and Yu (2015) proposed semiparametric models which can be applied to spatio-temporal regression, but both emphasized modeling the spatio-temporal dependence and the sample size needs to be kept at a modest size (in the thousands, not millions) for the methods to be computationally feasible. Alternatively, statistical modeling and inference can be carried out under a Bayesian framework and the computational challenges are addressed by, for example, dimension reduction (Brynjarsdóttir and Berliner 2014), predictive processes (Banerjee et al. 2008; Finley et al. 2012), latent Gaussian Markov random field on an auxiliary lattice (Xu et al. 2015), and Laplace approximation (Rue et al. 2009, 2017).

Here, we take a frequentist approach and aim to develop a novel computationally scalable procedure that enables the regression analysis of large datasets, while guided by asymptotic theory and computational complexity analysis. Our proposed method is semiparametric in the sense that no explicit distributional assumption is made about the regression error and we estimate the model parameters by maximizing a pseudolikelihood. In addition, we model the spatio-temporal dependence by autoregression. While the autoregression modeling idea

is widely used particularly in econometrics, most existing methods are not computationally scalable to the size of our LST data (see, e.g., Yu et al. 2008; Mariella and Tarantino 2010; Lee and Yu 2015; Shi and Lee 2017; Chi and Zhu 2019; Li and Yang 2021). Although Guo et al. (2016) and Gao et al. (2019) considered autoregressive models in a high-dimensional setting, their analyses focus on the estimation of coefficient matrices for zero-mean autoregressive processes without addressing the regression or computational complexity in detail. Furthermore, we adopt advanced computational techniques including efficient data preprocessing, constrained sequential quadratic programming (SQP), and implicit parallel computing. These computational innovations enable the computation to be linear in the sample size NT and thus are feasible for the LST data example.

The remainder of this paper is organized as follows. Section 2 presents the model and its estimation. Section 3 establishes the asymptotic properties of the maximum pseudolikelihood estimates of the model parameters. Section 4 provides a fast computational procedure for estimation and inference. The finite-sample properties of the estimators are assessed by simulation studies in Sect. 5, and the LST data example is given in Sect. 6. Section 7 concludes the paper with a discussion of possible avenues for future research. Proofs of the theoretical results and other technical details including additional computational aspects, tables, and figures are provided in Supplementary Materials. Data and code for this research are publicly available and can be downloaded from https://doi.org/10.17605/OSF.IO/WT84X.

2. MODEL AND ESTIMATION

2.1. Model Specification

At time $t \in \mathbb{Z}$, let $\boldsymbol{Y}_t = (Y_{1,t}, \ldots, Y_{N,t})'$ denote an N-dimensional real-valued vector that contains the response variables from N cells that partition the study region of interest in $\mathbb{R}^2$. Let $\boldsymbol{X}_t$ denote an $N \times k$ design matrix of k nonstochastic predictor variables. We model the spatio-temporal evolution of $\boldsymbol{Y}_t$ in relation to $\boldsymbol{X}_t$ through the following spatio-temporal regression model

$$\boldsymbol{Y}_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{U}_t,\ t \in \mathbb{Z}, \tag{1}$$

where $\boldsymbol{\beta}$ denotes a $k \times 1$ vector of regression coefficients. The spatio-temporal error $\boldsymbol{U}_t$ is stochastic and modeled by a spatio-temporal dynamic process such that

$$\boldsymbol{U}_t = \lambda \boldsymbol{W}\boldsymbol{U}_t + \rho \boldsymbol{W}\boldsymbol{U}_{t-1} + \gamma \boldsymbol{U}_{t-1} + \boldsymbol{V}_t, \tag{2}$$

where $\boldsymbol{V}_t = (v_{1,t}, \ldots, v_{N,t})'$ is an $N \times 1$ vector of real-valued innovations that are assumed to be *iid*, not necessarily Gaussian, with mean zero and variance $\sigma^2\boldsymbol{\mathcal{I}}_N$ and $\boldsymbol{\mathcal{I}}_N$ is the $N \times N$ identity matrix. The spatio-temporal dependence parameters include the conventional temporal lag effect γ, the contemporaneous spatial interactions effect λ, and the effect of spatial diffusion that takes place over time ρ (see, e.g., Anselin 2013; Lee and Yu 2015; Chi and Zhu 2019).

Finally, the spatial weight matrix $\boldsymbol{W}$ is an $N \times N$ nonstochastic symmetric matrix with zero diagonals for a given spatial neighborhood structure (Cressie 1993). The symmetry of $\boldsymbol{W}$ has important implications on computation, which will be elaborated in later sections. Special cases of the spatial weight matrix $\boldsymbol{W}$ include the block-diagonal structure and commonly assumed first- or second-order neighborhood structures. For a block-diagonal structure, $\boldsymbol{W} = \text{Diag}\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p\}$, where $\boldsymbol{w}_i$ is a $n_i \times n_i$ matrix, with $N = \sum_{i=1}^{p} n_i$ (Case 1991). On a regular square grid, the first-order neighbors are the four nearest cells whereas the second-order neighbors are the eight nearest cells (Cressie 1993). In addition, the spatial weight matrix could be used to construct the design matrix $\boldsymbol{X}_t$ in order to capture the spatial neighboring effects; for instance, let $\boldsymbol{X}_t = (\boldsymbol{1}_N, \boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t})$, where $\boldsymbol{Z}_{2t} = \tilde{\boldsymbol{W}}\boldsymbol{Z}_{1t}$ and $\tilde{\boldsymbol{W}}$ is a spatial weight matrix defined above.

Let $\boldsymbol{\theta} = (\lambda, \gamma, \rho)'$ denote the vector of the spatio-temporal dependence parameters. We define $\boldsymbol{R}(\boldsymbol{\theta}) = \rho\boldsymbol{W} + \gamma\boldsymbol{\mathcal{I}}_N$, $\boldsymbol{S}(\lambda) = \boldsymbol{\mathcal{I}}_N - \lambda\boldsymbol{W}$, and $\boldsymbol{A}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{S}(\lambda)^{-1}$. We may then rewrite the spatio-temporal dynamic process (2) as $\boldsymbol{S}(\lambda)\boldsymbol{U}_t = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{S}(\lambda)\boldsymbol{U}_{t-1} + \boldsymbol{V}_t$. That is, the spatio-temporal error $\boldsymbol{U}_t$ follows a vector autoregression model of order one and can be shown to be weakly stationary under the assumption that $\boldsymbol{S}(\lambda)$ is nonsingular and the eigenvalues of $A(\boldsymbol{\theta})$ are all strictly less than one in magnitude.

2.2. PARAMETER ESTIMATION

For the observed response vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ modeled by (1), we define the vector of all the spatio-temporal errors $\boldsymbol{U} = (\boldsymbol{U}_1', \boldsymbol{U}_2', \ldots, \boldsymbol{U}_T')'$ and its matrix operator

$$\boldsymbol{B}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{S}(\lambda) & 0 & \cdots & 0 & 0 \\ -\boldsymbol{R}(\boldsymbol{\theta}) & \boldsymbol{S}(\lambda) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -\boldsymbol{R}(\boldsymbol{\theta}) & \boldsymbol{S}(\lambda) \end{pmatrix}_{NT \times NT} \tag{3}$$

such that $\boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{U} = ((\boldsymbol{S}(\lambda)\boldsymbol{U}_1)', \boldsymbol{V}_2', \ldots, \boldsymbol{V}_T')'$. The covariance matrix of $\boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{U}$ is $\sigma^2\boldsymbol{\Omega}(\boldsymbol{\theta})$, where $\boldsymbol{\Omega}(\boldsymbol{\theta}) = \text{Diag}(\boldsymbol{K}(\boldsymbol{\theta}), \boldsymbol{\mathcal{I}}_N, \ldots, \boldsymbol{\mathcal{I}}_N)$ and $\boldsymbol{K}(\boldsymbol{\theta}) = \sum_{j=0}^{\infty} \boldsymbol{A}(\boldsymbol{\theta})^j \boldsymbol{A}(\boldsymbol{\theta})'^j$.

Let $\boldsymbol{\delta} = (\boldsymbol{\beta}', \boldsymbol{\theta}', \sigma^2)'$ denote the vector of all the model parameters. Recall that the distribution of the innovation $\boldsymbol{V}_t$ is not necessarily Gaussian. However, to estimate $\boldsymbol{\delta}$, we proceed as if $\boldsymbol{V}_t$ followed a Gaussian distribution, i.e., $N(\boldsymbol{0}, \sigma^2\boldsymbol{\mathcal{I}}_N)$, which yields the following log pseudolikelihood function,

$$\begin{aligned} \log L_{NT}(\boldsymbol{\delta}) = &- \frac{NT}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log\det(\boldsymbol{K}(\boldsymbol{\theta})) + T\log|\det(\boldsymbol{S}(\lambda))| \\ &- \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \end{aligned} \tag{4}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_T')'$ denotes the $NT \times 1$ vector of all the response variables, $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_T')'$ is the corresponding $NT \times k$ design matrix, and $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} = \boldsymbol{B}(\boldsymbol{\theta})'(\boldsymbol{\Omega}(\boldsymbol{\theta}))^{-1}\boldsymbol{B}(\boldsymbol{\theta})$ is the precision matrix. Denote by $\widehat{\boldsymbol{\delta}}$ the maximizer of the log pseudolikelihood function $\log L_{NT}(\boldsymbol{\delta})$; that is, $\widehat{\boldsymbol{\delta}} = \arg\max_{\boldsymbol{\delta} \in \Theta_{\boldsymbol{\delta}}} \log L_{NT}(\boldsymbol{\delta})$, where $\Theta_{\boldsymbol{\delta}}$ is the

parameter space specified in Appendix A. Throughout this paper, we refer to $\widehat{\boldsymbol{\delta}}$ as our maximum pseudolikelihood estimator (PMLE) of the model parameters $\boldsymbol{\delta}$.

As illustrated by the LST data example in Sect. 1, our primary interest is statistical inference of the regression coefficients $\boldsymbol{\beta}$, while the estimation of the spatio-temporal dependence parameters $\boldsymbol{\theta}$ is of secondary interest intended to account for spatio-temporal correlation when drawing the inference about $\boldsymbol{\beta}$.

3. INFERENCE

Under suitable regularity conditions, we may establish the asymptotic properties of the PMLE $\widehat{\boldsymbol{\delta}}$, as the number of cells $N \to \infty$ while the number of time points T can be either fixed or $T \to \infty$. Denote by $\boldsymbol{\delta}_0 = (\boldsymbol{\beta}_0', \boldsymbol{\theta}_0', \sigma_0^2)'$ the vector of true model parameters. We first consider the case that T is fixed.

Theorem 1. *Suppose that $\boldsymbol{W}$ has more than two distinct eigenvalues, and Assumptions (A.1)–(A.6) hold. Then, $\boldsymbol{\delta}_0$ is identifiably unique and $\widehat{\boldsymbol{\delta}} \xrightarrow{p} \boldsymbol{\delta}_0$ as $N \to \infty$.*

Theorem 1 establishes that the PMLE $\widehat{\boldsymbol{\delta}}$ is a consistent estimator of the true parameter vector $\boldsymbol{\delta}_0$ in the sense that $\widehat{\boldsymbol{\delta}}$ converges to $\boldsymbol{\delta}_0$ in probability, when $N \to \infty$. The condition that $\boldsymbol{W}$ has more than two distinct eigenvalues, along with Assumptions (A.1) and (A.2), ensures that $(\boldsymbol{\theta}, \sigma^2)$ can be uniquely identified from $\sigma^2(\boldsymbol{B}(\boldsymbol{\theta})'(\boldsymbol{\Omega}(\boldsymbol{\theta}))^{-1}\boldsymbol{B}(\boldsymbol{\theta}))^{-1}$; that is, $\sigma^2(\boldsymbol{B}(\boldsymbol{\theta})'(\boldsymbol{\Omega}(\boldsymbol{\theta}))^{-1}\boldsymbol{B}(\boldsymbol{\theta}))^{-1} = \sigma_0^2(\boldsymbol{B}(\boldsymbol{\theta}_0)'(\boldsymbol{\Omega}(\boldsymbol{\theta}_0))^{-1}\boldsymbol{B}(\boldsymbol{\theta}_0))^{-1}$ if and only if $\sigma^2 = \sigma_0^2$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ (see Lemma S.4 in Supplementary Materials).

Next, under additional conditions about the higher-order properties of the log pseudolikelihood function, we derive the asymptotic distribution of the PMLE $\widehat{\boldsymbol{\delta}}$.

Theorem 2. *Suppose that the conditions in Theorem 1 and additional Assumptions (A.7) and (A.8) are fulfilled. Then,*

$$\sqrt{N}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\left\{\overline{\boldsymbol{\Sigma}}_1^{-1} + \overline{\boldsymbol{\Sigma}}_1^{-1}\overline{\boldsymbol{\Sigma}}_2\overline{\boldsymbol{\Sigma}}_1^{-1}\right\}\right), \tag{5}$$

where $\overline{\boldsymbol{\Sigma}}_1 = \lim_{N\to\infty} N^{-1}\boldsymbol{\Sigma}_{1,N}$, $\overline{\boldsymbol{\Sigma}}_2 = \lim_{N\to\infty} N^{-1}\boldsymbol{\Sigma}_{2,N}$, $\boldsymbol{\Sigma}_{1,N} = \mathrm{Diag}(4\sigma_0^{-2}\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{X}, 2\boldsymbol{\Omega}_N)$ *with* $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1} = \boldsymbol{B}(\boldsymbol{\theta}_0)'(\boldsymbol{\Omega}(\boldsymbol{\theta}_0))^{-1}\boldsymbol{B}(\boldsymbol{\theta}_0)$ *and* $\boldsymbol{\Omega}_N$ *defined in* (11)*, and* $\boldsymbol{\Sigma}_{2,N}$ *is defined in (S.16) in Supplementary Materials. In particular, we have*

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_0}^{-1}\right), \tag{6}$$

where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_0} = \sigma_0^{-2}\lim_{N\to\infty} N^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{X}$. *Under the additional assumption that* $\mu_3 = \mathrm{E}(v_{j,t}^3) = 0$, $\widehat{\boldsymbol{\beta}}$ *is asymptotically independent of* $\widehat{\boldsymbol{\theta}}$ *and* $\hat{\sigma}^2$.

Theorem 2 establishes that $\widehat{\boldsymbol{\delta}}$ converges to a multivariate Gaussian distribution at the rate of $\sqrt{N}$. The asymptotic covariance matrix involves two matrices $\overline{\boldsymbol{\Sigma}}_1$ and $\overline{\boldsymbol{\Sigma}}_2$, which can be replaced by their consistent estimators for evaluating the asymptotic distribution of

$\widehat{\boldsymbol{\delta}}$ in practice. Since the primary interest is in the statistical inference about the regression coefficients $\boldsymbol{\beta}$, we present in (6) the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and its relationship to the other parameter estimators $\widehat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$, which provides the basis for a computationally efficient approach to the statistical inference about $\boldsymbol{\beta}$, as we will detail in Sect. 4.2. Moreover, the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ remains unchanged regardless of the distribution of the spatio-temporal innovations. In particular, when the innovation is symmetric (i.e., $\mu_3 = 0$, which is satisfied by many commonly used distributions including Gaussian and Student-T distributions), (6) establishes that $\widehat{\boldsymbol{\beta}}$ is asymptotically independent of the spatio-temporal dependence parameter estimators $\widehat{\boldsymbol{\theta}}$ and the variance component estimator $\hat{\sigma}^2$.

By Theorems 1 and 2, $N^{-1}\boldsymbol{\Sigma}_{1,N}$ and $N^{-1}\boldsymbol{\Sigma}_{2,N}$ with $\boldsymbol{\delta}_0$ replaced by $\widehat{\boldsymbol{\delta}}$ converge in probability to $\overline{\boldsymbol{\Sigma}}_1$ and $\overline{\boldsymbol{\Sigma}}_2$, respectively, as $N \to \infty$. Thus, a consistent estimator of the asymptotic covariance matrix of $\widehat{\boldsymbol{\delta}}$ can be obtained from

$$4\left\{\boldsymbol{\Sigma}_{1,N}^{-1} + (\boldsymbol{\Sigma}_{1,N})^{-1}(\boldsymbol{\Sigma}_{2,N})(\boldsymbol{\Sigma}_{1,N})^{-1}\right\} \tag{7}$$

evaluated at the PMLE $\widehat{\boldsymbol{\delta}}$. However, $\boldsymbol{\Sigma}_{1,N}$ and $\boldsymbol{\Sigma}_{2,N}$ are both challenging to compute when N is large. One major challenge is that the calculation of $\boldsymbol{\Sigma}_{1,N}$ and $\boldsymbol{\Sigma}_{2,N}$ requires solving large linear systems, or equivalently inverting large matrices, which is computationally expensive. On the other hand, the upper left block of $\boldsymbol{\Sigma}_{1,N}$ is $\sigma_0^{-2}\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{X}$, which can be consistently estimated by $\hat{\sigma}^{-2}\boldsymbol{X}'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}\boldsymbol{X}$. Thus, a consistent estimator of the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{X})^{-1}$, and its computation can in fact be made scalable (see Sect. 4).

In addition, the asymptotic results hold when T is either fixed or tends to infinity with N at an arbitrary rate. That is, Theorems 1 and 2 can be readily extended to the case when N and T both tend to infinity, in which case the rate of convergence in Theorem 2 becomes $\sqrt{NT}$ instead of $\sqrt{N}$, with corresponding adjustment of Assumptions (A.5), (A.6), and (A.8), and the asymptotic covariance matrices.

Before closing this section, we remark on Assumption (A.2) provided in Appendix, while discussions on the other assumptions are given in Supplementary Materials. A sufficient condition for the matrix $\boldsymbol{S}(\lambda) = \boldsymbol{\mathcal{I}}_N - \lambda\boldsymbol{W}$ being nonsingular and the eigenvalues of $A(\boldsymbol{\theta})$ being less than one in magnitude is that the parameters λ, γ, ρ satisfy the following inequality:

$$(\lambda^2 - \rho^2)\mathrm{d}_j^2 - 2(\lambda + \gamma\rho)\mathrm{d}_j + (1 - \gamma^2) > 0,\ j = 1, \ldots, r, \tag{8}$$

where $\{\mathrm{d}_i, i = 1, \ldots, r\}$ are the nonzero eigenvalues of $\boldsymbol{W}$ with the smallest eigenvalue (d_1) and the largest eigenvalue (d_r) of $\boldsymbol{W}$ having opposite signs. For (8) to hold, it is sufficient to consider the following set,

$$\left\{(\lambda, \gamma, \rho) : -1 < \gamma < 1, \frac{1-\gamma}{\mathrm{d}_1} < \lambda + \rho < \frac{1-\gamma}{\mathrm{d}_r}, \frac{1+\gamma}{\mathrm{d}_1} < \lambda - \rho < \frac{1+\gamma}{\mathrm{d}_r}\right\}. \tag{9}$$

In practice, we choose $\Theta_{\boldsymbol{\theta}}$ as a compact subset of the above set.

4. COMPUTATION

In this section, we will develop a novel fast computation procedure and show that its computational complexity is on the order of $\mathcal{O}(NT)$ for obtaining the PMLE $\widehat{\boldsymbol{\delta}}$ and the variance estimate of $\widehat{\boldsymbol{\beta}}$, which is linear to the sample size and thus is scalable to the size of the LST dataset. The existing state-of-the-art methodology generally approximates the dependence structure for computational ease, while our approach does not require an approximation of the spatio-temporal dependence. Thus, our computational procedure provides a novel and scalable alternative to the existing spatio-temporal modeling and inference without approximating the likelihood function.

4.1. Computational Procedure

We obtain the PMLE, $\widehat{\boldsymbol{\delta}}$, and an estimate of its variance $\text{Var}(\widehat{\boldsymbol{\delta}})$ by bringing together a set of computational techniques for nonlinear optimization and sparse matrix operations. An overview of the procedure is visualized by a flowchart in Fig. 1. Specifically, the procedure starts with the input of the spatial weight matrix $\boldsymbol{W}$, the response variable $\boldsymbol{Y}$, and the design matrix $\boldsymbol{X}$. We then preprocess the data by applying the reverse Cuthill–McKee (RCM) algorithm (Gilbert et al. 1992). In particular, the RCM algorithm permutes the rows and columns of $\boldsymbol{W}$, which is a symmetric, generally sparse matrix, into a symmetric sparse banded matrix with a small bandwidth. This effectively moves the nonzero elements of $\boldsymbol{W}$ toward the diagonal while preserving the spatial neighborhood structure. The underlying graph theory for the RCM algorithm views the spatial weight matrix as a graph with vertices (of spatial locations) and edges that connect spatial neighbors specified in $\boldsymbol{W}$. We then reorder the rows of $\boldsymbol{Y}_t$ and $\boldsymbol{X}_t$ according to the Cuthill–McKee ordering of $\boldsymbol{W}$ for $t = 1, \ldots, T$. For a given spatial weight matrix $\boldsymbol{W}$, it is always possible to convert it into a sparse banded matrix, without distorting the pre-specified spatial neighboring structure (Mafteiu-Scai 2015). Thus henceforth we assume that $\boldsymbol{W}$ is a pre-specified symmetric sparse banded matrix with bandwidth b, which eases the implementation of computational techniques for banded matrices and enables a more precise account of computational complexity.

Next, the parameter vector $\boldsymbol{\delta}$ is estimated by maximizing the log pseudolikelihood using an iterative SQP (i.e., fmincon() in MATLAB) (see Chapter 18 of Nocedal and Wright 2006). At each iteration, the log pseudolikelihood function (4) and its gradient functions are evaluated for optimizing (4) subject to a set of constraints on the parameter space $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$. To ensure the scalability of SQP, however, care is needed in the evaluation of the log pseudolikelihood function, as we will show in the next subsection. In addition, the constraints on the parameters need to be checked, which we will refer to as feasibility check. The standard feasibility check would require computational cost on the order of $\mathcal{O}(N^{2.4})$. Here we apply the sufficient condition (9) developed in Sect. 3, which requires solving for the smallest (d_1) and largest (d_r) eigenvalues of $\boldsymbol{W}$. We thus preprocess $\boldsymbol{W}$ by the Krylov–Schur algorithm, which is an iterative method for solving eigenproblems with sparsity and belongs to the class of Krylov subspace methods (Stewart 2002). The Krylov–Schur algorithm first generates a sequence of subspaces containing the approximations of a subset of eigenvectors and eigenvalues of $\boldsymbol{W}$. Then these approximations are extracted

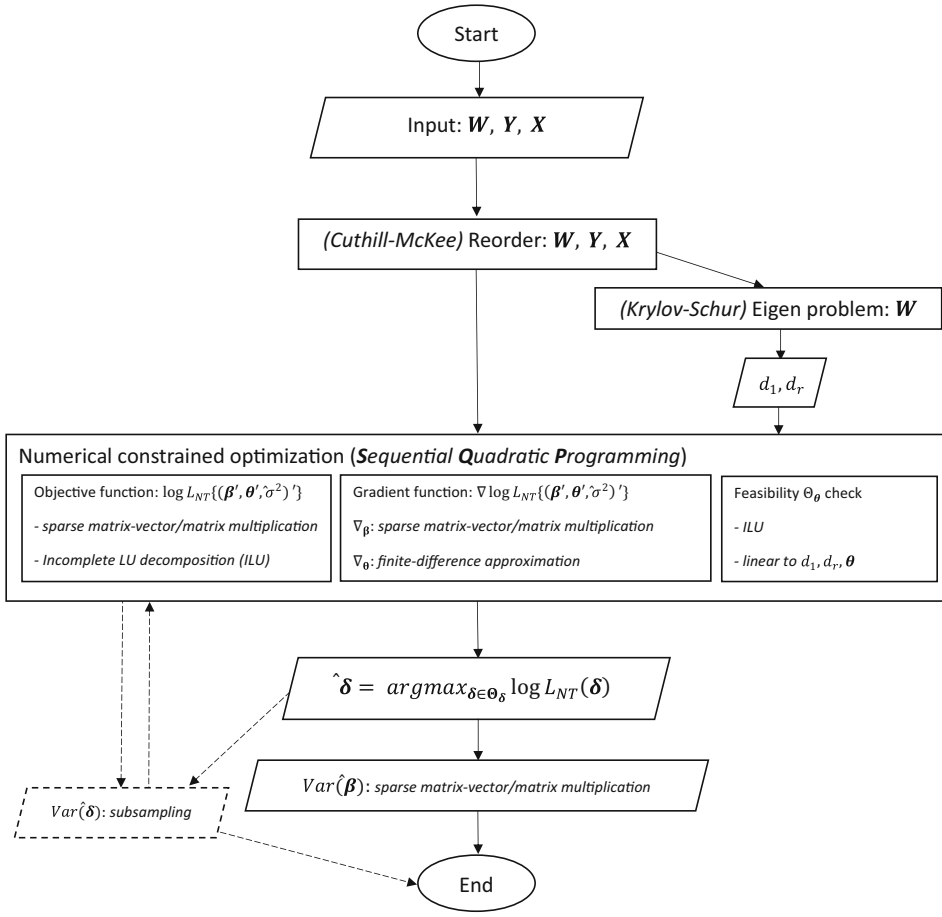

Figure 1. Flowchart for carrying out the proposed spatio-temporal regression and inference.

by applying the QR algorithm to the projection of $\boldsymbol{W}$ onto the subspaces and the subset of eigenvalues is approximated iteratively through the Arnoldi method. A reordering of the Schur decomposition in the previous step is considered to improve the standard Arnoldi method (see Chapter 3 of Kressner 2005). Based on d_1, d_r, and the sufficient condition (9), our feasibility check requires $\mathcal{O}(1)$ operations, which is a significant improvement over the $\mathcal{O}(N^{2.4})$ operations and the $\mathcal{O}(N)$ memory usage when a full eigendecomposition of $\boldsymbol{W}$ is used for (8).

In addition, most of the computations can be parallelized and in particular, we enable the implicit parallelism through maxNumCompThreads(), which distributes the computation in multiple cores and utilizes the sparsity of matrices in our MATLAB code (Luszczek 2009). Similar techniques can also be implemented in R and Python, for example through the Basic Linear Algebra Subroutines (BLAS) or Linear Algebra Package (LAPACK) (Anderson et al. 1999; Blackford et al. 2002; Buluc and Gilbert 2011).

4.2. COMPUTATIONAL COMPLEXITY

Direct evaluation of the log pseudolikelihood function (4) requires $\mathcal{O}(N^3T^3)$ operations and is computationally infeasible when NT is large. In the following, we show that our computational procedure has the computational complexity of $\mathcal{O}(NT)$.

The matrix operator $\boldsymbol{B}(\boldsymbol{\theta})$ in (3) and the covariance matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$ are involved in the log pseudolikelihood function (4). Storing the entirety of $\boldsymbol{B}(\boldsymbol{\theta})$ and $\boldsymbol{\Omega}(\boldsymbol{\theta})$ during the process of computing the precision matrix would require standard memory usage and operations to be on the order of $\mathcal{O}(N^2T)$. Instead, we partition the matrix operator $\boldsymbol{B}(\boldsymbol{\theta})$ and the covariance matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$ in such a way that we only store the unique nonzero blocks of quadratic terms. More specially, we rewrite the last term of the log pseudolikelihood function (4) as

$$
\begin{aligned}
&(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{B}(\boldsymbol{\theta})'(\boldsymbol{\Omega}(\boldsymbol{\theta}))^{-1}\boldsymbol{B}(\boldsymbol{\theta})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) \\
&\quad = (\boldsymbol{Y}_1-\boldsymbol{X}_1\boldsymbol{\beta})'\boldsymbol{S}(\lambda)\boldsymbol{K}(\boldsymbol{\theta})^{-1}\boldsymbol{S}(\lambda)(\boldsymbol{Y}_1-\boldsymbol{X}_1\boldsymbol{\beta}) + \sum_{t=2}^{T}(\boldsymbol{Y}_t-\boldsymbol{X}_t\boldsymbol{\beta})'\boldsymbol{S}(\lambda)^2(\boldsymbol{Y}_t-\boldsymbol{X}_t\boldsymbol{\beta}) \\
&\qquad + \sum_{t=1}^{T-1}(\boldsymbol{Y}_t-\boldsymbol{X}_t\boldsymbol{\beta})'\boldsymbol{R}(\boldsymbol{\theta})^2(\boldsymbol{Y}_t-\boldsymbol{X}_t\boldsymbol{\beta}) - 2\sum_{t=1}^{T-1}(\boldsymbol{Y}_t-\boldsymbol{X}_t\boldsymbol{\beta})'\boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{S}(\lambda)(\boldsymbol{Y}_{t+1}-\boldsymbol{X}_{t+1}\boldsymbol{\beta}).
\end{aligned}
\tag{10}
$$

The total number of nonzero elements (nnz) of $\boldsymbol{W}$ is $\mathcal{O}(bN)$. Since the product of two $N\times N$ banded matrices each with bandwidth $\mathcal{O}(b)$ is still banded with bandwidth $\mathcal{O}(b)$, it follows that $\boldsymbol{S}(\lambda)\boldsymbol{K}(\boldsymbol{\theta})^{-1}\boldsymbol{S}(\lambda)$, $\boldsymbol{S}(\lambda)^2$, $\boldsymbol{R}(\boldsymbol{\theta})^2$, and $\boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{S}(\lambda)$ in (10) are all banded matrices with bandwidth $\mathcal{O}(b)$. Thus, the computation of each quadratic form in the summand of (10) involves sparse matrix–vector multiplications and requires $\mathcal{O}(nnz)=\mathcal{O}(bN)$ operations. As a result, the computation of (10) has complexity $\mathcal{O}(bNT+kNT)$.

The second and third terms of the log pseudolikelihood function (4) involve the evaluation of two log determinants, $\log\det(\boldsymbol{K}(\boldsymbol{\theta}))$ and $\log|\det(\boldsymbol{S}(\lambda))|$, which is in general numerically unstable and computationally infeasible when the sample size NT is large. To overcome such challenges, we utilize the relationship between an LU decomposition and the determinant. Recall that $\boldsymbol{K}(\boldsymbol{\theta})$, given by $\sum_{j=0}^{\infty}\boldsymbol{A}(\boldsymbol{\theta})^j\boldsymbol{A}(\boldsymbol{\theta})'^j$, is dense in general. Thus, it is computationally challenging to compute its log determinant and invert the matrix, as these operations involve solving large linear systems and infinite sum of matrices. Here, we overcome the difficulty by taking full advantage of the symmetric spatial weight matrix and noting the following identity: $\boldsymbol{S}(\lambda)\boldsymbol{K}(\boldsymbol{\theta})^{-1}\boldsymbol{S}(\lambda)=\boldsymbol{S}(\lambda)^2-\boldsymbol{R}(\boldsymbol{\theta})^2$. After some algebra, we have $\log(\det(\boldsymbol{K}(\boldsymbol{\theta})))=\log\det(\boldsymbol{S}(\lambda)^2)-\log\det(\boldsymbol{S}(\lambda)^2-\boldsymbol{R}(\boldsymbol{\theta})^2)$, which converts the computationally intensive task into sparse matrix multiplication and calculation of the (log-)determinant of two positive definite matrices with bandwidth $\mathcal{O}(b)$. Furthermore, incomplete LU (ILU) decomposition of banded matrix takes advantage of the sparsity pattern to speed up the LU factorization without compromising the accuracy (Saad 2003). This reduces the computational cost from the standard $\mathcal{O}(N^{2.4})$ to $\mathcal{O}(b^2N)$ (see, e.g., Section 2 of Kilic and Stanica 2013) and ensures the numerical stability of the calculation of log determinant of banded matrices during the evaluation of log pseudolikelihood.

The first term of the log pseudolikelihood function (4) would require $\mathcal{O}(N)$ operations after profiling out σ^2 in (4). That is, by setting $\partial_{\sigma^2} \log L_{NT}(\boldsymbol{\delta}) = -(2\sigma^2)^{-1}NT + (2\sigma^4)^{-1}\boldsymbol{H}(\boldsymbol{\beta},\boldsymbol{\theta})$ to zero, we have $\hat{\sigma}^2 = (NT)^{-1}\boldsymbol{H}(\boldsymbol{\beta},\boldsymbol{\theta})$, where $\boldsymbol{H}(\boldsymbol{\beta},\boldsymbol{\theta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. Combining the results above, the overall computational complexity for evaluating the log pseudolikelihood function (4) is $\mathcal{O}(\text{bNT} + \text{kNT} + b^2N)$.

To compute the gradient of (4), the computational cost is on the order $\mathcal{O}(\text{kbNT} + b^2N)$, because the partial derivative of (4) with respect to $\boldsymbol{\beta}$ has a closed form

$$\frac{\partial \log L_{\text{NT}}(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}} = 2\Big[\sum_{t=1}^{T} \boldsymbol{X}_t'\boldsymbol{S}(\lambda)^2(\boldsymbol{X}_t\boldsymbol{\beta} - \boldsymbol{Y}_t) + \sum_{t=2}^{T-1} \boldsymbol{X}_t'\boldsymbol{R}(\boldsymbol{\theta})^2(\boldsymbol{X}_t\boldsymbol{\beta} - \boldsymbol{Y}_t) + \sum_{t=1}^{T-1} \left\{\boldsymbol{X}_t'\boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{S}(\lambda)(\boldsymbol{Y}_{t+1} - \boldsymbol{X}_{t+1}\boldsymbol{\beta}) + \boldsymbol{X}_{t+1}'\boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{S}(\lambda)(\boldsymbol{Y}_t - \boldsymbol{X}_t\boldsymbol{\beta})\right\}\Big]$$

and requires $\mathcal{O}(\text{kbNT})$ operations, due to the multiplication of sparse matrices. The computational complexity of calculating the partial derivative of (4) with respect to $\boldsymbol{\theta}$ using the analytical form remains computationally expensive as it involves solving large linear system requiring $\mathcal{O}(N^{2.4})$ operations. Thus, we use finite difference approximations in the gradient calculation, which reduce the computational cost from $\mathcal{O}(N^{2.4} + \text{kbNT})$ to $\mathcal{O}(b^2N + \text{kbNT})$.

With the results above combined, the estimation of $\widehat{\boldsymbol{\delta}}$ through numerical constrained optimization would require $\mathcal{O}(\text{kbNT} + b^2N)$ operations. In other words, the computational complexity of our method is linear to the total sample size (NT) when k and b are fixed and hence, is computationally feasible for large datasets even on the order of millions.

Last but not least, we turn to the computational cost involved in evaluating the estimate of $\text{Var}(\widehat{\boldsymbol{\delta}})$. By a similar argument in the evaluation of the log pseudolikelihood, computing $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{X})^{-1}$ requires only $\mathcal{O}(\text{kbNT})$ operations and $\mathcal{O}(\text{kNT} + bN)$ memory usage, as opposed to $\mathcal{O}(N^2T)$ operations and an extra $\mathcal{O}(N^2T)$ memory usage with the standard computation. Thus, our procedure facilitates the statistical inference about $\boldsymbol{\beta}$ with large sample size. However, the computation of (7) is dominated by solving a large linear system (Gilbert et al. 1992) in the calculation of $\boldsymbol{\Sigma}_{1,N}$ and $\boldsymbol{\Sigma}_{2,N}$, which requires at most $\mathcal{O}(N^{2.4}T^{2.4})$ computations using the Coppersmith–Winograd algorithm (Coppersmith and Winograd 1990). As such, for practical applications, it may be prudent to apply resampling to compute the standard errors of the spatio-temporal dependence parameter estimates in $\widehat{\boldsymbol{\theta}}$. For example, spatial subsampling may be applied to overlapping or nonoverlapping spatial blocks and provide replications of $\widehat{\boldsymbol{\theta}}$ for estimating the asymptotic covariance matrix (see, e.g., Sherman 1996; Nordman and Lahiri 2004).

5. SIMULATION STUDY

5.1. SIMULATION SETUP

We conduct simulation experiments to assess the finite-sample properties of our proposed methodology and evaluate its computational efficiency. For the design matrix $\boldsymbol{X}$, we let $k = 2$ including the intercept and a covariate sampled from the standard Gaussian

Table 1. Sample average bias ($\times 10^{-4}$) and mean squared error (MSE, $\times 10^{-4}$) of $\widehat{\boldsymbol{\delta}}$ based on 1000 simulations, and average computational time (in second) per simulation, with Gaussian innovations

N	T	Average bias $\times 10^{-4}$						Sample MSE $\times 10^{-4}$						Average time
		β_0	β_1	λ	γ	ρ	σ^2	β_0	β_1	λ	γ	ρ	σ^2	
10^2	5	−22.80	9.21	−5.24	−50.05	−20.98	− 105.50	68.32	16.42	2.60	17.73	5.20	40.92	0.05
	10	−14.23	−10.61	−11.03	−35.19	−6.29	− 56.64	34.28	8.32	1.51	8.42	2.51	20.72	0.05
	20	−17.43	12.26	1.40	−15.47	−7.07	− 45.87	18.00	4.35	0.66	3.90	1.38	9.99	0.07
	50	−9.01	−1.41	−2.64	1.08	−1.06	− 10.59	7.99	1.51	0.27	1.46	0.48	3.99	0.11
20^2	5	−23.22	14.89	2.23	3.27	−12.63	− 45.95	15.87	4.49	0.63	4.00	1.38	10.66	0.10
	10	16.95	0.09	−0.69	−10.73	−1.35	− 6.48	8.68	2.09	0.32	1.90	0.67	5.25	0.12
	20	−0.21	1.18	1.11	−9.17	−1.73	− 8.68	4.99	1.01	0.16	0.93	0.29	2.58	0.17
	50	3.58	1.33	−0.19	−1.55	0.24	− 2.55	1.89	0.41	0.07	0.40	0.10	0.94	0.24
50^2	5	−0.55	2.01	−0.32	0.82	0.00	− 4.31	2.57	0.65	0.11	0.61	0.22	1.69	0.43
	10	−0.85	−0.58	0.94	−0.26	−2.54	− 9.74	1.34	0.34	0.05	0.31	0.09	0.83	0.54
	20	−3.11	0.41	1.16	0.14	−1.53	− 0.69	0.75	0.16	0.02	0.15	0.05	0.40	0.65
	50	−1.87	−0.64	2.06	−0.45	−1.31	− 1.51	0.29	0.06	0.01	0.06	0.02	0.16	0.96
100^2	5	5.11	1.19	1.16	−1.89	−0.57	− 0.08	0.65	0.16	0.03	0.17	0.05	0.42	1.77
	10	−0.85	0.27	1.86	0.73	−0.65	− 1.17	0.34	0.08	0.01	0.08	0.02	0.19	2.12
	20	1.58	1.03	1.52	0.21	−0.98	− 0.60	0.19	0.04	0.01	0.04	0.01	0.10	2.67
	50	1.81	0.43	1.78	0.59	−1.20	− 1.27	0.08	0.02	0.00	0.01	0.00	0.04	4.41
200^2	5	3.33	−0.71	1.52	−0.53	−1.50	− 0.42	0.17	0.04	0.01	0.04	0.01	0.11	10.13
	10	−0.50	0.63	1.58	0.05	−1.35	− 1.80	0.09	0.02	0.00	0.02	0.01	0.06	11.67
	20	0.13	−0.37	1.75	0.15	−0.89	− 1.61	0.05	0.01	0.00	0.01	0.00	0.02	13.84
	50	−1.18	0.02	1.84	0.13	−0.88	− 0.82	0.02	0.00	0.00	0.00	0.00	0.01	20.29

distribution $\mathcal{N}(0, 1)$. Once generated, $\boldsymbol{X}$ is kept fixed. The random innovations $\boldsymbol{V}_t$ are sampled independently from $\mathcal{N}(0, 1)$, $t = 1, \ldots, T$. The true parameter vector $\boldsymbol{\delta}_0$ is set at $(1, 0.5, 0.05, 0.5, -0.05, 1)'$; that is, $\boldsymbol{\beta}_0 = (1, 0.5)'$, $\lambda_0 = 0.05$, $\gamma_0 = 0.5$, $\rho_0 = -0.05$, and $\sigma_0^2 = 1$. We also consider a two-dimensional spatial domain with the data taken at spatial coordinates $\{(1, 1), \ldots, (1, n), \ldots, (n, n)\}$ and the spatial weight matrix $\boldsymbol{W}$ is under a first-order spatial neighborhood structure. To examine the effect of sample sizes, we consider $N = n^2 \in \{10^2, 20^2, 50^2, 100^2, 200^2\}$ and $T \in \{5, 10, 20, 50\}$. For each combination of N and T, 1000 simulations are generated.

The core computation is executed on an application server with dual Intel Xeon Silver 4116 2.1GHz 12-core (24 thread) processors and 512GB of RAM, running MATLAB R2020a.

5.2. SIMULATION RESULTS

The PMLE $\widehat{\boldsymbol{\delta}}$ of the model parameter vector is obtained from maximizing the log pseudolikelihood (4). To evaluate the finite-sample properties of the parameter estimates, we compute the bias and mean squared error (MSE) by taking the sample average of the differences and the squared differences between the estimate $\widehat{\boldsymbol{\delta}}$ and the true value $\boldsymbol{\delta}_0$ over the 1000 simulations for different combinations of N and T (Table 1). Overall, both the bias and the MSE decrease gradually as N or T increases for each of the parameters in $\boldsymbol{\delta}_0$.

Next, we compare various estimates of the standard deviations of the regression coefficients $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$, which are of primary interest. Table 2 shows the sample standard deviation (SD) of the estimates among 1000 simulations, the asymptotic SD approximated

Table 2. Sample standard deviation (Sample SD), asymptotic standard deviation (Asy SD) at $\boldsymbol{\beta}_0$, plug-in standard error (Plug-in SE) by (6) at $\widehat{\boldsymbol{\beta}}$, average computational time (in second) per simulation for Plug-in SE, and coverage probabilities of the confidence intervals for β_0 and β_1 under the nominal level of 95% using Asy SD and Plug-in SE

N	T	Standard deviation/error							Coverage probability			
		Sample SD $\times 10^{-2}$		Asy SD $\times 10^{-2}$		Plug-in SE $\times 10^{-2}$		Average time	Asy SD		Plug-in SE	
		β_0	β_1	β_0	β_1	β_0	β_1		β_0	β_1	β_0	β_1
10^2	5	8.266	4.053	7.979	4.067	7.907	4.044	0.002	94.2	94.9	93.1	94.7
	10	5.856	2.884	5.954	2.936	5.897	2.930	0.003	95.6	95.6	95.3	95.3
	20	4.241	2.083	4.331	2.097	4.323	2.094	0.004	95.3	95.2	95.4	95.1
	50	2.826	1.228	2.793	1.244	2.792	1.243	0.008	95.3	94.5	95.3	94.5
20^2	5	3.978	2.116	4.001	2.085	3.992	2.080	0.005	95.8	94.1	95.1	94.0
	10	2.943	1.447	2.984	1.424	2.980	1.423	0.006	95.9	93.7	96.2	93.7
	20	2.235	1.005	2.170	0.999	2.168	0.999	0.011	94.4	95.0	94.2	95.1
	50	1.373	0.638	1.397	0.634	1.397	0.634	0.021	94.7	95.1	94.8	95.1
50^2	5	1.603	0.805	1.604	0.825	1.605	0.825	0.023	94.0	95.8	94.0	95.8
	10	1.157	0.581	1.194	0.570	1.193	0.570	0.031	96.1	94.5	96.1	94.6
	20	0.866	0.394	0.868	0.401	0.868	0.401	0.047	94.7	96.2	94.5	96.2
	50	0.538	0.246	0.559	0.252	0.559	0.252	0.083	95.5	96.2	95.5	96.2
100^2	5	0.807	0.397	0.803	0.414	0.803	0.414	0.088	94.6	96.5	94.5	96.5
	10	0.584	0.281	0.597	0.286	0.598	0.286	0.118	95.3	95.4	95.4	95.4
	20	0.436	0.194	0.434	0.201	0.435	0.201	0.186	95.9	96.1	95.9	96.1
	50	0.277	0.126	0.279	0.126	0.28	0.126	0.373	94.9	95.1	95.0	95.1
200^2	5	0.406	0.208	0.401	0.207	0.402	0.207	0.520	94.6	95.0	94.6	94.9
	10	0.301	0.143	0.299	0.143	0.299	0.143	0.636	94.5	96.0	94.6	96.0
	20	0.218	0.099	0.217	0.100	0.217	0.100	0.778	94.9	94.9	94.9	94.9
	50	0.140	0.063	0.140	0.063	0.140	0.063	1.465	94.9	94.6	94.9	94.6

by $\sigma_0^{-2} N^{-1} \boldsymbol{X}' \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{X}$, and the plug-in standard error (SE) developed in (6) evaluated $\widehat{\boldsymbol{\beta}}$. The sample SD can be viewed as the gold standard. Both the asymptotic SD and the plug-in SE are close to the sample SD for different combinations of N and T, supporting the results of (6).

We also evaluate the distributions of the estimated regression coefficients $\widehat{\boldsymbol{\beta}}$. Both $\left(\sigma_0^{-2} \boldsymbol{X}' \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{X}\right)^{1/2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\left(\hat{\sigma}^{-2} \boldsymbol{X}' \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1} \boldsymbol{X}\right)^{1/2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converge in distribution to the standard bivariate Gaussian distribution by (6) and the Slutsky's theorem. The last four columns of Table 2 report the coverage probabilities of the confidence intervals for β_0 and β_1 under the nominal level of 95% using the asymptotic SD and the plug-in SE. The confidence intervals for β_0 and β_1 achieve the nominal coverage well for different combinations of N and T. In addition, for different $\boldsymbol{\delta}_0$ and $\boldsymbol{W}$, the results are similar and not shown here to save space.

The last column of Table 1 and the seventh column of Table 2 report the average time (in second) required to obtain the PMLE $\widehat{\boldsymbol{\delta}}$ and the various measures of the variation of $\widehat{\boldsymbol{\beta}}$. The computation is reasonably fast. For example, when N is large (e.g., 200^2), the parameter estimation takes less than one minute per simulation. Moreover, the computation time is empirically linear to the spatial dimension N as the length T of the time series is relatively small. It is worthwhile to point out that the memory usage remains low (e.g., around 2GB when $N = 200^2$ and $T = 50$) in the computation.

Assumption (A.3) only requires the innovations to have finite higher-order moments; in other words, $\boldsymbol{V}_t = (v_{1,t}, \ldots, v_{N,t})'$ is allowed to have heavy tails as long as $E(|v_{j,t}|^{4+\eta}) < \infty$ for some $\eta > 0$. Thus, we also sample $\boldsymbol{V}_t, t = 1, \ldots, T$, independently from the skewed t_5 distribution to check the robustness of our method against heavy-tailed distribution (Fernández and Steel 1998). In addition, we include a comparison with the naive ordinary least squares (OLS) and the state-of-the-art GpGp (Guinness 2021) in estimating the regression coefficients and drawing inference. The results are summarized in Tables S.1 and S.3 for the Gaussian innovations, and in Tables S.2 and S.4 for the skewed t_5 distribution. Our proposed method achieves a smaller mean squared error than OLS, especially for the skewed t_5 innovations. In terms of the coverage probability of the confidence intervals for β_0 and β_1, the OLS has substantial under-coverage. Although GpGp produces more accurate estimators than our method, the difference between the two methods becomes less pronounced when N or T increases. On the other hand, the computation time for GpGp is much longer than that of our method. For example, when $N = 50^2$ and $T = 50$, our method gives results comparable to GpGp and takes 0.96 s to compute, whereas GpGp takes over 265 s.

Overall, the simulation experiments corroborate the theoretical properties of $\widehat{\boldsymbol{\delta}}$ and the computational complexity shown in Sects. 3 and 4, respectively.

6. DATA EXAMPLE: LAND SURFACE TEMPERATURE

As described in Sect. 1, to capture spatial variation in the LST arising from vegetation zonation, amount of the incoming solar radiation energy, and regional differences in environmental conditions, we regress the response variable of LST on the predictor variables of time trend, ecoregion classes, and interactions between the time trend and ecoregions, as well as the environmental covariates of elevation and latitude over $T = 19$ years and $N = 155,900$ image pixels per year. Thus, there are a total of $k = 171$ regression coefficients. To implement the proposed spatio-temporal regression method, we construct a binary spatial weight matrix, $\boldsymbol{W} = (w_{ii'})_{N\times N}$, such that $w_{ii'} = 1$ if cell i' is a first-order neighbor of cell i, and 0 otherwise. We then apply the computational procedure described in Sect. 4.

The majority of the regression coefficients are significant after false discovery rate adjustments, suggesting that, as expected, the mean LST values are different among different ecoregion classes and the time trend in LST varies among ecoregions (Fig. S.6). The upper left panel of Fig. 2 maps the estimated time trend across ecoregions for the LST. Overall, there is an increasing time trend, especially in the southern and southeastern parts of the USA, indicating that these regions are subject to more rapid increases in LST than the rest of the continental US. This finding is consistent with previous findings that South and Southeast United States have warmed up the most in recent decades (Vose et al. 2017; Yan et al. 2020). Tables S.5 and S.6 give the estimated regression coefficients of elevation, latitude, and the intercept (with respect to water), as well as the time trend of the five largest and smallest ecoregions, respectively. The LST tends to decrease with elevation and latitude, which are as expected. The estimates for σ^2, λ, γ, and ρ are 0.6061, 0.0360, 0.7273, and -0.0247, respectively.

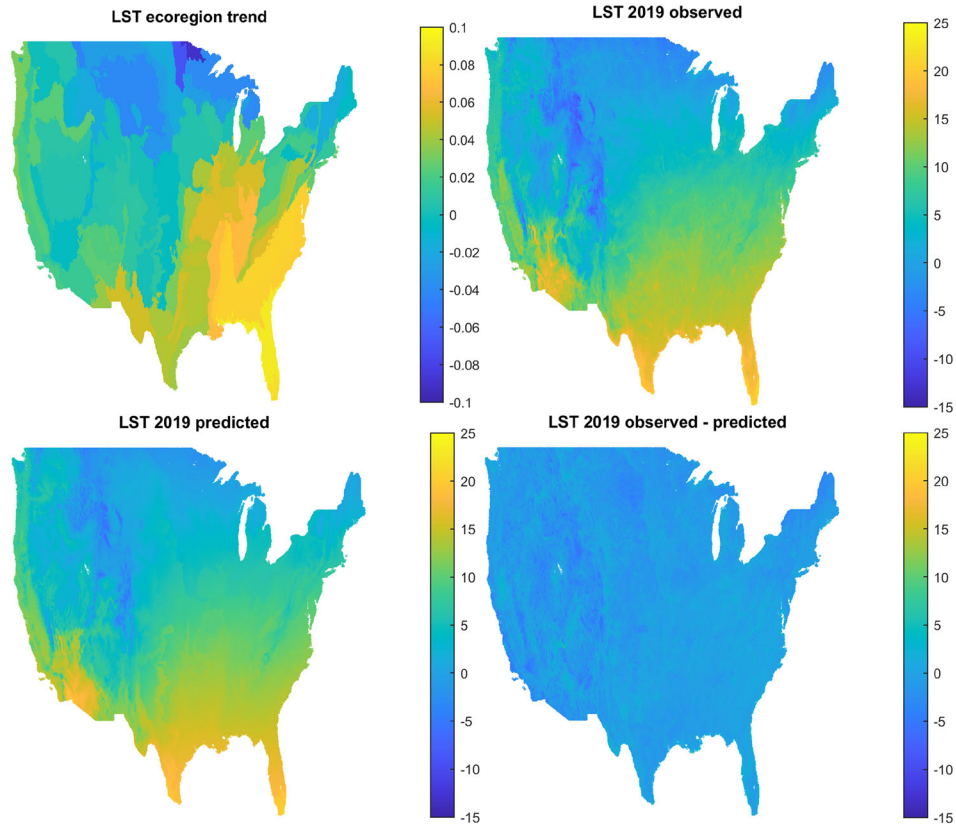

Figure 2. Estimated regression coefficients by ecoregion using data from 2001 to 2019 (upper left panel); Observed land surface temperature (LST) in degree Celsius in 2019 (upper right panel); Predicted LST in 2019 based on model fitting with data from 2001 to 2018 (lower left panel); and the prediction errors of LST in 2019 (lower right panel).

For model diagnostics, we first assess the in-sample model fit. From Figs. S.1 and S.2, the estimated LST in 2001 and 2019 as well as their difference are similar to those in the observations, indicating that our method can well recover the mean function using the covariates. We then evaluate the out-of-sample prediction by fitting the 2001–2018 data and predicting the LST in 2019. Figure 2 suggests that the predicted LST for 2019 match up with the actual observations.

Finally, we compare our method with GpGp with the same set of covariates. The default neighborhood structure and the exponential space-time covariance function are adopted for fitting models using the R package `GpGp`. The estimated LST values from GpGp seem to be quite different from the observed values (right panels of Fig. S.1), possibly due to numerical instability with the large sample size NT. The root mean squared error for the in-sample validation for GpGp is 2.02 and is 1.30 for our method. As for the out-of-sample prediction, the root mean squared prediction error of GpGp and our method are 1.98 and 1.41, respectively. While the computational complexity and the programming languages are

not directly comparable between GpGp and our method, it took GpGp more than four days and our method within two hours to perform the regression analysis.

7. CONCLUSIONS AND DISCUSSION

In this article, we have developed a novel computationally scalable procedure that enables regression analysis of large amounts of spatio-temporal data. We have estimated the model parameters by maximizing a pseudolikelihood function. Asymptotic properties under suitable regularity conditions have been established that enable the computation to be efficient and scalable for parameter estimation and inference.

In spatio-temporal statistics, it is popular to consider using spatio-temporal random effects to account for spatio-temporal dependence (see, e.g., Wikle et al. 2019). The autoregressive approach taken here offers an alternative and it would be interesting to explore the connections between these two seemingly distinct approaches. Our spatio-temporal model is semiparametric in the sense that no explicit distributional assumption is made about the regression innovation. However, the proposed model is not able to readily handle binary observations or count observations. One possible solution is to replace Model (1) with an equation that links the logit (or logarithm) transformation of the expectation of $\boldsymbol{Y}_t$ with $\boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{U}_t$, if $\boldsymbol{Y}_t$ is binary (or count). This is not a trivial task and we will leave it for future research. Moreover, our proposed framework is not designed to model observations whose locations change over time. Although we do not encounter such an issue when analyzing the satellite data, when this framework is applied to data that are not observed at each location across all time points, one could interpolate the "missing" observations via Kalman filter (Katzfuss et al. 2016). In addition, the spatial weight matrix in Model (2) is specified to be time invariant, which is reasonable when the spatial weight matrix is based on the geographical information. However, when one constructs a spatial weight matrix according to certain demographic characteristics or socioeconomic information, the matrix may well change over time. Thus, a possible extension of the current framework is to make the spatial weight matrix time varying. Finally, it is worth investigating the problem of nonconvex minimization of our PMLE. We leave this and other possible extensions for future investigation.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Aeronautics and Space Administration (NASA) under AIST-80NSSC20K0282 and the National Science Foundation (NSF) under DMS-2245906. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA and NSF.

[Received March 2022. Revised November 2022. Accepted November 2022.]

APPENDIX A: NOTATION AND ASSUMPTIONS

We first introduce some notation and conventions. Given an $n \times n$ matrix $\boldsymbol{P} = (p_{ij})_{n\times n}$, we use $\mathrm{tr}(\boldsymbol{P})$ and $\det(\boldsymbol{P})$ to denote the trace and determinant of a square matrix $\boldsymbol{P}$, and we let $vec_D(\boldsymbol{P})$ denote the column vector formed by the diagonal elements of $\boldsymbol{P}$. The (i, j)th element of a matrix $\boldsymbol{P}$ is denoted by $\mathrm{ent}_{ij}(\boldsymbol{P})$. We define $\|\boldsymbol{P}\|_1 = \max_{1\le j\le n} \sum_{i=1}^n |p_{ij}|$ and $\|\boldsymbol{P}\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |p_{ij}|$. We also let $\|\boldsymbol{P}\|_2 = \{\lambda_{\max}(\boldsymbol{P}'\boldsymbol{P})\}^{1/2}$ and $\|\boldsymbol{P}\|_F = \{\mathrm{tr}(\boldsymbol{P}'\boldsymbol{P})\}^{1/2}$ denote the spectral norm and the Frobenius norm, respectively. Let $abs(\boldsymbol{P}) = (|p_{i,j}|)_{n\times n}$. A sequence of $n \times n$ matrix $\boldsymbol{P}_n$ is said to be uniformly bounded in row and column sums (UB), if $\sup_{n\ge1} \|\boldsymbol{P}_n\|_1 < \infty$ and $\sup_{n\ge1} \|\boldsymbol{P}_n\|_\infty < \infty$. We also use $\mathbf{0}$ and $\mathbf{1}$ to denote a matrix or a vector with all elements equal zero and one, respectively. For a real-valued function $f(\boldsymbol{x})$, $\boldsymbol{x} = (\boldsymbol{X}_1, \ldots, x_k)' \in \mathbb{R}^k$, we let $\nabla f(\boldsymbol{x})$ denote the gradient vector and let $\nabla^2 f(\boldsymbol{x})$ denote the Hessian matrix. The partial derivative of f with respect to x_j is denoted by $\partial_{x_j} f(\boldsymbol{x})$ or $\frac{\partial f(\boldsymbol{x})}{\partial x_j}$, whereas the second partial derivative with respect to x_j is denoted as $\partial_{x_j x_j} f(\boldsymbol{x})$ (or $\frac{\partial^2 f(\boldsymbol{x})}{\partial x_j^2}$).

In the following, we provide the regularity conditions for establishing the large sample properties of the PMLE $\widehat{\boldsymbol{\delta}}$.

A.1 The spatial weight matrix $\boldsymbol{W}$ is nonstochastic and symmetric, with zero diagonal elements.

A.2 The parameter space $\Theta_{\boldsymbol{\delta}}$ of $\boldsymbol{\delta} = (\boldsymbol{\beta}', \boldsymbol{\theta}', \sigma^2)'$ is compact and is the product space of $\Theta_{\boldsymbol{\beta}}$, $\Theta_{\boldsymbol{\theta}}$, and $[\underline{\sigma}^2, \bar{\sigma}^2]$, where $\Theta_{\boldsymbol{\theta}}$ is a compact set such that the matrix $\boldsymbol{\mathcal{I}}_N - \lambda \boldsymbol{W}$ is nonsingular and the eigenvalues of $A(\boldsymbol{\theta})$ are less than one in magnitude, while $\Theta_{\boldsymbol{\beta}}$ is a compact subset of $\mathbb{R}^k$. The true value $\boldsymbol{\delta}_0 = (\boldsymbol{\beta}_0', \boldsymbol{\theta}_0', \sigma_0^2)'$ lies in the interior of $\Theta_{\boldsymbol{\delta}}$.

A.3 The vector of innovations $\boldsymbol{V}_t = (v_{1,t}, \ldots, v_{N,t})' \sim iid(0, \sigma_0^2 \boldsymbol{\mathcal{I}}_N)$ and $E(|v_{j,t}|^{4+\eta}) < \infty$ for some $\eta > 0$ for all j, t.

A.4 The precision matrix, infinite sum of power of $\boldsymbol{A}(\boldsymbol{\theta}_0)$, and the design matrix are UB. Namely,

(i) $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} = \boldsymbol{B}(\boldsymbol{\theta})'(\boldsymbol{\Omega}(\boldsymbol{\theta}))^{-1}\boldsymbol{B}(\boldsymbol{\theta})$ and $\boldsymbol{S}(\lambda)^{-1}$ are UB, $\forall \boldsymbol{\theta} \in \Theta$.

(ii) $\sum_{h=1}^\infty abs(\boldsymbol{A}(\boldsymbol{\theta}_0)^h)$ is UB.

(iii) The $N \times k$ design matrix $\boldsymbol{X}_t$ is nonstochastic with elements UB in N and t.

A.5 $\lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X} = \lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}'\boldsymbol{B}(\boldsymbol{\theta})'(\boldsymbol{\Omega}(\boldsymbol{\theta}))^{-1}\boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{X}$ is nonsingular, $\forall \boldsymbol{\theta} \in \Theta$.

A.6 $\liminf_{N\to\infty} N^{-1} \sum_{j=1}^{NT} \nabla^2 f_j(\boldsymbol{\alpha})$ is nonsingular, where $f_j(\boldsymbol{\alpha}) = -\log(\lambda_j(\boldsymbol{\theta})\sigma^{-2}\sigma_0^2) + \lambda_j(\boldsymbol{\theta})\sigma^{-2}\sigma_0^2$, $\boldsymbol{\alpha} = (\boldsymbol{\theta}', \sigma^2)'$, and $\lambda_j(\boldsymbol{\theta})$, $j = 1, \ldots, NT$, are the distinct eigenvalues of $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ in nonincreasing order.

A.7 $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, $\partial_{\theta_i}(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1})$, $\partial^2_{\theta_i\theta_j}(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1})$, and $\partial^3_{\theta_i\theta_j\theta_k}(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1})$ are UB in $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)' \in \Theta$.

A.8 $\lim_{N\to\infty} N^{-1}\boldsymbol{\Omega}_N$ is nonsingular, where

$$\boldsymbol{\Omega}_N = \begin{pmatrix} \mathrm{tr}(\boldsymbol{m}_\lambda^2) & \mathrm{tr}(\boldsymbol{m}_\lambda\boldsymbol{m}_\gamma) & \mathrm{tr}(\boldsymbol{m}_\lambda\boldsymbol{m}_\rho) & -\frac{1}{\sigma_0^2}\mathrm{tr}(\boldsymbol{m}_\lambda) \\ * & \mathrm{tr}(\boldsymbol{m}_\gamma^2) & \mathrm{tr}(\boldsymbol{m}_\gamma\boldsymbol{m}_\rho) & -\frac{1}{\sigma_0^2}\mathrm{tr}(\boldsymbol{m}_\gamma) \\ * & * & \mathrm{tr}(\boldsymbol{m}_\rho^2) & -\frac{1}{\sigma_0^2}\mathrm{tr}(\boldsymbol{m}_\rho) \\ * & * & * & \frac{NT}{\sigma_0^4} \end{pmatrix}, \tag{11}$$

with $\boldsymbol{m}_\lambda$, $\boldsymbol{m}_\gamma$, and $\boldsymbol{m}_\rho$ defined in (S.14) in Supplementary Materials.

REFERENCES

Agirbas E, Koca L, Aytan U (2017) Spatio-temporal pattern of phytoplankton and pigment composition in surface waters of south-eastern Black Sea. Oceanologia 59(3):283–299

Anderson E et al (1999) LAPACK users' guide, 3rd edn. SIAM, Philadelphia

Anselin L (2013) Spatial econometrics: methods and models. Springer, Cham

Asseng S et al (2015) Rising temperatures reduce global wheat production. Nat Clim Chang 5(2):143–147

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. J Roy Stat Soc B 70(4):825–848

Belgiu M, Stein A (2019) Spatiotemporal image fusion in remote sensing. Remote Sens 11(7):818

Blackford LS et al (2002) An updated set of basic linear algebra subprograms (blas). ACM Trans Math Softw 28(2):135–151

Brynjarsdóttir J, Berliner LM (2014) Dimension-reduced modeling of spatio-temporal processes. J Am Stat Assoc 109(508):1647–1659

Buluc A, Gilbert JR (2011) The combinatorial BLAS: design, implementation, and applications. Int J High Perform Comput Appl 25(4):496–509

Case AC (1991) Spatial patterns in household demand. Econometrica 59(4):953–965

Chakraborty T, Hsu A, Manya D, Sheriff G (2020) A spatially explicit surface urban heat island database for the United States: characterization, uncertainties, and possible applications. ISPRS J Photogramm Remote Sens 168:74–88

Chi G, Zhu J (2019) Spatial regression models for the social sciences. SAGE, New York

Chu T, Zhu J, Wang H (2019) Semiparametric modeling with nonseparable and nonstationary spatio-temporal covariance functions and its inference. Stat Sin 29(3):1233–1252

Coppersmith D, Winograd S (1990) Matrix multiplication via arithmetic progressions. J Symb Comput 9(3):251–280

Cressie N (1993) Statistics for spatial data. Wiley, New York

Cressie N, Shi T, Kang EL (2010) Fixed rank filtering for spatio-temporal data. J Comput Graph Stat 19(3):724–745

Cressie N, Wikle CK (2011) Statistics for spatio-temporal data. Wiley, New York

Diffenbaugh NS, Davenport FV, Burke M (2021) Historical warming has increased U.S. crop insurance losses. Environ Res Lett 16(8):084025

Dutilleul PRL (2011) Spatio-temporal heterogeneity: concepts and analyses. Cambridge University Press, Cambridge

Fernández C, Steel MF (1998) On Bayesian modeling of fat tails and skewness. J Am Stat Assoc 93(441):359–371

Finley AO, Banerjee S, Gelfand AE (2012) Bayesian dynamic modeling for large space-time datasets using gaussian predictive processes. J Geogr Syst 14(1):29–47

Fu P, Weng Q (2016) A time series analysis of urbanization induced land use and land cover change and its impact on land surface temperature with landsat imagery. Remote Sens Environ 175:205–214

Gao Z, Ma Y, Wang H, Yao Q (2019) Banded spatio-temporal autoregressions. J Econom 208(1):211–230

Gasparrini A et al (2017) Projections of temperature-related excess mortality under climate change scenarios. Lancet Planet Health 1(9):e360–e367

Gilbert JR, Moler C, Schreiber R (1992) Sparse matrices in MATLAB: design and implementation. SIAM J Matrix Anal Appl 13(1):333–356

Guinness J (2018) Permutation and grouping methods for sharpening Gaussian process approximations. Technometrics 60(4):415–429

Guinness J (2021) Gaussian process learning via Fisher scoring of Vecchia's approximation. Stat Comput 31(3):25

Guo S, Wang Y, Yao Q (2016) High-dimensional and banded vector autoregressions. Biometrika 103(4):889–903

Hanewinkel M et al (2013) Climate change may cause severe loss in the economic value of european forest land. Nat Clim Chang 3(3):203–207

Hu H-W et al (2016) Effects of climate warming and elevated CO2 on autotrophic nitrification and nitrifiers in dryland ecosystems. Soil Biol Biochem 92:1–15

Huang H-C, Cressie N (1996) Spatio-temporal prediction of snow water equivalent using the Kalman filter. Comput Stat Data Anal 22(2):159–175

IPCC (2021) Climate Change 2021: the physical science basis. Contribution of Working Group I to the Sixth assessment report of the intergovernmental panel on climate change, Volume In Press. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press

Johannesson G, Cressie N, Huang H-C (2007) Dynamic multi-resolution spatial models. Environ Ecol Stat 14(1):5–25

Katzfuss M, Guinness J (2021) A general framework for vecchia approximations of gaussian processes. Stat Sci 36(1):124–141

Katzfuss M, Stroud JR, Wikle CK (2016) Understanding the ensemble Kalman filter. Am Stat 70(4):350–357

Kilic E, Stanica P (2013) The inverse of banded matrices. J Comput Appl Math 237(1):126–135

Kressner D (2005) Numerical methods for general and structured eigenvalue problems. Springer, Cham

Lee L-F, Yu J (2015) Estimation of fixed effects panel regression models with separable and nonseparable space-time filters. J Econ 184(1):174–192

Lesk C et al (2017) Threats to North American forests from southern pine beetle with warming winters. Nat Clim Chang 7(10):713–717

Li L, Yang Z (2021) Spatial dynamic panel data models with correlated random effects. J Econ 221(2):424–454

Lu Z, Steinskog DJ, Tjøstheim D, Yao Q (2009) Adaptively varying-coefficient spatiotemporal models. J Roy Stat Soc B 71(4):859–880

Luszczek P (2009) Parallel programming in MATLAB. Int J High Perform Comput Appl 23(3):277–283

Mafteiu-Scai LO (2015) The bandwidths of a matrix: a survey of algorithms. Ann West Univ Timisoara Math Comput Sci 52(2):183–223

Mariella L, Tarantino M (2010) Spatial temporal conditional auto-regressive model: a new autoregressive matrix. Aust J Stat 39(3):223–244

Mueller SE et al (2020) Climate relationships with increasing wildfire in the southwestern US from 1984 to 2015. For Ecol Manag 460:117861

NOAA (2021). State of the climate: global climate report for annual 2020

Nocedal J, Wright SJ (2006) Numerical optimization, 2nd edn. Springer, Cham

Nordman DJ, Lahiri SN (2004) On optimal spatial subsample size for variance estimation. Ann Stat 32(5):1981–2027

Oleson KW et al (2018) Avoided climate impacts of urban and rural heat and cold waves over the U.S. using large climate model ensembles for RCP85 and RCP45. Clim Change 146(3):377–392

Omernik JM, Griffith GE (2014) Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. Environ Manag 54(6):1249–1266

Rue H et al (2017) Bayesian computing with INLA: a review. Ann Rev Stat Appl 4(1):395–421

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc B 71(2):319–392

Saad Y (2003) Iterative methods for sparse linear systems (Second ed.). SIAM

Sarangi C et al (2021) Urbanization amplifies nighttime heat stress on warmer days over the US. Geophys Res Lett 48(24):e2021GL095678

Shen X et al (2022) Effect of shrub encroachment on land surface temperature in semi-arid areas of temperate regions of the northern hemisphere. Agric For Meteorol 320:108943

Sherman M (1996) Variance estimation for statistics computed from spatial lattice data. J Roy Stat Soc B 58(3):509–523

Shi W, Lee L-F (2017) Spatial dynamic panel data models with interactive fixed effects. J Econom 197(2):323–347

Stewart GW (2002) A Krylov-Schur algorithm for large eigenproblems. SIAM J Matrix Anal Appl 23(3):601–614

Thompson R, Hornigold R, Page L, Waite T (2018) Associations between high ambient temperatures and heat waves with mental health outcomes: a systematic review. Public Health 161:171–191

Vecchia AV (1988) Estimation and model identification for continuous spatial processes. J Roy Stat Soc B 50(2):297–312

Vose RS et al (2017) Temperature changes in the United States. pp. 185–206. Climate Science Special Report: Fourth National Climate Assessment, Volume I. U.S. Global Change Research Program

Wan Z (2014) New refinements and validation of the collection-6 MODIS land-surface temperature/emissivity product. Remote sensing of Environment **140**:(36–45)

Westerling AL, Hidalgo HG, Cayan DR, Swetnam TW (2006) Warming and earlier spring increase western U.S. forest wildfire activity. Science 313(5789):940–943

Wikle CK, Zammit-Mangion A, Cressie N (2019) Spatio-temporal statistics with R. Chapman and Hall/CRC, London

Xu G, Liang F, Genton MG (2015) A Bayesian spatio-temporal geostatistical model with an auxiliary lattice for large datasets. Stat Sin 25(1):61–79

Yan Y et al (2020) Driving forces of land surface temperature anomalous changes in North America in 2002–2018. Sci Rep 10(1):6931

Yu J, De Jong R, Lee L-F (2008) Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and t are large. J Econom 146(1):118–134

Zhang B, Cressie N (2020) Bayesian inference of spatio-temporal changes of arctic sea ice. Bayesian Anal 15(2):605–631

Zhang B, Sang H, Huang JZ (2015) Full-scale approximations of spatio-temporal covariance models for large datasets. Stat Sin 25(1):99–114

Zhang W, Yao Q, Tong H, Stenseth NC (2003) Smoothing for spatiotemporal models and its application to modeling muskrat-mink interaction. Biometrics 59(4):813–821

Zhao C et al (2017) Temperature increase reduces global yields of major crops in four independent estimates. Proc Natl Acad Sci 114(35):9326–9331