The New England Journal of Statistics in Data Science Volume 0, 1–15 (2025)
DOI: https://doi.org/10.51387/25-NEJSDS78

Computationally Scalable Bayesian SPDE Modeling for Censored Spatial Responses

Indranil SAHOO*, Suman MAJUMDER, Arnab HAZRA, Ana G. RAPPOLD, and Dipankar BANDYOPADHYAY

Abstract

Observations of groundwater pollutants, such as arsenic or Perfluorooctane sulfonate (PFOS), are riddled with left censoring. These measurements have an impact on the health and lifestyle of the populace. Left censoring of these spatially correlated observations is usually addressed by applying Gaussian processes (GPs), which have theoretical advantages. However, this comes with a challenging computational complexity of $\mathcal{O}(n^3)$, impractical for large datasets. Additionally, a sizable proportion of the left-censored data creates further bottlenecks since the likelihood computation now involves an intractable high-dimensional integral of the multivariate Gaussian density. In this article, we tackle these two problems simultaneously by approximating the GP with a Gaussian Markov random field (GMRF) approach that exploits an explicit link between a GP with Matérn correlation function and a GMRF using stochastic partial differential equations (SPDEs). We introduce a GMRF-based measurement error into the model, which alleviates the likelihood computation for the censored data, drastically improving the computational speed while maintaining admirable accuracy. Our approach demonstrates robustness and substantial computational scalability compared to state-of-the-art methods for censored spatial responses across various simulation settings. Finally, the fit of this fully Bayesian model to the concentration of PFOS in groundwater available at 24,959 sites across California, where 46.62% responses are censored, produces prediction surface and uncertainty quantification in real-time, thereby substantiating the applicability and scalability of the proposed method. Code for implementation is made available via `GitHub`.

keywords and phrases: Censored high-dimensional spatial data, Gaussian Markov random field, Markov chain Monte Carlo, Measurement error model, Perfluorooctane sulfonate, Stochastic partial differential equation.

1. INTRODUCTION

The statistical literature has regularly analyzed censored data since the late 1900s, with the earliest instance of a statistical analysis of censored data being in 1766 [32]. Measurements are often censored due to limitations of measuring instruments, physical inability to acquire data, human error, or similar.

While most analyses on censored data focus on right-censoring [20], some applications like analyzing the concentration of contaminants such as arsenic or per-and polyfluoroalkyl substances (PFAS) in groundwater call for utilizing methods involving left-censored data. This kind of censoring is prevalent in environmental monitoring and has applications in environmental and public health, epidemiology, hydrology, agriculture, and more. Typically, these applications involve geostatistical data, where the measurements are censored because they fall below the minimum detection limit (MDL) of the measuring instrument. Early applications either remove censored observations, replace them with makeshift values such as MDL or MDL/2, or impute them with the mean or median of observed responses. Such ad-hoc imputations can result in biased estimates of the overall spatial variability, as demonstrated by [13].

Recent approaches for statistical inference of spatially distributed censored data overwhelmingly considered the Expectation-Maximization (EM) algorithm [25]. [28] proposed an exact maximum likelihood (ML) estimation of model parameters under censoring, called 'CensSpatial', using the Stochastic Approximation of the Expectation Maximization [SAEM; 11] algorithm. To tackle the computational complexities arising from censored likelihoods in correlated data, Monte Carlo approximations have been employed, both within the classical framework [40, 30], and the Bayesian paradigm [10, 41, 34]. For example, [36] introduced a semi-naive approach that utilizes an iterative algorithm and variogram estimation to determine imputed values at locations where data are censored. Finally, various data augmentation techniques have been proposed to facilitate analysis of spatially correlated censored data [1, 19, 13, 38]. However, the scalability of the suggested approaches is restricted, rendering them unsuitable for analyzing large spatial datasets featuring censoring, a common occurrence in contemporary scientific research.

[*] Corresponding author.

Gaussian processes [GPs; 37] are heavily used for modeling continuous spatial data due to their several theoretical and computational advantages: the likelihood involves only the first two moments, conditional independence and zeros in the underlying precision matrix are equivalent, and various linear algebraic results are well-known in the literature that are required for computing covariance matrices [14]. However, once the number of spatial sites is large and data at a large proportion of sites are censored, likelihoods based on the underlying GPs involve an intractable high-dimensional integral of a multivariate Gaussian density. This paper aims to overcome the computational challenges inherent to censored likelihoods for high-dimensional spatial settings through a combined application of two key steps:

1. We focus on a fully Bayesian method for censored point referenced data, where the underlying GP is approximated as a Matérn-like Gaussian Markov random field [GMRF, 33]. The GMRF is obtained as the solution of a stochastic partial differential equation [SPDE, 24] on a fine mesh, which yields a sparse precision matrix of the underlying basis function coefficients. This sparse spatial structure then allows for fast and scalable Bayesian computations.
2. We consider a GMRF-based measurement error model incorporating a nugget effect in formulating the underlying GMRF that expedites the imputation process for the censored observations. This inclusion effectively reduces the computational burden associated with censored likelihoods [17, 45, 46].

We draw inferences regarding model parameters using an adaptive Markov Chain Monte Carlo (MCMC) sampling approach, where we use random walk Metropolis-Hastings (MH) steps within Gibbs sampling. Extensive simulations demonstrate the scalability and performance of the proposed methodology in comparison to the 'CensSpatial' algorithm and a 2-dimensional B-splines basis function model across varying degrees of censoring and varying grid sizes. While traditional local likelihood methods [44, 35], when applied with Vecchia's approximation [42], are often considered ideal for handling high-dimensional spatial datasets, their implementation becomes computationally challenging in the presence of censoring. Specifically, they also require the evaluation of high-dimensional integrals, which renders these methods infeasible for large spatial datasets with censoring. The idea of a GMRF-based measurement error model has been explored in the context of spatial extremes [9, 16], where replications of the underlying spatial processes are available and censoring a portion of the data is artificial. However, per our knowledge this modeling strategy has not been explored yet for high-dimensional censored spatial data without replications. Although the lack of temporal replications typically leads to unstable computations, the proposed stable and scalable computational framework is tailored explicitly for handling censored spatial data without requiring temporal replications. Furthermore, unlike previous studies involving the GMRF-based measurement error model, a novel feature of the proposed approach is the inclusion of spatial predictions.

PFAS constitute a substantial group of synthetic compounds absent in natural environments, notable for their resistance to heat, water, and oil. PFAS are persistent in the environment, can accumulate within the human body over time, and are toxic at relatively low concentrations [43]. Exposure to elevated levels of PFAS can lead to various adverse health outcomes, including developmental issues during pregnancy, cancer, liver impairment, immune system dysfunction, thyroid disruption, and alterations in cholesterol levels [12]. Due to their chemical robustness, PFAS endure in the environment and are resistant to degradation. Contamination of drinking water with PFAS occurs through the use or accidental spillage of products containing these substances onto land or into waterbodies [18]. PFAS present a significant public health risk, with elevated concentrations identified at 3,186 locations across the United States as of August 2023. In response, the U.S. Environmental Protection Agency (EPA) introduced new safety standards on April 10, 2024, setting permissible limits between 4.0 and 10.0 parts per trillion (ppt; also expressed as nanograms per litre (ng/L)) for six specific PFAS chemicals (including PFOS) in drinking water [31]. A recent study [2] estimated that PFAS in publicly accessible drinking water could affect as many as 200 million people across the United States. Along similar lines, a robust Bayesian hierarchical approach was proposed [39] to accommodate left-censored PFAS responses; however, the model was implemented on a limited number of sample site locations. As such, the review of existing literature highlights the necessity for further investigation into PFAS occurrences in groundwater, alongside the development of fast and efficient approaches to handle large-scale left-censored spatial data in real-time. Motivated by data on PFAS concentrations collected by the Groundwater Ambient Monitoring and Assessment (GAMA) program [26] across the state of California, we develop our Bayesian scalable model for spatially-referenced left-censored PFAS responses in an attempt to provide a more accurate quantification of the groundwater contamination within the state. These data, collected by GAMA since 2019, allow thorough quality assessments of water sources and establish safety thresholds for select PFAS constituents. Thus, our analysis can identify possible hotspots of higher PFAS concentration, providing insights for further study of impacts on public health.

The subsequent sections of the paper are organized as follows. In Section 2, we provide details regarding the dataset on the groundwater levels of PFAS within California, along with some exploratory analyses. We outline our methodology and related computational details in Section 3 and test its scalability and predictive performance on simulated datasets in Section 4. In Section 5, we apply the proposed methodology to the PFAS dataset and report the findings. We conclude with a brief discussion in Section 6.

2. MOTIVATING PFAS DATA

The groundwater PFAS data for the state of California are available online at the website GAMA Groundwater Information Systems under the label `Statewide PFOS Data`. In this paper, we focus specifically on the measurements of the chemical substance known as Perfluorooctane sulfonate (PFOS). The dataset contains 24,959 measurements (in ng/L) of PFOS concentration and their locations (in longitudes and latitudes), as well as indicators of whether the observations are censored and the corresponding censoring limits within California. Almost half of the measurements (46.62%) are censored observations, with varying degrees of censoring limits.

Figure 1 shows transformed PFOS concentration measurements after transforming the raw PFOS by $g(\text{PFOS}) = \log(1 + \log(1 + \text{PFOS}))$ at the 24,959 irregularly sampled spatial locations across the state of California, prompting an approximate spatial inference model to be employed, which also accounts for the considerable proportion of censored observations. Most observation sites are located in densely populated areas on the coast. The censored observations are presented as tiny black dots in Fig 1. The censored observations are distributed across the entire spatial domain, rather than being concentrated in a specific area. While the majority of measurements are below 150 ng/L, some values reach as high as 1,330,000 ng/L, and approximately 47% of the observed concentrations exceed the new safety limit established by the EPA.

The histogram of the raw non-censored PFOS observations is presented in the left panel of Figure 2. The raw data exhibit a highly positively skewed nature. Thus, a stationary Gaussian process assumption naturally becomes questionable, even after considering a spatially-smooth mean surface with covariates like longitude and latitude, commonly used for estimating spatial trends [15, 34]. Following an exploration of different transformations of the raw data such that the histogram behaves in an approximately bell-shaped fashion, we identify that the iterated log-transformation $g(\text{PFOS}) = \log(1+\log(1+\text{PFOS}))$ performs reasonably well; the histogram of the transformed PFOS data is presented in the middle panel of Figure 2. We further explore the effects of the natural covariates longitude and latitude on the transformed data; following a simple linear regression, we obtain the residuals, and their histogram is presented in the right panel of Figure 2. This histogram is reasonably bell-shaped, and thus, we model this transformed PFOS data using a Gaussian process framework with a regression structure for the mean process where we allow longitude and latitude as covariates.

We further explore spatial correlation using a variogram analysis of the residuals (scaled by their sample standard deviation) discussed above. The sample semivariogram at distance d is defined as

$$\widehat{\gamma}(d) = \frac{1}{2N(d)} \sum_{i=1}^{n} \sum_{j=1}^{i} w_{ij}(d)(R(\boldsymbol{s}_i) - R(\boldsymbol{s}_j))^2,$$

where, $R(\boldsymbol{s}_i)$ and $R(\boldsymbol{s}_j)$ are the residuals at spatial sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, $w_{ij}(d) = 1$ if $d_{ij} \in (d - h, d + h)$ and $w_{ij} = 0$ otherwise, d_{ij} being the distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. Also, $N(d)$ is the number of pairs with $w_{ij}(d) = 1$. The sample semivariogram, presented in Figure 3, indicates the presence of spatial correlation and possible nugget effects [4]. We fit an isotropic Matérn spatial correlation function, with its smoothness parameter set to one, plus a nugget effect, given by

$$\rho(\boldsymbol{s}_i, \boldsymbol{s}_j) \equiv \rho(d) = \gamma \frac{d}{\phi} \kappa_1 \left(\frac{d}{\phi} \right) + (1 - \gamma) \mathbb{1}(\boldsymbol{s}_i = \boldsymbol{s}_j), \quad (2.1)$$

where d is the Euclidean distance between locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j, \phi > 0$ is the range parameter, $\gamma \in [0, 1]$ is the ratio of spatial to total variation, $\kappa_1(\cdot)$ is the modified Bessel function of second kind with degree 1, and $\mathbb{1}(\cdot)$ is the indicator function. The fitted population semivariance indicates a reasonable fit to the sample semivariogram. While these exploratory analyses are based on non-censored observations only, they indicate a need for proper spatial modeling after considering the censored nature of a large proportion of the data. Specifically, most of the observations near the eastern regions of the study domain are censored, and ignoring them in the spatial prediction would lead to poor spatial prediction for the nearby regions.

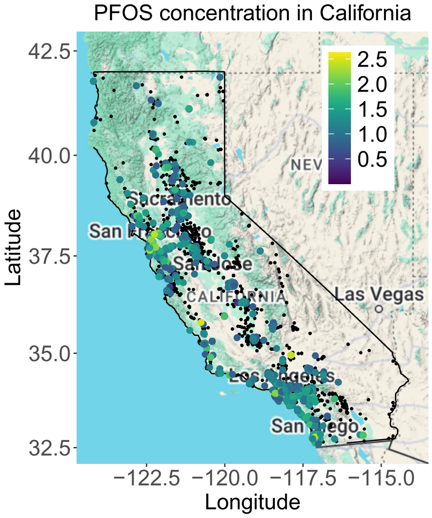

Figure 1: Concentrations of (transformed) PFOS, using the transformation $g(\text{PFOS}) = \log(1 + \log(1 + \text{PFOS}))$, measured at 24,959 irregularly-sampled spatial locations across the state of California (in ng/L). The tiny black dots indicate the sites with censored data.

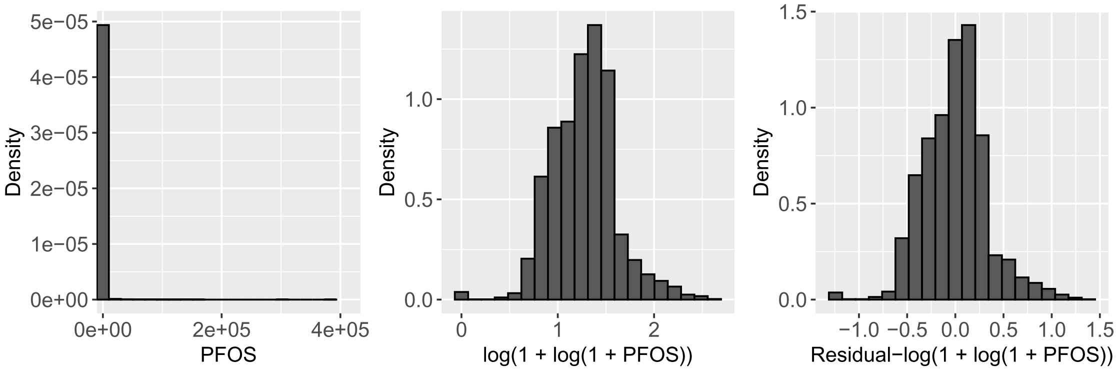

Figure 2: Pictoral representation of the raw PFOS responses. *Left panel:* Histogram of raw concentrations of PFOS across sites where the data are not censored. *Middle:* Histogram of non-censored PFOS concentrations after the transformation $g(\text{PFOS}) = \log(1 + \log(1 + \text{PFOS}))$. *Right panel:* Histogram of the residuals obtained after regressing non-censored transformed PFOS observations to longitude and latitude via a linear model.

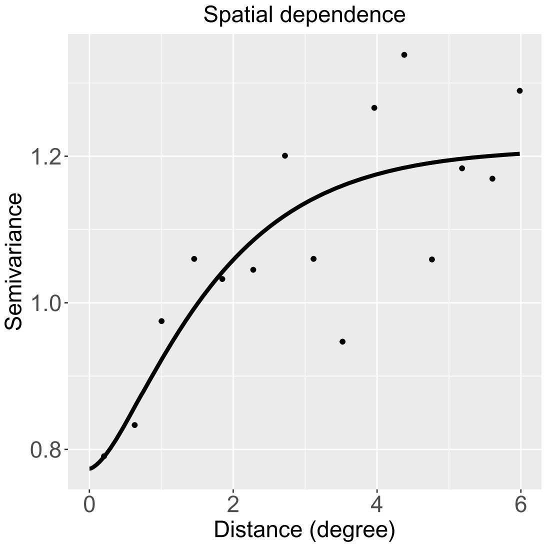

Figure 3: Sample semivariogram of residuals obtained after regressing non-censored transformed PFOS observations to longitude and latitude as a function of distance (dots). The overlapped solid line represents the fitted population semivariance obtained from (2.1).

3. METHODOLOGY

Let $Y(\boldsymbol{s})$ represent transformed PFOS concentration at a spatial location $\boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^2$, where $\mathcal{D}$ represents the spatial domain of interest, i.e., the entire state of California in our case. We model $Y(\boldsymbol{s})$ as

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})^T\boldsymbol{\beta} + \tau^{-1/2}Z(\boldsymbol{s}),$$

where, $X(\boldsymbol{s}) = [X_1(\boldsymbol{s}), \ldots, X_p(\boldsymbol{s})]^T$ denotes the vector of p covariates at location $\boldsymbol{s}$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ is a vector of unknown regression parameters and $\tau > 0$ is a spatially-constant precision parameter. Given the absence of meaningful covariates in our data, we choose $\boldsymbol{X}(\boldsymbol{s}) = [1, \text{longitude}(\boldsymbol{s}), \text{latitude}(\boldsymbol{s})]^T$ for our analysis. We assume that $Z(\cdot)$ is a standard (mean zero and variance one at each site) GP with an isotropic Matérn spatial correlation plus a nugget effect, given by (2.1). While the incorporation of the nugget effect is justified by our exploratory analysis (nonzero semivariance at origin), it effectively addresses the issue of censoring in the response, thereby circumnavigating the computational burden occurring due to censored likelihoods [17, 45, 46]. Here, we fix the smoothness parameter of the Matérn correlation of the purely spatial component of (2.1) to one. In practice, it is difficult to estimate the smoothness parameter from the data; hence, it is generally fixed. Besides, we later build a stochastic partial differential equation-based approximation of the Matérn correlation structure, where fixing the smoothness parameter to one is a standard choice [9, 16].

Suppose the data are observed (either censored or non-censored) at the set of sites $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$. In matrix notations, the spatial linear model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \tau^{-1/2}\boldsymbol{Z}, \tag{3.1}$$

where $\boldsymbol{Y}_{(n\times1)}$ is the response vector, $X_{(n\times p)}$ is the matrix of covariates, $\boldsymbol{\beta}_{(p\times1)}$ is the vector of regression coefficients, and $\boldsymbol{Z}_{(n\times1)} \sim \text{MVN}(\boldsymbol{0}, \gamma\boldsymbol{\Sigma} + (1-\gamma)\boldsymbol{I}_n)$, where $\boldsymbol{\Sigma}$ is the Matérn correlation matrix, and $\boldsymbol{I}_n$ denotes the identity matrix of order n. By construction and the PFAS dataset, $\boldsymbol{\Sigma}$ is non-singular, and $\boldsymbol{X}$ has full rank.

In a spatial censored linear (SCL) model, it is further assumed that $Y(\boldsymbol{s})$ is not fully observed at all spatial locations. Motivated by the dataset considered, we assume $Y(\cdot)$ to be left-censored at sites $\mathcal{S}^{(c)} = \{\boldsymbol{s}_1^{(c)}, \ldots, \boldsymbol{s}_{n_c}^{(c)}\} \subset \mathcal{S}$ and the corresponding censoring levels be $\mathcal{U} = \{u_1, \ldots, u_{n_c}\}$. However, a similar approach can be applied if the response is right or interval-censored. We define the censoring indicator $\delta(\boldsymbol{s})$ as

$$\delta(\boldsymbol{s}) = \begin{cases} 1, & \text{if } Y(\boldsymbol{s}) \text{ is censored at site } \boldsymbol{s}, \\ 0, & \text{otherwise,} \end{cases}$$

and the vector of censored observations as $\boldsymbol{v} = [Y(\boldsymbol{s}_i) : \delta(\boldsymbol{s}_i) = 1]^T \equiv [Y(\boldsymbol{s}_1^{(c)}), \ldots, Y(\boldsymbol{s}_{n_c}^{(c)})]^T$. Then, for censored spatial data, the likelihood is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \int_{\boldsymbol{v} \le \boldsymbol{u}} f_{\mathrm{MVN}}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}[\gamma\boldsymbol{\Sigma} + (1-\gamma)\boldsymbol{I}_n])\, d\boldsymbol{v}, \quad (3.2)$$

where the integral is over the censored responses $\{\boldsymbol{y} : y(\boldsymbol{s}_i) \le u_i \text{ if } \boldsymbol{s}_i \in \mathcal{S}^{(c)}\}$ and $f_{\mathrm{MVN}}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A version of this likelihood has been studied in [34].

We can rewrite the model (3.1) as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \tau^{-1/2}\boldsymbol{W} + \tau^{-1/2}\boldsymbol{\varepsilon}$, where $\boldsymbol{W} \sim \mathrm{MVN}(\boldsymbol{0}, \gamma\boldsymbol{\Sigma})$ and $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\boldsymbol{0}, (1-\gamma)\boldsymbol{I}_n)$, i.e., the components of $\boldsymbol{\varepsilon}$ are independently and identically distributed as $\mathrm{N}(0, (1-\gamma))$. Hence, given $\boldsymbol{W}$, the components of $\boldsymbol{Y}$ are independent and follow univariate normal distributions. Thus, (3.2) simplifies to

$$\mathcal{L}(\boldsymbol{\theta}) = \int \prod_{i:\boldsymbol{s}_i \notin \mathcal{S}^{(c)}} \frac{1}{\tau^{-1/2}(1-\gamma)^{1/2}} \phi\left(\frac{y_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} - w_i}{\tau^{-1/2}(1-\gamma)^{1/2}}\right) \times \prod_{i:\boldsymbol{s}_i \in \mathcal{S}^{(c)}} \Phi\left(\frac{u_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} - w_i}{\tau^{-1/2}(1-\gamma)^{1/2}}\right) f_{\mathrm{MVN}}(\boldsymbol{w}; \boldsymbol{0}, \tau^{-1}\gamma\boldsymbol{\Sigma}) d\boldsymbol{w} \quad (3.3)$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal distribution and density functions, respectively. Exploiting the conditional independence structure and the univariate normal distribution structure, the censored components can be easily imputed using sampling from truncated univariate normal distributions, bypassing computationally expensive multivariate imputations. The nugget parameter plays a critical role in this framework by allowing a hierarchical representation of the proposed model, where the data layer becomes conditionally independent across spatial locations. When $\gamma = 1$, the model loses this hierarchical representation, as the data, conditioned on the latent process $W(\cdot)$, are no longer independent across the space. Hence, the simplification of (3.2) using (3.3) is not possible. In that case, the multivariate imputation cannot be replaced with univariate imputations. For real datasets where γ is unknown, restricting the parameter space to $\gamma \in [0, 1)$ leads to equivalent Bayesian inferences in case of a continuous prior; hence, we assume this restriction throughout the rest of the paper.

3.1 Approximation of the Matérn Gaussian Process

Although we bypass the problem of evaluating high-dimensional integrals in (3.2) by introducing a latent variable and exploiting conditional independence, evaluating the likelihood in (3.3) remains a computationally taxing problem since $\boldsymbol{W}$ is a vector of large dimension (same as $\boldsymbol{Y}$). For that purpose, we take cues from [16] for an approximation strategy of the Matérn GP. To ensure computational efficiency, we choose to approximate the Gaussian process $W(\cdot)$ with a GP $\tilde{W}(\cdot)$, constructed from a Gaussian Markov random field (GMRF) defined on a finite mesh, thereby circumventing the computational overhead associated with the dense correlation matrix inherent in the exact Matérn GP defined by (2.1). This strategy capitalizes on the direct correspondence between continuous-space Matérn GP with dense covariance matrices and GMRFs with sparse precision matrices [24], which yields an approximate data process

$$Y(\boldsymbol{s}) \approx \boldsymbol{X}(\boldsymbol{s})^T\boldsymbol{\beta} + \tau^{-1/2}\tilde{W}(\boldsymbol{s}) + \tau^{-1/2}\varepsilon(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}.$$

For $\gamma = 1$, the Gaussian Matérn process $Z(\cdot)$ can be obtained as a solution to the linear Stochastic Differential Equation (SDE)

$$(\phi^{-2} - \Delta)W(\boldsymbol{s}) = 4\pi\phi^{-2}\mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathbb{R}^2, \quad (3.4)$$

where $\mathcal{W}(\cdot)$ is a Gaussian white noise process, and Δ is the Laplacian. The solution $W(\boldsymbol{s})$ to the SPDE can be effectively approximated through finite-element methods [8] applied to a triangulated mesh defined within a bounded region of $\mathbb{R}^2$, where the triangulation is formed through a refined Delaunay triangulation process [7]. In practical applications, the mesh can be easily constructed using the (currently depreciated) `inla.mesh.2d` function, implemented in the `R` package `INLA` (www.r-inla.org) or the `fm_mesh_2d_inla` function in the `R` package `fmesher` (https://cran.r-project.org/package=fmesher); see [22] for more details. The left panel of Figure 4 depicts the mesh utilized in the data application discussed in Section 5.

We choose the inner extension distance of 0.15° and the outer extension distance of 2.5°; these choices are set using the argument `offset`. We set the minimum allowed distance between points (`cutoff`) to 0.25°. The largest permitted triangle edge lengths in the inner and outer extensions (`max.edge`) are set to 0.25° and 1°. While any specific choices are not listed in the current literature, choices of the arguments are problem-specific. Some discussions are in [21]. In our case, these choices reasonably approximate the true Matérn correlation function, as seen in the right panel of Figure 4. We set the smoothness parameter to one as it is difficult to estimate this parameter from purely spatial data (where no time replications are available), like in our case. Setting the smoothness parameter to one reduces the stochastic partial differential equation (SPDE) introduced in [24] to a stochastic differential equation (SDE), as shown in (3.4). However, since the method is commonly referred to as the SPDE approach, we continue to use this terminology for consistency, even in this simplified case.

Let $\mathcal{S}^* = \{\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_N^*\}$ denote the set of mesh nodes. We construct a finite-element solution by writing $\tilde{W}(\boldsymbol{s}) =$

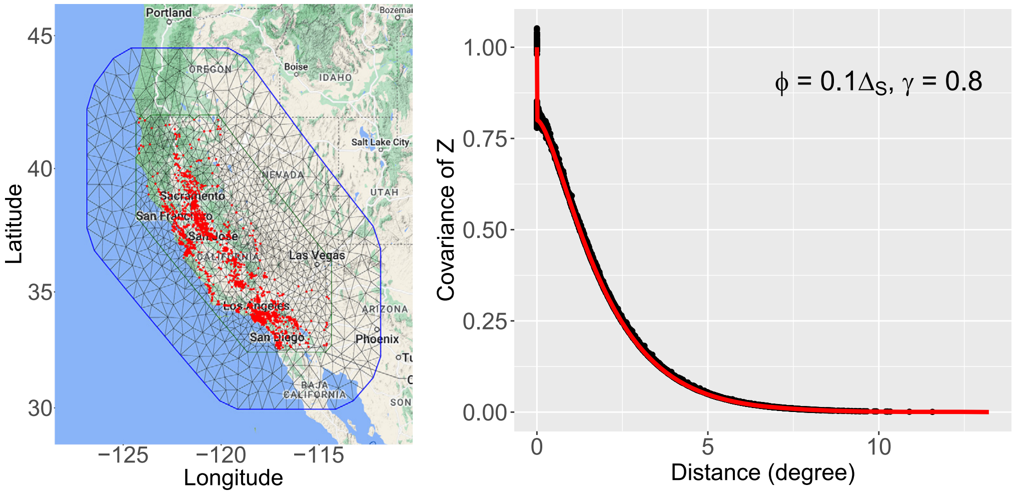

Figure 4: Triangulated mesh over California that approximates the spatial SDE process $Z(\cdot)$, where the black lines represent the edges, along with their corners representing the mesh nodes, and the red dots mark the observation locations (left panel); Comparison of the true Matérn correlation (solid line) and the pairwise covariances between two spatial locations obtained by the SDE approximation (points) as a function of distance (right panel). The parameters are set to $\phi = 0.1\Delta_{\mathcal{S}}$, where $\Delta_{\mathcal{S}}$ is the maximum spatial distance between two locations in the domain, and $\gamma = 0.8$.

$\sum_{j=1}^{N} \zeta_j(\boldsymbol{s})W_j^*$, and plugging it in (3.4) in place of $W(\cdot)$. Here, $\{\zeta_j(\cdot)\}$ are piecewise linear and compactly-supported "hat" basis functions defined over the mesh, and $\{W_j^*\}$ are normally distributed weights defined for each basis function (that is, one for each mesh node in $\mathcal{S}^*$). Then, $\boldsymbol{W}^* = [W_1^*, \ldots, W_N^*]^T \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{Q}_\phi^{-1})$, where the $(N \times N)$-dimensional precision matrix $\boldsymbol{Q}_\phi$ can be written as

$$\boldsymbol{Q}_\phi = \frac{\phi^2}{4\pi}\left[\frac{1}{\phi^4}\boldsymbol{D} + \frac{2}{\phi^2}\boldsymbol{G}_1 + \boldsymbol{G}_2\right], \tag{3.5}$$

where $\boldsymbol{D}$, $\boldsymbol{G}_1$, and $\boldsymbol{G}_2$ are sparse $(N \times N)$-dimensional finite-element matrices that can be obtained as follows. The matrix $\boldsymbol{D}$ is diagonal with its jth diagonal entry $D_{j,j} = \langle \zeta_j(\cdot), 1\rangle$, where $\langle f, g\rangle = \int f(\boldsymbol{s})g(\boldsymbol{s})d\boldsymbol{s}$ denotes an inner product. Similarly, $\boldsymbol{G}_1$ has the elements $G_{1;j,k} = \langle \nabla\zeta_j(\cdot), \nabla\zeta_k(\cdot)\rangle$ and $\boldsymbol{G}_2 = \boldsymbol{G}_1\boldsymbol{D}^{-1}\boldsymbol{G}_1$. Efficient computation of these sparse matrices is implemented using the function `inla.mesh.fem` from the `R` package `INLA`. For further theoretical details, see [3] and [23].

In order to map the spatial random effects $\boldsymbol{W}^*$ (defined across mesh nodes) back to the observation locations $\mathcal{S}$, we use an $(n \times N)$-dimensional projection matrix $\boldsymbol{A}$. The $(i, j)^{th}$ element of this matrix corresponds to $\zeta_j(\boldsymbol{s}_i)$ for every spatial location $\boldsymbol{s}_i \in \mathcal{S}$ and mesh node $\boldsymbol{s}_j^* \in \mathcal{S}^*$, allowing us to compute $\boldsymbol{A}\boldsymbol{W}^*$, the projection of $\boldsymbol{W}^*$ at the data locations. The generation of the matrix $\boldsymbol{A}$ is carried out through the function `inla.spde.make.A` within the `R` package `INLA`.

We now approximate $\boldsymbol{W}$ to include the nugget effect $\gamma \in [0, 1]$ as

$$\tilde{W} = \sqrt{\gamma}\boldsymbol{A}\boldsymbol{W}^*,$$

which has the covariance matrix

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{W}}} = \gamma \boldsymbol{A}\boldsymbol{Q}_\phi\boldsymbol{A}^\intercal.$$

This subsequently leads to the approximation of (3.3) to

$$\mathcal{L}(\boldsymbol{\theta}) = \int \prod_{i:\boldsymbol{s}_i \notin \mathcal{S}^{(c)}} \frac{1}{\tau^{-1/2}(1-\gamma)^{1/2}}\phi\left(\frac{y_i - \boldsymbol{x}_i^\intercal\boldsymbol{\beta} - w_i}{\tau^{-1/2}(1-\gamma)^{1/2}}\right) \times \prod_{i:\boldsymbol{s}_i \in \mathcal{S}^{(c)}} \Phi\left(\frac{u_i - \boldsymbol{x}_i^\intercal\boldsymbol{\beta} - w_i}{\tau^{-1/2}(1-\gamma)^{1/2}}\right) f_{\text{MVN}}(\boldsymbol{w}; \boldsymbol{0}, \tau^{-1}\boldsymbol{\Sigma}_{\tilde{\boldsymbol{W}}})d\boldsymbol{w}. \tag{3.6}$$

The SPDE approach yields a precise approximation of the true correlation structure; see the right panel of Figure 4. Leveraging the sparsity of the matrix $\boldsymbol{Q}_\phi^{-1}$, we can facilitate rapid Bayesian computations.

3.2 Final Hierarchical Model

We write the vector of the final approximate data process, $\tilde{Y}(\cdot)$, evaluated at $\mathcal{S}$ by $\tilde{\boldsymbol{Y}} = [\tilde{Y}(\boldsymbol{s}_1), \ldots, \tilde{Y}(\boldsymbol{s}_n)]^T$ and define a rescaled random effects vector defined at mesh nodes by $\tilde{\boldsymbol{W}}^* = \sqrt{\gamma/\tau}\boldsymbol{W}^*$. By introducing a latent process $W(\boldsymbol{s})$ (Section 3) and approximating it by $\tilde{\boldsymbol{W}} = \sqrt{\gamma}\boldsymbol{A}\boldsymbol{W}^*$ (Section 3.1), we avoid the need for multiple imputations, and the hierarchical model for $\tilde{\boldsymbol{Y}}$ can then be written as

$$\begin{aligned} \tilde{\boldsymbol{Y}} | \tilde{\boldsymbol{W}}^* &\sim \text{MVN}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\tilde{\boldsymbol{W}}^*, \tau^{-1}(1-\gamma)\boldsymbol{I}_n\right), \\ \tilde{\boldsymbol{W}}^* &\sim \text{MVN}(\boldsymbol{0}, \gamma\tau^{-1}\boldsymbol{Q}_\phi^{-1}), \\ \{\boldsymbol{\beta}, \tau, \phi, \gamma\} &\sim \pi(\boldsymbol{\beta}|\tau) \times \pi(\tau) \times \pi(\phi) \times \pi(\gamma). \end{aligned} \tag{3.7}$$

The first layer of (3.7) models the "true" data, which, when observed, is the same as the observed data. When the observed data is censored, the only information we have about the "true" data is that it is smaller than the censoring limit. These two situations are represented by the first two terms within the integral in (3.6), where the first term (normal density, except the product) conveys the contribution of the "true" data that is observed, and the second term (normal distribution function, except the product) presents the contribution of the censored observations to the data likelihood. The second layer of (3.7) corresponds to the latent spatial random effect, whose contribution is represented by the third term in (3.6). The integral in (3.6) comes from integrating out the spatial random effect to get the observed data likelihood solely, which isn't necessary under a hierarchical modeling. Here, the last layer of the model indicates prior choices for the model parameters that we discuss in Section 3.4. Instead of the likelihood based on (3.1), we fit the approximate data process (3.7) to the actual observation process with the likelihood function in (3.2) approximated as the one in (3.6).

3.3 Prediction

Let $\mathcal{S}^{(0)} = \{\boldsymbol{s}_1^{(0)}, \ldots, \boldsymbol{s}_{n_0}^{(0)}\} \subset \mathcal{D}$ denote a set of n_0 prediction sites, and define $\tilde{\boldsymbol{Y}}^{(0)} = [\tilde{Y}(\boldsymbol{s}_1^{(0)}), \ldots, \tilde{Y}(\boldsymbol{s}_{n_0}^{(0)})]^T$. Also, let $\boldsymbol{X}^{(0)}$ denote the $(n_0 \times p)$-dimensional design matrix, with its ith row $\boldsymbol{X}(\boldsymbol{s}_i^{(0)}), i = 1, \ldots, n_0$ denoting the vector of covariates at prediction location $\boldsymbol{s}_i^{(0)}$. For mapping the (scaled) spatial random effects $\tilde{\boldsymbol{W}}^*$ (defined across mesh nodes) to the prediction locations $\mathcal{S}^{(0)}$, we use an $(n_0 \times N)$-dimensional projection matrix $\boldsymbol{A}^{(0)}$. The $(i,j)^{th}$ element of this matrix corresponds to $\zeta_j(\boldsymbol{s}_i^{(0)})$ for every spatial location $\boldsymbol{s}_i^{(0)} \in \mathcal{S}^{(0)}$ and mesh node $\boldsymbol{s}_j^* \in \mathcal{S}^*$, allowing us to compute $\boldsymbol{A}^{(0)}\tilde{\boldsymbol{Z}}^*$, the projection of $\tilde{\boldsymbol{W}}^*$ at $\mathcal{S}^{(0)}$. Then, given $\tilde{\boldsymbol{W}}^*$, the conditional distribution of $\tilde{\boldsymbol{Y}}^{(0)}$ is

$$\tilde{\boldsymbol{Y}}^{(0)}|\tilde{\boldsymbol{W}}^* \sim \text{MVN}\left(\boldsymbol{X}^{(0)}\boldsymbol{\beta} + \boldsymbol{A}^{(0)}\tilde{\boldsymbol{W}}^*, \tau^{-1}(1-\gamma)\boldsymbol{I}_n\right) \tag{3.8}$$

3.4 Computational Details

Inference concerning the model parameters is conducted through Markov chain Monte Carlo (MCMC) sampling, implemented in R. Given the computational dependence on prior selections for the model parameters, we first specify these priors. Whenever feasible, we opt for conjugate priors and employ Gibbs sampling to update them iteratively. When prior conjugacy is unavailable, we resort to random walk Metropolis-Hastings (MH) steps for parameter updates. During the burn-in period, we adjust the candidate distributions within the MH steps to ensure that the acceptance rate throughout the post-burn-in period remains within the range of 0.3 to 0.5.

Here we draw samples from the full posterior

$$\pi(\boldsymbol{\beta}, \tau, \phi, \gamma, \tilde{\boldsymbol{W}}^*, \tilde{\boldsymbol{Y}}^{(c)}|\tilde{\boldsymbol{Y}}^{(nc)}),$$

where $\tilde{\boldsymbol{Y}}^{(c)}$ is the vector of censored data vector and $\tilde{\boldsymbol{Y}}^{(nc)}$ is the vector of non-censored data vector. For the vector of regression coefficients $\boldsymbol{\beta}$, we consider weakly-informative conjugate prior $\boldsymbol{\beta}|\tau \sim \text{MVN}(\boldsymbol{0}, 100^2\tau^{-1}\boldsymbol{I}_p)$. The full conditional posterior of $\boldsymbol{\beta}$ is then multivariate normal and updated using direct sampling within Gibbs steps. Due to the strong posterior correlation between $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{W}}^*$, they are updated jointly within each Gibbs sampling step. We also consider the non-informative priors for the hyperparameters involved in the correlation function in (2.1). Specifically, we choose $\phi \sim \text{Uniform}(0, 0.5\Delta_{\mathcal{S}})$ for the spatial range parameter, where $\Delta_{\mathcal{S}}$ is the largest Euclidean distance between two data locations, and $\gamma \sim \text{Uniform}(0,1)$ for the nugget effect γ. We further designate a non-informative conjugate prior for the spatially-constant precision parameter τ in the process model, namely $\tau \sim \text{Gamma}(0.1, 0.1)$. The full conditional posterior distribution of $\tilde{\boldsymbol{Y}}^{(c)}$ is $\text{MVN}\left(\boldsymbol{X}^{(c)}\boldsymbol{\beta} + \boldsymbol{A}^{(c)}\tilde{\boldsymbol{W}}^*, \tau^{-1}(1-\gamma)\boldsymbol{I}_n\right)$, where $\boldsymbol{X}^{(c)}$ and $\boldsymbol{A}^{(c)}$ are design matrix and SPDE projection matrix (from mesh nodes), respectively, corresponding to the locations with censored data, i.e., they comprise of the rows of $\boldsymbol{X}$ and $\boldsymbol{A}$ that correspond to the censored entries of $\boldsymbol{Y}$.

3.5 Software

We have developed an open-source R package, called CensSpBayes, which implements the proposed approximate Matérn GP model for large left-censored spatial data. Implementation code, along with details of execution using simulated data, are made available at https://github.com/SumanM47/CensSpBayes.

4. SIMULATION STUDY

In this section, we conduct simulation studies using synthetic data to assess the efficacy of our proposed scalable modeling framework in terms of spatial prediction while imputing censored values. We simulate 100 datasets over grids $\mathcal{D}^* = \{(i,j) : i,j \in \{1/K, 2/K, \ldots, 1\}\}$ of varying sizes within a spatial domain $[0,1]^2$. We consider $K \times K$ grids with $K = 20, 50, 100,$ and 200 to demonstrate the computational power and scalability inherent to our proposed methodology.

For simulating the datasets, we consider a model with two covariates and an intercept term. The values of the covariates are randomly generated from $N(0,1)$ and $N(5, 0.49)$ respectively, and we assume the true value of the regression coefficient to be $\boldsymbol{\beta}^{\text{true}} = (3, 1.2, 0.5)^{\mathsf{T}}$. The true value of the range parameter of the spatial Matérn correlation is chosen to be $\phi^{\text{true}} = 0.15 \times \Delta^*$, where $\Delta^* = \sqrt{2}$, the maximum spatial distance between two locations in $[0,1]^2$. The smoothness parameter is set to one and not estimated while fitting our proposed model. The true ratio of partial sill to total variation is chosen to be $\gamma^{\text{true}} = 0.9$, and the true precision parameter is chosen to be $\tau^{\text{true}} = 1/5$. Exact simulations are conducted to generate datasets from a GP with Matérn correlation, as given in (2.1).

Once the datasets are generated, we divide each dataset randomly into 80% training and 20% test datasets. Within each training set, we consider two different levels of censoring (denoted by L1 and L2) for the response by setting different values of the minimum detection limit (MDL):

L1 Low censoring: The MDL is at the 15^{th} percentile point of observations and thus 15% data are censored.

L2 High censoring: The MDL is at the 45^{th} percentile point of observations and thus 45% data are censored.

For each of the two levels of censoring, we implement our proposed approximate Matérn GP model under three different settings, denoted by S1, S2, and S3:

S1 We selectively exclude spatial locations where observations are censored and apply the spatial model approximated via SPDE, as elaborated in Section 3.1, exclusively to the observed locations. This does not require any imputation of the censored observations.

S2 We impute the censored observations using the mean of the observed data and employ the SPDE-approximated Matérn GP model, as detailed in Section 3.1. This once again circumvents the need for imputing the censored observations.

S3 We fit the full proposed model, treating the observations below MDL as censored observations, and implement the SPDE-approximated Matérn GP model, as in Section 3.1, while simultaneously performing imputations for the censored observations.

In each of the three scenarios, the approximated spatial process using SPDE has been fitted to assess the effects of removing or considering ad hoc imputations of censored observations in terms of mean squared prediction error (MSPE), in contrast to treating them as genuinely censored. The prior distributions for $\boldsymbol{\beta}$ and γ as described in Section 3.4 remain unchanged in the simulation study. However, for the range parameter, we assume $\phi \sim \text{Uniform}(0, 0.25\Delta^*)$.

For comparison, we consider two additional models, S4 and S5:

S4 We use a 2-dimensional B-spline estimation, along with covariates, for mean modeling ignoring the covariance structure while excluding the censored value, i.e.,

$$Y(\boldsymbol{s})=\boldsymbol{x}(\boldsymbol{s})^{\top}\boldsymbol{\beta}+\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\alpha_{ij}B(\boldsymbol{s}_1;\boldsymbol{s}^0_{ij,1})B(\boldsymbol{s}_2;\boldsymbol{s}^0_{ij,2})+\varepsilon(\boldsymbol{s}),$$

where $\varepsilon(\boldsymbol{s})$ are i.i.d. with zero mean and variance τ^2 and $B(\boldsymbol{s}_k;\boldsymbol{s}^0_{ij,k})$ is the evaluation of the one dimensional B-spline at the k-th coordinate of $\boldsymbol{s}$ and the k-th coordinate of the (i,j)th knot, $k = 1, 2$. Defining a model that models only the mean structure enables us to demonstrate the impact of disregarding both the covariance structure and the presence of censoring in the data. We use an ML estimation scheme here with the number of knots for the 2-dimensional B-splines set at 20% of the grid size.

S5 We use the 'CensSpatial' algorithm to perform an exact ML estimation of model parameters, which implements the SAEM algorithm of [28] with the package defaults used for the optimization procedures and standard settings for initial values and search limits. The covariance model was set as Matérn covariance with smoothness parameter of 1.

Table 1 presents the median (across 100 simulated datasets) mean squared prediction errors (MSPE), along with corresponding median standard errors, obtained from fitting the models S1-S5 to data in test sets that vary according to censoring levels and grid sizes. Notably, under low levels of censoring (L1), the proposed model in scenario S3 yields better results compared to situations where censored observations are either excluded from the analysis or imputed using the mean of observed values. However, ignoring spatial locations with censored observations entirely leads to unreliable estimates, particularly for the covariance parameters. Similarly, in high data censoring (L2) instances, the proposed model, along with the imputation of censored observations (S3), outperforms all other models. It is noteworthy that the 'CensSpatial' method also demonstrates relatively favorable performance for a grid size of 20 × 20 when the data-generating model is Matérn GP; however, its computational inefficiency and inadequate scaling impeded our ability to apply the method to the higher-dimensional simulated datasets. In fact, [28] showcased the efficacy of the 'CensSpatial' algorithm through simulations involving only 50 and 200 spatial locations, clearly indicating its inadequacy regarding scalability. Ignoring the covariance structure does affect the performance here as the MSPE for the B-spline based method (S4) is consistently higher than those for S1. Moreover, S4 often fails to produce a respectable MSPE likely because of ignoring both the censoring and the covariance structure. These conclusions are further supported by the boxplots of log (MSPE) values shown in Figure 5. Specifically, under the high data censoring scenario (100 × 100 in row 2), the method S4 displays a notably tall boxplot, suggesting that it produces more unstable results compared to the other methods.

Table 2 presents the median computation time corresponding to fitting the models S1-S5 to data in test sets that vary according to censoring levels and grid size. The computation times for S1, S2, and S3 are comparable, as they all employ the SPDE-approximated Matérn GP, with runtime approximately proportional to the size of the INLA mesh used for process approximation (between 557 and 673 nodes for different datasets of different sizes). As anticipated, the runtime for the local likelihood approach utilizing Vecchia's

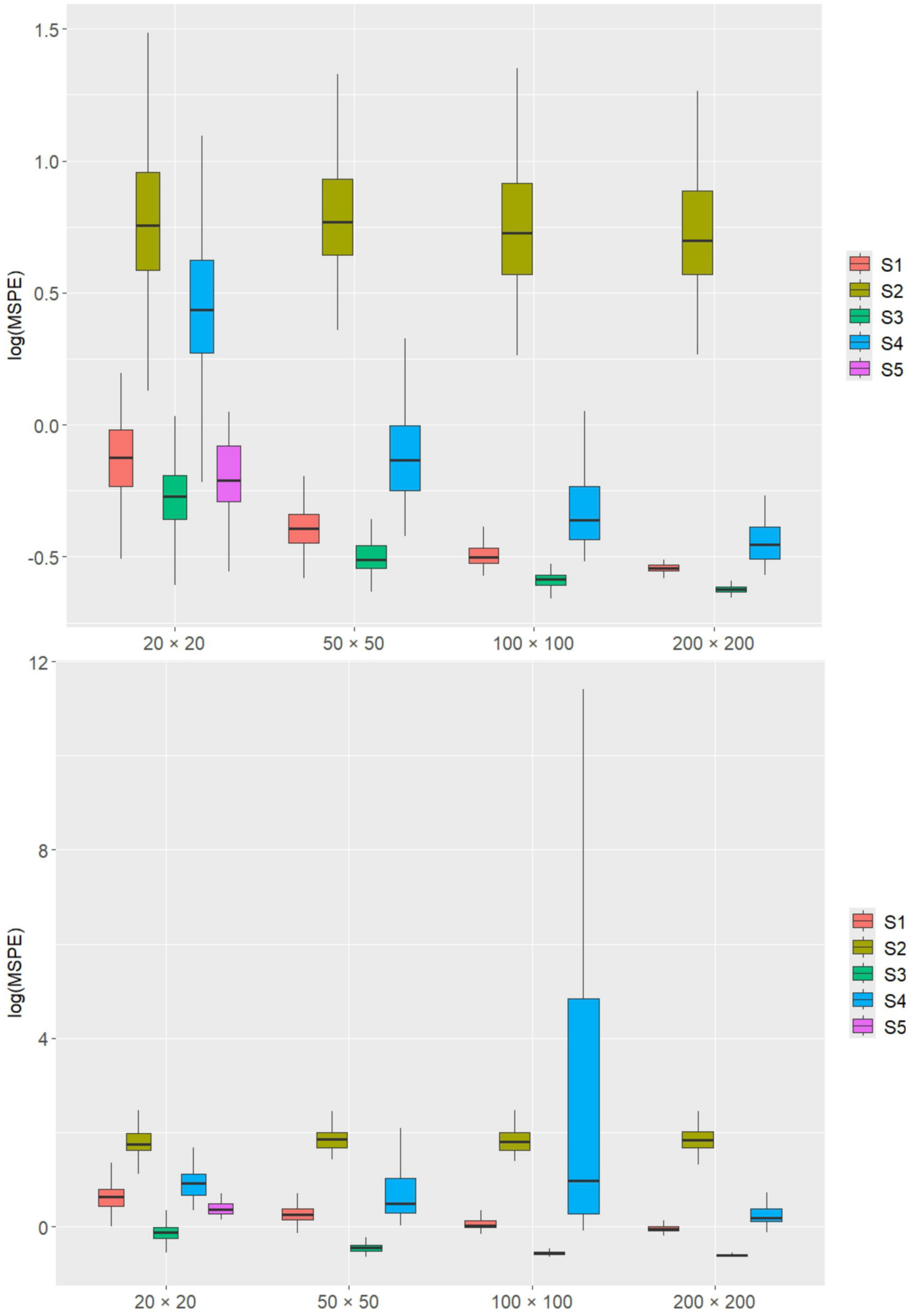

Figure 5: Boxplots of log (MSPE) values for S1 (censored data removed, GP approximated), S2 (censored data mean imputed, GP approximated), S3 (proposed methodology), S4 (censored data removed, 2D B-spline for mean modeling and no covariance modeling), and S5 (CensSpatial, only for 20 × 20) on the simulated datasets grouped by different gridsizes for 15% (top) and 45% (bottom) censoring scenarios. Outliers were not reported to make the boxplots easier to see.

Table 1. Median mean squared prediction error (MSPE) corresponding to model fitting under the five different settings (S1-S5) based on 100 simulated data sets from a Matérn GP that varies with censoring levels L1 (low-censoring; 15%) and L2 (high-censoring; 45%), and grid sizes. The values in parenthesis represent the corresponding median prediction standard errors. The lowest median MSPE in each row is in bold. Since model S5 ('CensSpatial') was infeasible for larger grids, table entries (median MSPE and standard errors) appear as '-'.

Censoring	Grid	Median Mean Squared Prediction Errors				
Level	size	S1	S2	S3	S4	S5
L1	20 × 20	0.88(0.83)	2.13(1.17)	**0.76**(0.84)	1.54(1.12)	0.81(0.94)
	50 × 50	0.67(0.75)	2.16(1.10)	**0.60**(0.78)	0.87(0.85)	-
	100 × 100	0.61(0.72)	2.07(1.07)	**0.56**(0.74)	0.70(0.76)	-
	200 × 200	0.58(0.71)	2.01(1.05)	**0.54**(0.73)	0.63(0.73)	-
L2	20 × 20	1.89(0.85)	5.77(0.92)	**0.88**(0.89)	2.50(1.04)	1.43(1.08)
	50 × 50	1.28(0.75)	6.34(0.85)	**0.64**(0.79)	1.63(0.90)	-
	100 × 100	1.03(0.70)	6.00(0.83)	**0.58**(0.76)	2.67(0.76)	-
	200 × 200	0.95(0.68)	6.26(0.82)	**0.54**(0.74)	1.20(0.08)	-

Table 2. Median computation time (in minutes) corresponding to model fitting under the five different settings (S1-S5) based on 100 simulated data sets from a Matérn GP that varies with censoring levels L1 (low-censoring; 15%) and L2 (high-censoring; 45%), and grid sizes. The values in parenthesis represent the median absolute deviation for the corresponding computing times. Since model S5 ('CensSpatial') was infeasible for larger grids, table entries appear as '-'.

Censoring	Grid	Median Computation Time (in minutes)				
Level	size	S1	S2	S3	S4	S5
L1	20 × 20	22.60(0.43)	23.41(0.38)	23.36(0.34)	< 0.01 (< 0.01)	42.65(1.01)
	50 × 50	19.43(0.33)	19.21(0.36)	19.32(0.60)	< 0.01 (< 0.01)	-
	100 × 100	25.45(0.41)	25.31(0.50)	25.76(0.32)	0.05(< 0.01)	-
	200 × 200	26.06(0.59)	26.30(0.49)	27.10(0.41)	2.20(0.08)	-
L2	20 × 20	21.06(0.38)	23.38(0.32)	23.69(0.29)	< 0.01 (< 0.01)	80.27(1.67)
	50 × 50	19.03(0.54)	19.29(0.36)	19.25(0.51)	< 0.01 (< 0.01)	-
	100 × 100	25.37(0.45)	25.57(0.39)	25.75(0.61)	0.04(< 0.01)	-
	200 × 200	26.35(0.81)	26.18(1.03)	27.04(0.60)	0.02(< 0.01)	-

approximation increases with larger grid sizes. As discussed earlier, the 'CensSpatial' algorithm was infeasible for larger grids. The initial four methods, S1–S4, were executed on SLURM clusters with one core per job and 8 GB RAM allocation. However, due to the current version of 'CensSpatial' on CRAN being incompatible with the UNIX system, the algorithm was implemented on a personal Dell 7210 computer featuring 16 GB RAM, an Intel Core i5 dual-core processor, and a Windows 11 Enterprise 64-bit operating system.

5. APPLICATION: CALIFORNIA PFAS DATA

5.1 Analysis Plan and Hyperparameters

We use the iterated log-transformed data (as described in Section 2) as our input to the proposed method. Since we have no covariates in this dataset, we use the coordinates of the locations (longitude, latitude) as covariates. In California, longitude reflects the distance from the Pacific Ocean, and the southern regions tend to have more desert-like areas compared to the northern regions. Consequently, incorporating geographic coordinates (longitude and latitude) can capture potential spatial trends in the data. Furthermore, recent news reports have suggested elevated PFAS concentrations in urban areas of Southern and Central California [27, 29]. Including longitude and latitude as covariates in our study enables us to corroborate these claims within a rigorous statistical framework. The hyperparameters for the priors are the same as mentioned in Section 3.4. We fit a variogram model on the non-censored observations to obtain an initial set of parameter estimates for $\boldsymbol{\beta}$, τ, ϕ, and γ. We run three chains with different starting values that are all close to the initial parameter estimates obtained by variogram fitting to allow for checking convergence and increasing the reliability of the model output. Each chain was run for 25, 000 iterations, with the first 15, 000 samples discarded as burn-in. We thinned the post-burn-in samples by 5 to obtain 2, 000 samples from the posterior distribution of the parameters.

5.2 Results

For the observed data comprising 24, 959 locations and the prediction grid of $0.1^\circ \times 0.1^\circ$, yielding 405, 893 prediction locations across California, each of the three MCMC chains completed in approximately 62 minutes. These com-

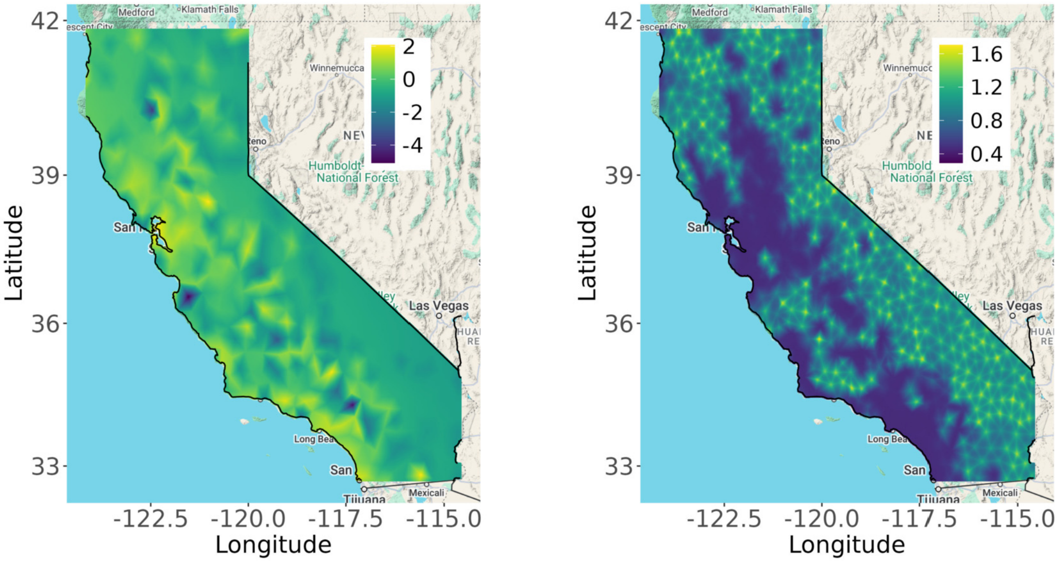

Figure 6: *Left*: The predicted surface map for $g(\text{PFOS}) = \log(1 + \log(1 + \text{PFOS}))$ concentration in the state of California. *Right*: The corresponding uncertainty estimates associated with the predictions for $g(\text{PFOS})$ across the pixels.

Table 3. Table of estimates (posterior mean) and posterior standard deviations corresponding to the parameters β_0, β_1 and β_2 denoting the intercept and the 2 covariates, respectively, ϕ (the spatial range), τ (the precision) and γ (ratio of partial sill to total variance).

	Estimates	Standard Deviation
β_0	-22.04	7.60
β_1	-0.23	0.08
β_2	-0.14	0.07
ϕ	0.07	0.02
τ	0.22	0.08
γ	0.95	0.02

putations were performed on an SLURM cluster with an 8GB RAM allocation. We observed a reasonable well-mixing of the three chains. The different parameter estimates, as calculated by the posterior means and the corresponding standard errors, as calculated by the posterior standard deviations, are presented in Table 3. Both estimated covariate effects are negative and are significant in our study. The estimated spatial range is relatively low ($\sim$7km). The posterior distribution of γ exhibits left skewness, suggesting that the majority of the variance in the data is attributable to the spatial structure rather than local noise, highlighting the significance of modeling the spatial covariance. The prediction surface is smooth at places (left panel of Figure 6) with higher detailing around the regions with observed data. The prediction standard deviation is low towards the western parts, where we have more observed data, but has an intriguing pattern on the east-southeastern parts (right panel of Figure 6). Our analysis confirms the findings of previous studies and news reports, as we observe significant negative effects of longitude and latitude. We predict elevated PFOS concentrations in urban areas along the western coast of Central and Southern California, with values exceeding the news EPA safety threshold of approximately 0.95 on the transformed scale. Notably high PFOS levels are detected in and around Sonoma, Napa, Solano, Contra Costa, Alameda, San Francisco, and Santa Clara counties in central-western California, as well as in Los Angeles, Orange, and San Diego counties in southwestern California.

6. DISCUSSION

We present a novel method to address the problem of modeling censored, spatially correlated outcomes in big data settings. We observe that the proposed model scales nicely with an increased number of observation locations and performs better than all other competing methods, even when nearly half of the observed data are censored. Despite being a fully Bayesian model, the runtime is moderate and at least comparable or better than the competing methods, highlighting its scalability, which, combined with its demonstrated accuracy, makes this method an efficient approximate method for modeling large spatial data in the presence of (left-) censoring. The real data analysis demonstrated this further as the model achieved satisfactory mixing of three chains for a large dataset in only an hour, producing sensible prediction surfaces and uncertainty quantification.

However, the data presents specific challenges during modeling, which may also be considered as limitations of the proposed method. The predicted surface in the left panel of Figure 6 is very smooth towards the east-southeast end of California. This is expected, as we have very few observations around that area to inform our spatial process. This, while being non-desirable, makes sense and is in line with what one would expect to happen for such a dataset. We can not hope to manufacture information in the absence of observations and we do not, reflecting the consistency of statistical principles being adhered to here in our analysis. The map of prediction uncertainty (right panel of Figure 6) also reflects this. We have nearly zero uncertainty for most of the region, where we observe numerous instances and have higher uncertainty whenever we move far from observations. Interestingly, we also notice a quilting pattern in the right panel of Figure 6. This is a byproduct of the mesh object and the lack of observations in the east-southeast region.

Further consideration is therefore needed to choose the mesh and smoothness parameters to fit the model. We explain our choice of mesh in Section 3.1. However, this process is ad-hoc, and a concrete workflow for selecting a mesh would greatly benefit users. We consider this to be a plausible future research direction. Another development on both the software and methodological fronts would include fractional smoothness parameters in the model, which is currently restricted to integer smoothness (we use $\nu = 1$ for all our analyses). One possible approach would be to use the fractional rational approximations to the SPDE model [6, 5]. Combining the theory for fractional approximations with the software should render additional model flexibility and be more well-suited to real data applications. Further developments for multivariate extensions of the model handling multiple spatial processes, simultaneously having a mix of censored and uncensored observations, are underway.

Another important component to consider here is the assumption of stationarity in the spatial covariance model, which may not be realistic in many real-world applications. To address potential non-stationarity, we incorporate geographical locations as covariates, aiming to ensure that the residuals are stationary and can be effectively modeled using the proposed approach, while accounting for data censoring. If the residuals exhibit non-stationary spatial dependence, it would be necessary to develop a model that accommodates non-stationary spatial structures. Given the large size of the dataset, a suitable approximation method for non-stationary spatial data would be required. Although such methods are relatively rare, some efforts have been made to address this challenge. In particular, we could extend our approach by utilizing the non-stationary version of the SPDE framework, as described by [24], to account for non-stationary spatial covariance in future work.

Other than proposing a novel, scalable spatial model in its own right, the implications of this study extend beyond academic interest. By elucidating the spatial distribution of PFAS/PFOS contamination and its associated factors, we can inform targeted interventions, policy recommendations, and resource allocation to mitigate the impact of PFAS exposure on public health. Additionally, our research provides a framework that can be adapted to analyze censored data in other environmental contexts, fostering a deeper understanding of complex contamination scenarios and enabling evidence-based decision-making.

APPENDIX A. ADDITIONAL ANALYSES FOR THE CALIFORNIA PFAS DATA

We provide additional analyses on PFOS concentration across California data by applying the two methods S1 and S2 (as detailed in Section 4) on the data along with the proposed method (S3) and doing a comparative analysis. Table 4 presents the different parameter estimates (posterior means) and the corresponding uncertainty (posterior standard deviation) for the covariate effects (β_0, β_1, β_2), the spatial range (ϕ) and precision (τ) parameters, and the ratio of partial sill to total variance (γ). These vary from analysis to analysis showing the different inference we can have by treating the censored observations differently (since that is the only difference between the methods). Unlike Section 4, the true values are unknown in the real data analysis, making it impossible to objectively determine which method is superior based on the provided estimates. However, the covariate effects of latitude and longitude are not statistically significant in both S1 and S2. This suggests that using methods S1 and S2 to pre-process censored data may result in the omission of important covariate effects, potentially leading to misleading conclusions in the analysis. Finally, the 2D B-splines method (S4) was not applied to the real data due to its inability to account for the spatial correlation structure. Similarly, the 'CensSpatial' method (S5) was excluded from this analysis because of its limitations in scalability.

Table 4. Posterior means (standard deviation) of the model parameters (covariate effects, spatial range, precision, and ratio of partial sill to total variance) in analyzing the California PFOS data using methods S1 through S3 (Approximate Gaussian process fit differing by how the censored data are treated: namely, left out, mean imputed or imputed by the proposed method).

	S1	S2	S3
β_0	−1.52(1.96)	−0.47(1.48)	−22.23(7.51)
β_1	−0.03(0.02)	−0.02(0.02)	−0.23(0.08)
β_2	−0.02(0.02)	−0.02(0.01)	−0.14(0.07)
ϕ	0.06(0.01)	0.05(0.01)	0.07(0.02)
τ	2.06(0.42)	2.18(0.51)	0.22(0.08)
γ	0.81(0.04)	0.8(0.05)	0.95(0.02)

Figure 7 presents a comparison between the predictions concentration maps (top) and their associated uncertainties (bottom) derived from methods S1 through S3 for estimating PFOS concentrations across a fine grid in California (displayed on the iterated logarithmic scale). The predictions from S1 and S2 show minimal spatial variation, and their

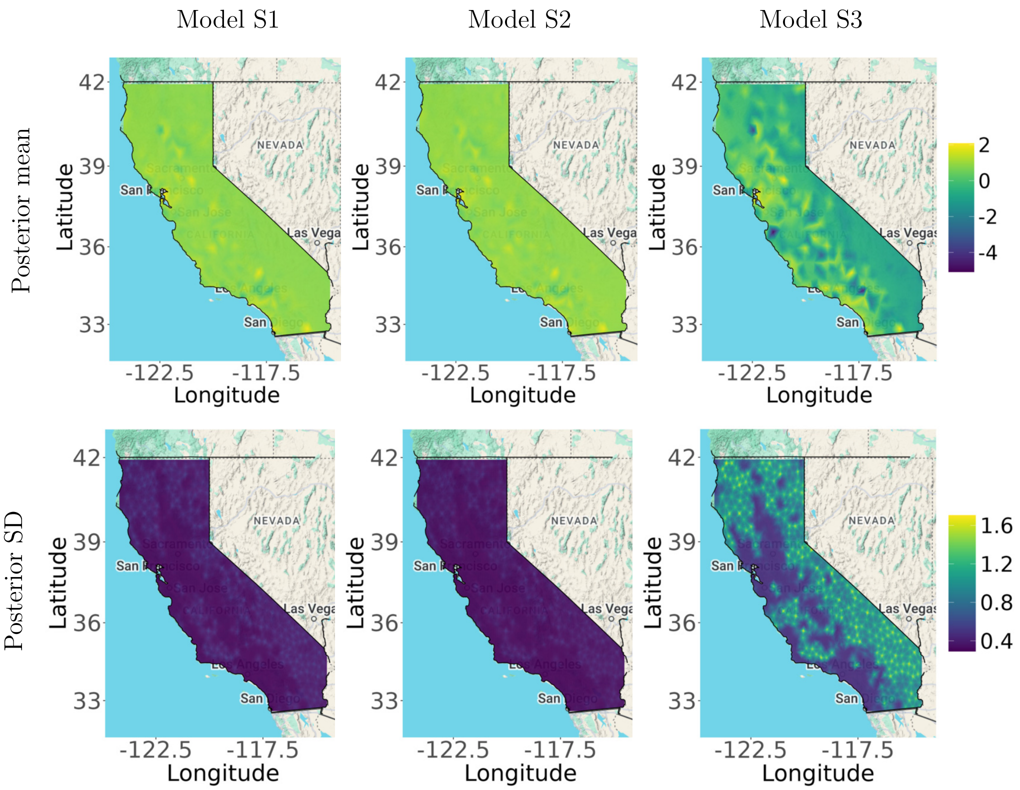

Figure 7: *Top*: Spatial surface maps of predicted PFOS concentrations (on the iterated log scale) based on Models S1, S2 and S3. *Bottom*: The corresponding uncertainty estimates.

uncertainty estimates reflect a similar homogeneity. This suggests that these methods may not adequately capture spatial variability. However, without comprehensive knowledge of the true PFAS levels across the region, it is difficult to accurately compare the performances of these methods.

ACKNOWLEDGEMENTS

Arnab Hazra is partially supported by the Indian Institute of Technology Kanpur and Rice University collaborative research grant under Award No. DOIR/2023246. Dipankar Bandyopadhyay acknowledges funding support from grant R21DE031879 from the US National Institutes of Health. This work does not represent views of the EPA.

Accepted 24 January 2025

REFERENCES

[1] Abrahamsen, P. and Benth, F. E. (2001). Kriging with inequality constraints. *Mathematical Geology* **33**(6) 719–744. https://doi.org/10.1023/A:1011078716252. MR1956391

[2] Andrews, D. Q. and Naidenko, O. V. (2020). Population-wide exposure to per-and polyfluoroalkyl substances from drinking water in the United States. *Environmental Science & Technology Letters* **7**(12) 931–936.

[3] Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D. and Lindgren, F. (2018). Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics* **10**(6) 1443. https://doi.org/10.1002/wics.1443. MR3873676

[4] Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V. and Pebesma, E. J. (2008) *Applied Spatial Data Analysis with R.* **747248717**. Springer. https://doi.org/10.1007/978-1-4614-7618-4. MR3099410

[5] Bolin, D. and Kirchner, K. (2020). The rational SPDE approach for Gaussian random fields with general smoothness. *Journal of Computational and Graphical Statistics* **29**(2) 274–285. https://doi.org/10.1080/10618600.2019.1665537. MR4116041

[6] Bolin, D., Simas, A. B. and Xiong, Z. (2024). Covariance–based rational approximations of fractional SPDEs for computationally efficient Bayesian inference. *Journal of Computational and Graphical Statistics* **33**(1) 64–74. https://doi.org/10.1080/10618600.2023.2231051. MR4713943

[7] Borouchaki, H. and Lo, S. (1995). Fast Delaunay triangulation in three dimensions. *Computer Methods in Applied Mechanics and Engineering* **128**(1-2) 153–167. https://doi.org/10.1016/0045-7825(95)00854-1. MR1376908

[8] Ciarlet, P. G. (2002) *The Finite Element Method for Elliptic Problems.* SIAM. https://doi.org/10.1137/1.9780898719208. MR1930132

[9] Cisneros, D., Gong, Y., Yadav, R., Hazra, A. and Huser, R. (2023). A combined statistical and machine learning approach for spatial prediction of extreme wildfire frequencies and sizes. *Extremes* **26**(2) 301–330. https://doi.org/10.1007/s10687-022-00460-8. MR4577409

[10] De Oliveira, V. (2005). Bayesian inference and prediction of Gaussian random fields based on censored data. *Journal of Computational and Graphical Statistics* **14**(1) 95–115. https://doi.org/10.1198/106186005X27518. MR2137892

[11] Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* **27**(1) 94–128. https://doi.org/10.1214/aos/1018031103. MR1701103

[12] for Toxic Substances, A. and (ATSDR), D. R. (2018). Toxicological profile for perfluoroalkyls (draft for public comment). *US Department of Health and Human Services, Public Health Service, Atlanta, GA*.

[13] Fridley, B. L. and Dixon, P. (2007). Data augmentation for a Bayesian spatial model involving censored observations. *Environmetrics: The Official Journal of the International Environmetrics Society* **18**(2) 107–123. https://doi.org/10.1002/env.806. MR2345649

[14] Gelfand, A. E. and Schliep, E. M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics* **18** 86–104. https://doi.org/10.1016/j.spasta.2016.03.006. MR3573271

[15] Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**(471) 1021–1035. https://doi.org/10.1198/016214504000002078. MR2201028

[16] Hazra, A., Huser, R. and Bolin, D. (2024). Efficient Modeling of Spatial Extremes over Large Geographical Domains. *Journal of Computational and Graphical Statistics*. **Just accepted**.

[17] Hazra, A., Reich, B. J., Shaby, B. A. and Staicu, A.-M. (2018). A semiparametric spatiotemporal Bayesian model for the bulk and extremes of the Fosberg Fire Weather Index. *arXiv preprint arXiv:1812.11699*.

[18] Hepburn, E., Madden, C., Szabo, D., Coggan, T. L., Clarke, B. and Currell, M. (2019). Contamination of groundwater with per-and polyfluoroalkyl substances (PFAS) from legacy landfills in an urban re-development precinct. *Environmental Pollution* **248** 101–113.

[19] Hopke, P. K., Liu, C. and Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the Arctic. *Biometrics* **57**(1) 22–33. https://doi.org/10.1111/j.0006-341X.2001.00022.x. MR1833288

[20] Hosmer, D. W., Lemeshow, S. and May, S. (2008). *Applied Survival Analysis*. *Wiley Series in Probability and Statistics* **60**. https://doi.org/10.1002/9780470258019. MR2383788

[21] Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F. and Rue, H. (2018) *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

[22] Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software* **63**(19).

[23] Lindgren, F., Bolin, D. and Rue, H. (2022). The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics* **50** 100599. https://doi.org/10.1016/j.spasta.2022.100599. MR4439328

[24] Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **73**(4) 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x. MR2853727

[25] Militino, A. F. and Ugarte, M. D. (1999). Analyzing censored spatial data. *Mathematical Geology* **31**(5) 551–561.

[26] Moran, J. E., Hudson, G. B., Eaton, G. F. and Leif, R. (2005). California GAMA Program: Groundwater Ambient Monitoring and Assessment Results for the Sacramento Valley and Volcanic Provinces of Northern California. Technical Report, Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States).

[27] Muigai, B. USC Researchers Assess Impact of PFAS in Drinking Water Systems in Southern California. *Keck School of Medicine of USC Newsroom*. Accessed 2024-09-10.

[28] Ordoñez, J. A., Bandyopadhyay, D., Lachos, V. H. and Cabral, C. R. (2018). Geostatistical estimation and prediction for censored responses. *Spatial Statistics* **23** 109–123. https://doi.org/10.1016/j.spasta.2017.12.001. MR3768178

[29] Pineda, D. Risk of tap water exposure to toxic PFAS chemicals higher in Southern California. *Los Angeles Times*. Accessed 2024-09-10.

[30] Rathbun, S. L. (2006). Spatial prediction with left-censored observations. *Journal of Agricultural, Biological, and Environmental Statistics* **11**(3) 317–336.

[31] Read, R., Briefs, C. B. and Cases, C. C. C. (2024). PFAS National Primary Drinking Water Regulation 89 Fed. Reg. 32532 (Apr. 26, 2024) Copy Cite. *Federal Register*.

[32] Rosen, G. (1955). Problems in the application of statistical analysis to questions of health: 1700–1880. *Bulletin of the History of Medicine* **29**(1) 27–45.

[33] Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. CRC Press. https://doi.org/10.1201/9780203492024. MR2130347

[34] Sahoo, I. and Hazra, A. (2021). Contamination mapping in Bangladesh using a multivariate spatial Bayesian model for left-censored data. *Journal of the Indian Statistical Association* **59**(2) 251–285. MR4810704

[35] Sahoo, I., Guinness, J. and Reich, B. J. (2023). Estimating atmospheric motion winds from satellite image data using space-time drift models. *Environmetrics* **34**(8) 2818. https://doi.org/10.1002/env.2818. MR4680778

[36] Schelin, L. and Sjöstedt-de Luna, S. (2014). Spatial prediction in the presence of left-censoring. *Computational Statistics & Data Analysis* **74** 125–141. https://doi.org/10.1016/j.csda.2014.01.004. MR3168965

[37] Schulz, E., Speekenbrink, M. and Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology* **85** 1–16. https://doi.org/10.1016/j.jmp.2018.03.001. MR3852577

[38] Sedda, L., Atkinson, P. M., Barca, E. and Passarella, G. (2012). Imputing censored data with desirable spatial covariance function properties using simulated annealing. *Journal of Geographical Systems* **14**(3) 265–282.

[39] Smalling, K. L., Romanok, K. M., Bradley, P. M., Morriss, M. C., Gray, J. L., Kanagy, L. K., Gordon, S. E., Williams, B. M., Breitmeyer, S. E., Jones, D. K. et al. (2023). Per-and polyfluoroalkyl substances (PFAS) in United States tapwater: Comparison of underserved private-well and public-supply exposures and associated health implications. *Environment International*, 108033.

[40] Stein, M. L. (1992). Prediction and inference for truncated spatial data. *Journal of Computational and Graphical Statistics* **1**(1) 91–110.

[41] Tadayon, V. (2017). Bayesian analysis of censored spatial data based on a non-Gaussian model. *arXiv preprint arXiv:1706.05717*.

[42] Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **50**(2) 297–312. MR0964183

[43] Wang, Z., DeWitt, J. C., Higgins, C. P. and Cousins, I. T. (2017). *A Never-Ending Story of Per- and Polyfluoroalkyl Substances (PFASs)?* ACS Publications.

[44] Wiens, A., Nychka, D. and Kleiber, W. (2020). Modeling spatial data using local likelihood estimation and a Matérn to spatial autoregressive translation. *Environmetrics* **31**(6) 2652. https://doi.org/10.1002/env.2652. MR4151871

[45] Yadav, R., Huser, R. and Opitz, T. (2019). Spatial hierarchical modeling of threshold exceedances using rate mixtures. *Environmetrics*, 2662. https://doi.org/10.1002/env.2662. MR4248739

[46] Zhang, L., Shaby, B. A. and Wadsworth, J. L. (2021). Hierarchical transformed scale mixtures for flexible modeling of spatial extremes on datasets with many locations. *Journal of the American Statistical Association* 1–13. https://doi.org/10.1080/01621459.2020.1858838. MR4480717

Indranil Sahoo. Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, USA.
E-mail address: sahooi@vcu.edu

Suman Majumder. Interdisciplinary Statistical Research Unit, Indian Statistical Institute, India.
E-mail address: smajumder@isical.ac.in

Arnab Hazra. Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, India.
E-mail address: ahazra@iitk.ac.in

Ana G. Rappold. United States Environmental Protection Agency, USA.
E-mail address: Rappold.Ana@epa.gov

Dipankar Bandyopadhyay. Department of Biostatistics, Virginia Commonwealth University, USA.
E-mail address: dbandyop@vcu.edu