Spatial Statistics 9 (2014) 166–179

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Exploration of the use of Bayesian modeling of gradients for censored spatiotemporal data from the *Deepwater Horizon* oil spill

Harrison Quick [a], Caroline Groth [b], Sudipto Banerjee [b,*], Bradley P. Carlin [b], Mark R. Stenzel [c], Patricia A. Stewart [d], Dale P. Sandler [e], Lawrence S. Engel [f], Richard K. Kwok [e]

[a] *Department of Statistics, University of Missouri, Columbia, MO 65211, United States*
[b] *Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, United States*
[c] *Exposure Assessments Applications, LLC, Arlington, VA 22207, United States*
[d] *Stewart Exposure Assessments, LLC, Arlington, VA 22207, United States*
[e] *National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, United States*
[f] *Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27599, United States*

ARTICLE INFO

Article history:
Received 10 September 2013
Accepted 7 March 2014
Available online 19 March 2014

Keywords:
Gaussian process
Gradients
Markov chain Monte Carlo
Censored data
Hierarchical modeling
Spatiotemporal data

ABSTRACT

This paper develops a hierarchical framework for identifying spatiotemporal patterns in data with a high degree of censoring using the gradient process. To do this, we impute censored values using a sampling-based inverse CDF method within our Markov chain Monte Carlo algorithm, thereby avoiding burdensome integration and facilitating efficient estimation of other model parameters. We illustrate use of our methodology using a simulated data example, and uncover the danger of simply substituting a space- and time-constant function of the level of detection for all missing values. We then fit our model to area measurement data of volatile organic compound (VOC) air concentrations collected on vessels supporting the response and clean-up efforts of the *Deepwater Horizon* oil release that occurred starting April 20, 2010. These data contained a high percentage of observations below the detectable limits of the measuring instrument. Despite this, we were still able to make some interesting discoveries, including elevated levels of VOC near the site of the oil well on June 26th. Using the results from

* Corresponding author. Tel.: +1 612 624 0624.
E-mail addresses: quickh@missouri.edu (H. Quick), baner009@umn.edu (S. Banerjee).

http://dx.doi.org/10.1016/j.spasta.2014.03.002

this preliminary analysis, we hope to inform future research on the *Deepwater Horizon* study, including the use of gradient methods for assigning workers to exposure categories.

1. Introduction

On April 20, 2010, an explosion occurred on the *Deepwater Horizon* oil drilling rig, located about 80 km off the coast of Louisiana, resulting in millions of barrels of oil being released into the Gulf of Mexico over a period of several months. After the search, rescue and recovery efforts ended, efforts concentrated on capping the wellhead, capturing released oil and drilling a new well to relieve the pressure on the well casing and blow-out-preventer. The well was "top capped" in mid-July 2010, a measure that stopped the release of the oil. Relief wells were completed and the well was plugged, or "bottom capped", at the depths of the oil formation, to relieve pressure on the original well casing, in early August. In addition to this work, efforts were made to collect or burn the released oil that was spreading across the Gulf primarily north and east of the well site and hitting the coastlines of Louisiana, Mississippi, Alabama and Florida.

As a result of this incident, the Gulf Long-term Follow-up Study (GuLF STUDY) was initiated by the US National Institute of Environmental Health Sciences (NIEHS) to investigate possible adverse health effects to the workers associated with exposure to these spill-related chemicals. Approximately, 33,000 workers participated in the study. These workers were involved in the response and cleanup efforts with possible exposure to a variety of harmful chemicals, including volatile organic compounds (VOCs) and specific chemicals including benzene, toluene, ethyl benzene and xylene. A crucial component of this study is to estimate the exposure levels of the workers to oil-related components.

During the response and clean-up, over 150,000 measurements were collected reflecting full work-shift, primarily oil-component, exposures experienced by workers (personal measurements). In addition, over 25,000,000 measurements of VOC were collected by direct-reading instruments that provided real time air concentrations every few minutes for a variety of locations on the 38 participating vessels (area – i.e., stationary – measurements). These personal and area measurements were only collected for a limited number of study subjects and vessels, respectively; thus some method must be used to assign exposure estimates to workers or vessels that were not measured. Furthermore, many of these measurements are below the measurement instrument's limit of detection (LOD); thus we face a situation where we have a high degree of censoring.

Due to the various challenges associated with data consisting of millions of observations (e.g., computational, administrative, etc.), we chose to conduct a preliminary analysis of a much smaller set of measurement data for the purposes of informing a more thorough investigation of the full data set. Our data consists of 447 area VOC air concentrations that were collected using an AreaRae direct-reading instrument (Rae Systems, San Jose, CA) often at short, regular intervals (e.g., once every 12 min), during the clean-up and response effort at varying times within a day from June 20 to July 20, 2010; these data were collected concurrently with the remaining data, but represent a complete set of data from a particular source. In contrast to the personal exposure measurements, where observations were collected on the individual workers, these data were collected on the *vessels* carrying the workers. As illustrated in Fig. 1, we may observe data from a vessel at a variety of locations throughout the day, and these locations may vary greatly between day; as such, our data are irregular in both space and time, requiring a continuous space, continuous time framework. Furthermore, over 83% of our VOC measurements are below the detectable limit of 0.1 ppm (parts per million); thus, we require techniques for handling censoring in our analysis.

In this paper, we aim to develop a framework for identifying spatiotemporal patterns in air concentrations of VOC, characterized by a high degree of censoring, from area (or stationary) measurement data collected on the vessels supporting the response and clean-up efforts of the oil

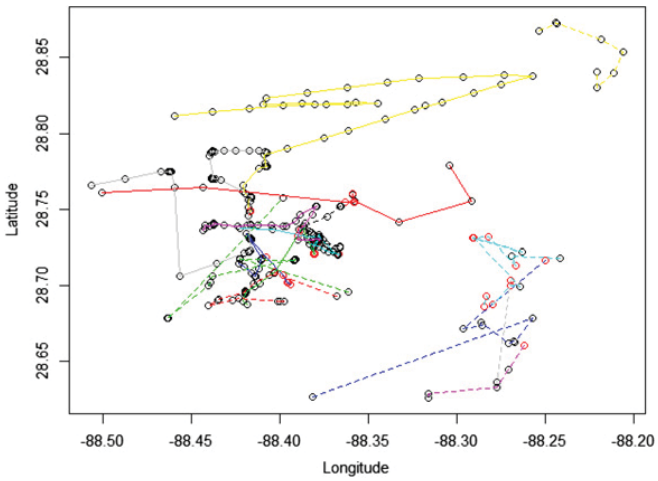


Fig. 1. Locations of the 447 observation sites from the *Deepwater Horizon* data set. Red circles denote VOC levels above the limit of detection (0.1 ppm), while black circles denote censored observations. Observations from each day are connected using lines, where each day has its own combination of color and line type. Each line represents a moving vessel. The *Deepwater Horizon* oil rig was located at 28.736667° N 88.386944° W. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

spill. To analyze these data, we employ a Bayesian hierarchical model that accounts for both spatial and temporal associations to investigate how the vapor cloud of VOC dispersed over space and time from the site of the spill. Specifically, we have used the spatiotemporal gradient process methods developed by Quick et al. (2013) to determine if there were sharp temporal gradients in air concentrations, whether the vapor cloud of VOC dispersed gradually (signified by the absence of significant spatial gradients), and if the size or shape of the VOC cloud changed during the course of a given time period (corresponding to significant "mixed" gradients). Our work here describes how to implement the spatiotemporal gradient process methodology in the presence of highly censored data and to gain insight into how to estimate air concentrations from the large number of area measurements available throughout the Gulf of Mexico measured vessels. The resultant estimates will provide information on air concentrations for study participants on vessels not measured.

The handling of missing or censored data in statistical methodology is not a new issue and dates back at least to Yates (1933), who proposed a method for replacing missing (though not censored) observations in such a way as to reduce the error sum of squares. Also see the excellent text by Little and Rubin (2002). Missing data problems in spatial statistics usually pertain to misalignment and change-of-support (see, e.g., Cressie, 1996; Mugglin et al., 2000; Banerjee and Gelfand, 2002; Ren and Banerjee, 2013; Gelfand et al., 2001). These refer to situations where not all spatially-referenced variables have been observed over the same spatial units. The review articles by Gotway and Young (2002) and Gelfand (2010) provide excellent overviews of the existing statistical literature on spatial misalignment problems.

Spatially censored data, on the other hand, have not attracted much attention. For general censored data analysis, researchers often *substitute* the values with some function of the limit of detection (e.g., $LOD/2$). Recent work by Ganser and Hewett (2010) has led to the development of the β-*substitution* method, a substitution method in which censored values are replaced by $\beta \cdot LOD$, where β is estimated to provide an unbiased estimate of the log-scale parameter of the log-Normal distribution. While these methods can be helpful and are simple to implement, using them in a spatiotemporal setting is not advisable. For instance, Fridley and Dixon (2007) showed that such substitution methods can lead to bias in the spatial variability in the purely spatial setting with only 20% censoring, suggesting that this issue may be compounded in the space–time setting with higher levels of censoring. Furthermore, substituting censored values with a number less than the LOD will lead to discontinuities in the response that may manifest as sharp changes in the residual surface; as our goal here is to conduct inference on the spatiotemporal gradient process, these discontinuities would likely result in bias in our gradient estimates, particularly around the boundaries between censored and

observed values. Rather than substitute our censored values with a fixed constant, we choose to use *data augmentation*, that is we impute values for our censored observations. Not only does this facilitate accurate estimation of our model parameters, the degree of variability, and gradient process, as we demonstrate in Section 4, but this also fits seamlessly into our Markov chain Monte Carlo (MCMC) algorithm. For a more detailed review of methods for analyzing censored data, see Fridley and Dixon (2007) and the references therein.

The remainder of this article will proceed as follows. Section 2 provides a brief overview of spatiotemporal gradients as developed in Quick et al. (2013), followed by an outline of how to implement our hierarchical model and incorporate censoring in Section 3. We analyze a highly censored simulated data set in Section 4, where we demonstrate the ability of our model to accurately estimate the underlying gradient process despite the level of censoring in the data set. Next, in Section 5 we turn our attention to data from the *Deepwater Horizon* oil spill and use our methods to investigate rates of change in VOC levels. Finally, we add some concluding thoughts in Section 6.

2. Review of spatiotemporal gradients

Quick et al. (2013) develop the calculus of spatiotemporal gradient processes by adapting the basic theory of random fields (e.g. Adler, 2009; Banerjee et al., 2003) and demonstrated that these can reveal valuable information about how an underlying space and time-continuous process evolves in both space and time. Such gradients can be used to identify potentially important predictors missing from the statistical model, or to locate boundaries in space or time at which significant changes in the residual surface occur. In order to model this process, the authors considered the spatiotemporal process model

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + Z(\mathbf{s}, t) + \epsilon(\mathbf{s}, t), \tag{1}$$

where $\mu(\mathbf{s}, t)$ captures overall trends (e.g., a regression model), $Z(\mathbf{s}, t)$ is an underlying spatiotemporal residual process, realizations of which are referred to as spatiotemporal random effects, and $\epsilon(\mathbf{s}, t)$ is a zero centered white noise process that captures micro-scale variability and other unstructured random deviations in the data. This model can be applied to settings where the response, $Y(\mathbf{s}, t)$, is assumed to exist for every space–time coordinate in $\Re^d \times \Re$, where d denotes the number of spatial dimensions. While in practice the data may only be able to be collected over a finite subset $\mathcal{S} \times \mathcal{T} \subset \Re^d \times \Re$, investigators may desire predictions at any given location or time point. Furthermore, although the methods presented here do extend to arbitrary d, our analyses here will be limited to the case where $d = 2$.

When modeling data arising from a model such as (1), where both space and time vary continuously, we may wish to make inference on the d directional spatial gradients, $\nabla_s Z(\mathbf{s}, t)^T$, the temporal gradient, $\nabla_t Z(\mathbf{s}, t)$, and the d mixed gradients, $\nabla_{s,t} Z(\mathbf{s}, t)^T$. We refer to the $(2d + 1) \times 1$ vector $\nabla Z(\mathbf{s}, t) = ((\nabla_s Z(\mathbf{s}, t))^T, \nabla_t Z(\mathbf{s}, t), (\nabla_t \nabla_s Z(\mathbf{s}, t))^T)^T$ as our spatiotemporal gradient process. In order for the spatiotemporal gradient process to exist, one must use a covariance structure which permits a mean-square differentiable process. To this end, we let $Z(\mathbf{s}, t)$ be a univariate stationary zero-mean Gaussian random field $GP(0, K(\cdot, \cdot))$, where $K(\cdot, \cdot)$ is a nonseparable and sufficiently smooth covariance function of the form

$$\begin{aligned} K(\boldsymbol{\Delta}, \delta) &= \mathrm{Cov}(Z(\mathbf{s}, t), Z(\mathbf{s} + \boldsymbol{\Delta}, t + \delta)) \\ &= \frac{\sigma^2}{\left(\phi_t^2 |\delta|^2 + 1\right)} \left(1 + \frac{\phi_s \|\boldsymbol{\Delta}\|}{\left(\phi_t^2 |\delta|^2 + 1\right)^{1/2}}\right) \exp\left[-\frac{\phi_s \|\boldsymbol{\Delta}\|}{\left(\phi_t^2 |\delta|^2 + 1\right)^{1/2}}\right], \end{aligned} \tag{2}$$

where σ^2 denotes the variance of the spatiotemporal process, ϕ_s and ϕ_t are the spatial and temporal range parameters, respectively, $\boldsymbol{\Delta}$ and δ denote spatial and temporal distances, respectively. Note that this is a special case of Eq. (11) from Gneiting (2002) where $\varphi(u)$ be the Matérn(3/2) correlation function and $\psi(\delta) = (\phi_t^2 |\delta|^2 + 1)$. Letting $\mathbf{Z}$ be the $N \times 1$ vector created by stacking our $Z(\mathbf{s}_i, t_j)$ (by time, in increasing order), we can construct Σ_Z to be the $N \times N$ variance–covariance matrix whose elements correspond to each pair of space–time coordinates; that is, the element associated with

$(\mathbf{s}_i, t_j)$ and $(\mathbf{s}_{i'}, t_{j'})$ is $K(\boldsymbol{\Delta}_{ii'}, \delta_{jj'})$ where $\boldsymbol{\Delta}_{ii'} = \mathbf{s}_{i'} - \mathbf{s}_i$ and $\delta_{jj'} = t_{j'} - t_j$. We can then define $R_Z(\phi_s, \phi_t)$ as the corresponding correlation matrix such that $\Sigma_Z = \sigma^2 R_Z(\phi_s, \phi_t)$.

To make inference on our gradient process at a given space–time coordinate, $\nabla Z(\mathbf{s}_0, t_0)$, we consider the joint distribution of $\mathbf{Z}$ and $\nabla Z(\mathbf{s}_0, t_0)$. For any arbitrary space–time coordinate $(\mathbf{s}_0, t_0)$, we can write $\text{Cov}\{\nabla Z(\mathbf{s}_0, t_0), \mathbf{Z}\}$ as the $(2d + 1) \times N$ matrix partitioned as $\nabla\mathbf{K}_0 = [\nabla\mathbf{K}_{0,1} : \nabla\mathbf{K}_{0,2} : \ldots : \nabla\mathbf{K}_{0,N}]$, where each column is the $(2d + 1) \times 1$ vector corresponding to $\text{Cov}\{\nabla Z(\mathbf{s}_0, t_0), Z(\mathbf{s}_i, t_j)\}$. Letting $C_{\nabla Z}(\mathbf{0}, 0)$ to be the cross-covariance matrix of $\nabla Z(\mathbf{s}, t)$ evaluated at $(\mathbf{0}, 0)$, the joint distribution of $\mathbf{Z}$ and $\nabla Z(\mathbf{s}_0, t_0)$ is

$$\begin{pmatrix} \mathbf{Z} \\ \nabla Z(\mathbf{s}_0, t_0) \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_Z & (\nabla\mathbf{K}_0)^T \\ \nabla\mathbf{K}_0 & \mathbf{C}_{\nabla Z}(\mathbf{0}, 0) \end{bmatrix} \right). \tag{3}$$

Upon estimating $\mathbf{Z}$ and the other model parameters, the gradient process, $\nabla Z(\mathbf{s}_0, t_0)$ can be estimated using the posterior predictive distribution. Full details regarding the derivation of the gradient process, including explicit expressions for $\nabla\mathbf{K}_0$ and $\mathbf{C}_{\nabla Z}(\mathbf{0}, 0)$, as well a thorough review of the gradient process literature, can be found in Quick et al. (2013).

3. Hierarchical modeling and inference

To construct our censored data model, we let $\boldsymbol{\theta}$ be a vector of model parameters and define an indicator variable, $D(\mathbf{s}, t)$, which equals 1 if $Y(\mathbf{s}, t)$ is above the detectable limit, C, and 0 otherwise. Then, we can define the conditional distribution of $Y(\mathbf{s}, t)$ given $\boldsymbol{\theta}, \mathbf{Z}$, and $D(\mathbf{s}, t)$ as a truncated Normal distribution; for instance, when $D(\mathbf{s}, t) = 0$, we have

$$p(Y(\mathbf{s}, t) \mid \boldsymbol{\theta}, \mathbf{Z}, D(\mathbf{s}, t) = 0) = \frac{\varphi\left(Y(\mathbf{s}, t) \mid \boldsymbol{\theta}, \mathbf{Z}\right)}{\Phi(C \mid \boldsymbol{\theta}, \mathbf{Z})} \times I\{Y(\mathbf{s}, t) < C\}\,, \tag{4}$$

where $\varphi(\cdot \mid \boldsymbol{\theta}, \mathbf{Z})$ and $\Phi(\cdot \mid \boldsymbol{\theta}, \mathbf{Z})$ denote the PDF and CDF of the Normal distribution, respectively, with mean $\mu(\mathbf{s}, t) + Z(\mathbf{s}, t)$ and variance τ^2. For the purposes of our hierarchical model and the subsequent development of our MCMC algorithm, however, it may be beneficial to consider the joint distribution of $Y(\mathbf{s}, t)$ and $D(\mathbf{s}, t)$ given $\boldsymbol{\theta}$ and $\mathbf{Z}$. Again considering the case of $D(\mathbf{s}, t) = 0$, we have

$$\begin{aligned} [Y(\mathbf{s}, t), D(\mathbf{s}, t) = 0 \mid \boldsymbol{\theta}, \mathbf{Z}] &\propto [Y(\mathbf{s}, t) \mid \boldsymbol{\theta}, \mathbf{Z}, D(\mathbf{s}, t) = 0] \times P[D(\mathbf{s}, t) = 0 \mid \boldsymbol{\theta}, \mathbf{Z}] \\ &\propto \left[\frac{N(Y(\mathbf{s}, t) \mid \mu(\mathbf{s}, t) + Z(\mathbf{s}, t), \tau^2)}{\Phi(C \mid \boldsymbol{\theta}, \mathbf{Z})} I\{Y(\mathbf{s}, t) < C\} \right] \times \Phi(C \mid \boldsymbol{\theta}, \mathbf{Z}) \\ &\propto N(Y(\mathbf{s}, t) \mid \mu(\mathbf{s}, t) + Z(\mathbf{s}, t), \tau^2) \times I\{Y(\mathbf{s}, t) < C\}. \end{aligned} \tag{5}$$

Here we see that the only place $\boldsymbol{\theta}$ or $\mathbf{Z}$ enters $[Y(\mathbf{s}, t), D(\mathbf{s}, t) \mid \boldsymbol{\theta}, \mathbf{Z}]$ is through the *unconstrained* Normal distribution, thus our estimation of $\boldsymbol{\theta}$ and $\mathbf{Z}$ through MCMC can proceed as though there is no censoring in our data, provided that we are able to impute values for these censored observations.

Embedding (5) in a Bayesian hierarchical framework, we model the trend using regressors indexed by both space and time, i.e., $\mu(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)^T\boldsymbol{\beta}$, and model $Z(\mathbf{s}, t)$ as a spatiotemporal process specified by the covariance kernel in (2). Specifications are completed by assigning prior distributions to unknown model parameters, thereby yielding the posterior distribution

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{Z} \mid \mathbf{Y}, \mathbf{D}) \propto\; & U(\phi_s \mid a_\phi, b_\phi) \times U(\phi_t \mid a_\phi, b_\phi) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times IG(\tau^2 \mid a_\tau, b_\tau) \\ & \times N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times N\left(\mathbf{Z} \mid \mathbf{0}, \sigma^2\mathbf{R}_Z(\phi_s, \phi_t)\right) \\ & \times \prod_{(i,j) \notin \mathcal{D}} N(Y(\mathbf{s}_i, t_j) \mid \mathbf{x}(\mathbf{s}_i, t_j)^T\boldsymbol{\beta} + Z(\mathbf{s}_i, t_j), \tau^2) \times I\{Y(\mathbf{s}_i, t_j) < C\} \\ & \times \prod_{(i,j) \in \mathcal{D}} N(Y(\mathbf{s}_i, t_j) \mid \mathbf{x}(\mathbf{s}_i, t_j)^T\boldsymbol{\beta} + Z(\mathbf{s}_i, t_j), \tau^2) \times I\{Y(\mathbf{s}, t) \geq C\}, \end{aligned} \tag{6}$$

where $\mathbf{Y}$ and $\mathbf{D}$ are the $N \times 1$ vectors constructed from $Y(\mathbf{s}_i, t_j)$ and $D(\mathbf{s}_i, t_j)$, respectively, $\mathcal{D} = \{(i, j) \mid D(\mathbf{s}_i, t_j) = 1\}$ is the collection of non-censored observations, and $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \phi_s, \phi_t, \sigma^2, \tau^2\}$ is the set

of process parameters in the spatiotemporal and the white-noise processes. The parametrizations for the standard densities are as in Carlin and Louis (2009). We assume all the other hyperparameters in (6) are known.

In the absence of censored data – that is, a "complete" data setting – implementing MCMC to draw samples from the posterior distribution in (6) using a combination of Gibbs and Metropolis steps would be straightforward. Thus, our goal here is to describe how one can impute a censored observation, $Y(\mathbf{s}_i, t_j)$. One can treat this unobserved value as yet another unknown parameter in our model, and then estimate $\boldsymbol{\theta}$ and $\mathbf{Z}$ in the usual fashion. As described by Gelfand et al. (1992), we have two primary methods for sampling censored $Y(\mathbf{s}_i, t_j)$. The first option would be to generate values from $[Y(\mathbf{s}, t) \mid \boldsymbol{\theta}, \mathbf{Z}]$, an unconstrained Normal distribution, and reject those which are above the threshold, C. While simple to implement, this option may be inefficient, particularly if the likelihood of sampling a value below the censoring threshold is small.

Instead, we use an inverse CDF method like those described in Section 2.2 of Devroye (1986). First, note that the full conditional for each censored $Y(\mathbf{s}_i, t_j)$ is simply

$$p(Y(\mathbf{s}_i, t_j) \mid \boldsymbol{\theta}, \mathbf{Z}, \mathbf{Y}_{(i,j)}, \mathbf{D}) \propto N(Y(\mathbf{s}_i, t_j) \mid \mathbf{x}(\mathbf{s}_i, t_j)^T \boldsymbol{\beta} + Z(\mathbf{s}_i, t_j), \tau^2) \times I\{Y(\mathbf{s}_i, t_j) < C\}, \tag{7}$$

a truncated Normal distribution, where $\mathbf{Y}_{(i,j)}$ denotes the vector $\mathbf{Y}$ with $Y(\mathbf{s}_i, t_j)$ removed. In general, one may not possess functionality to generate random variates from a given truncated distribution; however, in these situations, a draw from the truncated distribution may be achieved by transforming a random uniform variate using its inverse CDF. Extending this to a collection of censored values (each potentially having a unique censoring threshold) requires additional steps in the MCMC algorithm.

Since our censored values can be imputed using the Gibbs sampler (allowing us to obtain a "complete" data set), our MCMC sampler can proceed as before for our other unknown parameters without sacrificing accuracy for computational ease. Furthermore, we can still take advantage of using sampling-based Bayesian inference by seamlessly obtaining inference on the residual space–time effects and prediction at arbitrary space–time coordinates, $(\mathbf{s}_0, t_0)$, by sampling from their respective posterior predictive distributions. Of course, prediction at these coordinates assumes that all covariates $\mathbf{x}(\mathbf{s}_0, t_0)$ are available. Similarly, inference on the spatiotemporal gradient process proceeds in posterior predictive fashion, requiring only the post-convergence MCMC samples for our spatiotemporal model parameters; thus, the methods for making inference on the gradients themselves are unaffected by the censoring present in the data.

4. Simulated data example

To demonstrate the effectiveness of our methods, we have conducted an experiment using data generated from a model on $\Re^2 \times \Re$ where the true underlying gradient process is known:

$$Y(\mathbf{s}_i, t_j) \sim N\left(5\left[\sin(s_{i1}3\pi) + \cos(s_{i2}3\pi)\cos(t_j\pi/7)\right], \tau^2\right). \tag{8}$$

In addition to allowing us to easily compute the true gradient process (using calculus), this expression also leads to some interesting features in the gradients that will be discussed later. Our data are generated from (8) where $\tau^2 = 1$ at $N_s = 100$ randomly placed locations on the unit square with coordinates (s_{i1}, s_{i2}), for $i = 1, \ldots, N_t$. Each site is observed $N_t = 15$ times at evenly spaced temporal increments, $t \in \{1, 2, \ldots, 15\}$, and all locations are observed at each time point, giving us a total of $N = 100 \times 15 = 1500$ observations. We then mimic the scenario in our real data by censoring half of our observations, enabling us to investigate the quality of our spatiotemporal gradient estimation in the presence of censored data. To analyze these data, we use an intercept-only regression model, allowing the Gaussian process to capture all spatiotemporal variability. Our results are based on 10,000 iterations of the MCMC sampler, discarding the first 5,000 as burn-in, with convergence diagnosed visually using trace plots.

In this example, we desire to censor 50% of our observations. Without loss of generality, we have centered our data such that a threshold of $C = 0$ will accomplish this goal. Fig. 2 displays the surface created by (8), where any shade of blue denotes a negative, and thus censored, value using

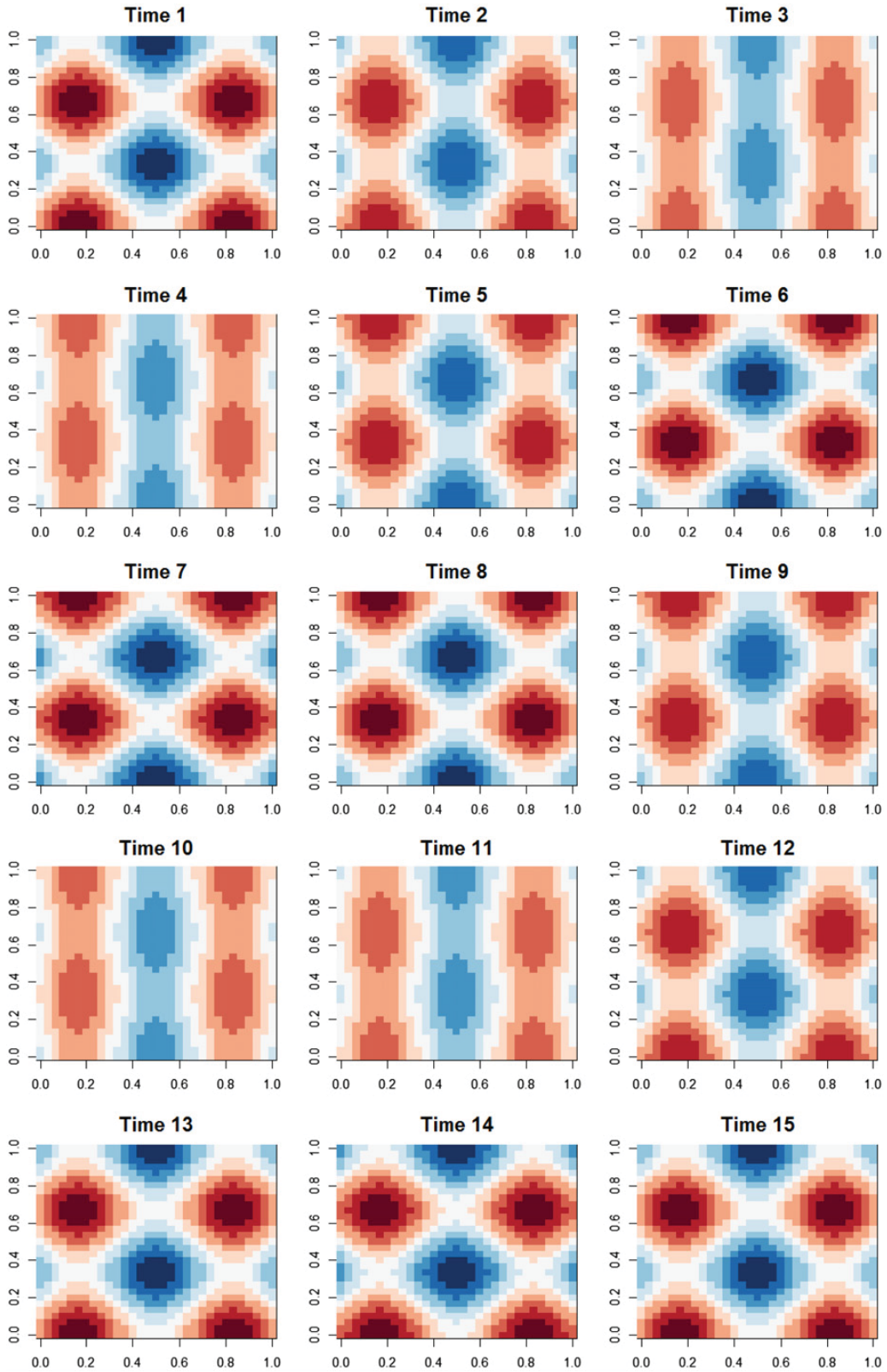
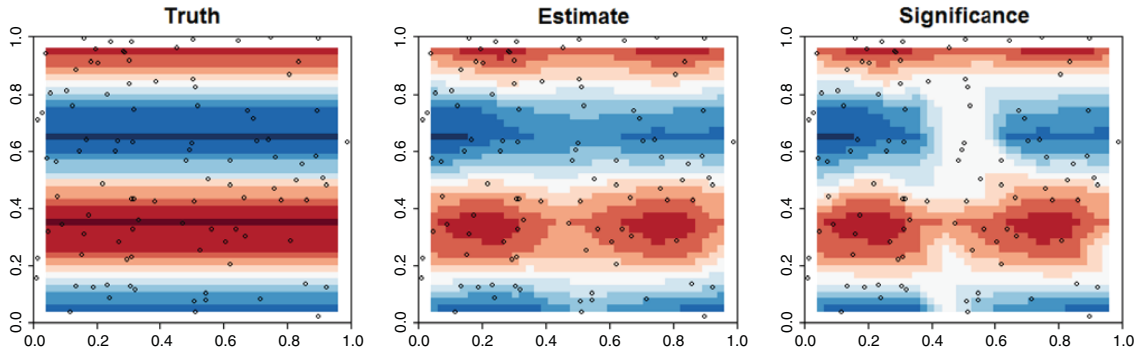
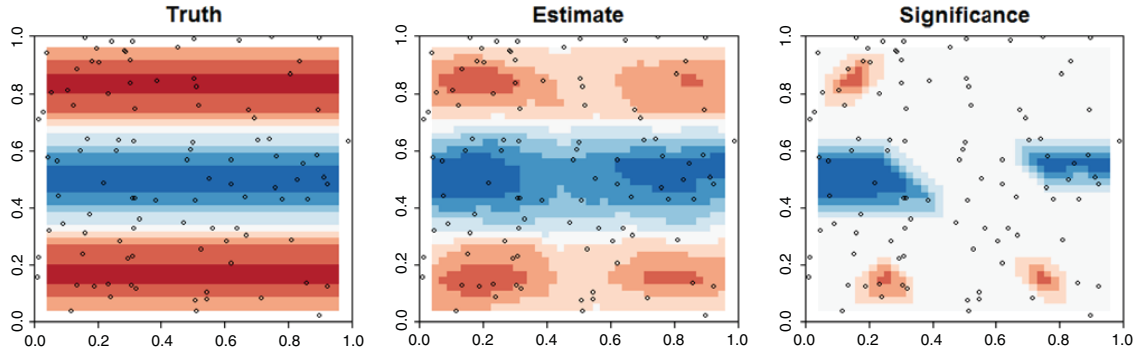


Fig. 2. Underlying process for the data generated by (8). Values range from −10 (dark blue) to 10 (dark red), where here any shade of blue denotes a negative, and thus censored, value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Temporal Gradient.



(b) s_{i2} Mixed Gradient.

Fig. 3. Comparison of the true temporal and the s_{i2} mixed gradients for time point 3 (see Fig. 2) and their posterior median values based on our gradient theory. The color scheme here goes from highly negative (blue) to highly positive (red), centered around 0 (light gray), with locations of observed locations plotted as open circles. The third panel highlights significant gradients; i.e., gradients whose 95% CIs contain 0 are assigned value 0, while gradients whose 95% CIs do not contain 0 are assigned their posterior medians. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

our threshold of $C = 0$. This figure is particularly helpful, as it helps us identify two features with which our spatial gradients will have to contend. First, we note that for values of our spatial coordinate $s_{i1} \in (0.4, 0.6)$, we are quite likely to encounter censored values, thus we should expect to learn very little information regarding our s_{i2} spatial gradients in this region. Coincidentally, this region corresponds to a relatively flat valley in the s_{i1} direction – that is, we *should not* have any interesting gradients here – so an investigation of these gradients should provide insight into how likely our model is to find significant gradients in a censored region when the underlying gradient process is quite flat. Second, the temporal gradients are constant across values of s_{i1}, so it will be easy to identify the effect censoring has on these gradients as well.

Our Bayes-MCMC procedure accurately estimates the posterior median of the error variance parameter as $\tau^2 = 0.95$ with a 95% Bayesian credible interval (CI) of (0.85, 1.06), covering its true value of 1 and providing nearly identical results to those obtained in a fully-observed data set. In fact, many of our findings are quite similar to those from the analysis of the uncensored data set, including the estimation and coverage of both our spatiotemporal random effects and gradients, where 96% of our random effects' true values and 99% of our true gradients are contained in their 95% CIs. In Fig. 3, we recreate the analogous figure from the uncensored example in Quick et al. (2013), in which the first column maps the true temporal and s_{i2} mixed gradients at time point 3, the second column maps the corresponding posterior median estimates, and the final column indicates the "posterior significance" of these gradients (as measured by the 95% Bayesian credible interval's exclusion of 0). The true horizontal bands are largely captured in our estimates and, to a lesser extent, their observed significance. Still, as compared to the uncensored case, the effect that censoring has on the interval

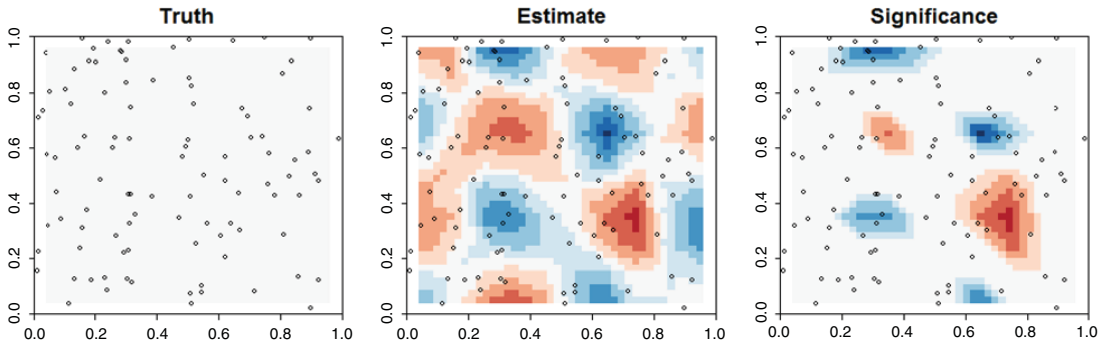


Fig. 4. Comparison of the true s_{i1} mixed gradients for time point 3 (see Fig. 2) and their posterior median values based on our gradient theory when using a substitution method. Values in the vertical band $s_{i1} \in (0.4, 0.6)$ have been substituted with $\log(1/2)$, resulting in biased estimates of the gradients in and around this region.

$s_{i1} \in (0.4, 0.6)$ is apparent, as these gradients are much less pronounced (and less likely to be significant) in this vertical band in the center of these plots.

In addition, we tested two additional examples, the first of which deals with the case of 80% censoring. This scenario is much more akin to the data we aim to analyze here, where 83.8% of our measurements are below detectable limits. In this case, the underlying process remains well estimated, with 92% of our random effects' true values being covered by their 95% CIs. Our gradients are also estimated well despite the extreme censoring, with 97% of our true gradients being covered by their 95% CIs. That being said, the impact of this extreme level of censoring relative to 50% censoring is felt when investigating *significance* in the gradients, particularly in the vertical band $s_{i1} \in (0.4, 0.6)$. In this region, we have no uncensored observations in any of our 15 time points, resulting in increased levels of variability in our random effects, which in turn increases the variability in the gradients. As this level of censoring is likely much more extreme than many researchers will encounter in practice, and as our results do not vary substantially between 50% and 80% censoring, we have chosen to focus our discussion here on the more informative scenario.

In our second additional example, we tested the claim proposed in the introduction that the use of a substitution method would result in poor estimation of the gradient surface. For this example, we returned to the 50% censoring case and substituted all negative values with $\log(1/2)$, a value determined by assuming our true data arose from a log-normal distribution with $\text{LOD} = \exp(0) = 1$ and substituted values equal to LOD/2. First and foremost, these substituted values led to substantial oversmoothing of our underlying process in which only 39% of our random effects' true values were covered by their 95% CIs. Estimation of our gradients was also hindered, with only 81% of the true gradients being covered by their 95% CI.

More concerning is that for the mixed gradients in the s_{i1} direction (i.e., the temporal change in the horizontal gradients), we find a number of *significant* gradients when the true gradient surface is flat, as seen in Fig. 4. In this figure, we observe clusters of significantly positive gradients around coordinates (0.3, 0.7) and (0.7, 0.3) and significantly negative gradients around (0.3, 0.3) and (0.7, 0.7) at time point 3. The locations correspond to boundaries between modest peaks and the vertical band of censored observations (see Fig. 2). In the true surface, the rate of decline/incline surrounding these peaks in the s_{i1} direction is the same for every time point (thus the lack of a mixed gradient). When we substitute all censored values with a function of the LOD, however, we have the potential to over- or underestimate the gradients at the boundary (e.g., if the true value is above the substituted value, we will overestimate the absolute gradient). In this case, detecting this bias is easier in the mixed gradient, as we find that our spatial gradients have changed over time. For instance, moving right from (0.3, 0.7) in this figure, we first encounter *positive* gradients—due to the substitution, our estimated spatial gradients at time point 4 is lower in magnitude (i.e., less negative) than it should be, and thus is lower in magnitude than the spatial gradient at time 3, resulting in an *increase* in the gradient.

Based on the results of these examples, it appears that even with high degree of censoring we are still able to make important inference on the spatiotemporal gradient process after imputing the censored values. While regions with a large amount of censoring are less likely to include *significant*

Table 1
Parameter estimates from the analysis of the BP data with 95% Bayesian credible intervals. ϕ_s is on a scale of 1 unit = 5 km, and ϕ_t is on a scale of 1 unit = 5 min.

Parameter	Median (95% CI)	Parameter	Median (95% CI)
β_0 (Intercept)	−5.324 (−5.934, −4.580)	ϕ_s (Spatial)	1.552 (0.484, 2.907)
τ^2 (Error Var)	0.729 (0.482, 1.093)	ϕ_t (Time)	0.043 (0.040, 0.062)
σ^2 (S–T Var)	6.424 (3.840, 9.906)		

gradients, our posterior estimates in these regions remain quite accurate, and there does not seem to be much loss of significance in the regions with values above the limit of detection. We have also shown that it is not advisable to simply replace censored observations with a space- and time-constant function of the LOD, as this can result in bias in the underlying residual process, as well as the gradient process.

5. Analysis of the *Deepwater Horizon* oil spill data

As previously described, the monitoring data we consider consist of $N = 447$ space- and time-specific measurements of VOC collected during the clean-up efforts of the *Deepwater Horizon* oil spill, 375 (83.8%) of which are below the limit of detection (0.1 ppm). The data were collected within 30 km of the site of the oil rig over a period of 30 days. Aside from the spatial location (in latitude and longitude) and time of observation (measured to the nearest second), we have no other covariates at the same level of spatial or temporal resolution as our response; as such, we will fit an intercept-only model to our data, allowing the Gaussian process to capture all spatiotemporal variability. Spatial distances are computed using a sinusoidal projection[1]: $s_{i1} = E\lambda\cos\theta$ and $s_{i2} = E\theta$, where $E = 6371$ km is the radius of the Earth, and θ and λ denote latitude and longitude, respectively, and are on a scale where one spatial unit equals 5 km. Temporal distances were computed by letting one temporal unit equal 5 min. Following the convention for air concentrations, we model our data on the log-scale, whence our censoring threshold is $C = \log(0.1) = -2.3$.

A summary of our model's parameter estimates based on 20,000 iterations of the MCMC sampler (10,000 used as burn-in) can be found in Table 1. Our variance parameters τ^2 and σ^2 can be used to estimate $\sigma^2/(\sigma^2+\tau^2)$ as 0.90, indicating that 90% of the variability in our data is being explained by the spatiotemporal process, which is not surprising given that we have no covariate information. Our spatial and temporal range parameters, ϕ_s and ϕ_t, can be interpreted as controlling how quickly the correlation between two observations drops off. For instance, plugging $\delta = 0$ into (2), we find that the spatial correlation between two observations at the same time point falls to less than 0.4 at a distance of roughly 6 km; similarly, the temporal correlation between two observations at the same location falls to less than 0.4 after roughly 2 h. One thing to note is that the covariance matrix in this analysis is sensitive to small values of ϕ_t; for values of $\phi_t < 0.04$, the covariance matrix is numerically singular. This is likely due to the extreme temporal extent of our data (30 days) compared to the temporal range (less than 2 h) and the minimum temporal distance of a few minutes. As a result, we have set our prior for ϕ_t such that it is restricted to values greater than 0.04. This phenomenon does not occur for the spatial component, though this could be due to the spatial extent (roughly 40 km) relative to the spatial range (6 km) or simply the difference between the Matérn structure used for spatial correlation and the alternative structure used for time. Regardless, more work is needed to determine the precise cause of this and, if necessary, identify a more appropriate correlation structure for analyzing data with a temporal extent that substantially dwarfs the important temporal range.

After fitting our model, we interpolated the hourly VOC surface for each day in our data set with a large number of observations. For instance, Fig. 5 displays 16 hourly surface plots for June 26, 2010, the day in our data set with the most non-censored observations. Here, "clouds" of red shading denote

[1] While a two-dimensional projection may induce some bias into our distance measurements, we believe that this is negligible due to our small spatial domain.

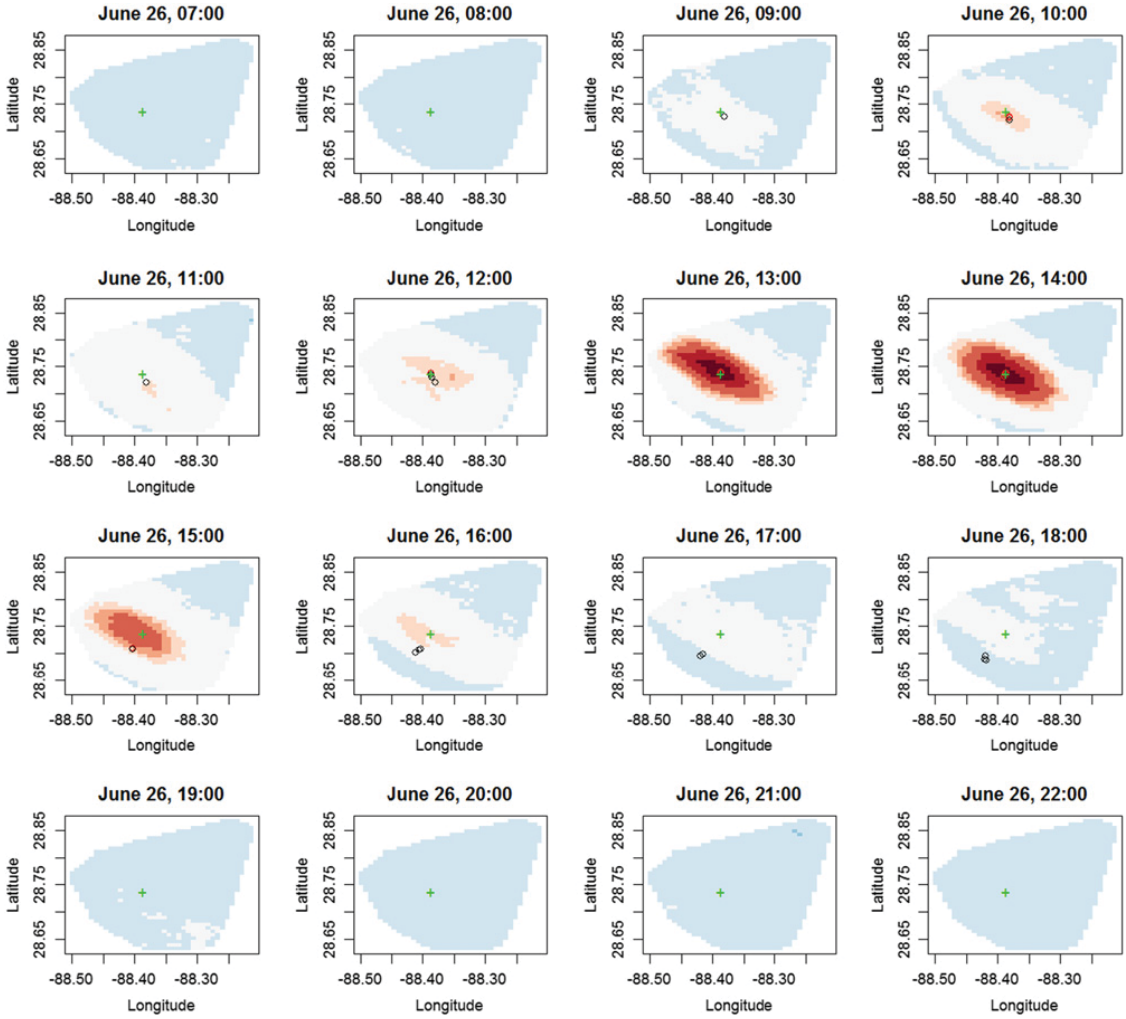


Fig. 5. Predicted VOC surface on June 26th for sixteen different time points, where blue and light gray shading denote levels below and near the limit of detection ($\log(0.1) = -2.3$), respectively, and darker shades of red indicate higher levels above detectable limits (up to 2 on the log-scale, or $\exp(2) = 7.39$ ppm). The area outside the observed area is shaded white and values have not been extrapolated. Here, circles denote observations within 30 min of the labeled times, where black circles denote censored observations and red denotes observations above 0.1 ppm. Green plus-sign (+) denotes the site of the oil rig. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

locations where we expect to observe levels of VOC above the limit of detection (darker shades of red indicating higher values) while light gray denotes values near the LOD and blue denotes censored values. These correspond to the red and black circles, which indicate values above and below the LOD, respectively. As we can see, the red cloud of high VOC levels appears to be centered around the site of the well, denoted here by a green plus-sign (+). Based on this figure, it would appear that something occurred around 12:00, which led to elevated concentrations of VOC over the next few hours. These data also reflect the challenge of using nonrandom data, as shown by the lack of observations on this day west of well. This figure can provide investigators an idea of how widespread the concentrations of VOC may have been, given the information that we have available.

In Fig. 6, we narrow our focus to the period from 12:00 to 15:00 and highlight the location of the ship on its 10 km trip in the affected area, during which concentrations above 2 ppm were observed from 12:40 until 14:19. Here, we can more clearly see that the dark shade of red in the 13:00 and 14:00 plots is the result of multiple uncensored observations, culminating in the recorded value of 8.6 ppm at 13:41.

In Fig. 7, we again focus on June 26, 2010, this time looking at the estimated spatial and temporal gradients at 13:00. Immediately, we find that the inflection point of both of our spatial gradients occurs

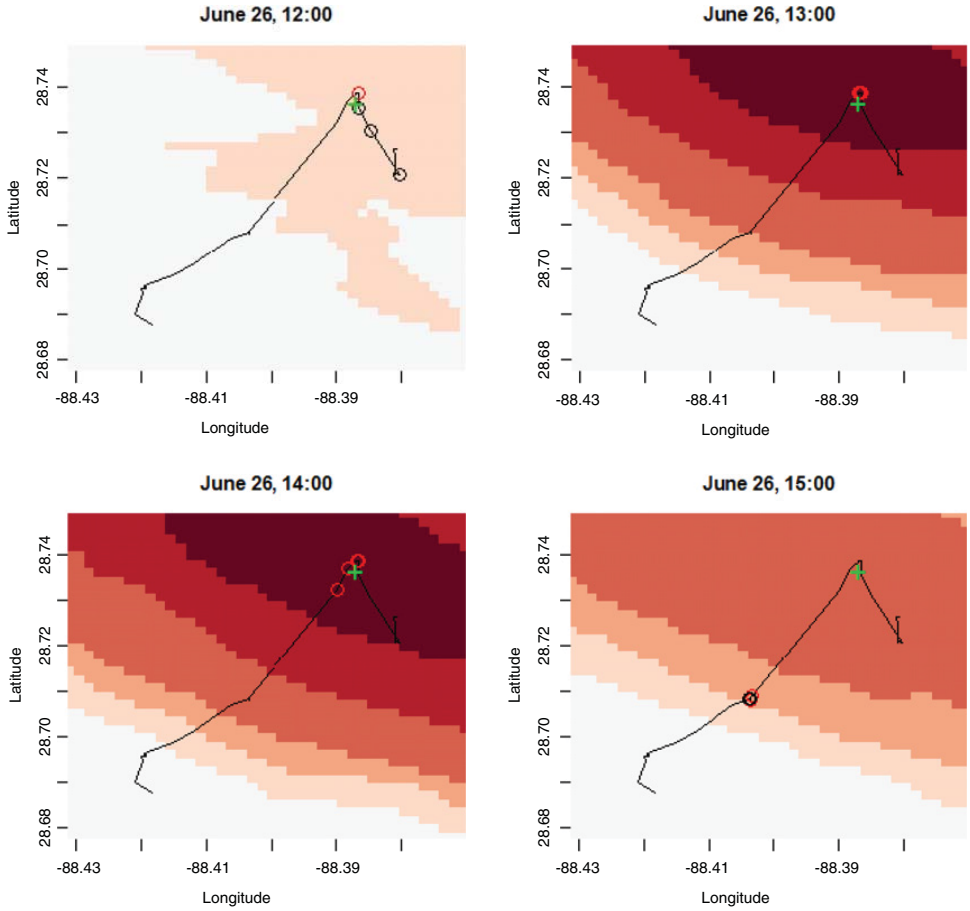


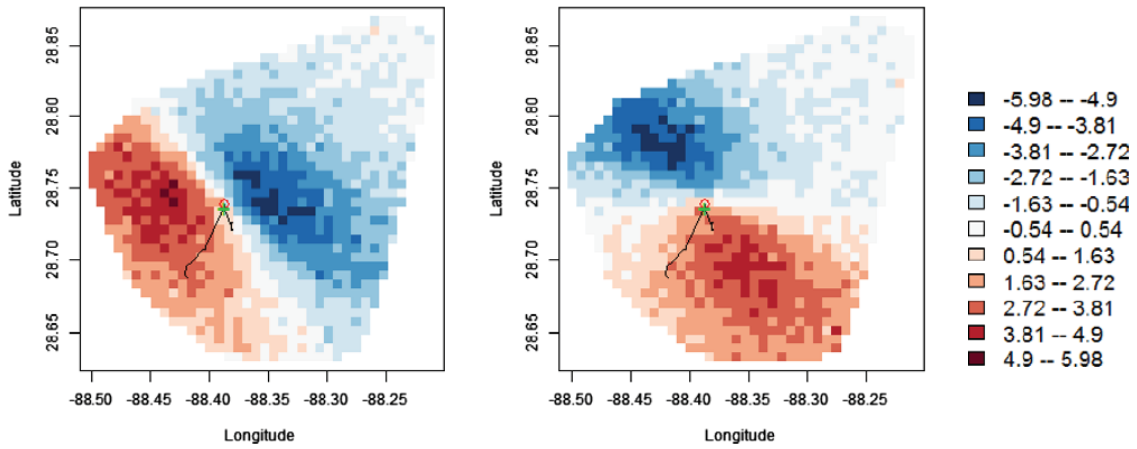
Fig. 6. Predicted VOC surface on June 26 for four different time points. Also plotted is a black line denoting the path of the ship collecting measurements at intervals of 12 min from 09:00 to 18:30. See Fig. 5 for information regarding shading and symbols used. For a sense of scale, the distance between the observations in the fourth panel is roughly 4 km away from the site of the spill.

at the site of the oil spill, as evidenced by the change from red to blue shading at this point. It is worth noting, however, that we may to some extent be picking up the boundary of our observations, as our observed data on this day are not dispersed evenly on all sides of the well site. As for changes over time, the area of positive temporal gradients centered around the well site indicate an increase in VOC concentrations (observed in Fig. 6), the largest observed value of VOC being the aforementioned concentration of 8.6 ppm at 13:41. Unfortunately, the amount of censoring in these data, coupled with the irregularity of the locations and times of the observations, appears to hinder our ability to identify *significant* gradients, and our mixed gradients (not shown) appear to capture little more than noise.

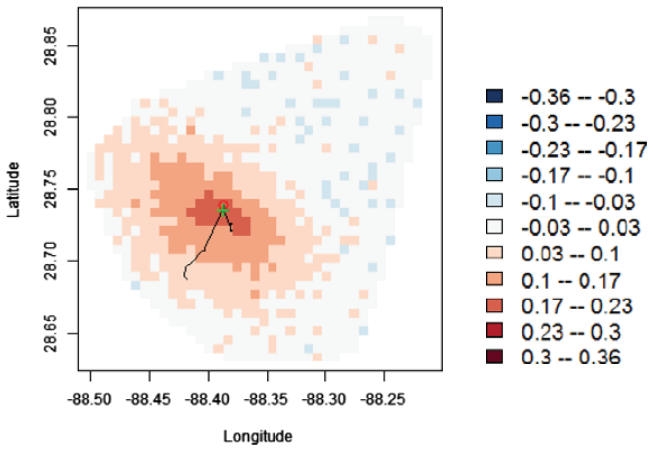
6. Discussion and conclusions

In this paper, we first described how one could fit a censored hierarchical model. Instead of imputing the censored values as described in Section 3, one may wish to integrate out the censored values and simply let their likelihood contribution be in the form of a Normal CDF. While equally correct to our approach, this method would have required Metropolis updates instead of Gibbs updates for many of our model parameters due to the lack of conjugacy, resulting in a dramatic increase in computational burden. By imputing the censored observations as part of our MCMC algorithm, we can fit highly censored data with only a moderate increase in computing time over a fully-observed data set.

In Section 4, we demonstrated the ability of our proposed methodology to estimate the underlying gradient process via a simulated data example. In addition to exploring the performance of our model



(a) E/W Spatial Gradient. (b) N/S Spatial Gradient. (c) Spatial Key.



(d) Temporal Gradient. (e) Temporal Key.

Fig. 7. Estimated east/west and north/south spatial gradients and temporal gradients for 13:00 on June 26 for the BP oil spill data. Negative predicted gradients are shaded in blue and positive gradients in red, with light gray shading corresponding to values near zero. See Figs. 5 and 6 for an explanation of the shading and symbols used. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

at two extreme levels of censoring, we also compared our method of imputing censored values to a substitution approach, and discussed the associated risks. Based on these results, it appears that the spatial and temporal gradients are not only well-estimated, but that the ability to identify significance in these gradients is not dramatically hindered by the presence of censored data. That said, the added uncertainty due to the censoring seems to make identifying significant mixed gradients more challenging, though they remain well-estimated by the model.

Having shown the ability to identify substantive findings in such a highly censored data setting, we proceeded to an analysis of data collected during the clean-up efforts of the *Deepwater Horizon* oil spill. Despite having less than 17% of our data above the LOD, our model was still able to deliver valuable insight into the air concentrations around the *Deepwater Horizon* well site. First and foremost, we provided estimates of the size and shape of the vapor clouds of VOCs by borrowing strength from other days in the study. Second, these results suggest that it is possible that the observed air concentrations may have been impacted by the clean-up efforts, such as the use of dispersants. Further investigation will be necessary to determine if information regarding the use of dispersants is sufficient for inclusion in our statistical model.

These data suggest, as exhibited on June 26 and on other days not discussed here, differences over time and space in VOC air concentrations. However, due to the nature of the data, i.e., a small data set, irregular in both space and time and highly censored, any substantive scientific conclusions based

on the results of this analysis would be pure speculation and thus have been avoided here. Using this methodology and the findings from this work, however, we hope to be able to analyze the larger data set, comprised of 25 million area measurements, in a much more efficient fashion and we may be able to identify short duration peak exposures occurring on the measured vessels. As the observations in the full data set encompass a larger, and certainly a more densely populated spatial region, we expect to observe significant gradients in both space and time. These significant gradients can then be used to construct zones of similar exposures, which can then be used to assign air concentration estimates to individual workers unobserved in our data.

Acknowledgments

We would like to thank Wendy McDowell and other members of the GuLF STUDY for their continued efforts. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences, as well as a contract with Social & Scientific Systems, Inc. (contract #S939000025). The opinions of this paper are those of the author and not necessarily of any funding body.

References

Adler, R.J., 2009. The Geometry of Random Fields. SIAM—Society for Industrial and Applied Mathematics, Philadelphia, PA.
Banerjee, S., Gelfand, A.E., 2002. Prediction, interpolation and regression for spatially misaligned datasets. Sankhya A 64, 227–245.
Banerjee, S., Gelfand, A.E., Sirmans, C.F., 2003. Directional rates of change under spatial process models. J. Amer. Statist. Assoc. 98, 946–954.
Carlin, B.P., Louis, T.A., 2009. Bayesian Methods for Data Analysis. Chapman and Hall/CRC Press, Boca Raton, FL.
Cressie, N.A.C., 1996. Change of support and the modifiable areal unit problem. Geogr. Syst. 3, 159–180.
Devroye, L., 1986. Non-Uniform Random Variate Generation. Springer-Verlag, New York.
Fridley, B.L., Dixon, P., 2007. Data augmentation for a Bayesian spatial model involving censored observations. Environmentrics 18, 107–123.
Ganser, G.H., Hewett, P., 2010. An accurate substitution method for analyzing censored data. J. Occup. Environ. Hygiene 7, 233–244.
Gelfand, A.E., 2010. Misaligned spatial data: The change of support problem. In: Gelfand, A.E., Diggle, P., Guttorp, P., Fuentes., M. (Eds.), Handbook of Spatial Statistics. CRC Press, Boca Raton, FL.
Gelfand, A.E., Smith, A.F.M., Lee, T.-M., 1992. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. J. Amer. Statist. Assoc. 87, 523–532.
Gelfand, A.E., Zhu, L., Carlin, B.P., 2001. On the change of support problem for spatio-temporal data. Biostatistics 2, 31–45.
Gneiting, T., 2002. Nonseparable, stationary covariance functions for space–time data. J. Amer. Statist. Assoc. 97, 590–600.
Gotway, C., Young, L., 2002. Combining incompatible spatial data. J. Amer. Statist. Assoc. 97, 632–648.
Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. Wiley-Interscience, New York, NY.
Mugglin, A.S., Carlin, B.P., Gelfand, A.E., 2000. Fully model-based approaches for spatially misaligned data. J. Amer. Statist. Assoc. 95, 877–887.
Quick, H., Banerjee, S., Carlin, B.P., 2013. On Bayesian inference for rates of change in spatiotemporal process settings. Research Report 2013–013, Division of Biostatistics, University of Minnesota.
Ren, Q., Banerjee, S., 2013. Hierarchical factor models for large spatially misaligned datasets: a low-rank predictive process approach. Biometrics 69, 19–30.
Yates, F., 1933. The analysis of replicated experiments when the field results are incomplete. Emp. J. Exp. Agric. 1, 129–142.