Bayesian Analysis (2006) **1**, Number 3, pp. 595–624

# Multivariate Spatiotemporal CDFs with Random Effects and Measurement Error

Margaret B. Short*, and Bradley P. Carlin†

**Abstract.**

Spatial cumulative distributions (SCDFs) are useful in environmental applications – for example, by helping assess the fraction of a region exposed to harmful pollutants. Data sets containing the requisite spatial information often contain temporal data as well. We therefore extend the notion of an SCDF to a *spatiotemporal cumulative distribution function* (STCDF), with the goal of increasing precision by making use of repeated measurements. Ours is a hierarchical Bayesian approach, with estimation carried out by Markov chain Monte Carlo (MCMC) methods. We develop linear algebra results and corresponding computational techniques to handle the difficulties in evaluating the likelihood wrought by the large data sets (due to the added temporal component), the inclusion of spatial and temporal random effects, the need to account for measurement error, and the handling of missing data. We illustrate the concepts in a univariate setting with an Atlanta ozone data set, and in a bivariate (two pollutant) setting with a California $NO/NO_2$ data set.

**Keywords:** Air pollution; Bayesian methods; Change of support problem; Kriging; Markov chain Monte Carlo (MCMC) methods

## 1 Introduction

Epidemiological data often and environmental data usually include a spatial component; such data may contain a temporal component as well. Statistical techniques that account for correlations, both spatial and temporal, are useful in providing a more comprehensive analysis than techniques that ignore them. For example, cancer control efforts can be greatly abetted by an understanding of the harmful exposures that may be associated with cancer risk. The spatial and temporal distribution of these exposures is critical.

In some cases, interest lies in the cumulative fraction of the spatial domain that is "exposed" at any given time. For instance, a study of childhood asthma might seek to determine the proportion of children exposed to harmful levels of one or more pollutants on each of many consecutive summer days. A concept that is useful in such settings is that of *spatiotemporal cumulative distribution functions*, or STCDFs for short. An STCDF is a function which, for a spatiotemporal process $Y(\mathbf{s}, t)$ with $\mathbf{s}$ in a spatial domain $D$ and $t$ in a temporal domain $D^*$, gives the fraction of $D$'s area for which

*Statistical Sciences Group, D-1, Los Alamos National Laboratory, Los Alamos, NM

†Division of Biostatistics, School of Public Health at the University of Minnesota, Minneapolis, MN, http://www.biostat.umn.edu/~brad

 

$Y(\mathbf{s},t)$ lies below a given value $y_0$ at time $t$. We define the STCDF $F_t(y_0)$ by

$$F_t(y_0) = \frac{1}{|D|}\int_D I(Y(\mathbf{s},t) \le y_0)d\mathbf{s}\ ,$$

where $|D|$ denotes the area of $D$. That is, we fix a time point $t$ and a cutoff value $y_0$ and define $F_t(y_0)$ as the fraction of the region $D$ which, at time $t$, has a process level below $y_0$. In our applications, $Y(\mathbf{s},t)$ is a point level process, the (possibly log-transformed) value of an air pollutant (e.g., ozone).

In a bivariate (e.g., two pollutant) setting, if $\{U(\mathbf{s},t), V(\mathbf{s},t)\}$ is a joint spatial process where again $\mathbf{s} \in D$ and $t \in D^*$, we define the *weighted* (by $V$) STCDF $F_{V,t}(u_0)$ by

$$F_{V,t}(u_0) = \frac{\int_D h(V(\mathbf{s},t))I(U(\mathbf{s},t) \le u_0)d\mathbf{s}}{\int_D h(V(\mathbf{s},t))d\mathbf{s}}\ . \tag{1}$$

The weighting function $h(V)$ must be nonnegative and strictly positive over the range of $V(\mathbf{s},t)$. A convenient and easy to interpret choice is $h(v) = I(v \le v_0)$. This choice implies that $F_{V,t}(v_0)$ is the answer to the question, "If at time $t$ we restrict our attention to the portion of $D$ which has (pollutant) $V$ at or below the threshold $v_0$, what fraction of the remaining area has (pollutant) $U$ at or below level $u_0$?" The choice $h \equiv 1$ results in an unweighted STCDF, and this is where we focus our attention in order to concentrate on the spatiotemporal issue.

The notion of spatial cumulative distribution functions (SCDFs), in which the temporal domain $D^*$ consists of a single time point, was pioneered by Overton (1989). Hierarchical Bayesian estimation of SCDFs using Monte Carlo methods was initially suggested by Handcock (1999). Previous related work in the area covers spatiotemporal prediction at a small number of site-time combinations (Gelfand et al. (2001)), spatial prediction for bivariate processes (Banerjee and Gelfand (2002)), and bivariate, conditional, and weighted SCDFs in the absence of a nugget (Short et al. (2005)).

In this paper, we extend the SCDF model considered in Short et al. (2005) to the spatiotemporal case. We also introduce spatial and temporal random effects, as well as a nugget term to account for measurement error. These new model developments in turn require us to develop more sophisticated computational techniques. We adopt a hierarchical Bayesian approach, assume an underlying Gaussian structure, and carry out prediction and estimation via Markov chain Monte Carlo (MCMC) integration. We impose a correlation structure that, were there no nugget term, would be separable (Banerjee et al. 2004, Ch. 8). The inclusion of a nugget term adds greatly to the computational burden, since our covariances are no longer Kronecker products but rather sums of Kronecker products of matrices. This complicates our MCMC calculations at both the parameter sampling and spatiotemporal prediction stages, primarily because the matrices are quite large due to the temporal component. Finally, we allow for missing data, which is likely to occur in spatiotemporal data sets. For example, a new monitoring site might be added to a network, or one pollutant might not be recorded at one or more time points due to transient equipment failure.

The remainder of our paper evolves as follows. Section 2 presents our motivating data sets. Section 3 then covers the basics of both univariate and bivariate STCDF prediction. In Section 4, we explore these concepts in two settings. The first of these contains information on a single pollutant (ozone) in the Atlanta area; this is a complete data set. In the other setting, we have NO and $NO_2$ measurements taken across central and southern California; this bivariate example involves missing data. Finally, Section 5 offers a brief discussion and directions for future research.

# 2 Motivating data sets

## 2.1 Atlanta ozone data

Our first illustrative data set consists of daily one hour peak ozone measurements taken at ten monitoring sites in the Atlanta metropolitan area during July, 1995. Figure 1 shows the locations of the ten sites. In this figure, the green outline encloses the domain $D$ of 36 zip codes on which we carry out prediction. Figure 2 shows a time series plot of the data for this 31-day period. The solid lines correspond to the four sites in or closest to the 36 zip codes of interest, and the dashed lines correspond to the remaining six sites; again, refer to Figure 1. A potentially helpful covariate, the daily high temperature, is also shown as a dotted line (with scale on the right vertical axis). The lag-1 temporal autocorrelations for the 10 sites range from .282 to .740, suggesting that a temporal component should be included in our model. For this time period, no data are missing.

The possible differential exposure of whites and blacks to the harmful effects of ozone, and how this might change over time, is of interest to researchers assessing environmental justice. This motivates estimation of both weighted and unweighted STCDFs in this problem. Here, however, the spatial misalignment between the point-level ozone observations and the zip-level racial population counts $n_j$, $j = 1, \ldots, 36$, requires an empirical Bayes-type approximation when computing the weighted SCDF (1); for details see (Short et al. 2005, Section 3).

## 2.2 California air pollution data

Our second data set features NO and $NO_2$ measurements at 82 monitoring sites in central and southern California taken at 3 hour intervals on July 12, 2001. These data are from the publicly accessible website `www.arb.ca.gov/aqd/aqdcd/aqdcddld.htm`. Figure 3 gives the locations of the sites, as well as a grid of 500 sites at which prediction is of interest. Figure 4 shows contour plots of NO and $NO_2$ at eight p.m. on the day in question. It appears from these graphs that NO and $NO_2$ are positively spatially correlated: both show a high peak down around Los Angeles and a local minimum to the north near Sacramento. In fact, the sample correlation between log(NO) and log($NO_2$) ranges between .497 and .885 over these eight time points.

Both NO and $NO_2$ are pollutants associated with lung dysfunction, so we require a simultaneous evaluation of the proportion of the region negatively impacted by each pollutant. The spatial and temporal correlations evident in Figure 4 as well as the time

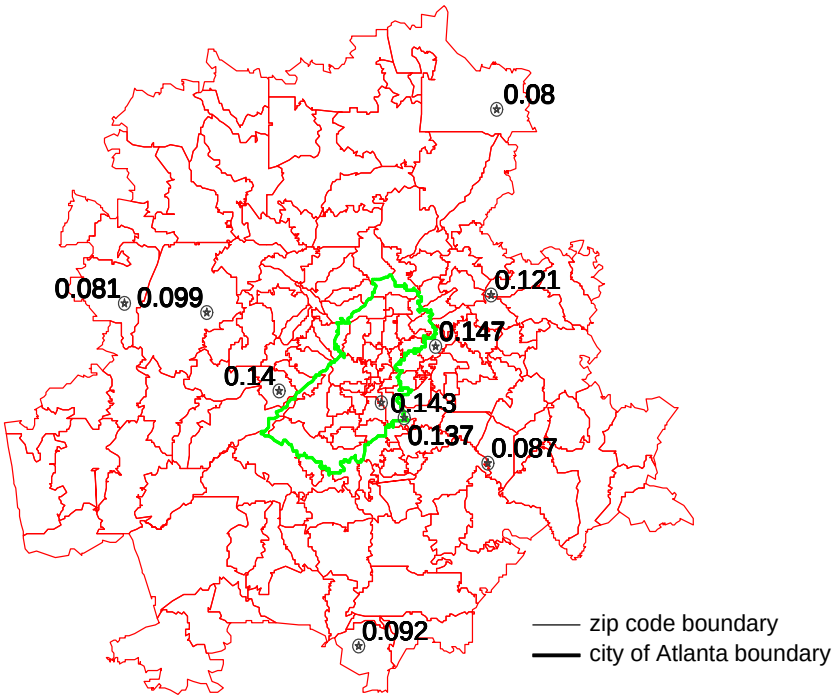


Figure 1: Zip code boundaries in the Atlanta metropolitan area and eight-hour maximum ozone levels (ppm) at the 10 monitoring sites for July 10, 1995.

series plot in Figure 5 (which for easier viewing shows just half of the sites, selected at random) suggest the use of bivariate STCDFs. In particular, there is a clear temporal trend evident in the time series plot of the two pollutants. This suggests we may also wish to include a time-dependent term in the mean structure.

## 3 Spatiotemporal CDF estimation

There are close analogies between STCDF prediction in the univariate and bivariate settings. Throughout this paper, we assume an underlying Gaussian process, construct a hierarchical model, and use Bayes-MCMC methods to estimate parameters, combining Gibbs steps with Metropolis steps when closed-form full conditional distributions are not available. We carry out spatial prediction at a fairly dense set of new spatial locations, distributed across the region of interest. We may thin the samples when carrying out prediction at new sites if sample autocorrelations are high. Sampling is quick for the univariate setting in which there are no missing data, but not for the bivariate setting with missing data and a nugget term. Lastly, we estimate SCDFs by (in essence) estimating an SCDF (or weighted SCDF) for each kriged prediction and then forming

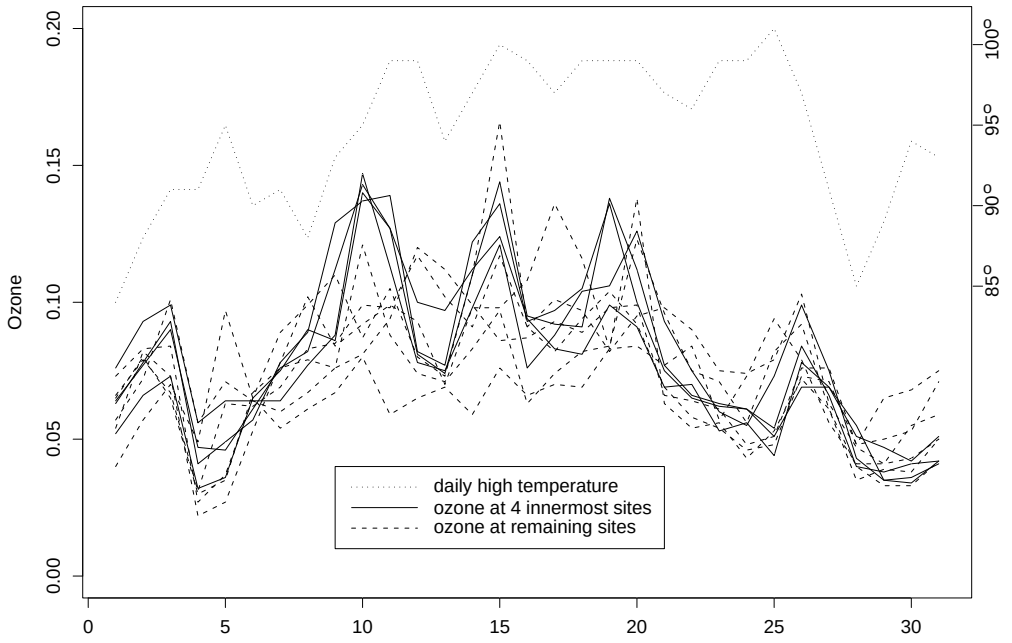


Figure 2: Atlanta ozone and temperature time series plot, July 1995. July 1 is a Saturday.

the pointwise average of the resulting curves.

## 3.1 Univariate setting

Let $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$ be the sites and $\mathbf{t} = (t_1, \ldots, t_N)$ be the discretely indexed times at which we have observations. Let $Y_{ij}$ be the observed measurement at site $\mathbf{s}_i$ at time $t_j$, and define $Y_{obs} = (Y_{11}, Y_{12}, \ldots, Y_{1N}, \ldots, Y_{n1}, Y_{n2}, \ldots, Y_{nN})^T$. We assume for the moment that all data are present and that $Y_{obs} \sim MVN(X\boldsymbol{\beta}, \Sigma)$, where $X$ is an $nN \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, and

$$\Sigma = \sigma^2 H(\lambda_1) \otimes C(\lambda_2) + \tau^2 I \,, \tag{2}$$

where $H(\lambda_1)$ is an $n \times n$ matrix corresponding to a two-dimensional spatial correlation function, $C(\lambda_2)$ is an $N \times N$ matrix corresponding to a one-dimensional temporal correlation function, and $\otimes$ denotes the Kronecker product. Were there no nugget term $\tau^2$, the Kronecker structure imposed on $\Sigma$ would imply that the space-time correlations were separable. Note that the spatial-only problem is a special case of the spatiotemporal problem in which $C = 1$. We assume for now that $h_{ij}$, the $(i, j)$ entry of $H$, is given by $h_{ij} = \exp(-\lambda_1 ||\mathbf{s}_i - \mathbf{s}_j||)$, where $|| \cdot ||$ is a suitable distance metric. We

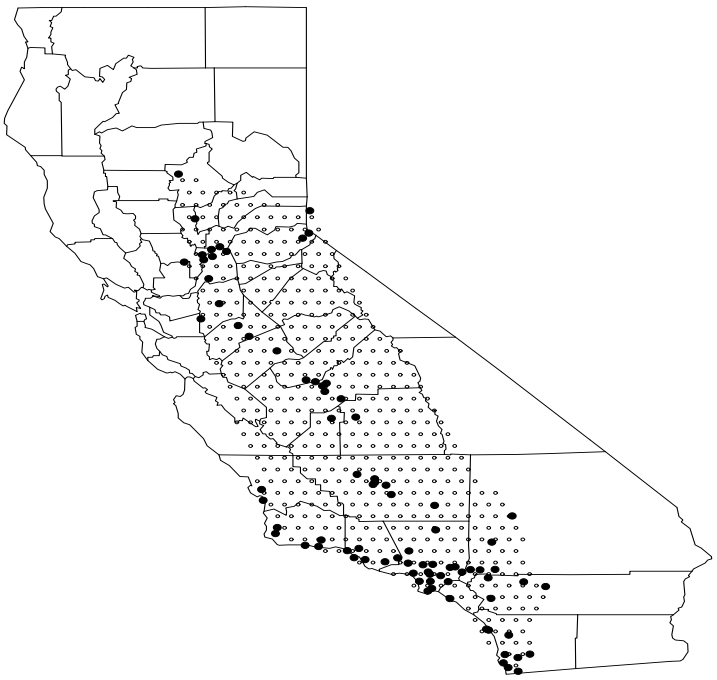

Figure 3: Locations of NO and $NO_2$ monitoring sites, California air quality data; 500 selected target locations are also shown.

further assume that $c_{ij}$, the $(i, j)$ entry of $C$, is given by $c_{ij} = \exp(-\lambda_2|t_i - t_j|)$. Thus correlations are assumed to decay exponentially in both time and in space. Other correlation structures, such as the spherical, Matérn, and so on, are certainly possible. We explore several mean structures, including ones that incorporate fixed or random effects for time, or a time-varying covariate.

### Sampling parameters

Consider first the inclusion of a temporal component in the separable (no nugget) form corresponding to a Kronecker product for the covariance structure. That is, for now we assume $Y_{obs} \sim MVN(X\boldsymbol{\beta}, \sigma^2 H(\lambda_1) \otimes C(\lambda_2))$. It is straightforward to show that the (possibly multivariate) Gaussian distribution offers a conjugate prior for $\boldsymbol{\beta}$, and an inverse gamma distribution is a conjugate prior for $\sigma^2$. The use of gamma priors for $\lambda_1$ and $\lambda_2$ results in nonstandard full conditional distributions for these parameters, so we use Metropolis steps to sample them.

When including a nugget term as in (2), the normal distribution again offers a

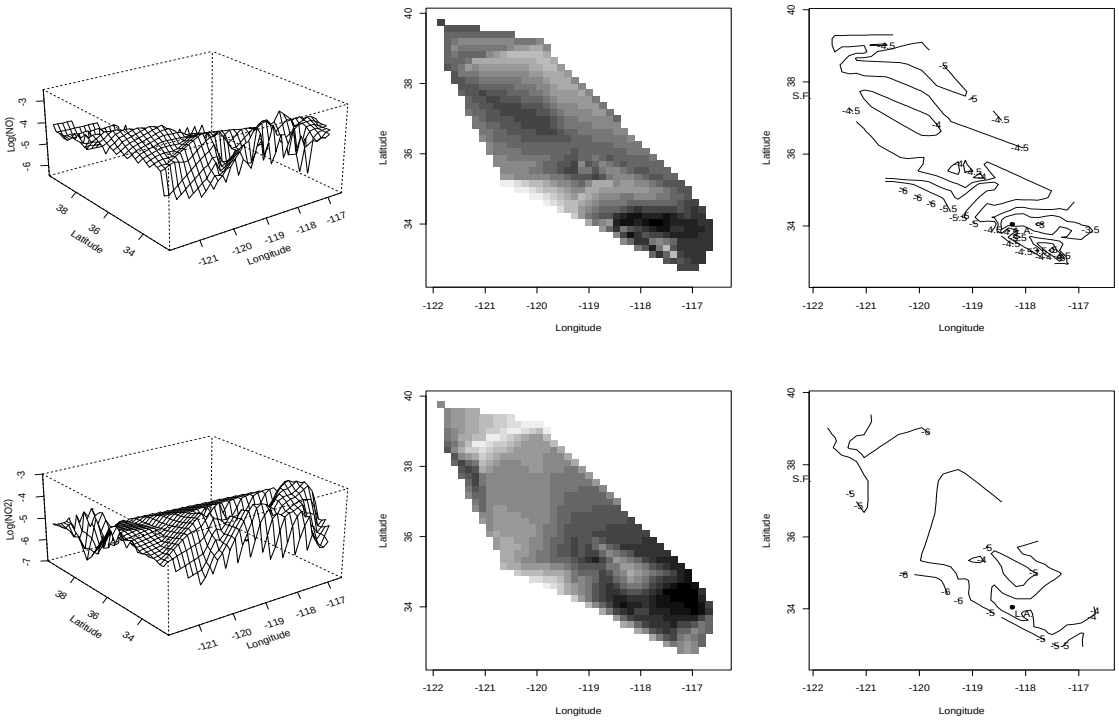


Figure 4: Interpolated perspective, image, and contour plots of the raw log-NO (first row) and log-$NO_2$ (second row), California air quality data, July 12, 2001, eight p.m.

conjugate prior for $\boldsymbol{\beta}$, but the conjugacy of the inverse gamma prior for $\sigma^2$ is lost; conjugate priors for $\lambda_1$ or $\lambda_2$ also do not exist. A possible approach is to reformulate these models adding random effects (Banerjee et al. 2004, Sec. 5.1). This approach allows inverse gamma distributions to emerge as conjugate priors for $\sigma^2$ and $\tau^2$. In our settings, test code for this approach ran very quickly, but the convergence of the chains was unacceptably slow. As such, we stayed with our original approach, using a vague normal (conjugate) prior for $\boldsymbol{\beta}$, $\text{IG}(a_1, b_1)$ and $\text{IG}(a_2, b_2)$ priors for $\sigma^2$ and $\tau^2$, and $\text{G}(c_1, d_1)$ and $\text{G}(c_2, d_2)$ priors for $\lambda_1$ and $\lambda_2$, where $\text{IG}(a, b)$ denotes the inverse gamma distribution with mean $1/[b(a-1)]$ and variance $1/[b^2(a-1)^2(a-2)]$, and $\text{G}(c, d)$ denotes the gamma distribution with mean $cd$ and variance $cd^2$. At first glance, this marginal model approach appears to impose the condition that $nN$ not be "too large", since we require several matrix inversions per iteration. However, eigenvalue-eigenvector decompositions (described in the appendices) allow us to work with moderately large data sets.

For all parameters except $\boldsymbol{\beta}$, we used Metropolis steps in which the proposal densities were lognormal, centered on the previous value of the chain. (Here and in all our other simulations, Metropolis proposal variances were fine-tuned so that the empirical acceptance rate fell between approximately 20% and 45%, as encouraged by

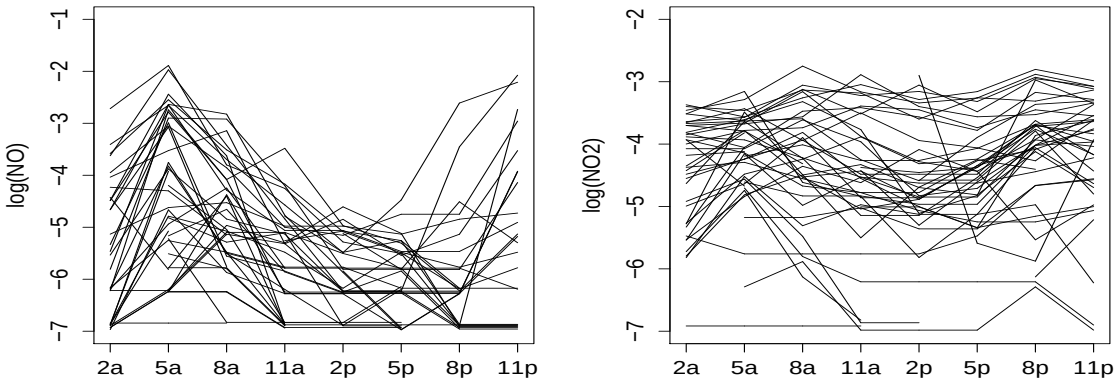


Figure 5: California air quality time series plots, selected sites, July 12, 2001.

Gelman et al. (1996). Convergence was diagnosed in several ways, including inspection of trace plots and Gelman-Rubin plots for multiple overlaid chains, and examination of autocorrelation plots.

A potential bottleneck in the parameter-sampling step lies in the computation of the log-likelihood, since it may be necessary to evaluate this quantity several times for every MCMC iteration. Eigenvector-eigenvalue approaches for speeding up this computation are described in Appendices A.1 and A.2 for the univariate and bivariate cases, respectively. The Atlanta data set was sufficiently small that we did not need these tricks, but the bivariate version was necessary to carry out our California data analysis, so we include both for completeness.

### Prediction at new sites

To carry out prediction at a collection of new sites $\tilde{\mathbf{s}} = (\tilde{\mathbf{s}}_1, \ldots, \tilde{\mathbf{s}}_L)^T$, at time points that are the same as those of the observed values, we must draw from the posterior predictive distribution,

$$f(Y_{pred}|Y_{obs}) = \int f(Y_{pred}|Y_{obs}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|Y_{obs}) d\boldsymbol{\theta}, \tag{3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \lambda_1, \lambda_2, \tau^2)$, $Y_{pred} = (\tilde{Y}_{11}, \ldots, \tilde{Y}_{1N}, \ldots, \tilde{Y}_{L1}, \ldots, \tilde{Y}_{LN})^T$, and $\tilde{Y}_{ij}$ corresponds to site $\tilde{s}_i$ at time $t_j$. We assume

$$\begin{pmatrix} Y_{obs} \\ Y_{pred} \end{pmatrix} \sim MVN\left( \begin{pmatrix} X_{obs} \\ X_{pred} \end{pmatrix} \boldsymbol{\beta}\ ,\ \sigma^2 H \otimes C + \tau^2 \begin{pmatrix} I_{n\times n} & 0 \\ 0 & 0_{L\times L} \end{pmatrix} \otimes I_{N\times N} \right),$$

where $n$ and $L$ are the number of old and new sites respectively, $N$ is the number of times, and $X_{obs}$ and $X_{pred}$ are $nN \times p$ and $LN \times p$ matrices of predictors, respectively. Here $H = \left(\begin{smallmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{smallmatrix}\right)$, with $H_{11}$ (respectively $H_{12}$ and $H_{22}$) of size $n \times n$ (respectively

$n \times L$ and $L \times L$) and having $(i, j)$ entry equal to $\text{cor}(Y_{ij}, Y_{ik})$ giving correlations between old sites (respectively $\text{cor}(Y_{ij}, \tilde{Y}_{ik})$ and $\text{cor}(\tilde{Y}_{ij}, \tilde{Y}_{ik})$) for $k = 1, \ldots, N$. $H$ is symmetric, so $H_{21} = H_{12}^T$. Thus the observations are modeled with measurement error, but we make predictions without measurement error (Cressie 1993, p. 128). The standard formula for the conditional distribution of Gaussian random variables gives us $Y_{pred} \mid Y_{obs}, \boldsymbol{\theta} \sim MVN(\boldsymbol{\mu}_{1\cdot 2}, \Sigma_{1\cdot 2})$, where

$$\boldsymbol{\mu}_{1\cdot 2} = X_{pred}\boldsymbol{\beta} + (\sigma^2 H_{21} \otimes C)(\sigma^2 H_{11} \otimes C + \tau^2 I_{nN \times nN})^{-1}(Y_{obs} - X_{obs}\boldsymbol{\beta}) \ ,$$

and

$$\Sigma_{1\cdot 2} = (\sigma^2 H_{22} \otimes C) - (\sigma^2 H_{21} \otimes C)(\sigma^2 H_{11} \otimes C + \tau^2 I_{nN \times nN})^{-1}(\sigma^2 H_{12} \otimes C) \ .$$

In principle, this distribution is easy to draw from since it is multivariate normal. However, to ensure accurate STCDF estimation, we wish to carry out spatial prediction at as many sites as is computationally feasible. For the Atlanta data set, with $N = 31$ days and $L = 1000$ sites, $\Sigma_{1\cdot 2}$ is a 31,000 $\times$ 31,000 matrix, which is too large to hold or manipulate easily in a desktop PC's memory, and also too large to Cholesky-factorize. The principal result of this section, spelled out in the following paragraphs, is a method that allows us to carry out simultaneous prediction under these circumstances.

First, if there were no nugget term, $\Sigma_{1\cdot 2}$ would simplify to

$$\Sigma_{1\cdot 2} = \sigma^2(H_{22} - H_{21}H_{11}^{-1}H_{12}) \otimes C,$$

which is a Kronecker product. Since, in general, $(H \otimes C)^{1/2} = H^{1/2} \otimes C^{1/2}$, it would now be straightforward to draw from $MVN(\boldsymbol{\mu}_{1\cdot 2}, \Sigma_{1\cdot 2})$. In the presence of a nugget, if $\Sigma_{1\cdot 2}$ could be written as the sum of (positive definite) matrices rather than the difference of matrices, we could feasibly (although not necessarily easily) draw from $MVN(\boldsymbol{\mu}_{1\cdot 2}, \Sigma_{1\cdot 2})$. The following linear algebra identity, which appears in a slightly different form in Golub and Van Loan (1996), enables us to rewrite $\Sigma_{1\cdot 2}$ as a sum and thus carry out prediction for our Atlanta data set, as described in the remainder of this section.

**Identity 1**: If $A^{-1}$ exists and $\lambda \in \mathbb{R}$, then $(A + \lambda I)^{-1} = A^{-1} - \lambda A^{-1}(A + \lambda I)^{-1}$.

We apply this identity in our setting with $\lambda = \tau^2$ and $A = \sigma^2 H_{11} \otimes C$, to obtain

$$\Sigma_{1\cdot 2} = (\sigma^2 H_{22} \otimes C) - (\sigma^2 H_{21} \otimes C)\left[(\sigma^2 H_{11} \otimes C)^{-1} - \tau^2(\sigma^2 H_{11} \otimes C)^{-1} A_1^{-1}\right](\sigma^2 H_{12} \otimes C) \ ,$$

where $A_1 = A + \lambda I = \sigma^2 H_{11} \otimes C + \tau^2 I_{nN \times nN}$. Upon rearranging, we find $\Sigma_{1\cdot 2} = \Sigma_1 + \Sigma_2$, where $\Sigma_1 = \sigma^2(H_{22} - H_{21}H_{11}^{-1}H_{12}) \otimes C$, $\Sigma_2 = \tau^2(\sigma^2 H_{21} \otimes C)A_2(\sigma^2 H_{12} \otimes C)$, and $A_2 = [A_1(\sigma^2 H_{11} \otimes C)]^{-1} = (\sigma^4 H_{11}^2 \otimes C^2 + \tau^2\sigma^2 H_{11} \otimes C)^{-1}$. Thus we can draw $Z \sim MVN(0, \Sigma_{1\cdot 2})$ by drawing $Z_i \overset{ind}{\sim} MVN(0, \Sigma_i)$, $i = 1, 2$, and obtaining $Z = Z_1 + Z_2$. It is easy to draw $Z_1 \sim MVN(0, \Sigma_1)$ since $\Sigma_1$ is a Kronecker product, and we can draw $Z_2$ from $MVN(0, \Sigma_2)$ by drawing $Z_2^* \sim MVN(0, A_2)$ and defining $Z_2 = \tau(\sigma^2 H_{21} \otimes C)Z_2^*$. Note that $A_2$ is positive definite and thus $\Sigma_2$ is positive definite. The key point concerning $\Sigma_2$ is that while $\Sigma_2$ is large ($NL \times NL$) in our case, $A_2$ is relatively small

($nL \times nL$, where $n << N$) and therefore it is easy to draw from $MVN(\cdot, A_2)$. Furthermore, the multiplication implied by $(\sigma^2 H_{21} \otimes C) Z_2^*$ can be greatly facilitated by means of the following:

**Identity 2**: If $H$ is an $L \times L$ matrix, $C$ is a $N \times N$ matrix, and $Z$ is a $LN \times 1$ column vector, then

$$(H \otimes C) Z = \sum_{i=1}^{L} H_{\cdot,i} \otimes (C Z_i), \tag{4}$$

where we have written $Z = (Z_1^T, \ldots, Z_L^T)^T$ (each $Z_i$ is $N \times 1$), and used $H_{\cdot,i}$ to denote the $i^{th}$ column of $H$. That is, we can evaluate the product $(H \otimes C)Z$ without ever forming the Kronecker product, by adding together the Kronecker products of column vectors $H_{\cdot,i}$ and $CZ_i$. This result is easy to establish via the definition of Kronecker product and block matrix multiplication.

This deceptively simple result has two important consequences. First, it enables us to evaluate the product without forming a potentially enormous Kronecker product matrix. Second, carrying out the computation by use of this identity requires fewer computer operations than the brute force approach, since we need to perform $2L^2N$ multiplications instead of $2L^2N^2$, and $LN(N+L-2)$ additions rather than $LN(LN-1)$. In the event that $C = I_{N \times N}$, the time savings are even greater.

Note also that Identity 2 does not require that $H$ or $C$ have any special structure. In fact, $H$ need not be square. If $H$ is $p \times q$, $C$ is $N \times N$ and $Z$ is $qN \times 1$, then $(H \otimes C)Z = \sum_{i=1}^{q} H_{\cdot,i} \otimes (CZ_i)$, where now $H_{\cdot,i}$ is $p \times 1$ and, as before, each $Z_i$ is $N \times 1$.

Prediction is accomplished by using values from the parameter-sampling step to obtain draws from the posterior predictive distribution. Because we assumed the underlying process is multivariate normal, the prediction step consists of drawing first from the posterior distribution, then from the full conditional for $Y_{pred}$ given the data $Y_{obs}$ and the posterior draws $\boldsymbol{\theta}^{(g)}$. By composition, (3) implies the result is a draw from $p(Y_{pred}|Y_{obs})$.

### Estimation of STCDFs

The basic method for finding the predictive mean of the SCDF is described in Short et al. (2005); what follows is the extension to the spatiotemporal case. Whereas in the spatial setting we create a table of $\widehat{F}^{(g)}(w_k)$ values containing $K$ rows (one for each gridded value $w_k$ of the pollutant) and $G$ columns (one for each Gibbs sample), we now create $N$ such $K \times G$ tables – one for each of the $N$ time points. Thus the entry for the unweighted STCDF corresponding to sample $g$ at time $t_i$ for ozone value $w_k$ is obtained by evaluating the expression,

$$\widehat{F}_{t_i}^{(g)}(w_k) = \frac{1}{L} \sum_{l=1}^{L} I(Y^{(g)}(\tilde{\mathbf{s}}_l, t_i) \leq w_k).$$

Our final STCDF predictive mean estimate for time $t_i$ is obtained via pointwise averaging; that is, $\widehat{F}_{t_i}(w_k) = \frac{1}{G}\sum_{g=1}^{G} \widehat{F}_{t_i}^{(g)}(w_k)$. In essence, we are obtaining an STCDF estimate for each Gibbs sample, and then taking the pointwise mean of these estimates as our final answer.

## 3.2 Bivariate setting

Let $U_{obs} = (U(\mathbf{s}_1)^T, \ldots, U(\mathbf{s}_n)^T)^T$, where $U(\mathbf{s}_i) = (U(\mathbf{s}_i, t_1) \ldots, U(\mathbf{s}_i, t_N))^T$; similarly let $V_{obs} = (V(\mathbf{s}_1)^T, \ldots, V(\mathbf{s}_n)^T)^T$. Define $Y_{obs} = (U_{obs}^T, V_{obs}^T)^T$ as the spatiotemporal processes we wish to model jointly, and assume

$$Y_{obs} \sim MVN\left(X\boldsymbol{\beta}\,,\; T \otimes H(\lambda_1) \otimes C(\lambda_2) + S \otimes I_{nN \times nN}\right)\,, \tag{5}$$

where

$$S = \begin{pmatrix} \tau_U^2 & 0 \\ 0 & \tau_V^2 \end{pmatrix}\,,$$

$X$ is a $2nN \times 2p$ matrix of predictors, and $\boldsymbol{\beta}$ is a $2p \times 1$ vector of parameters. As in the univariate case, we explored several forms for $X$; these are discussed in Section 4.2.

Here, as before, $H(\lambda_1)$ and $C(\lambda_2)$ correspond to 2-dimensional and 1-dimensional correlation functions, respectively. The California data set to which our methods will be applied covers a large geographic region (unlike the Atlanta data set). Thus it would be at best questionable to use Euclidean distance in our calculations. Instead, we use great circle distance (*gcd*), which is the length of the shortest path that is constrained to lie on the surface of a sphere; see, for example, Banerjee (2005). As before, we assume exponentially decaying spatial correlations – i.e., $h_{ij} = \exp(-\lambda_1 gcd(\mathbf{s}_i, \mathbf{s}_j))$. The $2 \times 2$ matrix $T$ is assumed to be positive definite. As before, it specifies $\{\mathrm{Cov}(U(\mathbf{s}_i, t_j), V(\mathbf{s}_i, t_j))\}$, the within-site covariances.

### Sampling parameters, including missing data

In the no-nugget case, we assume $Y_{obs} \sim MVN(X\boldsymbol{\beta}\,,\; T \otimes H(\lambda_1) \otimes C(\lambda_2))$. Generalizing the spatial-only case, it is straightforward to show that a bivariate normal distribution offers a conjugate prior for $\boldsymbol{\beta}$ and that an inverse Wishart (IW) distribution is a conjugate prior for $T$.

When we include a nugget term, however, our spatiotemporal model becomes (5). As in the univariate settings and the bivariate no-nugget settings, the normal prior remains conjugate for $\boldsymbol{\beta}$, but the conjugacy of the inverse Wishart prior for $T$ is destroyed. As such, we again place inverse gamma priors on $\tau_U^2$ and $\tau_V^2$ and gamma priors on $\lambda_1$ and $\lambda_2$, and use Metropolis steps to update all model parameters other than $\boldsymbol{\beta}$.

The California $NO/NO_2$ data set is incomplete, in that at several site-time combinations, both NO and $NO_2$ values are missing. Although (only) entire $NO/NO_2$ pairs were missing in our California data set, at no point do our methods require this to be the form of the missingness. In general terms, if our modeling assumption for $n$ observations

is $Y \sim MVN(X\boldsymbol{\beta}, \Sigma)$, and $Y_{i_1}, \ldots, Y_{i_r}$ are missing where $1 \le i_1 < i_2 < \cdots < i_r \le n$, then there exists a permutation matrix $A_{n\times n}$ such that the final $r$ components of $AY$ are $(Y_{i_1}, \ldots, Y_{i_r})$. That is, if $Y_{mis} = (Y_{i_1}, \ldots, Y_{i_r})^T$, and if $Y_{obs} = (Y_{j_1}, Y_{j_2}, \ldots, Y_{j_{n-r}})^T$ corresponds to actual observations, (so that $\{i_1, \ldots, i_r\} \cup \{j_1, \ldots, j_{n-r}\} = \{1, \ldots, n\}$ and $\{i_1, \ldots, i_r\} \cap \{j_1, \ldots, j_{n-r}\} = \emptyset$), then

$$AY_{full} = \begin{pmatrix} Y_{obs} \\ Y_{mis} \end{pmatrix} \sim MVN(AX\boldsymbol{\beta}\,,\; A\Sigma A^T).$$

We insert into the parameter-sampling stage an additional step to draw $Y_{mis}|Y_{obs}$ by applying the standard formulae for drawing from a conditional normal distribution. Note that multiplication of a matrix (respectively, vector) on the left by a permutation matrix is equivalent to a series of row (respectively, coordinate) swapping operations; multiplication on the right by a permutation matrix is equivalent to a series of column swaps. In software it is far more efficient to swap rows and columns than to perform matrix multiplication, so this is the route we chose.

Two bottlenecks arose in the parameter-sampling step, and we later describe the eigenvalue-eigenvector approaches we devised to hasten the computations. Appendix A.2 describes a method for speeding up the calculation of the full-data log-likelihood calculations; Appendix A.3 shows how to reduce the time needed to draw from the conditional distribution $Y_{mis}|Y_{obs}$. Without these methods, it would not be possible to carry out the MCMC for our California data set in a reasonable amount of time. Even with these improvements, it took approximately two hours for every 1000 iterations using a 3gHz desktop PC.

### Prediction at new sites

We wish to carry out prediction for a large number of sites, which means we again encounter difficulties in handling large matrices. For the California data set, with 82 monitoring sites, 500 prediction sites and 8 time points, the associated covariance matrices are $9312 \times 9312$. Without the method we are about to describe, a generalization of the result in Section 3.1, we would be unable to carry out the prediction step.

Again, let $S = \left(\begin{smallmatrix} \tau_U^2 & 0 \\ 0 & \tau_V^2 \end{smallmatrix}\right)$. Our modeling assumption is that

$$\begin{pmatrix} Y_{full} \\ Y_{pred} \end{pmatrix} \sim MVN\left(\begin{pmatrix} X_{obs} \\ X_{pred} \end{pmatrix}\boldsymbol{\beta}\,,\; H\otimes T\otimes C + \begin{pmatrix} I_{n\times n} & 0 \\ 0 & 0_{L\times L} \end{pmatrix}\otimes S\otimes I_{N\times N}\right),$$

where $X_{obs}$ (respectively $X_{pred}$) is a $2nN \times p$ (respectively $2LN \times p$) matrix of predictors and $L$ is the number of sites at which we wish to make predictions. Then $Y_{pred}|Y_{full}, \boldsymbol{\theta} \sim MVN(\boldsymbol{\mu}_{1\cdot 2}, \Sigma_{1\cdot 2})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda_1, \lambda_2, T, \tau_U^2, \tau_V^2)$,

$$\boldsymbol{\mu}_{1\cdot 2} = X_{pred}\boldsymbol{\beta} + (H_{21}\otimes T\otimes C)(H_{11}\otimes T\otimes C + I_n\otimes S\otimes I_N)^{-1}(Y_{obs} - X_{obs}\boldsymbol{\beta}),$$

and

$$\Sigma_{1\cdot 2} = H_{22}\otimes T\otimes C - (H_{21}\otimes T\otimes C)(H_{11}\otimes T\otimes C + I_n\otimes S\otimes I_N)^{-1}(H_{12}\otimes T\otimes C)\,.$$

Here we are partitioning $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$, as in Section 3.1, where $H_{11}$ is $n \times n$, $H_{22}$ is $L \times L$, $H_{12}$ is $n \times L$, and $H_{21} = H_{12}^T$. As with the univariate case, it can be shown that $\Sigma_{1\cdot 2} = \Sigma_1 + \Sigma_2$, where $\Sigma_1$ and $\Sigma_2$ are positive definite matrices given by $\Sigma_1 = (H_{22} - H_{21}H_{11}^{-1}H_{12}) \otimes T \otimes C$, $\Sigma_2 = (H_{21} \otimes T \otimes C) A_2 (H_{12} \otimes T \otimes C)$, and

$$\begin{aligned} A_2 &= [(H_{11} \otimes T \otimes C + I_n \otimes S \otimes I_N)(H_{11} \otimes (S^{-1}T) \otimes C)]^{-1} \\ &= [H_{11}^2 \otimes (TS^{-1}T) \otimes C^2 + H_{11} \otimes T \otimes C]^{-1}. \end{aligned}$$

Thus we can draw $Z \sim MVN(\boldsymbol{\mu}_{1\cdot 2}, \Sigma_{1\cdot 2})$ by drawing $Z_1 \sim MVN(0, \Sigma_1)$ and $Z_2^* \sim MVN(0, A_2)$, defining $Z_2 = (H_{21} \otimes T \otimes C) Z_2^*$, and setting $Z = \boldsymbol{\mu}_{1\cdot 2} + Z_1 + Z_2$. We emphasize that although $\Sigma_2$ is quite large in our data setting ($2NL \times 2NL$), $A_2$ is relatively small ($2nL \times 2nL$) and therefore of manageable size.

The incorporation of missing data adds no appreciable difficulties to the prediction step. We merely use the values $Y_{mis}$ that were drawn in the sampling stage and overwrite them in the correct positions of $Y_{full}$.

We remark that for the parameter sampling stage, it was easier to calculate the full conditional distribution for $T$ when the data was ordered as $T \otimes H \otimes C$. For prediction, the order $H \otimes T \otimes C$ was easier to handle since all observation sites precede all prediction sites in the components of $Y$. The change in the order of the factors $T$, $H$ and $C$ corresponds to multiplication of $Y$ by a permutation matrix, and this multiplication adds no appreciable computational burden.

### Estimation of STCDFs

Our primary quantities of interest are the STCDF $F_{t_i}(u_k)$ and the weighted STCDF $F_{V,t_i}(u_k)$, both evaluated over a grid of values $U = u_k$ for each time point $t_i$, and in the latter case weighted by a function $V$. Working in the slightly more general weighted case, we estimate this first by obtaining the Monte Carlo approximant of equation (1), given by

$$F_{V,t_i}^{(g)}(u_0) = \frac{\frac{1}{L}\sum_{l=1}^{L} h(V^{(g)}(\tilde{\mathbf{s}}_l, t_i)) I(U^{(g)}(\tilde{\mathbf{s}}_l, t_i) \le u_0)}{\frac{1}{L}\sum_{l=1}^{L} h(V^{(g)}(\tilde{\mathbf{s}}_l, t_i))} \tag{6}$$

for each Gibbs sample $g$. We then take the sample mean of this predictive distribution, $\widehat{F}_{V,t_i}(u_0) = \frac{1}{G}\sum_{g=1}^{G} F_{V,t_i}^{(g)}(u_0)$, as our final estimate of the weighted STCDF. In (6), we might take $h(V) = I(V \le v_0)$ (*indicator* weighting); $h(V) = 1$ of course delivers the unweighted STCDF estimate. Note that missing data plays no direct role at this stage; this has already been accounted for in the earlier parameter sampling and spatiotemporal prediction stages.

# 4 Examples

## 4.1 Atlanta ozone data

We apply the methods of Subsection 3.1 to our Atlanta ozone data set, which consists of daily ozone measurements at ten monitoring sites for July, 1995. Refer to Figures 1

and 2, which show the locations of the monitoring sites and a time series plot of the raw data. As mentioned previously, we explore models containing several mean structures, which we now describe. In each of these, the $nN \times p$ matrix of predictors, $X$, consists of $n$ vertically stacked copies of some $N \times p$ matrix $X_0$. This special form is a consequence of the fact that while we may include time-specific terms in the mean structure, we have not included any site-specific terms. Recall that $Y_{ij}$ corresponds to site $\mathbf{s}_i$ at time $t_j$.

- Model A: $E(Y_{ij}) = \mu_0$, $i = 1, \ldots, n$, $j = 1, \ldots, N$. This is the simplest model, and it assumes an overall mean only. We place a vague normal prior on $\mu_0$.
- Model B: $E(Y_{ij}) = \mu_j$, with independent vague normal priors on the $\mu_j$. This corresponds to fixed effects for time. Thus $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_N)^T$, $p = N$, and $X_0 = I_{N \times N}$.
- Model C: $E(Y_{ij}) = \mu_0 + \mu_j$, with a vague normal prior on $\mu_0$, $\mu_j \overset{iid}{\sim} N(0, \sigma_\mu^2)$, and a vague inverse gamma prior on $\sigma_\mu^2$. This results in random effects for time. Thus $\boldsymbol{\beta} = (\mu_0, \mu_1, \ldots, \mu_N)^T$, $p = N + 1$, and $X_0 = [\mathbf{1}_N, I_{N \times N}]$.
- Model D: $E(Y_{ij}) = \mu_0 + k_j \beta_0$, where $k_j$ is a covariate, the daily high temperature for the Atlanta area, and $\beta_0 \in I\!R^1$. We place vague normal priors on $\mu_0$ and $\beta_0$.
- Model E: This contains the covariate from Model D along with the time random effects of Model C.
- Model F: $E(Y_{ij}) = \mu_0 + \mu_j$, corresponding to correlated random effects for time. We place a vague normal prior on $\mu_0$ and an AR(1) prior on the $\mu_j$ by letting $\mu_{j+1} - \mu_0 = \rho * (\mu_j - \mu_0) + \epsilon_j$, with $\rho \sim$ Uniform(–1,1) and $\epsilon \sim N(0, \sigma_\epsilon^2)$.
- Model G: This model consists of separate spatial-only models for each time point. That is, the software is run separately for each time point, each with its own overall mean $\mu_j$.

For the spatiotemporal model with exponential correlations (in both space and time) and a nugget term, we specified the gamma prior distribution for $\lambda_1$ by setting the prior mean of the effective spatial range equal to one half the maximum diagonal distance and prior standard deviation equal to half the prior mean. This resulted in a G(4, 0.109) prior for $\lambda_1$. A similar rationale led to a $G(4, 0.05)$ prior distribution for $\lambda_2$, which controls temporal correlations. Independent inverse gamma priors with mean 0.05 and variance 10 were placed on $\sigma^2$ and $\tau^2$. The sum of these means was chosen to be approximately the median sample variance over the 31 daily values of log-ozone. The full conditionals for these four parameters do not correspond to familiar distributions, thus we used Metropolis updates for them. In each case, proposal densities were log-normal, centered on the previous value of the chain (recall that $\lambda_1, \lambda_2, \sigma^2$, and $\tau^2$ must all be positive). In each of the models listed above, the stated prior distributions are conjugate for $\boldsymbol{\beta}$.

Convergence assessment consisted of examination of overlaid trace plots (three chains, each of length 1000), Gelman-Rubin plots, and lag-1 autocorrelations; convergence appeared to be rapid. We then generated a production chain of 55,000 iterations, discarded the first 5000 iterations as burn-in, and retained every fiftieth iteration thereafter for prediction, resulting in a total posterior sample size of 1000. We view the use of burn-in plus thinning as a fairly minor inconvenience, since parameter sampling ran quickly, and in any case very large output files are undesirable since their overconsumption of RAM slows STCDF estimation.

For the prediction step, we used 1000 sites that were approximately uniformly distributed across the 36 Atlanta zip codes. We used the prediction method described in Subsection 3.1, which allowed us to predict simultaneously the true underlying value of ozone (as opposed to observed ozone) at each of the 1000 sites on the 31 days of July, 1995. We then predicted unweighted STCDFs using $K = 100$ grid points equally spaced between .02 and .20, which was roughly the spread of ozone in the raw data set, by exponentiating the log-ozone values that were predicted in the preceding step. Results from several days for Models A, D, E, and G (all with nugget) appear in Figure 6. For all intents and purposes, the STCDFs corresponding to the models other than Model G are identical; this remained unchanged with or without the nugget. Among the selected days, July 15 appears to have the highest average ozone exposure, and July 4 the lowest. This is generally consistent with Figure 2, the time series plot.

We carried out model selection using the DIC statistic Spiegelhalter et al. (2002). Smaller values of DIC correspond to models with a better combination of fit and parsimony, while $p_D$ is the model's effective number of parameters. Results for models with and without a nugget appear in Table 1. In addition to the models just described, we considered versions of Models A, D, and F with simplified (but still exponentially decaying) spatial correlations. One set assumed compound symmetry for temporal correlations (i.e., $c_{ij} = \lambda_2$ if $i \neq j$, and $c_{ii} \equiv 1$). For these models, we used a Uniform(0,1) prior on $\lambda_2$, which precluded negative temporal correlations. The other set assumed independent times – that is, $C = I_{N \times N}$. Model D, the spatiotemporal model with exponential correlations (in both space and time), a nugget term, and an overall mean term and a single covariate (daily high temperature), emerges as the preferred model.

Before settling on Model D as our final choice based on its DIC score, we examined the residual plots shown in Figure 7. In each of these model-specific plots, fitted values were positively correlated with residual values. Sample correlations $r$ (see plot titles) are smallest for Models B and D. Note that going from uncorrelated random effects for time (Model C) to correlated random effects for time (Model F) greatly improves the appearance of the residual plot.

We have already noted that our computations are feasible largely thanks to the Kronecker product (or products, when a nugget is present) form of the covariance matrix, which allows us to take computational "short cuts." This means that we are depending on the assumption of space-time separability. To check this assumption, rather than adopt a formal testing framework (as in e.g. Fuentes (2003)), we follow the idea of a *separability plot* (Banerjee et al. 2004, p. 298). Briefly, these authors suggest that if

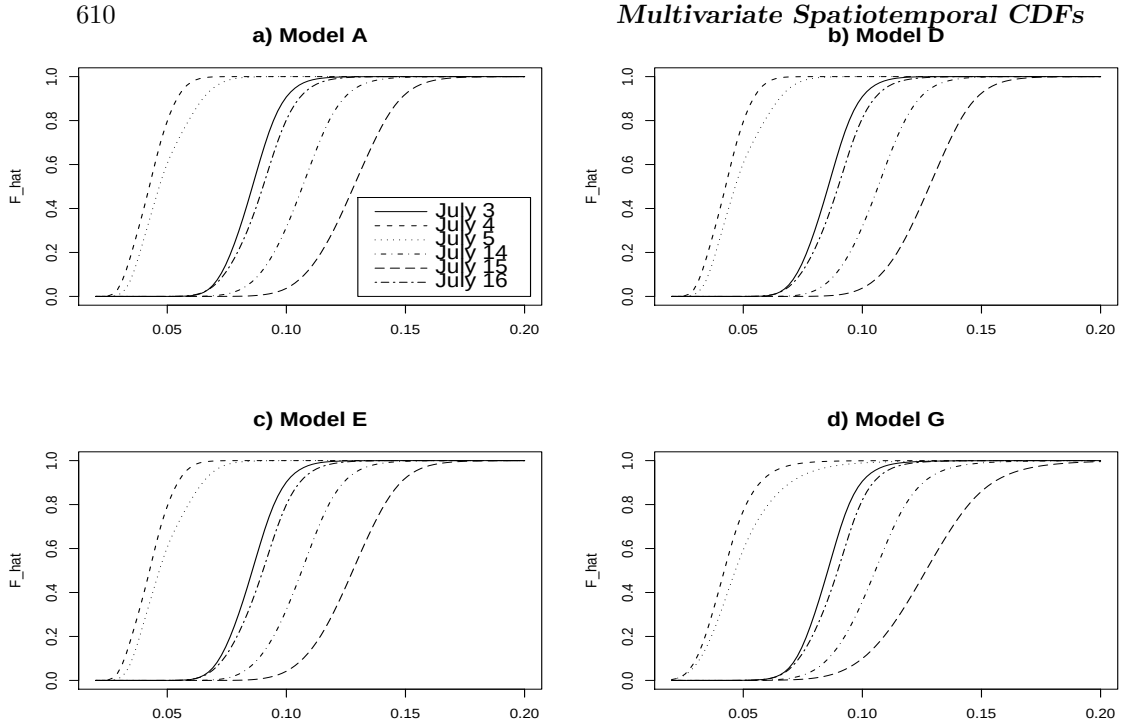


Figure 6: Atlanta unweighted STCDFs for July, 1995, all with nugget: (a) Model A ($\mu_0$ only); (b) Model D ($\mu_0$ plus temperature covariate); (c) Model E (temperature covariate plus random effects for time); (d) Model G (separate spatial model for each time).

$Y(\mathbf{s}_i, t_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ arise from a mean-zero stationary spatiotemporal process, we let $a_{ii'} = \sum_{j=1}^{m} Y(\mathbf{s}_i, t_j)Y(\mathbf{s}_{i'}, t_j)/m$, $b_{jj'} = \sum_{i=1}^{n} Y(\mathbf{s}_i, t_j)Y(\mathbf{s}_i, t_{j'})/n$, and $c_{ii',jj'} = Y(\mathbf{s}_i, t_j)Y(\mathbf{s}_{i'}, t_{j'})$. Then under separability, a plot of $c_{ii',jj'}$ versus $a_{ii'} \cdot b_{jj'}$ should lie roughly along a straight line; departures from linearity thus indicate a problem with our separability assumption. Figure 8 shows separability plots for the raw data (after centering around the grand mean) and for Models A, D and E (after subtracting their fitted values from the raw data); plots are again annotated with sample correlations $r$. The plots for Models A and D look reasonably good, while that of Model E suggests a possible problem in this case.

Finally, we performed a mild check on the robustness of our model to changes in the prior distribution by rerunning Models A, D and E (each with nugget) using weaker priors on $\lambda_1$ and $\lambda_2$. Using the same prior mean but ten times the prior variance for these parameters did not affect the appearance of the final STCDFs (plots omitted).

| Model | Temporal Correlation | Nugget DIC ($p_D$) | No nugget DIC ($p_D$) |
|---|---|---|---|
| A ($\mu_0$ only) | Exponential | 86.6 (4.16) | 89.2 (3.56) |
| B (f.e. for time) | Exponential | 104.0 (31.13) | 101.1 (31.12) |
| C (r.e. for time) | Exponential | 91.7 (25.66) | 91.5 (26.03) |
| D ($\mu_0$ + temp.) | Exponential | 83.5 (5.11) | 85.4 (4.53) |
| E (temp + r.e.(time)) | Exponential | 89.8 (26.30) | 89.7 (26.44) |
| F (AR(1) for r.e.(time)) | Exponential | 100.5 (30.59) | 99.2 (30.95) |
| G (separate times) | NA | 136.0 (30.39) | 131.5 (30.71) |
| A ($\mu_0$ only) | C.S. | 138.8 (3.22) | 137.8 (3.11) |
| A ($\mu_0$ only) | Identity | 144.6 (2.72) | 143.5 (2.72) |
| D ($\mu_0$ + temp.) | C.S. | 131.9 (4.23) | 131.0 (4.16) |
| D ($\mu_0$ + temp.) | Identity | 138.1 (3.75) | 136.9 (3.73) |
| F (AR(1) for r.e.(time)) | C.S. | 154.6 (29.8) | 160.9 (29.41) |
| F (AR(1) for r.e.(time)) | Identity | 162.4 (29.15) | 160.9 (29.04) |

Table 1: DIC for univariate models, Atlanta ozone data.

## 4.2 California air quality data

In this second example, California air pollution observations from July 12, 2001 were selected for inclusion if they were taken at any of the time points 2 a.m., 5 a.m., ..., 11 p.m., except that an entire site was omitted if both pollutant measurements were missing at all eight time points. Values of 0.0 corresponded to readings that were below the threshold for detection and were replaced with a value of .001, the smallest value found elsewhere in the data set. (Smaller replacement values were also investigated; we discuss this briefly below.)

As mentioned earlier, Figure 5 shows a time series plot of this data. Note that NO peaks in early morning and bottoms out in mid-afternoon. $NO_2$ appears to dip somewhat in midday, but this is less clear from the plot. In constructing this plot, a distinct random jitter was added to the log(NO) value at each site so the curves would separate visually. In fact, the resolution of the data was quite coarse; all values were multiples of .001, and there were just two digits of accuracy. For this data set, 35 values of NO were missing, as were the values of $NO_2$ at these same sites.

In the bivariate setting, we explore the following four mean structures:

- Model A: The model contains an overall mean for each type of pollutant. That is, $E(U_{ij}) = \mu_{U0}$ and $E(V_{ij}) = \mu_{V0}$. We place a vague normal independent priors on $\mu_{U0}$ and $\mu_{V0}$.

- Model B: This model contains uncorrelated random effects for site. That is, $E(U_{ij}) = \mu_{U0} + \mu_{U,i}$ and $E(V_{ij}) = \mu_{V0} + \mu_{V,i}$. Priors and hyperpriors for $\mu_{U,i}$ and $\mu_{V,i}$ are analogous to those used for time random effects in the univariate model.

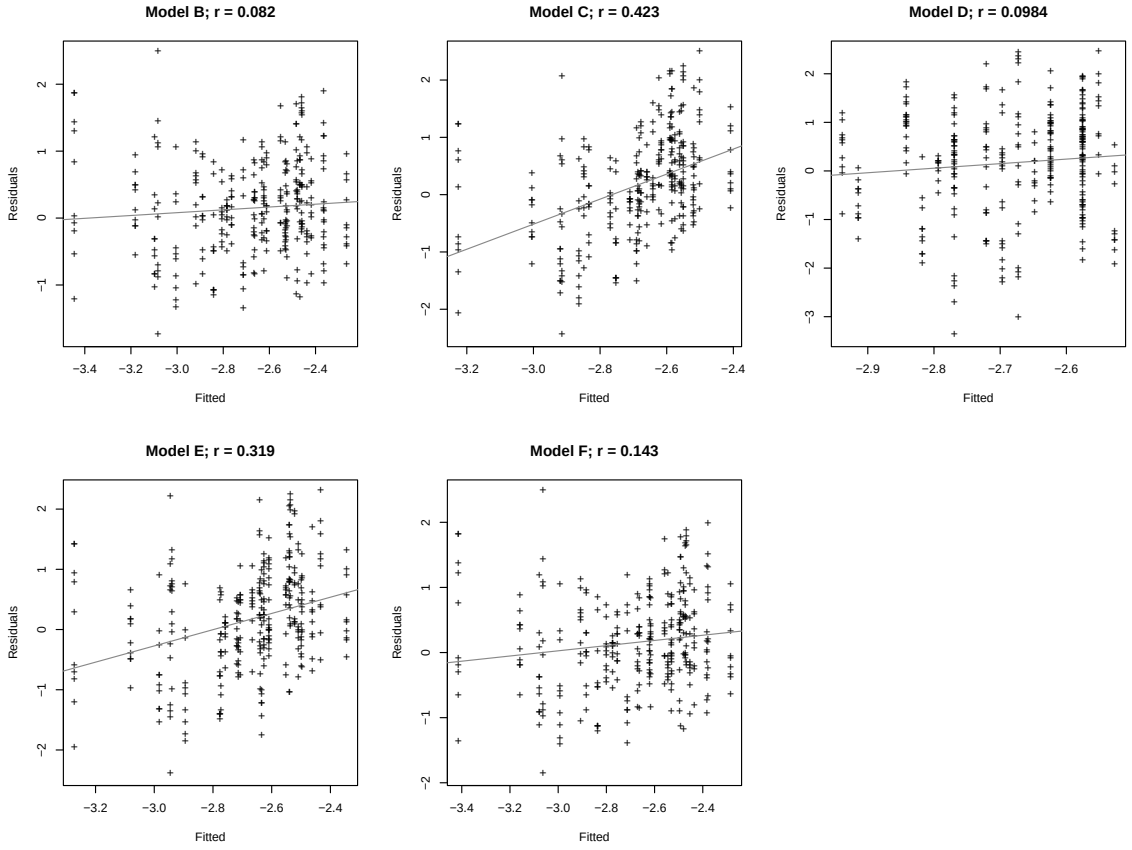


Figure 7: Atlanta residual plots.

- Model C: This model contains uncorrelated random effects for time.
- Model D: This model includes uncorrelated random effects for both site and time.
- Model E: This model consists of separate spatial-only models for each of the eight time points.

Gaussian conjugate priors were used on all mean parameters. Gamma(4, 0.0165) and Gamma(4, 0.214) priors were chosen for $\lambda_1$ and $\lambda_2$, respectively, using reasoning as in Section 4.1. (Note that the maximum distance between any pair of monitoring sites was 911 km, and that the eight chosen times were coded as $t = 0, \ldots, 7$.) The prior distribution for $T$ was taken to be $IW(2, Diag(1.1, 0.9))$; the values 1.1 and 0.9 are approximately the median of the sample variances (across the eight time points) for NO and $NO_2$, respectively. Independent inverse gamma priors with means 1.1 and 1.2 and variance 10 were used for $\tau_U^2$ and $\tau_V^2$.

For parameter sampling, independent lognormal proposal densities were used for $\lambda_1$

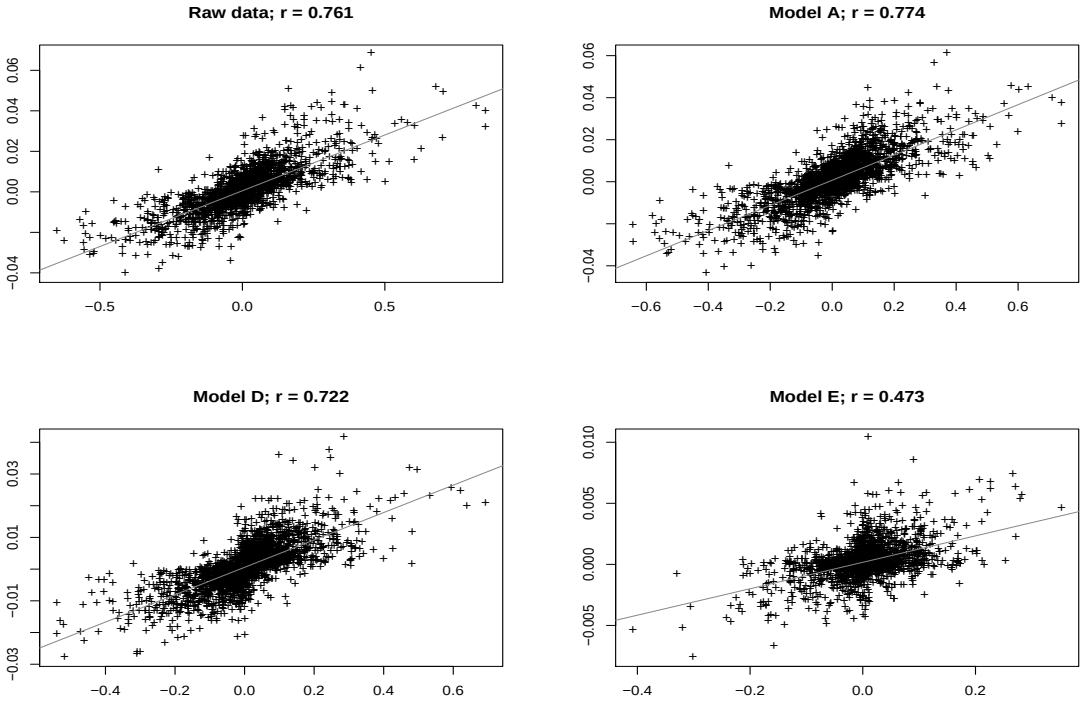


Figure 8: Atlanta separability plots.

and $\lambda_2$, centered on these parameters. The nugget term, $(\tau_U^2, \tau_V^2)$, was drawn as a pair, using a correlated bivariate normal distribution centered on the log of the parameters. The proposal density for $T$ used independent normal distributions, centered on $\log(T_{11})$, $\log(T_{22})$ and $\rho = \text{cor}(T_{11}, T_{22})$, automatically rejecting proposed values for $T$ that were negative definite. The methods described in Subsection 3.2 were used to impute missing data values at each iteration.

Allowing for missing data slowed the sampling step considerably. In our initial attempt at implementing these models, in order to draw filler values at each iteration for the missing data, we were required to invert a matrix whose dimension is equal to the number of observations, in this case 1312 × 1312 (82 sites × 8 time points × 2 pollutants). After running three chains of length 1000 to assess convergence, 6000 samples were generated as a production run. Again, convergence was rapid, within approximately 50 iterations. After discarding the first 1000 samples as burn-in, we retained every fifth sample for prediction at the 500 sites shown in Figure 3. Prediction of unweighted STCDFs was carried out using these 1000 predictions and a grid of $K = 200$ points evenly spaced across the spread of values of log(NO), approximately –10 to –2.

The results from our five models are shown in Figure 9. The left column corresponds

| Model | Temporal Correlation | Nugget DIC ($p_D$) | No nugget DIC ($p_D$) |
|---|---|---|---|
| A (mean only) | Exp | 62.0 (41.6) | 213.5 (40.9) |
| B (r.e. for site) | Exp | 3.77 (148.5) | 172.1 (146.5) |
| C (r.e. for time) | Exp | 6.07 (51.9) | 104.8 (53.6) |
| D (r.e. for site,time) | Exp | –24.3 (155.8) | 25.0 (161.7) |
| E (separate times) | NA | 579.0 (54.1) | 705.6 (65.5) |
| A (mean only) | C.S. | 190.2 (40.9) | – |
| A (mean only) | Identity | 560.5 (26.8) | – |
| D (r.e. for site,time) | C.S. | 120.6 (152.3) | – |
| D (r.e. for site,time) | Identity | 17.0 (180.5) | – |

Table 2: DIC for bivariate models, California data.

to models that contain a nugget, the right column for those without a nugget. This figure suggests a fair amount of Bayesian shrinkage in going from a spatial-only to a spatiotemporal model. Models without a nugget show considerably more shrinkage than their nuggeted counterparts. As might be expected, the general appearances of the STCDFs within a given spatio-temporal model are more consistent with one another than those in the spatial-only models. That is, the STCDFs for the various time points appear to be horizontal translates of each other for the spatiotemporal models, but not for the spatial-only models.

Model selection via DIC can again be carried out, albeit somewhat more arduously due to the missing data; see e.g. Celeux et al. (2005) in this regard. Results appear in Table 2. Again we considered four models in addition to the four already described above: all assumed exponentially decaying spatial correlations and a nugget term. Two assumed a compound-symmetry form for the temporal correlation matrix (as in the Atlanta analysis subsection), and the last two assumed $C = I_{N\times N}$ (i.e., no temporal correlation). The best (i.e., smallest DIC) model is Model D. This model has exponentially decaying correlations in both space and time, and a mean structure that includes random effects for both site and time. We note that in all cases, the model with nugget has considerably better DIC value than its no-nugget counterpart. These differences appear to be due primarily to better fit, rather than to a smaller number of parameters $p_D$.

We briefly address the separability issue in Figure 10, which features panels corresponding to each of the pollutants for both the centered raw data and the chosen model (Model D). The plots for Model D, while not outstanding, suggest the separability assumption is not totally unjustified.

As pointed out by a referee, our choice of 0.001 as the replacement value for 0.0's in the data set is somewhat questionable. There were 112 values of 0.0 for $NO$, and 10 values of 0.0 for $NO_2$. We briefly explored the effect of replacing 0.0 by 0.0005 and 0.0001 instead of 0.001, and show the results in Figure 11. As can be seen from

the graph, the effect was not extreme when the replacement value is .0005, but can be appreciable for the value .0001, especially at the low end (left side) of the STCDFs.

We close this section by noting some important differences between the Atlanta and California data sets. There are far fewer sites in the Atlanta data set, and few of the sites were particularly close to each other. In contrast, many of the California sites appeared in clusters. Additionally, the data were spaced a full day apart in the Atlanta data set, but a mere three hours apart in the California data set. Both of these aspects lead us to believe that our California air quality data analyses may be somewhat more reliable than those arising from the Atlanta data set, since the former are more likely to take advantage of the potential borrowing of strength across both space and time offered by our STCDFs.

# 5 Discussion

In this paper we developed STCDFs in a Bayesian setting for both univariate and bivariate models using MCMC methods. We briefly discussed methods for weighting the STCDF (which changes the interpretation of the quantities we predict), but focused our efforts on unweighted STCDFs. We developed methods for inclusion of a nugget term in both univariate and bivariate models, and we carried out prediction for the true underlying process (rather than for the observed process). Finally, we incorporated the ability to handle missing data, including modification to the model choice process using DIC. STCDFs were shown to take advantage of the borrowing of strength across days, which is important in settings where we have several time periods' worth of information, but perhaps only a small amount of information for any time period. Implementing the Bayesian analytic engine using fairly vague priors then allows the data to tell us how much shrinkage in the STCDFs is appropriate over the time periods, and how it is affected by the addition of random effects or a nugget term to the model.

An objective of our work was to see whether we could carry out spatial CDF computations on the decidedly larger spatiotemporal data sets, given that the relevant matrix operations increase at rate $N^3$, where $N$ is the number of time points in our data. In this, we were fairly successful at the prediction stage; we found techniques that allowed us to draw from the much larger dimensional distributions (by making copious use of the sum of Kroneckers structure imposed on the covariances) and also carry out computations without ever requiring that we hold the large matrices (e.g., $31{,}000 \times 31{,}000$) in computer memory. This moved the bottleneck to the initial parameter-sampling step. A suggestion by Dr. David Higdon led us to a much faster approach to the matrix inversion and related computations for the parameter-sampling step. The idea is to avoid a brute force approach (or Cholesky factorization) by using an eigenvalue-eigenvector approach. This decomposition approach handles sum-of-Kroneckers matrices, and is described in the univariate and bivariate settings in the appendices. Our computing approach might also be abetted by the use of coarse-fine grid methods as encouraged for example by Higdon et al. (2003). Such methods may help speed mixing of the MCMC algorithm, although they may also complicate STCDF evaluation somewhat.

There are several directions in which this work could be expanded. First, our approach to developing the STCDF has been to add a temporal dimension to the notion of an SCDF. But a referee has pointed out that the other obvious approach would have been to first define the notion of a "TCDF" as the proportion of time for which a pollution surface lies below a given value $y_0$, and then extend *this* definition by adding the spatial dimension. Such an approach might be useful if our interest lay in, say, EPA regulations regarding cumulative levels of a pollutant over time, rather than the proportion of a spatial region that is exposed.

Returning to our problem formulation, our data analysis could be expanded to incorporate more weather-related covariates. Our Atlanta analysis considered the simplest case, incorporating a time-varying covariate that did not depend on spatial location. The more general situation, with covariates depending on space or on both space and time, might be considerably more difficult, since such measurements are unlikely to have been collected at the same spatial locations where the pollutants were monitored. This would in turn mean imputing a great deal of missing (or misaligned) data ( c.f. Zhu et al. (2003)). Among other things, this implies a rapid escalation in the size of the matrices we must handle.

Finally, modeling in the paper has been of point-level covariates. As mentioned above, the issue of incorporating block-level covariates (such as race or income, collected at, say, the zip code or census tract level) was partly addressed in Short et al. (2005) in a spatial-only setting. The incorporation of such information as weights in the STCDFs allows the necessary change of interpretation in the face of the misaligned data: rather than describing what fraction of the area of $D$ is exposed to a given level of pollutant, the realigned weighted STCDF can address the question of what fraction of the *people* are exposed. The resulting space-time solution to the problem of joint modeling of point- and block-level data would be valuable, especially in studies comparing human exposure in different sociodemographic groups.

# A Appendices

## A.1 Full-data likelihood calculations in the univariate setting

One computational bottleneck in the parameter sampling stage is in the evaluation of the log-likelihood, which is used in forming the Metropolis ratio for several parameters (and is subsequently used for DIC calculations). We can avoid a brute force (Cholesky factorization) approach to the required matrix inversion by means of an eigenvalue-eigenvector decomposition, as follows.

Let $\Sigma = H \otimes C + \tau^2 I$, where $H$ and $C$ are symmetric, positive definite matrices of size $n \times n$ and $p \times p$, respectively. Our objective is to evaluate the quantity $\mathbf{z}^T \Sigma^{-1} \mathbf{z}$, where $\mathbf{z}$ is a $np \times 1$ column vector. Let $(P_H, \Lambda_H)$ be an eigenvalue-eigenvector decomposition for $H$ – that is, the columns of $P_H$ are eigenvectors of $H$ with corresponding eigenvalues along the main diagonal of the (diagonal) matrix $\Lambda_H$. Since $H$ is symmetric, we may assume $P_H$ is an orthogonal matrix. Similarly, let $(P_C, \Lambda_C)$ be an eigenvalue-eigenvector decomposition for $C$ with $P_C^T P_C = I_p$.

Define $\Lambda = \Lambda_H \otimes \Lambda_C$ and $P = P_H \otimes P_C$. Per Schott (1997), the diagonal elements of $\Lambda$ are the eigenvalues of $H \otimes C$ (with correct multiplicities) and the columns of $P$ are eigenvectors of $H \otimes C$. For general non-symmetric matrices $H$ and $C$, $P$ need not have full rank; in our setup, however, it does: $P^T P = (P_H \otimes P_C)^T (P_H \otimes P_C) = (P_H^T P_H) \otimes (P_C^T P_C) = I_n \otimes I_p$. Thus

$$\Sigma^{-1} = (P \Lambda P^T + \tau^2 I)^{-1} = \cdots = P \tilde{\Lambda} P^T \,, \tag{7}$$

where $\tilde{\Lambda} = (\Lambda + \tau^2 I)^{-1} = \mathrm{Diag}(a_1, \ldots, a_{np})$. Finally, $\mathbf{z}^T \Sigma^{-1} \mathbf{z} = \mathbf{z}^T P \tilde{\Lambda} P^T \mathbf{z} = ||\tilde{\Lambda}^{1/2} P^T \mathbf{z}||^2$. Lastly, we note that $\tilde{\Lambda}^{1/2} P^T \mathbf{z} = (\sqrt{a_1}(P^T \mathbf{z})_1, \ldots, \sqrt{a_{np}}(P^T \mathbf{z})_{np})^T$.

In summary, the approach we have just described allows us to calculate $\mathbf{z}^T \Sigma^{-1} \mathbf{z}$ by calculating eigenvalue-eigenvector decompositions for $H$ and $C$, evaluating $P^T \mathbf{z} = (P_H^T \otimes P_C^T)\mathbf{z}$ by applying the identity in Equation refidentity2, and evaluating $\sum_{i=1}^{np} a_i (P^T \mathbf{z})_i$. In particular, we did not need to find a Cholesky decomposition of $\Sigma = H \otimes C + \tau^2 I$.

## A.2 Full-data likelihood calculations in the bivariate setting

Suppose $\Sigma = T \otimes G + S \otimes I$, where $T$ is a $2 \times 2$ symmetric, positive definite matrix, $S = \left(\begin{smallmatrix} \tau_1^2 & 0 \\ 0 & \tau_2^2 \end{smallmatrix}\right)$, and $G$ is a $n \times n$ symmetric positive definite matrix, which we assume can be decomposed in the usual way as $P \Lambda P^T$ where $\Lambda$ is diagonal and $P$ is an orthogonal matrix. Our definition of $\Sigma$ means $\tau_1^2$ is added to the first $n$ diagonal entries of $T \otimes G$ and $\tau_2^2$ is added to the remaining diagonal entries. If we wish to evaluate $\mathbf{z}^T \Sigma^{-1} \mathbf{z}$, where $\mathbf{z}$ is a $2n \times 1$ column vector, the method of the preceding section no longer applies, because the calculations in equation (7) break down.

Let us write $T = \left(\begin{smallmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{smallmatrix}\right)$. Applying the formula for the inverse of a partitioned matrix, we find that

$$\Sigma^{-1} = \begin{pmatrix} P^T \Lambda_1^{-1} P & -P^T \Lambda_2 P \\ -P^T \Lambda_2 P & P^T \Lambda_3^{-1} P \end{pmatrix},$$

where $\Lambda_1$, $\Lambda_2$, and $\Lambda_3$ are diagonal matrices given by

$$\Lambda_1 = t_{11}\Lambda + \tau_1^2 I - t_{12} t_{21} \Lambda (t_{22}\Lambda + \tau_2^2 I)^{-1} \Lambda,$$

$$\Lambda_2 = t_{21} (t_{22}\Lambda + \tau_2^2 I)^{-1} \Lambda \Lambda_1^{-1},$$

and

$$\Lambda_3 = (t_{22}\Lambda + \tau_2^2 I) - t_{21} t_{12} \Lambda (t_{11}\Lambda + \tau_1^2 I)^{-1} \Lambda.$$

We partition $\mathbf{z} = \left(\begin{smallmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{smallmatrix}\right)$, where $\mathbf{z}_1$ and $\mathbf{z}_2$ are $n \times 1$ column vectors. Then

$$\begin{aligned} \mathbf{z}^T \Sigma^{-1} \mathbf{z} &= \mathbf{z}_1^T [P \Lambda_1^{-1} P^T] \mathbf{z}_1 + 2 \mathbf{z}_1^T [-P \Lambda_2^{-1} P^T] \mathbf{z}_2 + \mathbf{z}_2^T [P \Lambda_3^{-1} P^T] \mathbf{z}_2 \\ &= ||(\Lambda_1^{-1})^{1/2} P^T \mathbf{z}_1||^2 - 2 (z_1^T P) \Lambda_2 (P^T \mathbf{z}_2) + ||(\Lambda_3^{-1})^{1/2} P^T \mathbf{z}_2||^2 \,. \end{aligned}$$

Recall that as part of our likelihood calculations we require the value of $\det \Sigma$. In the univariate setting, this is an easy computation since we have the eigenvalues of $\Sigma$ available

as part of the decomposition of $\Sigma$. However, in the bivariate setting, we do not decompose $\Sigma$ in the form $\Sigma = P\Lambda P^T$ for diagonal $\Lambda$ and orthogonal $P$; thus, the eigenvalues of $\Sigma$ are not readily apparent. Fortunately, there is a formula for the determinant of a block partitioned matrix (see Schott (1997)); we have $\det\Sigma = \det\Sigma_{22}\det(\Sigma_{11}-\Sigma_{12}\Sigma_{22}\Sigma_{21}) = \det((t_{22}\Lambda + \tau_2^2 I)\Lambda_1)$, where we have written $\Sigma = \left(\begin{smallmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{smallmatrix}\right)$, and $\Sigma_{ij}$ $(i,j=1,2)$ are $n\times n$ matrices. Up to this point, we have not achieved great computer time savings. In calculating $\mathbf{z}^T\Sigma^{-1}\mathbf{z}$, we have replaced a $2n\times 2n$ matrix inversion with an eigenvalue-eigenvector decomposition of a $n\times n$ matrix, a saving of at best a factor of $2^3$ (and simulations show the actual saving is even more modest). However, if we replace $G$ by $H\otimes C$, we reap large benefits. Assuming $H$ is $n\times n$ and $C$ is $p\times p$, we need only calculate eigenvalues and eigenvectors for $n\times n$ and $p\times p$ matrices (rather than $np\times np$ matrices). Additionally, we are able to use Identity 4 in evaluating $P^T\mathbf{z}_1$ and $P^T\mathbf{z}_2$, speeding up the computations and avoiding the need to store large matrices in memory.

## A.3 Missing-data likelihood and DIC calculations

When carrying out MCMC with incomplete data, bottlenecks appear when we attempt to draw from the conditional distribution $Y_{mis}|Y_{obs}$ and again in the DIC calculations, at which point we are trying to evaluate $p(Y_{obs}|Y_{mis})$. Here we use the notation of Section 3.2: $Y_{mis}$ and $Y_{obs}$ correspond to the components of $Y_{full}$ after re-ordering, so that the "missing" components appear at the end of the vector. Although we discussed the handling of missing data for just the bivariate setting, what we say here applies equally well, with obvious parallels in the notation, for the univariate setting. In the "tricks" we outline here, we make use of the eigenvalue-eigenvector notions and notations of the preceding appendices. The key point relies in the manner in which we apply the formula for the inverse and the determinant of a block partitioned matrix; rather than building up a result about the larger matrix in terms of expressions involving its constituent sub-matrices, we now obtain results about a selected, difficult-to-handle submatrix in terms of the larger matrix.

The following notation applies throughout the remainder of this section. $Y_{mis}$ and $Y_{obs}$ have lengths $r$ and $n-r$, respectively. We assume $A$ is an $n\times n$ permutation matrix such that $AY_{full} = \left(\begin{smallmatrix}Y_{obs}\\ Y_{mis}\end{smallmatrix}\right)$. When applied as subscripts to a vector of length $n$, the words "*obs*" and "*mis*" refer to the first $n-r$ and last $r$ components, respectively. When applied as subscripts to an $n\times n$ block-partitioned matrix, the subscripts 11, 12, 21, and 22 refer to the upper left, upper right, lower left and lower right blocks, respectively, and these blocks have sizes $(n-r)\times(n-r)$, $(n-r)\times r$, $r\times(n-r)$ and $r\times r$, respectively.

One of our objectives is to obtain draws from the conditional distribution of $Y_{mis}|Y_{obs}$. We assume that $\left(\begin{smallmatrix}Y_{obs}\\ Y_{mis}\end{smallmatrix}\right) \sim MVN(AX\boldsymbol{\beta}, A\Sigma A^T)$, which implies that

$$Y_{mis}|Y_{obs} \sim MVN(\tilde{\mu}, \tilde{\Sigma}),$$

where

$$\tilde{\mu} = (AX\boldsymbol{\beta})_{mis} + (A\Sigma A^T)_{21}[(A\Sigma A^T)_{11}]^{-1}(Y_{obs} - (AX\boldsymbol{\beta})_{obs})$$

and

$$\tilde{\Sigma} = (A\Sigma A^T)_{22} - (A\Sigma A^T)_{21}[(A\Sigma A^T)_{11}]^{-1}(A\Sigma A^T)_{12}.$$

We note that $\tilde{\Sigma}$ is fairly small, just $r \times r$. So if we can calculate $\tilde{\Sigma}$, it will be easy to get a draw from $MVN(0, \tilde{\Sigma})$. The difficulty lies in evaluating for every MCMC iteration $[(A\Sigma A^T)_{11}]^{-1}$, which is moderately large – perhaps somewhat larger than $1000 \times 1000$.

We note that it is easy to calculate $(A\Sigma A^T)^{-1} = (A^T)^{-1}\Sigma^{-1}A^{-1}$, using the eigenvalue-eigenvector decompositions outlined in the earlier subsections of the appendix. If we write $(A\Sigma A^T)^{-1}$ as a partitioned matrix, $(A\Sigma A^T)^{-1} = \binom{C \quad D}{E \quad F}$, where $C$, $D$, $E$, and $F$ are of sizes $(n-r) \times (n-r)$, $(n-r) \times r$, $r \times (n-r)$ and $r \times r$ respectively, then the formula for the inverse of a partitioned matrix tells us immediately that $F = \tilde{\Sigma}$. That is, we simply read off the $r \times r$ quantity we are interested in, namely $\tilde{\Sigma}$. We also require $\tilde{\mu}$, which means we need to evaluate $(A\Sigma A^T)_{21}[(A\Sigma A^T)_{11}]^{-1}$; but this expression is precisely $-DF^{-1}$.

In order to carry out DIC calculations, we need to evaluate

$$\log p(Y_{pres}|Y_{mis}) \propto -\frac{1}{2}\{(Y_{pres} - \hat{\mu})^T\hat{\Sigma}^{-1}(Y_{pres} - \hat{\mu})) + \log(det(\tilde{\Sigma}))\},$$

where

$$\hat{\mu} = (AX\boldsymbol{\beta})_{pres} + (A\Sigma A^T)_{12}[(A\Sigma A^T)_{22}]^{-1}(Y_{mis} - (AX\boldsymbol{\beta})_{mis})$$

and

$$\hat{\Sigma} = (A\Sigma A^T)_{11} - (A\Sigma A^T)_{12}[(A\Sigma A^T)_{22}]^{-1}(A\Sigma A^T)_{21}.$$

This time, the primary difficulty lies in the repeated evaluations of $\hat{\Sigma}^{-1}$, since it is a moderately large matrix. However, it is easy to evaluate $(A\Sigma A^T)^{-1}$, and $\hat{\Sigma}^{-1}$ is $C$ from the partition of $(A\Sigma A^T)^{-1}$ described earlier in this section.

Lastly, we need to calculate $\log(det(\tilde{\Sigma}))$. To do this, we use the formula for the determinant of a partitioned matrix and find that:

$$det(\tilde{\Sigma}) = \frac{det(A\Sigma A^T)}{det((A\Sigma A^T)_{22})} = \frac{det(\Sigma)}{det((A\Sigma A^T)_{22})}\ .$$

The denominator of this fraction is easy to calculate, since $(A\Sigma A^T)_{22}$ is just $r \times r$, and the numerator is readily available as a by-product of the eigenvalue-eigenvector decomposition of $\Sigma$.

## References

Banerjee, S. (2005). "On geodetic distance computations in spatial modelling." *Biometrics*, 61: 617–625. 605

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, Florida: Chapman and Hall/CRC Press. 596, 601, 609

Banerjee, S. and Gelfand, A. (2002). "Prediction, interpolation and regression for spatially misaligned data." *Sankhya*, 64: 227–245. 596

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2005). "Deviance Information Criteria for missing data models." *Bayesian Analysis*. (To appear, this issue.). 614

Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley, second edition. 603

Fuentes, M. (2003). "Testing for separability of spatial-temporal covariance functions." Mimeo Series Report #2545, Department of Statistics, North Carolina State University. 609

Gelfand, A., Zhu, L., and Carlin, B.P. (2001). "On the change of support problem for spatio-temporal data." *Biostatistics*, 2: 31–45. 596

Gelman, A., Roberts, G., and Gilks, W. (1996). "Efficient Metropolis jumping rules." In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics 5*, 599–607. Oxford: Oxford University Press. 602

Golub, G. and Van Loan, C. (1996). *Matrix Computations*. Baltimore, MD: Johns Hopkins University Press, third edition. 603

Handcock, M. (1999). (Comment on "Prediction of spatial cumulative distribution functions using subsampling."). 596

Higdon, D., Holloman, C., and Lee, H. (2003). "Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems (with discussion)." In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics 7*, 181–197. Oxford: Oxford University Press. 615

Overton, W. (1989). "Effects of measurements and other extraneous errors on estimated distribution functions in the National Surface Water Surveys." Technical Report 129, Department of Statistics, Oregon State University. 596

Schott, J. (1997). *Matrix Analysis for Statistics*. New York: Wiley. 617, 618

Short, M., Carlin, B., and Gelfand, A. (2005). "Bivariate spatial process modeling for constructing indicator or intensity weighted spatial CDFs." *Journal of Agricultural, Biological, and Environmental Statistics*, 10: 259–275. 596, 597, 604, 616

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). "Bayesian measures of model complexity and fit (with discussion)." *J. Roy. Statist. Soc., Ser. B*, 64: 583–639. 609

Zhu, L., Carlin, B., and Gelfand, A. (2003). "Hierarchical regression with misaligned spatial data: Relating ambient ozone and pediatric asthma ER visits in Atlanta." *Environmetrics*, 14: 537–557. 616

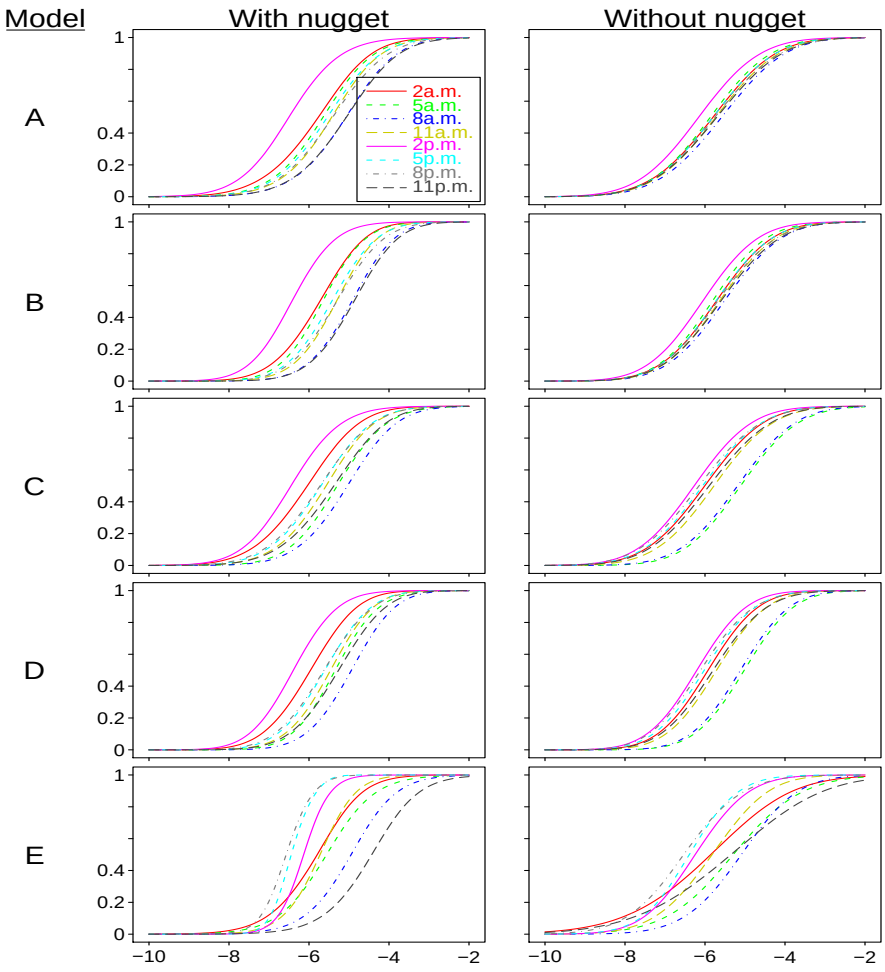


Figure 9: California unweighted STCDFs for July 12, 2001.

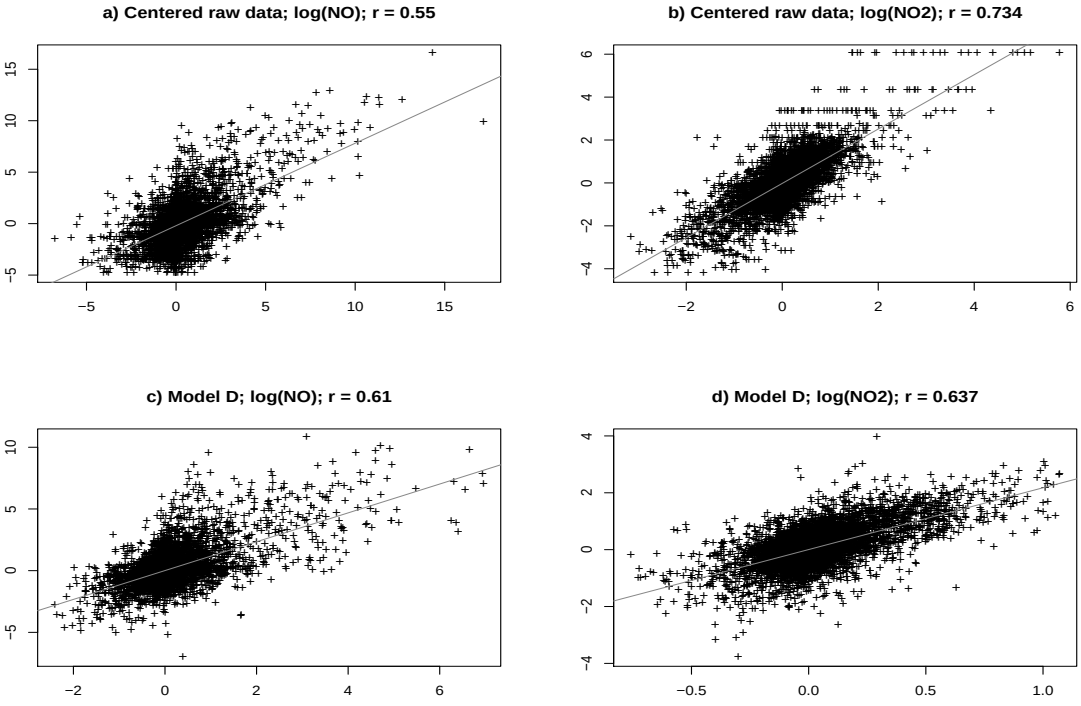


Figure 10: California separability plots.

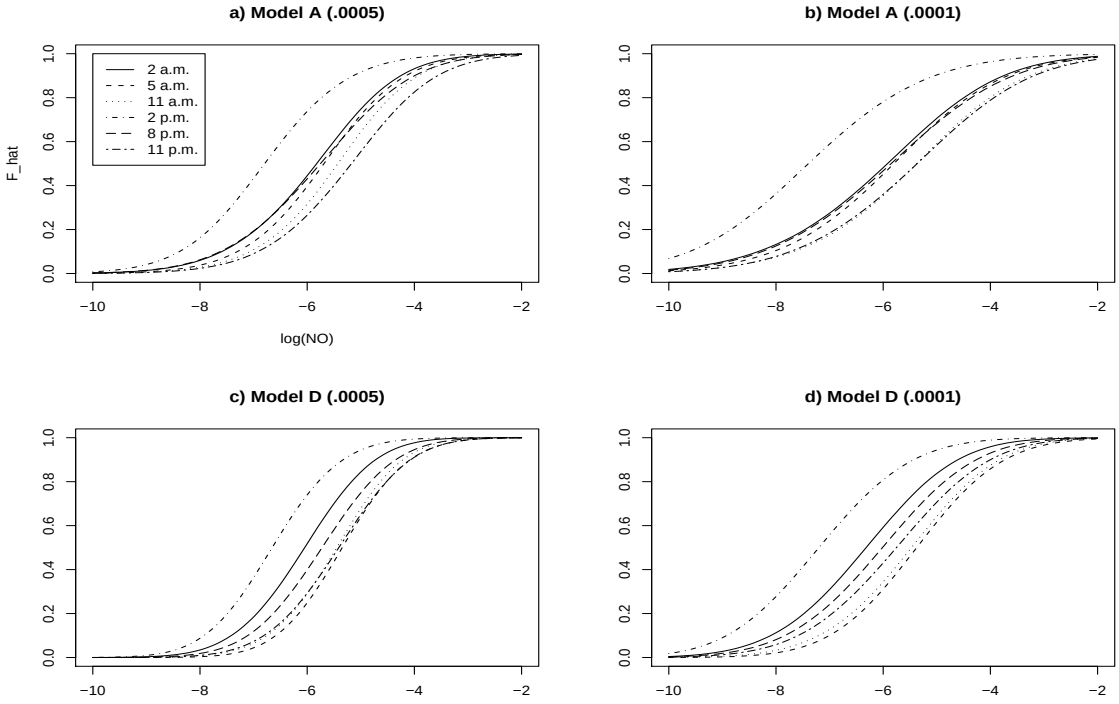


Figure 11: California STCDFs corresponding to varying replacement values for 0.0.