ENVIRONMETRICS
Environmetrics 2010; **21**: 48–65
Published online 18 August 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.984

Combining numerical model output and particulate data using Bayesian space–time modeling

Nancy J. McMillan[1*,†], David M. Holland[2], Michele Morara[1] and Jingyu Feng[1]

[1] *Statistics and Information Analysis, Battelle, 505 King Avenue, Columbus, OH 43201, U.S.A.*
[2] *U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC 27711-0111, U.S.A.*

SUMMARY

Over the past few years, Bayesian models for combining output from numerical models and air monitoring data have been applied to environmental data sets to improve spatial prediction. This paper develops a new hierarchical Bayesian model (HBM) for fine particulate matter ($PM_{2.5}$) that combines U. S. EPA Federal Reference Method (FRM) $PM_{2.5}$ monitoring data and Community Multi-scale Air Quality (CMAQ) numerical model output. The model is specified in a Bayesian framework and fitted using Markov Chain Monte Carlo (MCMC) techniques. We find that the statistical model combining monitoring data and CMAQ output provides reliable information about the true underlying $PM_{2.5}$ process over time and space. We base these conclusions on results of a validation exercise in which independent monitoring data were compared with predicted values from the HBM and predictions from a standard kriging model based solely on the monitoring data. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: hierarchical Bayesian; space–time modeling; data fusion

1. INTRODUCTION

In recent years, the focus of environmental management has shifted to regional-scale strategies that require the accurate spatial characterization of ground-level air pollution levels for successive time periods. The most direct way to obtain accurate air quality information is from measurements made at surface air monitoring stations. However, many areas of the U.S. are not monitored and, typically, air monitoring sites are sparsely and irregularly spaced over large areas. As the need for spatial prediction has become reality in the regulatory environment, it is now important to combine air monitoring data and numerical model output in a coherent way for better prediction of air pollution over short time periods. High spatial resolution numerical model output from deterministic simulation models such as the Community Multi-Scale Air Quality Model (CMAQ; http://www.epa.gov/asmdnerl/CMAQ) are now available over a 12 km (or less) grid. This expanded coverage not only helps to identify local and

*Correspondence to: N. J. McMillan, Statistics and Information Analysis, Battelle, 505 King Avenue, Columbus, OH 43201, U.S.A.
†E-mail: McMillanN@battelle.org

Received 21 December 2007
Accepted 30 November 2008

regional sources of particulates, but also provides insight on the role of transboundary transport on U.S. air quality.

Given the extensive and continuing public concern over adverse health effects from exposure to fine particulate matter ($PM_{2.5}$) concentrations, public health officials need high resolution, in terms of space and time, predictions of $PM_{2.5}$. Recently, the U.S. Environmental Protection Agency (EPA) and the Center for Disease Control (CDC) collaborated in the Public Health Air Surveillance Evaluation (PHASE) project to identify spatial-temporal interpolation tools that can be used to generate daily surrogate measures of exposure to ambient air pollution and relate those measures to available public health data. This paper describes a new hierarchical Bayesian modeling approach for modeling combined or fused sources of data that was developed for the PHASE program.

We propose a hierarchical spatial-temporal model that draws strength from $PM_{2.5}$ monitoring data from the U.S. EPA's Federal Reference Method (FRM) $PM_{2.5}$ monitoring network and the CMAQ numerical model output to predict pollutant levels at daily time scales for use in modeling public health–air quality relationships. The model assumes that both monitoring data and CMAQ output provide good information about the same underlying pollutant surface, but with different measurement error structures. It gives more weight to accurate monitoring data in areas where monitoring data exists, and relies on bias-adjusted model output in non-monitored areas. The spatial domain of interest includes the eastern United States for the time period 1 January 2001–31 December 2001. The results of this work provide a clearer picture of the spatial extent of successive daily $PM_{2.5}$ concentrations. This is a particularly important result for evaluating potential relationships with daily public health data given that most of the FRM $PM_{2.5}$ monitoring data are collected every 3 days. One additional benefit of modeling daily concentrations is the ability to easily aggregate to annual or any other temporal summary of $PM_{2.5}$. This information contributes to regulatory efforts that focus on defining areas in attainment or non-attainment with the annual $PM_{2.5}$ national air quality standard.

Although spatial prediction with combined data is a relatively new field, several papers have appeared in the literature on this topic. Fuentes and Raftery (2005) developed a hierarchical statistical framework to model the ''true'' pollutant process as jointly Gaussian random fields. They estimate the parameters for the bias of CMAQ output and the parameters of the covariance structure for CMAQ and measurement error processes, and then simulate the conditional distribution of the ''true'' process given both sources of spatial information. This methodology applies to spatial processes at a fixed time point, without evaluation of the space–time dependence structure. Zimmerman and Holland (2005) consider the problem of optimal spatial prediction of wet deposition data using data from two monitoring networks with network-specific biases and variances. Cowles and Zimmerman (2003) use a Bayesian modeling approach for spatial-temporal data from two acid deposition monitoring networks that accounts for possible differences in network measurement error bias and variances. Jun and Stein (2004) suggest new ways of comparing space–time correlation structure of monitoring observations with CMAQ numerical model output. Non-combined modeling of particulate matter has been addressed by several researchers. Zidek *et al*. (2002) developed predictive distributions of non-monitored PM_{10} concentrations in Vancouver, Canada. Cressie *et al*. (1999) compared classical kriging and Markov-random field models for predicting PM_{10} concentrations in Pittsburgh, Pennsylvania. Smith *et al*. (2003) proposed a spatial-temporal model for predicting weekly averages of $PM_{2.5}$ and Sahu and Mardia (2005) presented a short-term forecasting analysis of $PM_{2.5}$ data in New York City during 2002. Sahu *et al*. (2006) proposed a spatial-temporal model for weekly values of $PM_{2.5}$ in the Midwest U.S. that allows for a shift in the mean level and a variability increase as you move from rural to urban sites. While these previous efforts have demonstrated the utility of combining monitoring data and CMAQ output for predicting air quality parameters, none have demonstrated an approach on the

DOI: 10.1002/env

scale we consider; we predict daily $PM_{2.5}$ levels for the entire eastern U.S. on a 12 km grid for 1 year. This results in a space–time predictive grid of over 9 million cells.

The remainder of this paper describes the data, the statistical model used to fit the data, and the results of the analysis. Section 2 describes the data. Section 3 describes the hierarchical Bayesian statistical model used to fit the data described in Section 2. Section 4 describes the results of fitting the statistical model to the $PM_{2.5}$ data and demonstrates the type of information that can be obtained by combining air monitoring and CMAQ output. Section 5 contains information on the utility of the statistical model by comparing its performance against a standard kriging approach for fitting spatial data. Section 6 briefly presents conclusions and future work.

2. DATA

Daily $PM_{2.5}$ concentration ($\mu g/m^3$) data (Figure 1) in the eastern U.S. were obtained from two distinct sources. First, monitoring data (24 h integrated samples) from EPA's $PM_{2.5}$ Federal Reference Method (FRM) [part of the NAMS/SLAMS network (U. S. EPA, 2003)] were used in this study. In Figure 1, the full $PM_{2.5}$ FRM network is shown that includes all sites sampling at frequencies of 1, 3, and 6 days. The second $PM_{2.5}$ data source is numerical model output from the CMAQ (http://www.epa.gov/asmdnerl/CMAQ) model converted to local standard time. CMAQ relies on emission estimates and meteorological predictions to simulate the physical and chemical processes in the atmosphere to provide gridded estimates of air pollutant concentrations. Typically, the model is used to predict regional air quality and evaluate the effects of projected changes in emission levels for input to making regional-scale environmental decisions. The grid resolution of these data are$12 \times 12\,km^2$. Figure 1 shows the monitoring data superimposed on the CMAQ 12 km grid and clearly illustrates the broad continuous CMAQ spatial coverage over the eastern U.S.

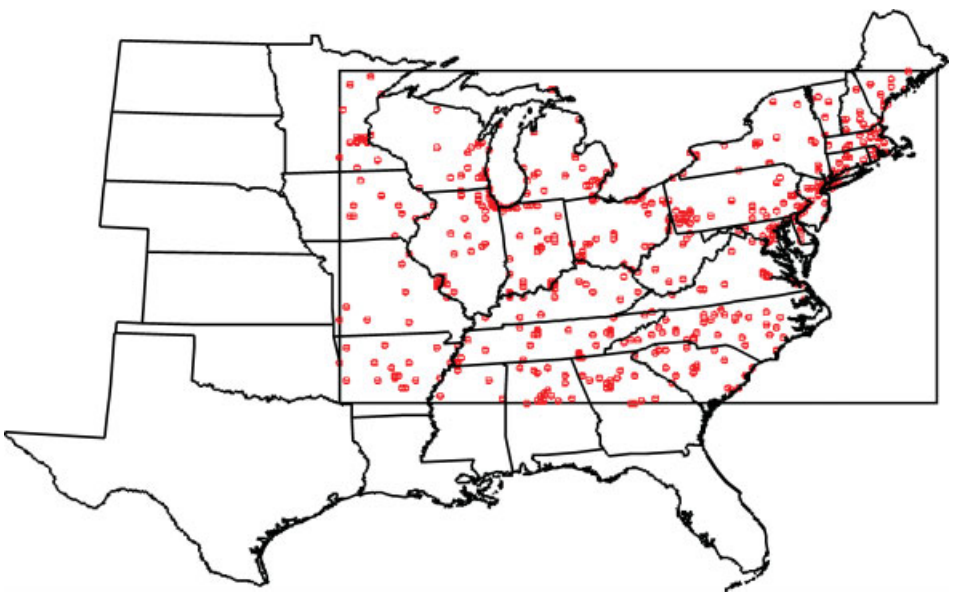

Figure 1. Modeling domain, $PM_{2.5}$ monitoring sites (red circles) superimposed on CMAQ grid

DOI: 10.1002/env

The Speciation Trends Network (STN, http://www.epa.gov/ttn/amtic/files/ambient/pm25/spec) and Interagency Monitoring of Protected Visual Environments (IMPROVE, http://vista.cira.colostate.edu/improve) gravimetric mass data were used to provide validation data for our proposed model. Similar to the EPA FRM network, both of these networks produce 24 h integrated $PM_{2.5}$ concentrations. The STN sites are mainly located in urban areas, and are sometimes collocated with an FRM monitor. In such cases, the STN data were not included in the validation. A total of 44 STN and IMPROVE sites were used in the validation analysis; none of these sites were located in a CMAQ grid cell that contains an FRM monitor.

3. HIERARCHICAL BAYESIAN SPACE–TIME MODEL

Statistical modeling of space–time $PM_{2.5}$ data at high levels of spatial and temporal resolution are described in this section. In this analysis, complexity from the spatial-temporal data comes from spatial and temporal misalignments, complicated underlying errors for each of the data sources, missing data, and the large size of the prediction grid. We placed all air monitoring data on the CMAQ predictive 12 km grid and the space–time $PM_{2.5}$ process was modeled as occurring on this grid. This choice allowed representation of the entire model in terms of grid cells that can be indexed by temporal and spatial indices. We divided the problem into hierarchical components and modeled each level in the hierarchy conditional on its preceding levels (Wikle *et al*., 1998). Before considering the specific hierarchical Bayesian model (HBM) model for combining monitoring data and CMAQ model output, we present an abstraction of the approach. Consider a series consisting of a latent variable and observed quantities all of which are indexed by the same space–time grid:

W: $PM_{2.5}$ (latent variable);
X: monitor data (observation);
Y: CMAQ output (observation);
D: CMAQ bias representation (known quantity).

X and Y are observations of the W process, and these observations are made with error (Figure 2). There are additional (non-grid) model parameters θ specified that govern the relationships among these space–time components. These will be identified in detail in the subsequent sections. Our statistical approach to linking the latent $PM_{2.5}$ surface W to the observed data starts with specifying a prior distribution for this unobserved surface using a set of unknown model parameters θ

$$[W, \theta] = [W|\theta] \times [\theta] \tag{1}$$

(We use the notation $[x]$ to denote the joint distribution for the multivariate quantity x and $[x|y]$ to denote the conditional distribution of x given another set of variables y.) This abstraction looks simple; however, in reality, the relationship between the components can be quite complicated and the individual components have large dimensions. Statistical algorithms to deal with these and other model complexities are relatively recent, the most notable of them being Markov chain Monte Carlo (MCMC) algorithms (e.g., Gilks *et al*., 1998; Chen *et al*., 2000; and Gelman *et al*., 2004). In our HBM, the prior (1), is updated with the monitoring X and CMAQ Y data. Conditional on the true process, W, and θ, the data sources are assumed to be independent. The CMAQ bias representation D is embedded within the CMAQ data model $[Y \mid W, \theta]$.

DOI: 10.1002/env

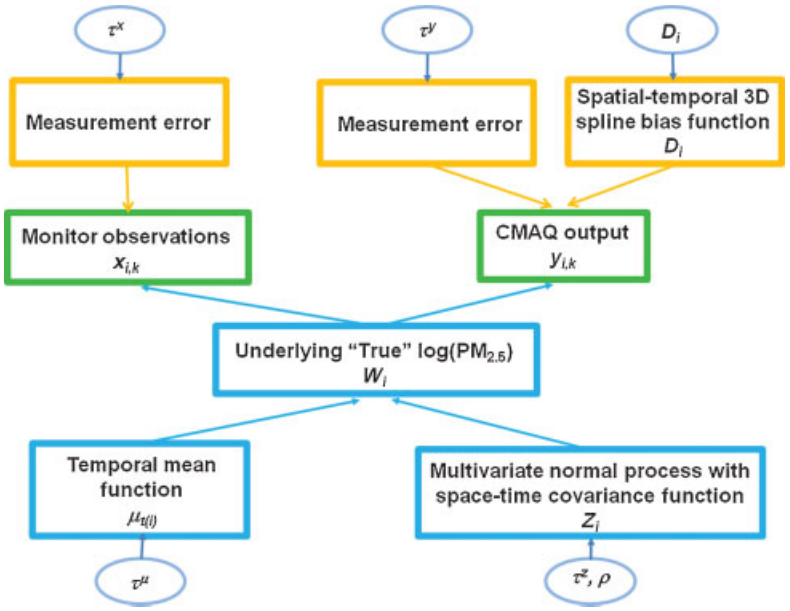

Figure 2. Bayesian modeling framework

The basis of inference from an HBM is posterior analysis. Thus, our analysis seeks to explore the posterior $[W, \theta | X, Y]$, which captures all available information about the true underlying $PM_{2.5}$ surface and the parameters governing the relationships among these surfaces and the observed quantities. The posterior is computed, via Bayes' Theorem

$$[W, \theta | X, Y] \propto [X, Y | W, \theta] \times [W, \theta] \tag{2}$$

An algorithm based on the Gibbs sampler (e.g., Gelfand and Smith, 1990; Robert and Casella, 1999) is used to fit the HBM with MCMC techniques. Briefly, to sample from the posterior distribution, $[W, \theta | X, Y]$, we simulate successively from the steps

$$\begin{matrix} [W|\theta, X, Y] \\ [\theta|W, X, Y] \end{matrix} \tag{3}$$

At each step we condition on the latest values we obtained from the previous step. These distributions are referred to as the full conditional distributions. In our HBM, the variables W and θ have large dimensions and in practice we sample them in a univariate manner. Thus, the set of full conditionals from which we must actually sample at each iteration of the Gibbs sampler is very large.

When full conditional distributions can only be calculated up to a normalizing constant, we carry out the simulation in that step by performing a Metropolis step (e.g., Tierney, 1994; Robert and Casella, 1999). The Metropolis within Gibbs approach retains the idea of sequential sampling, but addresses parameters for which the full conditional does not simplify to a known distributional form. This tends to slow up the MCMC procedure, but is necessary when a full conditional distribution is not in a recognizable form. There is much judgment involved in constructing an MCMC algorithm that converges quickly to the target stationary distribution. As well as examining the iteration history of several model parameters, we monitor the acceptance rate of the step involving Metropolis-type

DOI: 10.1002/env

draws. A good proposal distribution in a Metropolis-type step is diffuse enough to move the Markov chain sufficiently from step to step (mixing) while minimizing the number of steps in which the proposed value is not accepted (acceptance rate). Our proposal distribution is discussed in Section 3.5.

3.1. Data distributions

$PM_{2.5}$ concentration values are modeled on the logarithmic (log) scale since their distribution tends to be positively skewed. Let N^T be the number of time points, N^P the number of space points, and $N = N^T \times N^P$ be the total number of grid cells (space–time points). The log of the kth CMAQ output in cell $i = \{1, \ldots, N\}$ is denoted y_{ik}. In practice, there is only one CMAQ output per grid cell. The log of the kth monitor observation in cell $i = \{1, \ldots, N\}$ is denoted x_{ik}, $k = 1, \ldots, N_i^x$, where N_i^x represents the number of monitor observations in the grid cell i, and ranges from zero to greater than one. Although monitor readings may not fall exactly at the center of a grid cell, a monitor reading is assigned to the cell with the closest center. In practice, multiple urban monitors can occur within the same $12 \times 12\,\text{km}^2$ grid cell.

The true underlying log-$PM_{2.5}$ surface in cell i is denoted by w_i. Both the monitor data and CMAQ output are realizations of the true underlying log-$PM_{2.5}$ surface, but the statistical model hypothesizes different forms for the ways in which each actually relates to the underlying process. For the monitor data, the monitors are assumed to measure the true ambient levels with some error, but no bias, and can be expressed as a probability distribution,

$$[x_{ik}|w_i, \tau^X] \sim N(w_i, \tau^X) \tag{4}$$

where $N(\mu, \tau)$ represents the univariate normal distribution with mean μ and variance τ. This error is normally distributed on the log scale, implying a multiplicative error on the ordinary scale. Each monitor observation is assumed to be conditionally independent of all others given the underlying $PM_{2.5}$ surface, $W = \{w_i : i = 1, ..., N\}$.

For CMAQ output, the log of the observation is assumed normally distributed around the sum of the true process w_i and a bias process, represented as a linear model

$$[y_{ik}|w_i, \beta^D, \tau^Y] \sim N(w_i + D_i\beta^D, \tau^Y) \tag{5}$$

where D_i is a vector of bias covariates and β^D is a vector of parameters to be estimated within the model. Each CMAQ grid cell is assumed to be independent of all others given the underlying $PM_{2.5}$ surface, W.

3.2. B-Spline model for CMAQ

The CMAQ bias structure D is evaluated as a linear combination of 2nd order uniform B-spline (Spiegel and Tiller, 1996) functions defined over a regular 3-dimensional lattice of knots. The coefficients of the linear combination, $\beta_j^D : j = 1, \ldots, N^D$, where N^D is the number of knots, represent 1-dimensional control points. The CMAQ bias for grid cell i is represented as

$$\sum_{j=1}^{N^D} D_{ij}\beta_j^D \tag{6}$$

DOI: 10.1002/env

Let N_1, N_2, and N_3 be the dimensions of the CMAQ grid (that is, $N_1 = N^T$, $N_2 \cdot N_3 = N^P$ and $N_1 \cdot N_2 \cdot N_3 = N$), and M_1, M_2, and M_3 the dimensions of the control-points grid (this defines the degrees of freedom of the bias, that is, $M_1 \cdot M_2 \cdot M_3 = N^D$). We decompose the indices as: $i = i_1 + N_1(i_2 + N_2 i_3)$, $j = j_i + M_1(j_2 + M_2 j_3)$, so that i_1, i_2, and i_3 indicate the location of grid cell i in temporal and spatial dimensions respectively and j_1, j_2, and j_3 indicate the location of knot j in the temporal and spatial dimensions of the lattice of knots. The bias matrix is then defined as

$$D_{ij} = b_{j_1}(i_1) b_{j_2}(i_2) b_{j_3}(i_3) \tag{7}$$

where $b_k(u)$ is the 2nd order kth B-spline basis function evaluated at the point u.

Setting $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ to be the space–time domain over which the bias is defined (such domain must be bigger than or equal to the CMAQ grid domain), the uniform knots vectors over which the B-spline basis functions are defined are respectively

$$U_r = \{a_r, a_r, a_r, a_r + s_r, a_r + 2s_r, \ldots, a_r + (M_r - 3)s_r, b_r, b_r, b_r\} \quad r = 1, 2, 3$$

where $s_r = \frac{b_r - a_r}{M_r - 2}$ $r = 1, 2, 3$. This is the standard uniform knot vector for 2nd order B-splines. The knot vectors U_r are used in the evaluation of the basis functions, $b_{j_r}(u)$. M_r basis functions are defined for each dimension $r = 1,2,3$, and each basis function $b_{j_r}(u)$ is non-zero over the interval $u \in [U_{r,j_r}, U_{r,j_r+3}]$. Thus, the order of the B-spline and the number of control points together define the locality of the bias surface. The order of the B-spline sets the number of knot intervals over which the basis functions are non-zero, and the number of control points determines the length of the intervals.

Using B-splines as basis functions for the bias allows controlling the degrees of freedom of the bias structure through the number of control points. Furthermore, the piece-wise nature of the B-spline functions respects the principle of locality; that is, local information does not affect regions far from where the local information is defined. On the numerical side, B-splines allow a tensor factorization of the bias matrix into three matrixes $B^r_{i_r j_r} = b_{j_r}(i_r)$, $r = 1, 2, 3$ for a total dimension of $N_1 M_1 + N_2 M_2 + N_3 M_3$, which is very much less than the total dimension of the full D-matrix, which is $N_1 M_1 \cdot N_2 M_2 \cdot N_3 M_3$.

3.3. Space–time process priors

The underlying log-$PM_{2.5}$ process in the model, W, is separated into two components: one representing the overall mean of the surface and one representing the spatially and temporally correlated variations from the mean

$$w_i = \mu_{t(i)} + Z_i \tag{8}$$

where $t(i)$ indicates the temporal index of the grid cell i. The mean of the surface, $\mu_{t(i)}$, is constant across space. Originally we introduced an extra regression term $C_i \beta^C$ in the definition of the mean (8) to account for extra covariate data C. We considered temperature data but they provided no predictive benefit. Independent, normal prior distributions for $\mu_{t(i)}$ are specified

$$\mu_{t(i)} \sim N(0, \tau^\mu) \tag{9}$$

The prior distribution for Z is a space–time multivariate normal prior that is characterized by an autoregressive prior in the time dimension and a conditional autoregressive (CAR) prior in the spatial dimension. For the current application, first, second, and third order neighborhood spatial structures

DOI: 10.1002/env

were investigated. The critical factor driving the error structure model choice was our decision to model the $PM_{2.5}$ process on a space–time grid rather than as a continuous surface. This decision was largely driven by our desire to side-step the change of support issues that would have arisen with a continuous model due to the different averaging domains of the monitor data and CMAQ output. (See Fuentes and Raftery (2005) who address the spatial dimension of this problem with a continuous model.) Specifically, monitor data are averaged over an entire day but measured at a single location in space; CMAQ output represents a volume average measurement over a $12 \times 12\,\text{km}^2$ grid cell, and we average 24 hourly values to produce a daily average. Thus, the change of support problem in our model is addressed by defining our underlying log-$PM_{2.5}$ process to be average values in each grid cell. Part of the measurement error associated with the monitoring data is implicitly the difference between a point measurement and the grid cell spatial average.

The temporal correlation matrix is modeled as an autoregressive process of order 1 AR(1). The spatial precision matrix was chosen to be

$$[\Lambda^P]_{kl} = \begin{cases} 1 & k = l \\ -\frac{1}{n^P} & l \in \partial k \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where ∂l indicates the spatial neighbors of site l and n^P is the size of the spatial neighborhood. Neighborhoods of order 1 (four adjacent grid cells), 2 (four adjacent plus four diagonal grid cells), and 3 (four adjacent, four diagonal, and four ''one step away'' grid cells) were considered. Notice that the choice of considering a constant neighborhood size (for both the inner sites and the sites at the border) corresponds to fixing the spatial boundary conditions to 0. In other words, the mean of border sites conditional on all other grid cell values is the average of n^P values where some of those values are not on the grid and, thus, are treated as 0. We write the space–time joint prior distribution for Z as a multivariate normal distribution with zero mean and precision matrix equal to the Kronecker product between the temporal AR(1) precision matrix $\Lambda^T(\rho) \equiv [\Sigma^{\text{AR}(1)}(\rho)]^{-1}$ and the spatial precision matrix (10), that is,

$$[Z|\tau^Z, \rho] \sim N(0, \tau^Z[(\Lambda^T(\rho) \otimes \Lambda^P)]^{-1}) \tag{11}$$

For positive values of ρ smaller than 1 this is a proper distribution.

3.4. Other prior distributions

To complete specification of the model, prior distributions must be specified for all of the remaining unknown parameters in the model. These parameters are:

- $\mu_{t(i)}$, the mean for underlying space–time log-$PM_{2.5}$ process;
- β^D, the covariates for CMAQ bias structure;
- τ^X, the variance of the measurement error in the monitor observations;
- τ^Y, the variance of the measurement error in the CMAQ output;
- τ^Z, the variance of the underlying space–time log-$PM_{2.5}$ process;
- ρ, the temporal autocorrelation parameter of the underlying space–time log-$PM_{2.5}$ process in the temporal AR(1) covariance matrix.

DOI: 10.1002/env

Table 1. Prior assumptions

Parameter	Prior	Mean	Variance
$\mu_{t(i)}$	$N(0,1 \times 10^3)$	0	1000
β^D	$N(0,1 \times 10^3)$	0	1000
τ^X	IG(25×10^6, 1×10^6)	0.04	6.4E-11
τ^Y	IG(2×10^9, 1×10^9)	0.5	1.25E-10
τ^Z	IG(1×10^{-3}, 1×10^{-3})	Not defined	Not defined
ρ	$U(0,1)$	0.5	0.083

Non-informative normal priors are assigned to $\mu_{t(i)}$ and β^D. Inverse gamma distributions are assigned to each τ. A uniform prior distribution between zero and one is defined for ρ. Exact hyper-parameters are provided in Table 1.

3.5. Full conditionals and Markov chain Monte Carlo

Computing posterior distributions for complex models such as the one proposed here in closed form is generally impossible. As previously stated, we will use a Gibbs sampler to sample from the joint posterior distribution, with a Metropolis–Hastings within Gibbs steps. To implement Gibbs sampling, the full conditional distribution for each random variable in the posterior must be determined. For one component of the posterior, ρ, the full conditional distribution is not in a recognizable form. Thus, a Metropolis–Hastings step for this variable is implemented within the MCMC scheme.

There are two types of full conditionals to be sampled in our Gibbs sampler: Gaussian distributions for the $PM_{2.5}$ grid surface parameters and regression parameters μs and βs; inverse gamma distributions for the variance parameters τs. The full conditional for the temporal correlation parameter ρ is not in recognizable form and ρ is drawn via Metropolis–Hasting sampling. Full conditionals for the Gaussian and inverse gamma distributions and the sampling equations for ρ are provided in Appendix A.

Setting

$$n_t^T = \begin{cases} 1 & t = 1, N^T \\ 1 + \rho^2 & 1 < t < N^T \end{cases} \tag{12}$$

and

$$\mu_i^Z(\rho, Z) = \frac{\rho}{n_{t(i)}^T} \sum_{j \in \partial_t i} Z_j + \frac{1}{n^P} \sum_{j \in \partial_p^r i} Z_j - \frac{\rho}{n_{t(i)}^T n^P} \sum_{j \in \partial^r i} Z_j \tag{13}$$

$$\tau_i^Z(\rho) = \tau^Z \frac{1 - \rho^2}{n_{t(i)}^T n^P} \tag{14}$$

$$I_{(a,b)}(\rho) = \begin{cases} 1 & a < \rho < b \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

DOI: 10.1002/env

a little algebra yields

$$[\rho|-] \propto I_{(a,b)}(\rho)(1-\rho^2)^{-\frac{1}{2}N^T(N^T-1)} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\frac{1}{\tau_i^Z(\rho)}Z_i(Z_i-\mu_i^Z(\rho,Z))\right\} \tag{16}$$

Notice that in Equations (13) and (14), $t(i)$ indicates the temporal index of the grid cell i, ∂_t^i indicates the temporal first nearest neighborhood of the grid cell i, $\partial_p^r i$ indicates the spatial neighborhood of order r of the grid cell i, and $\partial^r i$ indicates the space–time neighborhood of order 1 in time and order r in space of the grid cell i.

A Metropolis–Hastings step is required for this parameter. To accomplish this, a proposal distribution is selected. In this case, we use

$$\rho' \sim N(\rho, \tau^\rho) \tag{17}$$

The variance of the proposal distribution, τ^ρ, is set equal to the maximum minus the minimum ρ observed in the chain divided by 100. The 100 factor was tuned manually to achieve an optimal acceptance. The proposed new value for ρ is accepted with probability

$$\min\left\{\frac{[\rho'|-]}{[\rho|-]}, 1\right\} \tag{18}$$

with $[\rho|-]$ being the full conditional probability defined by Equation (16).

4. MODEL FITTING RESULTS

When data for a full year (2001) on the $12 \times 12\,\text{km}^2$ grid pictured in Figure 1 are analyzed, the computational burden is large. There are $213 \times 188 \times 365$ grid cells (over 9 million!) for which $PM_{2.5}$ concentration must be inferred. In the context of a Bayesian model, this implies that our posterior is extremely high dimensional. As with many HBMs, the posterior is sampled using an MCMC algorithm based on the Gibbs sampler (e.g., Gelfand and Smith, 1990; Roberts and Casella, 1999). The algorithm generates a Markov chain consisting of realizations of each posterior parameter. The distribution of the realizations from the Markov chain converges to the posterior distribution as the number of steps in the chain increases. Thus, after a sufficient ''burn-in'' period, observations from the Markov chain are approximately distributed according to the posterior. For our simulations, the burn-in period consisted of 1000 draws. After the burn-in period, 5000 samples from the Markov chain were used to characterize posterior distributions. Convergence was assessed by plotting chains of the model parameters. However, it was not possible to store and evaluate the posterior distributions of Z given the space–time dimensions of this analysis. To store values for 9 million grid cells (as 4 byte floating point numbers) for even 10 iterations would require nearly half of a gigabyte of memory.

Table 1 defines the prior assumptions. With the exception of τ^X and τ^Y, the prior assumptions were selected as non-informative. Our approach for defining prior distributions for τ^X and τ^Y was based on FRM quality assurance data and sensitivity analyses of CMAQ model runs. We used very small prior variances to define highly informative prior distributions for both of these variables. The prior on τ^X was chosen to correspond to a monitoring data coefficient of variation (CV) of 20%. While a 20% CV for the FRM data is slightly high, this was chosen as appropriate due to the hidden change of support issue underlying this component of the model; i.e., the mean of the monitoring data distribution is

DOI: 10.1002/env

constant across $12 \times 12\,\text{km}^2$ grid cells. A CV of 80% was assigned to the CMAQ output. Because there is such an overwhelming quantity of CMAQ output available (over 9 million data points), it took quite a strong prior on τ^Y for the prior to have much of an effect on the posterior. The underlying lesson from this analysis is that perception of prior strength must be adjusted when such overwhelming quantities of numerical model output are being analyzed. For this analysis, a neighborhood of size 1 was used. Little difference in the predictive results was found using a larger neighborhood of order 2.

Figure 3 shows the daily mean levels for the predicted $PM_{2.5}$ surface, the monitor data, and the CMAQ results. The daily means include only CMAQ grid cells for which monitoring data are available in order to make each time series equally representative of the spatial domain. Since monitors are placed to protect public health, the monitors are much more representative of the surface extremes than an average over the entire surface would be. We immediately see that the temporal pattern is common to each of these even though they are based on very different sources. The fact that there appear to be some biases among them should be expected. The CMAQ results are lower than the other levels in the spring and summer months and higher in the fall and winter months.

4.1. Primary model parameters

The estimates and credible intervals of the primary model and variance parameters are shown in Table 2. Note that $\mu_{t(i)}$, the daily means, and β_t^D, the bias coefficients, are not provided in this table for brevity.

The model variance estimates in Table 2 are dependent on the priors chosen for them and are also dependent on each other. Because the $PM_{2.5}$ measurement data were log-transformed prior to modeling, the variance parameters are best interpreted as coefficients of variation on the $PM_{2.5}$ scale using the transformation $\text{CV} = \sqrt{e^{\tau} - 1} \times 100$. Thus, the posterior credible interval for the monitoring

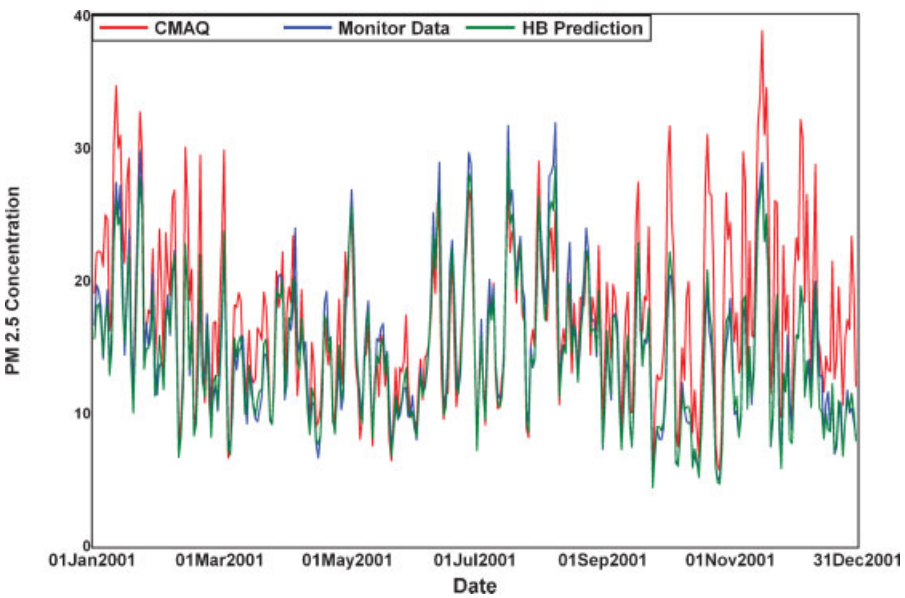

Figure 3. Daily mean levels for: predicted surface, monitoring data, and CMAQ

DOI: 10.1002/env

Table 2. Marginal posterior distributions of non-spatial parameters

Parameter	Mean	95% Credible Interval
τ^X	4.0027E-2	(4.0012E-2, 4.0043E-2)
τ^Y	4.9909E-1	(4.9907E-1, 4.9911E-1)
τ^Z	8.2179E-2	(8.1724E-2, 8.2961E-2)
ρ	0.40	(0.40, 0.41)

data coefficient of variation is quite tight, about 20.21%. The exponentiated posterior expectation of the predicted log-transformed surface is used to predict daily $PM_{2.5}$ levels. Finally, the model $PM_{2.5}$ autocorrelation estimate of 0.40 in Table 2 indicates considerable auto-correlation from one day to the next. The autocorrelation is a measure of how quickly the concentrations change.

Daily predicted $PM_{2.5}$ surfaces and estimates of uncertainty in these predictions as measured by the coefficient of variation associated with each grid cell prediction are shown in Figure 4. The days chosen for these spot predictions are 4 July 2001 and 24 December 2001. On 4 July 2001, 97 monitoring sites collected data; on 24 December 2001, 513 monitoring sites collected data. Consistent with the quantity of monitoring data available on each day, the coefficient of variation (CV) is higher on 4 July as compared to 24 December in most grid cells. The CV range in Figure 4 seems consistent with our prior assumptions for CMAQ (CV$\sim$80%) and monitoring data (CV$\sim$20%). On 4 July a large scale $PM_{2.5}$ event was occurring over the northeastern U.S. On 24 December, $PM_{2.5}$ was much lower with high

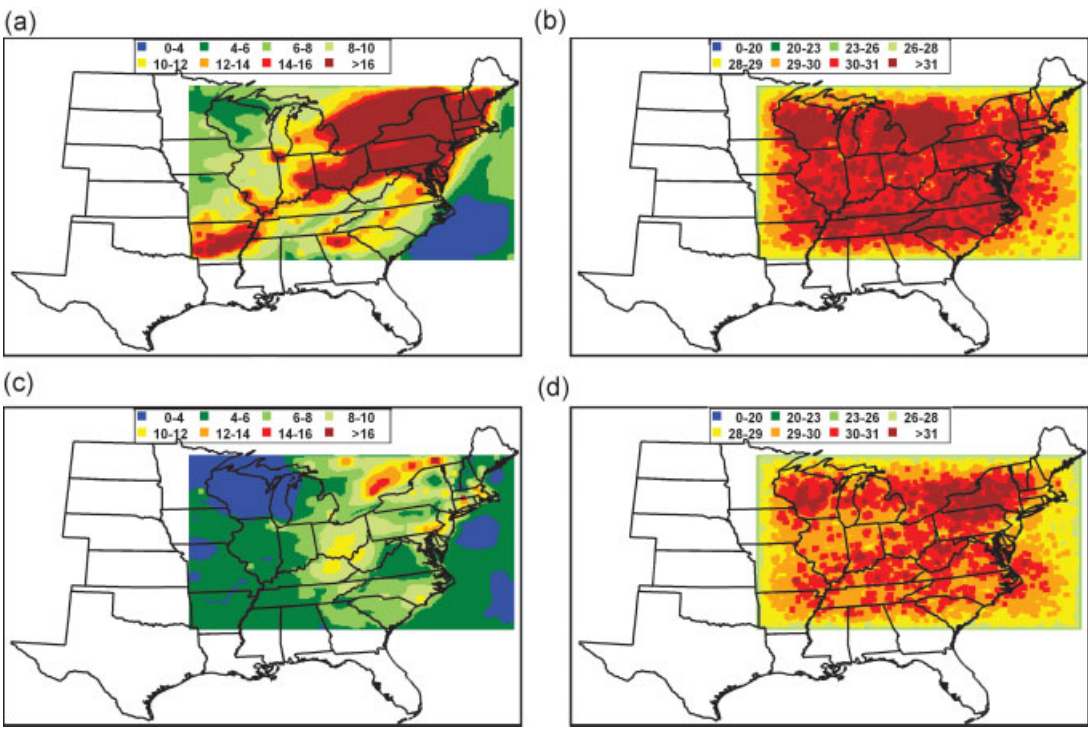

Figure 4. (a and b) 4 July 2001 and (c and d) 24 December 2001 daily $PM_{2.5}$ predictive surfaces and coefficients of variation for daily predictive surfaces

DOI: 10.1002/env

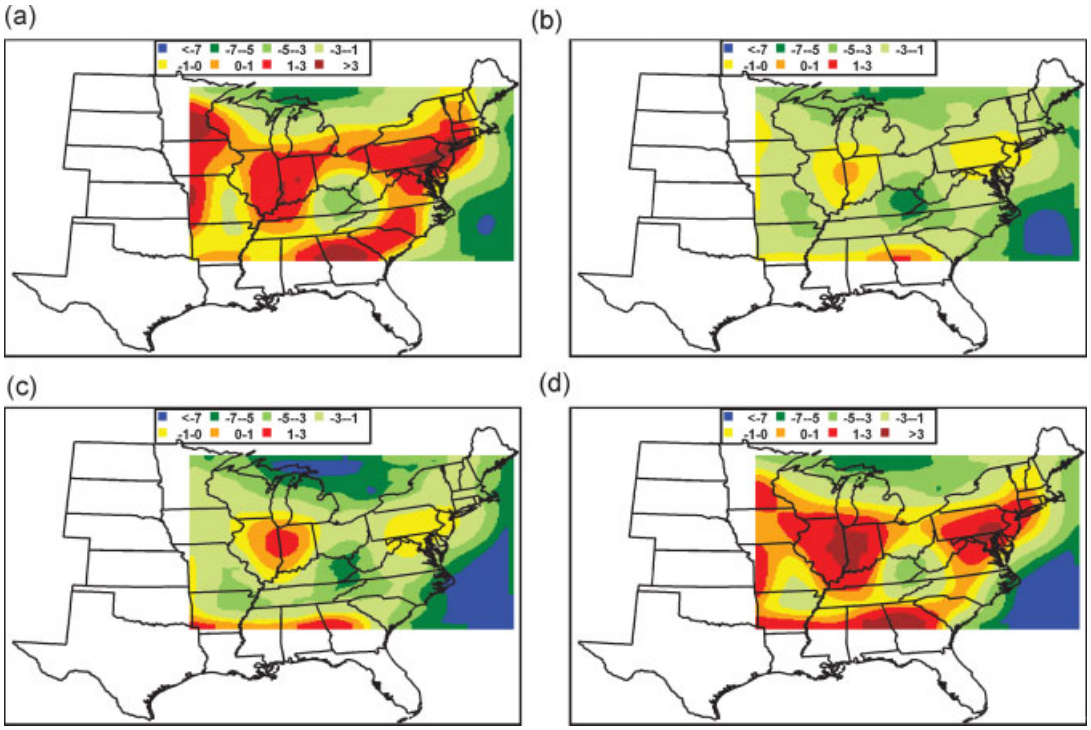

Figure 5. CMAQ $PM_{2.5}$ seasonal bias patterns: (a) December–February, (b) March–May, (c) June–August, and (d) September–November

levels appearing predominantly in Toronto, Ottawa, and Quebec. A software tool for visualizing all the daily predictive surfaces and the input datasets is available from the authors.

4.2. Spatial-temporal bias results

The spline spatial-temporal bias function was fitted through the posterior estimation of a vector of β^D coefficients. These coefficients allowed the expected posterior $PM_{2.5}$ value in each grid cell to be different than the CMAQ value in that grid cell in a systematic manner. Figure 5 demonstrates the estimated seasonal spatial bias of CMAQ and is calculated by averaging the daily bias estimates. When the bias surface is negative, the observed CMAQ values are assessed to be under predictions of the true $PM_{2.5}$ levels. Positive bias indicates that CMAQ is over-predicting $PM_{2.5}$. Thus, Figure 5 indicates that CMAQ is over-predicting in a number of metropolitan areas, including Chicago, Atlanta, Indianapolis, Philadelphia, and Washington, D.C. Examination of bias plots over time indicates that CMAQ has a tendency to under-predict more in the summer than the winter.

5. MODEL VALIDATION RESULTS

We performed a model validation analysis to compare the HB predictive results at 2001 STN/IMPROVE monitoring sites to predictions at those locations from two other approaches: (1) traditional

DOI: 10.1002/env

kriging predictions based solely on the FRM monitoring data and (2) CMAQ output at these locations. We assumed the STN/IMPROVE measurements to be representative of truth, and did not consider potential bias in either the STN or IMPROVE gravimetric mass measurements. STN data collocated with FRM monitoring sites used in fitting the HB model were eliminated from the validation data set, leaving 44 sites for the validation analysis.

Mean squared prediction error (MSE) and bias are calculated to evaluate the predictive capability of these three different models. To assess the ability of the Bayesian model to accurately characterize prediction uncertainty, the percentage of validation data within the 95% prediction credible interval was calculated. We performed a similar analysis for the kriging model by calculating 95% confidence intervals at the validation sites. We used an exponential variogram model for the kriging model. The exponential parameters were estimated by fitting this model to an empirical variogram based on combining the daily empirical variograms. We decided against fitting daily variogram models due to the sparsity of FRM data for 2 of every 3 days within the year. For these days, we would not be able to obtain good estimates of small-scale variability. For each day, predictions were obtained for the STN/IMPROVE site locations from the three modeling approaches and the validation statistics were calculated across all days and sites. Our validation only occurs every third day, according to the sampling schedule of STN/IMPROVE. This corresponds to the full network FRM schedule. Thus, we are unable to evaluate sparse monitoring days where we expect data fusion to outperform interpolation techniques based solely on the monitoring data. We did consider using a few every day FRM monitoring sites for validation, but decided that they were more important for estimating temporal structure in the model. However, future analyses should give some attention to defining every day FRM sites as validation sites to evaluate the benefit of including CMAQ output in fusion analyses.

We fitted the HBM several times using a range of reasonable priors for τ^X and τ^Y while always assuming τ^X to follow a non-informative prior. Then we performed a validation analysis to assess the relative predictive performance of the HBM, traditional kriging, and CMAQ as described above. In terms of MSE, the HBM and kriging approaches provided similar results across all HBM runs. For bias, the HBM outperformed kriging by 10–15% depending on the prior assumptions for τ^X and τ^Y. CMAQ was nearly unbiased for this analysis.

Kriging uncertainties were found to be quite small and this result is reflected in the small percentage (59%) of kriging prediction intervals capturing the validation data. This compares to HBM predictive interval results of 80–90% depending on the HBM run. We attribute the difference between the HBM results and the 95% nominal rate to the difference in the measurement errors in the validation to those in the FRM data used in fitting the HBM model. Unfortunately, error-free $PM_{2.5}$ monitoring data are not available with current $PM_{2.5}$ monitoring approaches.

6. CONCLUSIONS

We have proposed a high resolution, flexible spatial-temporal model for daily $PM_{2.5}$ concentrations for most of the eastern U.S. The HBM approach provides a coherent framework for combining monitoring data with numerical mode output. The primary advantages are increased model flexibility and the ability to predict pollution gradients and uncertainties for successive days that might otherwise be unknown using interpolation results from $PM_{2.5}$ monitoring data with varying sampling frequencies. This model provides daily spatial surfaces of CMAQ bias that can be used to guide future research for improving CMAQ output. In comparison to interpolation of the monitoring data with ordinary kriging, the combined space–time model outperforms kriging in terms of bias and prediction intervals. $PM_{2.5}$

DOI: 10.1002/env

predictions from this combined modeling approach will be useful for developing environmental public health indicators and linking $PM_{2.5}$ with public health data. Future analyses will consider the use of continuous Geostationary Operational Environmental Satellites (GOES) satellite data that can be averaged over daily time periods. (http://www.nesdis.noaa.gov/satellites.html) These data will be treated as another data source providing information on the underlying space–time log-$PM_{2.5}$ process.

ACKNOWLEDGEMENTS

The authors thank Jenise Swall, Kristen Foley, and Fred Dimmick for their many helpful comments and suggestions. The research described in this article has been funded in part by the U.S. EPA through Contract Number 68-D-02-061 to Battelle. Although it has been reviewed by the U.S. EPA, it does not necessarily reflect the Agency's policies or views.

REFERENCES

Chen M-H, Shao Q-M, Ibrahim JG. 2000. *Monte Carlo Methods in Bayesian Computation*. Springer: NewYork, NY.

Cowles MK, Zimmerman DL. 2003. A Bayesian space-time analysis of acid deposition data combined from two monitoring networks. *Journal of Geophysical Research* **108**: 90–106.

Cressie N, Kaiser MS, Daniels MJ, Aldworth J, Lee J, Lahiri SN, Cox L. 1999. Spatial analysis of particulate matter in an urban environment. In *GeoEnvII: Geostatistics for Environmental Applications*, Gomez-Hernandez J, Soares A, Froidevaux R (eds). Kluwer: Dordrecht; 41–52.

Fuentes M, Raftery A. 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61**: 36–45.

Gelfand AE, Smith AFM. 1990. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. *Bayesian Data Analysis*, (2nd edn). Chapman and Hall/CRC: Boca Raton, FL.

Gilks WR, Richardson S, Speigelhalter DJ. 1998. *Markov Chain Monte Carlo in Practice*. Chapman & Hall: London.

Jun M, Stein ML. 2004. Statistical comparison of observed and CMAQ modeled daily sulfate levels. *Atmospheric Environment* **38**: 4427–4436.

Robert CP, Casella G. 1999. *Monte Carlo Statistical Methods*. Springer-Verlag: New York.

Sahu S, Mardia KV. 2005. A Bayesian kriged-kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society* Series C, **54**: 223–244.

Sahu S, Gelfand A, Holland DM. 2006. Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics* **11**: 61–86.

Smith RL, Kolenikov S, Cox LH. 2003. Spatio-temporal modeling of $PM_{2.5}$ data with missing values. *Journal of Geophysical Research-Atmospheres* **108**(D24): 9004. DOI: 10.1029/2002JD002914

Spiegel L, Tiller W. 1996. *The NURBS book*, (2nd edn). Springer-Verlag: Berlin Heidelberg.

Tierney L. 1994. Markov chains for exploring posterior distributions. *The Annals of Statistics* **22**: 1701–1728.

U.S. Environmental Protection Agency. 2003. *National Air Quality and Emission Trends Report, 2003 Special Studies Edition*. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, EPA 454/R-03-005.

Wikle CK, Berliner M, Cressie N. 1998. Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics* **5**: 117–154.

Zidek JV, Sun L, Le N, Ozkaynak H. 2002. Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM10 field. *Environmetrics* **13**: 595–613.

Zimmerman DL, Holland DM. 2005. Complementary co-kriging: spatial prediction using data combined from several environmental monitoring networks. *Environmetrics* **16**: 219–234.

APPENDIX A: SAMPLING EQUATIONS

In the following formulae $N(\mu, \tau)$ indicates a univariate normal distribution with mean μ and variance τ, $N(\mu, \Sigma)$ indicates a multivariate normal distribution with mean μ and covariance matrix Σ, and $IG(\gamma, \delta)$ indicates an inverse gamma distribution with shape γ and scale δ.

DOI: 10.1002/env

In the sampling equation for Z we define

$$\mu_i^Z(\rho, Z) = \frac{\rho}{n^T_{t(i)}} \sum_{j \in \partial_t i} Z_j + \frac{1}{n^P} \sum_{j \in \partial_p^r i} Z_j - \frac{\rho}{n^T_{t(i)} n^P} \sum_{j \in \partial^r i} Z_j$$

$$\tau_i^Z(\rho) = \tau^Z \frac{1 - \rho^2}{n^T_{t(i)} n^P}$$

where $t(i)$ indicates the temporal index of the grid cell i, $\partial_t i$ denotes the temporal first nearest neighborhood of the grid cell i, $\partial_p^r i$ is the spatial r-nearest neighborhood of the grid cell i, and $\partial^r i$ denotes the space–time neighborhood of order 1 in time and order r in space.

$\theta^\mu, \tau^\mu, \theta^D, \tau^D$ are mean and variance hyper-parameters coming from the normal priors for μ and β^D; $\gamma^X, \delta^X, \gamma^Y, \delta^Y, \gamma^Z, \delta^Z$ are shape and scale hyper-parameters coming from the inverse gamma priors for τ^X,τ^Y, and τ^Z; and a,b are the hyper-parameters coming from the uniform prior for ρ. In the following, N_i^X represents the number of monitor observations in the space–time grid cell i and I_t represents the set of indexes of all the space–time grid cells having temporal component equal to t.

Variable: $Z_i \in \mathbf{R}$

$$[Z_i|-] = N(\tau \cdot b, \tau)$$

where

$$\frac{1}{\tau} = \frac{1}{\tau^X} N_i^X + \frac{1}{\tau^Y} + \frac{1}{\tau_i^z}$$

$$b = \frac{1}{\tau^X}[X_i - N_i^X \mu_{t(i)}] + \frac{1}{\tau^Y}[y_i - (\mu_{t(i)} + \beta^D D_i)] + \frac{1}{\tau_i^Z(\rho)} \mu_i^Z(\rho, Z)$$

Variable: $\mu_t \in \mathbf{R}$

$$[\mu_t|-] = N(\tau \cdot b, \tau)$$

where

$$\frac{1}{\tau} = \frac{1}{\tau^X} \sum_{i \in I_t} N_i^X + \frac{1}{\tau^Y} N^T + \frac{1}{\tau^\mu}$$

$$b = \tfrac{1}{\tau^X} \sum_{i \in I_t} [X_i - N_i^X Z_i] + \tfrac{1}{\tau^Y} \sum_{i \in I_t} [y_i - (Z_i + \beta^D D_i)] + \tfrac{1}{\tau^\mu} \theta^\mu$$

Variable: $\beta^D \in \mathbf{R}^{N^D}$

$$[\beta^D|-] = N(\Sigma \cdot b, \Sigma)$$

where

$$(\Sigma^{-1})_{kl} = \frac{1}{\tau^Y} \sum_{i=1}^{N} N_i^Y D_{ik} D_{il} + \delta_{kl} \frac{1}{\tau^D}$$

$$b_k = \frac{1}{\tau^Y} \sum_{i=1}^{N} D_{ik}[y_i - (\mu_{t(i)} + Z_i)] + \frac{1}{\tau^D} \theta^D$$

DOI: 10.1002/env

Variable: $\tau^X \in \mathbf{R}$

$$[\tau^X|-] = \mathrm{IG}(\gamma, \delta)$$

where

$$\gamma = \frac{1}{2}\sum_{i=1}^{N} N_i^X + \gamma^X$$

$$\delta = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N_i^X} [x_{ik} - (\mu_{i(t)} + Z_i)]^2 + \delta^X$$

Variable: $\tau^Y \in \mathbf{R}$

$$[\tau^Y|-] = \mathrm{IG}(\gamma, \delta)$$

where

$$\gamma = \frac{1}{2}N + \gamma^Y$$

$$\delta = \frac{1}{2}\sum_{i=1}^{N} [y_i - (\mu_{t(i)} + Z_i + \beta^D D_i)]^2 + \delta^Y$$

Variable: $\tau^Z \in \mathbf{R}$

$$[\tau^Z|-] = G(\gamma, \delta)$$

where

$$\gamma = \frac{1}{2}N + \gamma^z$$

$$\delta = \frac{1}{2}\sum_{i=1}^{N} \frac{\tau^Z}{\tau_i^Z(\rho)} Z_i(Z_i - \mu_i^z) + \delta^Z$$

Variable: $\rho \in \mathbf{R}$

$$[\rho|-] \propto I_{(a,b)}(\rho)(1-\rho^2)^{-\frac{1}{2}N^P\left(N^T-1\right)} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} \frac{1}{\tau_i^Z(\rho)} Z_i(Z_i - \mu_i^Z(\rho, Z))\right\}$$

where $I_{(a,b)}$ is the indicator function of the interval $[a,b]$. This full conditional is not a recognized form, so it has to be sampled using a Metropolis–Hastings step.

Jump:

$$\rho' \sim N(\rho, \tau^\rho)$$

DOI: 10.1002/env

Acceptance:

$$\min\left\{\frac{I_{(a,b)}(\rho')(1-(\rho')^2)^{-\frac{1}{2}N^P\left(N^T-1\right)}\exp\left\{-\sum_{i=1}^{N}\frac{1}{2\tau_i^Z(\rho')}Z_i(Z_i-\mu_i^Z(\rho',Z))\right\}}{(1-\rho^2)^{-\frac{1}{2}N^P(N^T-1)}\exp\left\{-\sum_{i=1}^{N}\frac{1}{2\tau_i^Z(\rho)}Z_i(Z_i-\mu_i^Z(\rho,Z))\right\}},1\right\}$$

The proposal variance τ^ρ was manually tuned to achieve optimal acceptance rate.

DOI: 10.1002/env