Fusing point and areal level space-time data with application to wet deposition

Sujit K. SAHU, Alan E. GELFAND and David M. HOLLAND *

Abstract

Motivated by the problem of predicting chemical deposition in the eastern United States at weekly, seasonal, and annual scales, this paper develops a framework for joint modeling of point and grid referenced spatio-temporal data in this context. The proposed hierarchical model is able to provide accurate spatial interpolation and temporal aggregation by combining information from observed point referenced monitoring data and gridded output from a numerical simulation model known as the Community Multi-scale Air Quality (CMAQ) model. The technique avoids the change of support problem which arises in other hierarchical models for data fusion settings to combine point and grid referenced data. The hierarchical space-time model is fitted to weekly wet sulfate and nitrate deposition data over the eastern United States. The model is validated with set-aside data from a number of monitoring sites. Predictive Bayesian methods are developed and illustrated for inference on aggregated summaries such as quarterly and annual deposition maps.

Key Words: Change of support problem; hierarchical model; Markov chain Monte Carlo; measurement error model; spatial interpolation; stochastic integrals.

*Sujit K. Sahu is senior lecturer, School of Mathematics, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK. (Email: S.K.Sahu@soton.ac.uk). Alan E. Gelfand is Professor, Department of Statistical Science, Duke University, Durham, NC, USA (Email: alan@stat.duke.edu). David M. Holland is senior statistician, U.S. Environmental Protection Agency, National Exposure Research Laboratory, Research Triangle Park, NC, USA (Email: holland.david@epa.gov).

1 Introduction

The combustion of fossil fuel produces a wide variety of chemicals, including such gases as sulfur dioxide and nitrogen oxides. These gases are emitted to the air, transformed to acidic compounds, and are then returned to the earth. Most of the acid deposition in the eastern U.S. can be attributed to the release of sulfur dioxide and nitrogen oxides from large fossil fueled power plants. When delivered by precipitation, such as rain, snow, or fog, the process is called wet sulfate and nitrate deposition. Wet deposition is responsible for damage to lakes, forests, and streams.

The primary objective of this study is to develop a high-resolution model for wet chemical deposition that offers better inference than is currently possible using just National Atmospheric Deposition Program (NADP, `nadp.sws.uiuc.edu`) wet deposition measurements and classical interpolation techniques. The proposed model uses deposition and precipitation data from NADP monitoring sites and output from a computer simulation model known as the Community Multi-Scale Air Quality Model (CMAQ, `epa.gov/asmdnerl/CMAQ`) on a 12 kilometer square grid. The CMAQ model uses variables, such as power station emission volumes, meteorological information, and land-use, to predict average deposition levels. However, it is well known that these predictions are biased; the monitoring data provide more accurate deposition information. The mismatch in the spatial domains for the point and grid referenced computer output is often alluded to as the 'change of support problem' and creates challenges in modeling and model fitting, see below for more details. Combining information from disparate sources is a relatively new activity in modeling air and deposition data, but fundamental to providing improved information for environmental decisions and enabling greater understanding of the processes that underlie deposition.

The contribution of this article is the development of a joint model by combining a conditionally auto-regressive (CAR) model for the gridded CMAQ data and a space-time process model for observed point level data. Model components are linked using latent space-time processes in a Bayesian hierarchical modeling setup. This modeling strategy yields a better model for high spatial resolution and time. All predictive inference is performed using the point level model. A key feature of our strategy is avoidance of stochastic integration of the observed point level monitoring process to a grid level process.

More precisely, the average deposition level in a grid cell A_j at time t, denoted by

$Z(A_j, t)$, need not be the level observed at any particular site $\mathbf{s}_i$ in A_j, denoted by $Z(\mathbf{s}_i, t)$. The change of support problem in this context addresses converting the point level $Z(\mathbf{s}_i, t)$ to the grid level $Z(A_j, t)$ through the stochastic integral,

$$Z(A_j, t) = \frac{1}{|A_j|} \int_{A_j} Z(\mathbf{s}, t)\ d\mathbf{s}, \tag{1}$$

where $|A_j|$ denote the area of the grid cell A_j. Fusion modeling, working with *block averaging* as in (1) has been considered by, e.g., Fuentes and Raftery (2005).

Our approach introduces a latent point level atmospheric process which is centered, in the form of a measurement error model (MEM), around a grid cell based latent atmospheric process. The latent processes are introduced to capture point masses at 0 with regard to deposition while the MEM circumvents the stochastic integration. In particular, the point level observed data represent 'ground truth' while gridded CMAQ output are anticipated to be biased. As a result, the MEM enables calibration of the CMAQ model. The opposite problem of disaggregation, i.e. converting the grid level computer output denoted by $Z(A_j, t)$ to point level ones, $Z(\mathbf{s}_i, t)$ is not required. The only assumption is that $Z(A_j, t)$ is a reasonable surrogate for $Z(\mathbf{s}_i, t)$ if the site $\mathbf{s}_i$ is within the grid cell A_j. (This is confirmed by empirical evidence, see the discussion about Figure 5 in Section 2.)

The amount of wet deposition is directly related to precipitation – there can be no deposition without precipitation. Hence, accurate predictions here require utilization of precipitation information. Note that both the precipitation and deposition data have atomic distributions, i.e., they are continuous random variables with positive mass at zero. Our proposal is to build a model for deposition based on precipitation which is able to handle these atoms. We introduce a conceptual latent space-time atmospheric process which drives both precipitation and deposition as assumed in the mercury deposition modeling of Rappold *et al.* (2008). However, Rappold *et al.* did not address the fusion problem with modeled output. Rather, they used a point level joint process model, specified conditionally for the atmospheric, precipitation and deposition processes. We incorporate a huge amount of CMAQ numerical model output data at 12km grid scale.

The wet deposition model is applied separately to the wet sulfate and wet nitrate deposition. There is high positive correlation between the compounds because of their dependence on precipitation but our interest here is to predict the sulfate and nitrate depositions separately. The model is fitted at point level spatial resolution and weekly temporal resolution,

enabling spatial interpolation and temporal prediction of deposition as well as aggregation in space or time to facilitate seeing patterns and trends in deposition.

Our fully model-based approach removes many of the shortcomings of inverse distance weighting (IDW)used by NADP to predict annual spatial patterns of wet deposition, see `nadp.sws.uiuc.edu/isopleths/annualmaps.asp`. The IDW method interpolates a deposition value at a new site by taking weighted means of depositions at data sites; the weights are inversely proportional to the square of the distance between the interpolation site and the data sites so these interpolations are most accurate near the data sites. However, IDW has serious limitations: (i) it cannot accommodate covariate information; (ii) handling of missing observations by simple averaging of available observations is ad-hoc and fails to take account of variability in these observations; (iii) it is not possible to associate any sort of uncertainty with estimated quarterly or annual totals, i.e. to provide uncertainty maps for the region. Lastly, it does not recognize the problem of point mass at 0. Ordinary kriging suffers similar problems and, by ignoring uncertainty in model parameters, tends to underestimate predictive variability; and (iv) modeling at annual scale is possible but sacrifices process understanding that is available at weekly resolution.

In recent years there has been a surge of interest in developing methods for modeling space-time data. Hierarchical Bayesian approaches for spatial prediction of air pollution have been developed, see e.g., Brown *et al.*, (1994), Huerta *et al.*, (2004), Le and Zidek (1992), Sahu and Mardia (2005), Sahu *et al.* (2006, 2007), and Wikle (2003). However, there are only a handful of papers which have discussed models for wet deposition. In a series of articles Haas (1990a, 1990b, 1995, 1996) used statistical methods including moving window regression, kriging and co-kriging and spatio-temporal modeling to study various aspects of depositions. Oehlert (1993) used a spatio-temporal model to estimate trend in annual sulfate depositions. Bilonick (1985) used classical geostatistical methods to model the space-time covariance structure of wet sulfate deposition. Grimm and Lynch (2004) developed a high-resolution model for wet sulfate and nitrate deposition using precipitation and many topographic variables observed over a dense grid. Rappold *et al.* (2008) modeled wet mercury deposition data. Fuentes and Raftery (2005) offer the only data fusion work in this area. However, their analysis is not dynamic and fitting their model with a large number of grid cells becomes computationally infeasible, as we clarify below.

The remainder of this article is organized as follows. In Section 2 we describe the

available data. Modeling developments are presented in Section 3. Prediction details are discussed in Section 4. Section 5 provides the modeling results and analyses. A few summary remarks are provided in Section 6 and an Appendix contains the computational details for Gibbs sampling.

2 Exploratory Analysis

We have weekly deposition and precipitation data for the 52 weeks in the year 2001 from 128 sites in the eastern U.S., see Figure 1 for their locations. We use data from $n = 120$ sites for model estimation and prediction and the data for remaining 8 sites are used for validation. The validation sites are marked as A–H in Figure 1 and have been chosen based on several considerations. The eight sites are spread across the study region without forming clusters. The validation sites are some distance away from nearest data sites, in fact the distance between a validation site and its nearest data site ranges from 40 kilometers to 186 kilometers. More specifically, sites A and D are chosen because they fall in a high precipitation area (see Figure 2 for the annual precipitation map). Site H is chosen because it is in an area where annual deposition is higher than the average. There are 9 missing observations (out of a total of 416 =8×52) in the validation data set most of which are for week 52, the end of the year holiday period in the USA.

There are 6240 (=120×52) total number of observations in our modeling data set. For precipitation, 536 ($\approx$ 8.6%) of these were missing. The deposition values in these 536 location-week combinations were also missing. The precipitation and deposition values in 507 location-week combinations (out of the remaining 5704) were zero. The deposition values in additional 119 location-week combinations were also recorded as missing. Hence, there are 655 location-week combinations ($\approx$ 10.5%) where deposition values were missing. We note that at each location and week combination either both the two types of deposition values are positive or both are zero, i.e. one cannot be zero without the other being zero as well. We also note that positive precipitation necessarily implies positive deposition (which sometimes can be very small).

The boxplots of the weekly sulfate and nitrate depositions in units of kilogram per hectare (kg/ha) are plotted in Figure 3. The labels on the horizontal axes are the last week of the months. The figure confirms the well known fact that depositions levels are higher

on the average for the wetter spring and summer months than the dryer winter months, see e.g. Brook *et al* (1995). Strong linear relationships between deposition and precipitation on the log scale are seen in Figure 4.

We model weekly CMAQ output from $J = 33,390$ grid cells covering our study region yielding 1,736,280 modeled values for the year. There is some evidence of linear relationship on the log scale between observed deposition and CMAQ model output for the cell containing the observation location, especially for higher values, see Figure 5. The association between the two is degraded toward the lower-end of the scale due to the presence of zero values which have been replaced by a small positive number to avoid taking logarithm of 0. This is done for data presentation purposes only.

For spatial prediction we have weekly precipitation data from 2827 predictive sites covering our study region. A map of the annual total precipitation (in centimeters) is provided in Figure 2. Areas in the south-west corner of the map received more precipitation than others. However, in the model fitting we used only the precipitation data from the 120 NADP monitoring sites where we have deposition data. In principle, we could attempt to introduce the full set of precipitation data into our modeling but this will add substantially to the computation (see expression (6) below) with little expected gain.

We have examined empirical variograms and their smoothed fits for many different versions of aggregated data as well as for residuals after fitting regression models for log deposition values on log precipitation and log CMAQ values. The variograms revealed clear evidence of spatial dependence and suggested ranges between 500 to 1500 kilometers. In our model fitting we choose optimal values of the range parameters using validation methods. Throughout the paper we use the geodetic distance, see, e.g. Banerjee *et al.* (2004, Chapter 1) between two locations with given latitudes and longitudes.

3 Modeling Wet Deposition

We develop the wet deposition model in two stages described in Sections 3.1 and 3.2 respectively and provide a directed acyclic graph (DAG) in Figure 6. At the end of Section 3.2 we briefly discuss what a fusion model using block averaging (as in (1)) would look like, with an associated DAG in Figure 7. Section 3.3 discusses the prior distributions and records the joint posterior distribution.

3.1 First Stage Specification

Let $P(\mathbf{s}_i, t)$ and $Z(\mathbf{s}_i, t)$ denote the observed precipitation and deposition (either sulfate or nitrate) respectively at a site $\mathbf{s}_i, i = 1, \ldots, n$ in week t, $t = 1 \ldots, T$. We suppose that $P(\mathbf{s}_i, t)$ and $Z(\mathbf{s}_i, t)$ are driven by a conceptual point level latent atmospheric process, denoted by $V(\mathbf{s}_i, t)$, and both take the value zero if $V(\mathbf{s}_i, t) < 0$ to reflect that there is no deposition without precipitation. That is,

$$P(\mathbf{s}_i, t) = \begin{cases} \exp\left(U(\mathbf{s}_i, t)\right) & \text{if } V(\mathbf{s}_i, t) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

and

$$Z(\mathbf{s}_i, t) = \begin{cases} \exp\left(Y(\mathbf{s}_i, t)\right) & \text{if } V(\mathbf{s}_i, t) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The random variables $U(\mathbf{s}_i, t)$ and $Y(\mathbf{s}_i, t)$ are thus taken as log observed precipitation and deposition respectively when $V(\mathbf{s}_i, t) > 0$. The models described below will specify their values when $V(\mathbf{s}_i, t) \leq 0$ and/or the corresponding $P(\mathbf{s}_i, t)$ or $Z(\mathbf{s}_i, t)$ are missing. Introduction of the $V(\mathbf{s}_i, t)$ process is made to accommodate the point masses; a model without the V's, e.g., setting $P(\mathbf{s}_i, t) = \exp(U(\mathbf{s}_i, t))\text{iff}U(\mathbf{s}_i, t) > 0$ implies a discontinuity in $P(\mathbf{s}_i, t)$ at $U(\mathbf{s}_i, t) = 0$.

Let $Q(A_j, t)$ denote the CMAQ model output at grid cell A_j for week t, $j = 1, \ldots, J$. Similar to (3) we suppose that $Q(A_j, t)$ is positive if a conceptual areal level latent atmospheric process, denoted by $\tilde{V}(A_j, t)$, is positive,

$$Q(A_j, t) = \begin{cases} \exp\left(X(A_j, t)\right) & \text{if } \tilde{V}(A_j, t) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The values of $X(A_j, t)$ when $\tilde{V}(A_j, t) \leq 0$ will be given by the model described below. As computer model output, there are no missing values in the $Q(A_j, t)$.

Let $\mathbf{P}$, $\mathbf{Z}$ and $\mathbf{Q}$ denote all the precipitation values, wet deposition values and the CMAQ model output, respectively. Similarly define the vectors $\mathbf{U}$, $\mathbf{Y}$, and $\mathbf{X}$ collecting all the elements of the corresponding random variable for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Let $\mathbf{V}$ and $\tilde{\mathbf{V}}$ denote the vectors collecting the elements $V(\mathbf{s}_i, t), i = 1, \ldots, n$ and $\tilde{V}(A_j, t), j = 1, \ldots, J$ respectively, for $t = 1, \ldots, T$.

The first stage likelihood implied by the definitions (2), (3) and (4) is given by:

$$f(\mathbf{P}, \mathbf{Z}, \mathbf{Q}|\mathbf{U}, \mathbf{Y}, \mathbf{X}, \mathbf{V}, \tilde{\mathbf{V}}) = f(\mathbf{P}|\mathbf{U}, \mathbf{V}) \times f(\mathbf{Z}|\mathbf{Y}, \mathbf{V}) \times f(\mathbf{Q}|\mathbf{X}, \tilde{\mathbf{V}}) \tag{5}$$

which takes the form

$$\prod_{t=1}^{T}\left[\prod_{i=1}^{n}\left\{1_{\exp\left(u(\mathbf{s}_i,t)\right)}{}^{1_{\exp\left(y(\mathbf{s}_i,t)\right)}}{}^{I(v(\mathbf{s}_i,t)>0)}\right\}\prod_{j=1}^{J}\left\{1_{\exp\left(x(A_j,t)\right)}{}^{I(\tilde{v}(A_j,t)>0)}\right\}\right]$$

where 1_x denotes a degenerate distribution with point mass at x and $I(\cdot)$ is the indicator function.

3.2 Second Stage Specification

In the second stage of modeling we begin by specifying a spatially-colored regression model for log-precipitation based on the latent process $V(\mathbf{s}_i,t)$. In particular, we assume the model:

$$U(\mathbf{s}_i,t)=\alpha_0+\alpha_1 V(\mathbf{s}_i,t)+\delta(\mathbf{s}_i,t), \tag{6}$$

where $\boldsymbol{\delta}_t=(\delta(\mathbf{s}_1,t),\ldots,\delta(\mathbf{s}_n,t))'$ for $t=1,\ldots,T$ is an independent Gaussian process following the $N(\mathbf{0},\Sigma_\delta)$ distribution; Σ_δ has elements $\sigma_\delta(i,j)=\sigma_\delta^2\exp(-\phi_\delta d_{ij})$, the usual exponential covariance function, where d_{ij} is the geodetic distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$. Using vector notation, the above specification is equivalently written as:

$$\mathbf{U}_t\sim N\left(\alpha_0\mathbf{1}+\alpha_1\mathbf{V}_t,\Sigma_\delta\right)$$

where $\mathbf{U}_t=(U(\mathbf{s}_1,t),\ldots,U(\mathbf{s}_n,t))'$ and $\mathbf{V}_t=(V(\mathbf{s}_1,t),\ldots,V(\mathbf{s}_n,t))'$ and $\mathbf{1}$ denotes a vector with all elements unity (of appropriate order).

To model $Y(\mathbf{s}_i,t)$, we assume that:

$$Y(\mathbf{s}_i,t)=\beta_0+\beta_1 U(\mathbf{s}_i,t)+\beta_2 V(\mathbf{s}_i,t)+\left(b_0+b(\mathbf{s}_i)\right)X(A_{k_i},t)+\eta(\mathbf{s}_i,t)+\epsilon(\mathbf{s}_i,t), \tag{7}$$

for $i=1,\ldots,n$ and $t=1,\ldots,T$ where, unless otherwise mentioned, A_{k_i} is the grid cell which contains the site $\mathbf{s}_i$.

The error terms $\epsilon(\mathbf{s}_i,t)$ are assumed to follow $N(0,\sigma_\epsilon^2)$ independently, providing the so-called nugget effect. The reasoning for the rest of the specification in (7) is as follows. The term $\beta_1 U(\mathbf{s}_i,t)$ is included because of the strong linear relationships between log-deposition and log-precipitation, see Figure 4. The term $\beta_2 V(\mathbf{s}_i,t)$ captures any direct influence of the atmospheric process $V(\mathbf{s}_i,t)$ on $Y(\mathbf{s}_i,t)$ in the presence of precipitation.

The exploratory analyses presented earlier also provided evidence for possible linear relationships between log deposition and log CMAQ values. To specify a rich class of *locally* linear models we may think of a spatially varying slope for the regression of $Y(\mathbf{s}_i,t)$

on log-CMAQ values $X(A_j,t)$, specified as $(b_0 + b(\mathbf{s}_i))\,X(A_{k_i},t)$ in (7). Writing $\mathbf{b} = (b(\mathbf{s}_1),\ldots,b(\mathbf{s}_n))'$ we propose a mean 0 Gaussian process for $\mathbf{b}$, i.e.

$$\mathbf{b} \sim N(\mathbf{0}, \Sigma_b)$$

where Σ_b has elements $\sigma_b(i,j) = \sigma_b^2 \exp(-\phi_b d_{ij})$.

The term $\eta(\mathbf{s}_i,t)$ provides a spatially varying intercept which can also be interpreted as a spatio-temporal adjustment to the overall intercept parameter β_0. We assume that

$$\boldsymbol{\eta}_t \sim N(\mathbf{0}, \Sigma_\eta), \quad t = 1,\ldots,T$$

independently where $\boldsymbol{\eta}_t = (\eta(\mathbf{s}_1,t),\ldots,\eta(\mathbf{s}_n,t))'$ and Σ_η has elements $\sigma_\eta(i,j) = \sigma_\eta^2 \exp(-\phi_\eta d_{ij})$. We can consider replacing $\eta(\mathbf{s}_i,t)$ with $\eta(\mathbf{s}_i)$. The pure spatial term will fail to capture the between week variability in the intercept, see Carroll *et al.* (1997) for a related discussion. However, it does provide a common term for all weekly predictions yielding possibly appropriate increased uncertainty in long-term averaging, see Stein's discussion to Carroll *et al.* (1997).

The regression model (7) is now equivalently written as:

$$\mathbf{Y}_t \sim N(\boldsymbol{\vartheta}_t, \sigma_\epsilon^2 I_n)$$

where $\mathbf{Y}_t = (Y(\mathbf{s}_1,t),\ldots,Y(\mathbf{s}_n,t))$ and $\boldsymbol{\vartheta}_t = \beta_0\mathbf{1} + \beta_1\mathbf{U}_t + \beta_2\mathbf{V}_t + b_0\mathbf{X}_t + X_t\mathbf{b} + \boldsymbol{\eta}_t$ where $\mathbf{X}_t$ is the n-dimensional vector with the ith element given by $X(A_{k_i},t)$ and X_t is a diagonal matrix whose ith diagonal entry is $X(A_{k_i},t)$, $i = 1,\ldots,n$ and I_n is the identity matrix of order n.

The CMAQ output $X(A_j,t)$ is modeled using the latent process $\tilde{V}(A_j,t)$ as follows:

$$X(A_j,t) = \gamma_0 + \gamma_1\tilde{V}(A_j,t) + \psi(A_j,t), \quad j = 1,\ldots,J. \tag{8}$$

where $\psi(A_j,t) \sim N(0,\sigma_\psi^2)$ independently for all $j = 1,\ldots,J$, $t = 1,\ldots,T$ and σ_ψ^2 is unknown. In vector notation, this is given by:

$$\mathbf{X}_t \sim N(\gamma_0\mathbf{1} + \gamma_1\tilde{\mathbf{V}}_t,\, \sigma_\psi^2 I_J)$$

where $\mathbf{X}_t = (X(A_1,t),\ldots X(A_J,t))'$ and $\tilde{\mathbf{V}}_t = (\tilde{V}(A_1,t),\ldots \tilde{V}(A_J,t))$, see the partitioning of $\tilde{\mathbf{V}}_t$ below Equation (9) regarding the order of the grid cell indices $1,\ldots,J$.

We now turn to specification of the latent processes $V(\mathbf{s}_i,t)$ and $\tilde{V}(A_j,t)$. Note that it is possible to have $Z(\mathbf{s}_i,t) > 0$ and $Q(A_{k_i},t) = 0$ and vice versa since $Q(A_{k_i},t)$ is the

output of a computer model which has not used the actual observation $Z(\mathbf{s}_i, t)$. This implies that $V(\mathbf{s}_i, t)$ and $\tilde{V}(A_{k_i}, t)$ can be of different signs. To accommodate this flexibility and to distinguish between the point and areal processes we assume the simple measurement error model:

$$V(\mathbf{s}_i, t) \sim N(\tilde{V}(A_{k_i}, t), \sigma_v^2), \tag{9}$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where σ_v^2 is unknown. Without loss of generality we write $\tilde{\mathbf{V}}_t = (\tilde{\mathbf{V}}_t^{(1)}, \tilde{\mathbf{V}}_t^{(2)})$ where the n-dimensional vector $\tilde{\mathbf{V}}_t^{(1)}$ contains the values for the grid cells where the n observation sites are located and $\tilde{\mathbf{V}}_t^{(2)}$ contains the values for the remaining $J - n$ grid cells. The specification (9) can now be written equivalently as

$$\mathbf{V}_t \sim N(\tilde{\mathbf{V}}_t^{(1)}, \sigma_v^2 I_n), \ t = 1, \ldots, T.$$

The latent process $\tilde{V}(A_j, t)$ is assumed to follow a first order auto-regressive process in time and a conditionally auto-regressive (CAR) process in space. That is,

$$\tilde{V}(A_j, t) = \rho \tilde{V}(A_j, t-1) + \zeta(A_j, t) \tag{10}$$

for $j = 1, \ldots, J$ and $t = 1, \ldots, T$. The $\zeta(A_j, t)$ are independent improper conditional autoregressive models (see e.g. Banerjee *et al.*, 2004) over t. That is,

$$\zeta(A_j, t) \sim N\left(\sum_{i=1}^{J} h_{ji}\zeta(A_i, t), \frac{\sigma_\zeta^2}{m_j}\right) \tag{11}$$

where

$$h_{ji} = \begin{cases} \frac{1}{m_j} & \text{if } i \in \partial_j \\ 0 & \text{otherwise} \end{cases}$$

and ∂_j defines the m_j neighboring grid cells of the cell A_j.

We initiate the process in (10) with $\tilde{V}(A_j, 0) = \frac{1}{T}\sum_{t=1}^{T} X(A_j, t)$, the mean of the observed $X(A_j, t)$ values. Now we have the temporally vectorized auto-regressive and spatially CAR specification:

$$f(\tilde{\mathbf{V}}_t | \tilde{\mathbf{V}}_{t-1}, \rho, \sigma_\zeta^2) \propto \exp\left\{-\frac{1}{2}\left(\tilde{\mathbf{V}}_t - \rho\tilde{\mathbf{V}}_{t-1}\right)' D^{-1}(I - H)\left(\tilde{\mathbf{V}}_t - \rho\tilde{\mathbf{V}}_{t-1}\right)\right\} \tag{12}$$

where D is diagonal with the jth diagonal entry given by σ_ζ^2/m_j. In summary, the second stage specification is given by:

$$\begin{aligned}\prod_{t=1}^{T} \big[& f(\mathbf{Y}_t | \mathbf{U}_t, \mathbf{V}_t, \mathbf{X}_t, \boldsymbol{\eta}_t, \mathbf{b}, \boldsymbol{\theta}) \times f(\boldsymbol{\eta}_t | \boldsymbol{\theta}) \\ & \times f(\mathbf{U}_t | \mathbf{V}_t, \boldsymbol{\theta}) \times f(\mathbf{V}_t | \tilde{\mathbf{V}}_t^{(1)}, \boldsymbol{\theta}) \times f(\mathbf{X}_t | \tilde{\mathbf{V}}_t, \boldsymbol{\theta}) \times f(\tilde{\mathbf{V}}_t | \tilde{\mathbf{V}}_{t-1}, \boldsymbol{\theta})\big] \, f(\mathbf{b} | \boldsymbol{\theta})\end{aligned}$$

where $\boldsymbol{\theta}$ denote the parameters α_0, α_1, β_0, β_1, β_2, b_0, γ_0, γ_1, ρ, σ^2_δ, σ^2_b, σ^2_η, σ^2_ϵ, σ^2_ψ, σ^2_v and σ^2_ζ. Figure 6 provides a directed graphical model for our entire specification, noting the measurement error specification.

As noted in the introduction, hierarchical modeling for fusion between monitoring data and model output data has been proposed in Fuentes and Raftery (2005). Their approach introduces a model for latent true deposition $Z^{(\text{true})}(\mathbf{s}_i, t)$. In our setting the true deposition would be driven by a regional atmospheric process as in Section 3.1. To connect the process to the grid cell model output data, block averaging is required. Figure 7 shows the analogue of our modeling using this approach. The key point is the direction arrow linking the $V(\mathbf{s}_i, t)$ and the $\tilde{V}(A_{k_i}, t)$, a MEM versus block averaging. Hence, the infeasibility of fitting the model in Figure 7 in the case of a large number of grid cells emerges. For the CMAQ output we use, 33,390 block averages are required, and, in fact, these are required for each $t = 1, ..., 52$ weeks. (We note further that the fusion model of Fuentes and Raftery, 2005, has not actually been implemented in a dynamic setting.) The advantage of the model in Figure 6 is clear. We only have to fit the measurement error model to the 120 monitoring sites while doing cheap CAR updates for the $\tilde{V}$'s. Generally, our approach will be preferred for environmental data settings since there will always be many more grid cells than monitoring stations, and this will be further exacerbated by computer models seeking higher spatial resolution.

We attempt further clarification of the V and $\tilde{V}$ processes as well as justification for the measurement model in (9). Again, our specification does not view $\tilde{V}(A_{k_i}, t)$ as a block average of $V(\mathbf{s}_i, t)$ over A_{k_i}. Rather, it views $V(\mathbf{s}_i, t) - \tilde{V}(A_{k_i}, t)$ as a deviation from the areal average and we assume that these are independent across the $\mathbf{s}_i$ where V and $\tilde{V}$ are two distinct mean 0 spatial processes operating at different spatial scales. Careful algebraic calculation, using the models in (6)-(10), shows that $(U(\mathbf{s}_i, t), Y(\mathbf{s}_i, t))$ given $V(\mathbf{s}_i, t)$ and $X(A_{k_i}, t)$ is a bivariate space-time Gaussian process which is captured through point level space-time random effects, $\delta(\mathbf{s}_i, t)$ and $\eta(\mathbf{s}_i, t)$. But, under the models for $V(\mathbf{s}_i, t)|\tilde{V}(A_{k_i}, t)$ and $X(A_{k_i}, t)|\tilde{V}(A_{k_i}, t)$, we can marginalize over V and X to obtain a marginal bivariate Gaussian process, $(U(\mathbf{s}_i, t), Y(\mathbf{s}_i, t))$ given $\tilde{V}$. In other words, the $\tilde{V}(A_{k_i}, t)$ introduce spatial random effects at the areal unit scale. So, the overall specification is a multi-scale space-time process with uncorrelated effects introduced in an additive manner. Such specifications have a long history in geostatistics (see, e.g., Goulard and Voltz (1992), and Gotway and

Young (2002). Adopting such specification is a familiar device for avoiding block averaging.

3.3 Prior and Posterior Distributions

We now complete the Bayesian model specification by assuming prior distributions for all the unknown parameters. We assume that, a priori, each of α_0, α_1, β_0, β_1, β_2, b_0, γ_0, γ_1, ρ is normally distributed with mean 0 and variance 10^4, essentially a flat prior specification. The inverse of the variance components $\frac{1}{\sigma_\delta^2}$, $\frac{1}{\sigma_b^2}$, $\frac{1}{\sigma_\eta^2}$, $\frac{1}{\sigma_\epsilon^2}$, $\frac{1}{\sigma_\psi^2}$, $\frac{1}{\sigma_v^2}$, and $\frac{1}{\sigma_\zeta^2}$, are all assumed to follow the gamma distribution $G(\nu, \lambda)$ having mean ν/λ. In our implementation we take $\nu = 2$ and $\lambda = 1$ implying that these variance components have prior mean 1 and infinite variance.

Ideally, $\boldsymbol{\phi} = (\phi_\delta, \phi_b, \phi_\eta)$ should be estimated within the Bayesian model as well. However, in a classical inference setting it is not possible to consistently estimate both ϕ and σ^2 in a typical model for spatial data with a covariance function belonging to the Matèrn family, see Zhang (2004). Moreover, Stein (1999) shows that spatial interpolation is sensitive to the product $\sigma^2\phi$ but not to either one individually. In our Bayesian inference setup using Gibbs sampling, joint estimation is often poorly behaved due to weak identifiability and extremely slow mixing of the associated Markov chains under vague prior distributions for $\boldsymbol{\phi}$. In addition, the full conditional distribution for each of the decay parameters is not conjugate so sampling them in a Gibbs sampler requires expensive likelihood evaluations in each iteration. These difficulties are exacerbated by the large volume of data we model. Instead, in Section 5 we shall choose *optimal* values of $\boldsymbol{\phi}$ using a validation mean square error criterion and fit the rest of the model conditional on those values.

Finally, the log of the likelihood times prior in the second stage up to an additive constant is given by:

$$
\begin{aligned}
&-\tfrac{nT}{2}\log(\sigma_\epsilon^2) - \tfrac{1}{2\sigma_\epsilon^2}\textstyle\sum_{t=1}^{T}(\mathbf{y}_t - \boldsymbol{\vartheta}_t)'(\mathbf{y}_t - \boldsymbol{\vartheta}_t) - \tfrac{nT}{2}\log(\sigma_\eta^2) - \tfrac{1}{2\sigma_\eta^2}\textstyle\sum_{t=1}^{T}\boldsymbol{\eta}_t' S_\eta^{-1}\boldsymbol{\eta}_t \\
&-\tfrac{nT}{2}\log(\sigma_\delta^2) - \tfrac{1}{2\sigma_\delta^2}\textstyle\sum_{t=1}^{T}(\mathbf{u}_t - \alpha_0\mathbf{1} - \alpha_1\mathbf{v}_t)' S_\delta^{-1}(\mathbf{u}_t - \alpha_0\mathbf{1} - \alpha_1\mathbf{v}_t) \\
&-\tfrac{nT}{2}\log(\sigma_v^2) - \tfrac{1}{2\sigma_v^2}\textstyle\sum_{t=1}^{T}(\mathbf{v}_t - \tilde{\mathbf{v}}_t^{(1)})'(\mathbf{v}_t - \tilde{\mathbf{v}}_t^{(1)})' \\
&-\tfrac{JT}{2}\log(\sigma_\psi^2) - \tfrac{1}{2\sigma_\psi^2}\textstyle\sum_{t=1}^{T}(\mathbf{x}_t - \gamma_0\mathbf{1} - \gamma_1\tilde{\mathbf{v}}_t)'(\mathbf{x}_t - \gamma_0\mathbf{1} - \gamma_1\tilde{\mathbf{v}}_t) \\
&-\tfrac{JT}{2}\log(\sigma_\zeta^2) - \tfrac{1}{2}\textstyle\sum_{t=1}^{T}(\tilde{\mathbf{v}}_t - \rho\tilde{\mathbf{v}}_{t-1})' D^{-1}(I - H)(\tilde{\mathbf{v}}_t - \rho\tilde{\mathbf{v}}_{t-1}) \\
&-\tfrac{n}{2}\log(\sigma_b^2) - \tfrac{1}{2\sigma_b^2}\mathbf{b}' S_b^{-1}\mathbf{b} + \log(f(\boldsymbol{\theta}))
\end{aligned}
$$

where $f(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$ and $\Sigma_\delta = \sigma_\delta^2 S_\delta$, $\Sigma_b = \sigma_b^2 S_b$, $\Sigma_\eta = \sigma_\eta^2 S_\eta$.

4 Predicting Deposition at a New Location

The models developed in Section 3 allow us to interpolate the spatial deposition surface at any given week in the year. Consider the problem of predicting $Z(\mathbf{s}', t')$ in week t' at any new location $\mathbf{s}'$ falling on the grid cell A'. The prediction is performed by constructing the posterior predictive distribution of $Z(\mathbf{s}', t')$ which in turn depends on the distribution of $Y(\mathbf{s}', t')$ as specified by Equation (7) along with the associated $V(\mathbf{s}', t')$. We estimate the posterior predictive distribution by drawing samples from it.

Several cases arise depending on the nature of information available at the new site $\mathbf{s}'$ at week t'. If precipitation information is available and there is no positive precipitation, i.e. $p(\mathbf{s}', t') = 0$, then we have $Z(\mathbf{s}', t') = 0$ and no further sampling is needed, since there can be no deposition without precipitation. Now suppose that there is positive precipitation, i.e. $p(\mathbf{s}', t') > 0$, then set $u(\mathbf{s}', t') = \log(p(\mathbf{s}', t'))$. We need to generate a sample $Y(\mathbf{s}', t')$. We first generate $V(\mathbf{s}', t') \sim N(\tilde{V}(A', t'), \sigma_v^2)$ following the measurement error model (9). Note that $\tilde{V}(A', t')$ is already available for any grid cell A' (within the study region) and week t' (in the current year) from model fitting, see Equation (10). Similarly, $X(A', t')$ is also available either as the log of the CMAQ output, $\log(Q(A', t'))$, if $Q(A', t') > 0$ or from the MCMC imputation when $Q(A', t') = 0$, see the Appendix. To sample $\eta(\mathbf{s}', t')$ we note that:

$$\begin{pmatrix} \eta(\mathbf{s}', t') \\ \boldsymbol{\eta}_{t'} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \sigma_\eta^2 \begin{pmatrix} 1 & S_{\eta,12} \\ S_{\eta,21} & S_\eta \end{pmatrix} \right],$$

where $S_{\eta,12}$ is $1 \times n$ with the ith entry given by $\exp(-\phi_\eta d(\mathbf{s}_i, \mathbf{s}'))$ and $S_{\eta,21} = S'_{\eta,12}$. Therefore,

$$\eta(\mathbf{s}', t') | \boldsymbol{\eta}_{t'}, \boldsymbol{\theta} \sim N\left[S_{\eta,12} S_\eta^{-1} \boldsymbol{\eta}_{t'},\ \sigma_\eta^2 \left(1 - S_{\eta,12} S_\eta^{-1} S_{\eta,21}\right) \right]. \tag{13}$$

If the term $b(\mathbf{s})$ is included in the model we need to simulate $b(\mathbf{s}')$ conditional on $\mathbf{b}$ and model parameters. To do this we note that:

$$\begin{pmatrix} b(\mathbf{s}') \\ \mathbf{b} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \sigma_b^2 \begin{pmatrix} 1 & S_{b,12} \\ S_{b,21} & S_b \end{pmatrix} \right],$$

where $S_{b,12}$ is $1 \times n$ with the ith entry given by $\exp(-\phi_\eta d(\mathbf{s}_i, \mathbf{s}'))$ and $S_{b,21} = S'_{b,12}$. Therefore,

$$b(\mathbf{s}') | \boldsymbol{\theta} \sim N\left[S_{b,12} S_b^{-1} \mathbf{b},\ \sigma_b^2 \left(1 - S_{b,12} S_b^{-1} S_{b,21}\right) \right]. \tag{14}$$

If it is desired to predict $Z(\mathbf{s}', t')$ where $P(\mathbf{s}', t')$ is not available, we proceed as follows. We generate $V(\mathbf{s}', t') \sim N(\tilde{V}(A', t'), \sigma_v^2)$ following the measurement error model (9). If this $V(\mathbf{s}', t') < 0$, then we set both $p(\mathbf{s}', t')$ and $Z(\mathbf{s}', t')$ to zero. If, however, $V(\mathbf{s}', t') > 0$ we need to additionally draw $U(\mathbf{s}', t')$ using the precipitation model (6). For this we note that,

$$\begin{pmatrix} U(\mathbf{s}', t') \\ \mathbf{U}_{t'} \end{pmatrix} \sim N \left[\begin{pmatrix} \alpha_0 + \alpha_1 V(\mathbf{s}', t') \\ \alpha_0 \mathbf{1} + \alpha_1 \mathbf{V}_{t'} \end{pmatrix}, \sigma_\delta^2 \begin{pmatrix} 1 & S_{\delta,12} \\ S_{\delta,21} & S_\delta \end{pmatrix} \right],$$

where $S_{\delta,12}$ is $1 \times n$ with the ith entry given by $\exp(-\phi_\delta d(\mathbf{s}_i, \mathbf{s}'))$ and $S_{\delta,21} = S'_{\delta,12}$. Therefore,

$$U(\mathbf{s}', t') | \mathbf{U}_{t'}, \boldsymbol{\theta} \sim N \left[\mu(\mathbf{s}', t'),\ \sigma_\delta^2 \left(1 - S_{\delta,12} S_\delta^{-1} S_{\delta,21}\right) \right], \tag{15}$$

where

$$\mu(\mathbf{s}', t') = \alpha_0 + \alpha_1 V(\mathbf{s}', t') + S_{\delta,12} S_\delta^{-1} (\mathbf{U}_{t'} - \alpha_0 \mathbf{1} - \alpha_1 \mathbf{v}_{t'}).$$

If $Z(\mathbf{s}', t')$ is not inferred to be zero then we set it to be $\exp(Y(\mathbf{s}', t'))$. If we want the predictions of the smooth deposition surface without the nugget term we simply ignore the nugget term $\epsilon(\mathbf{s}', t')$ in generating $Y(\mathbf{s}', t')$. Annual and quarterly predictions at a location $\mathbf{s}'$ are obtained by forming sums of $Z(\mathbf{s}', t')$ appropriately, e.g. the annual deposition is $g(\mathbf{s}') = \sum_{t'=1}^{T} Z(\mathbf{s}', t')$. Thus at each MCMC iteration j we have $Z^{(j)}(\mathbf{s}', t')$ and $g^{(j)}(\mathbf{s}')$. We use the median of the accumulated MCMC samples and the lengths of the 95% intervals to summarize the predictions. The median as a summary measure preserves the one-to-one relationships between summaries for Y and Z. Exploratory data analyses of the MCMC output showed rapid convergence for the adopted models. For making inference, we used 10,000 MCMC iterations after discarding the first 5,000 iterations.

5 Analysis

5.1 Model Checking

As noted in Section 3.3, under weak prior distributions it is not possible to estimate all the parameters in the covariance structure consistently, see e.g. Zhang (2004), Sahu *et al.* (2006, 2007) and the references therein. Hence, we use the set-aside validation data from 8 stations to select the three decay parameters ϕ_η and ϕ_δ and ϕ_b. The variance components are estimated using MCMC. Let $\hat{Z}(\mathbf{s}_i^*, t)$ denote the model based validation estimate for $Z(\mathbf{s}_i^*, t)$ where $\mathbf{s}_i^*$ denote the ith validation site. The validation mean-square error is given

by

$$\text{VMSE} = \frac{1}{n_v} \sum_{i=1}^{8} \sum_{t=1}^{T} \left(Z(\mathbf{s}_i^*, t) - \hat{Z}(\mathbf{s}_i^*, t) \right)^2 I(Z(\mathbf{s}_i^*, t))$$

where $I(Z(\mathbf{s}_i^*, t)) = 1$ if $Z(\mathbf{s}_i^*, t)$ has been observed and 0 otherwise, and n_v is the total number of available observations at the 8 validation sites. For our data set $n_v = 407$ since there were 9 missing observations, see Section 2. We searched for the optimal values of ϕ_η, ϕ_δ and ϕ_b in a three dimensional grid formed of the values 0.002, 0.003, 0.006, 0.012 and 0.06 corresponding to spatial ranges of 1500, 1000, 500, 250 and 50 kilometers, separately for the sulfate and nitrate deposition models. (The data can not be expected to inform about the range to a finer resolution.) The combination of values $\phi_\eta = 0.006$, $\phi_\delta = 0.003$ and $\phi_b = 0.006$ provided the best VMSE values both for the sulfate and nitrate model. The corresponding optimal ranges are 500, 1000 and 500 kilometers respectively. The VMSE is not at all sensitive to the choice of the decay parameters near these best values. As a result, although it is possible to further refine the grid in a neighborhood of the best value we do not explore beyond our grid here. In fact, this insensitivity is also supported by our investigation of the empirical variograms discussed in Section 2.

We compared several possible models using the predictive Bayesian model selection criterion of Gelfand and Ghosh (1998). The additional term $b(\mathbf{s}_i)X(A_{k_i}, t)$ did not improve model fitting a great deal. Only a few $b(\mathbf{s}_i)$ were significant, see Figure 8 where the estimated $b(\mathbf{s})$ surfaces along with their standard deviation surfaces have been plotted. The figure show that the $b(\mathbf{s})$ values are very small in absolute value relative to the standard deviations. Moreover, the Gelfand and Ghosh criterion was much smaller for the model without the $b(\mathbf{s}_i)X(A_{k_i}, t)$ term. Henceforth, we worked with the sub-model corresponding to $b(\mathbf{s}) = 0$. This is also explained by the fact that after accounting for the very large influence of precipitation and a spatio-temporally varying intercept term, the model is not able to detect a significant spatially varying contribution of the CMAQ output towards explaining deposition. This, however, *does not* mean that there is no spatio-temporal bias in the CMAQ output – such biases can simply be recovered by the differences between the model based predictions and the CMAQ output. If the intention is to recover the bias using a parametric form then a model omitting the most significant regressor, viz., precipitation must be specified.

We have examined the classic residuals versus fitted values plots both for sulfate and

nitrate. The plots (not included) illustrate only a few outlying values, the models fitted well otherwise. Figure 9 provides predictions at the validation sites versus the observed values along with the validation prediction intervals on the original scale for all the 407 available observations in the eight validation sites. Note that more than one observation can assume the same value due to truncation. The overall VMSE is 0.035 for sulfate and 0.015 for nitrate, and 95% and 96% of the nominal 95% validation prediction intervals contain the true sulfate and nitrate depositions respectively. Overall, the validation analysis indicates that the model does not appear to introduce any bias in prediction and performs very well for out of sample predictions.

5.2 Results and Interpretation

Table 1 provides the parameter estimates. There is very strong effect of precipitation since the parameter β_1 is significant both for sulfate and nitrate. Note that β_2 is not significant (for both models), likely attributable to the fact that the regional atmospheric driver process influences deposition directly through precipitation. Although small, the significant estimates of b_0 indicate that the point level data and the gridded CMAQ output are strongly correlated corroborating the exploratory analysis in Figure 5. As expected the parameters α_1 and γ_1 are significant showing that the point level atmospheric process is strongly related to precipitation and the areal level atmospheric process is a very good predictor of CMAQ output. There is strong temporal dependence between the CMAQ output in successive weeks (estimate of $\rho = 0.7688$ and 0.7492 for sulfate and nitrate, respectively with standard deviation 0.0012 and 0.0013). The estimates of the variance components show that the magnitude of the nugget effect σ^2_ϵ is the smallest. Hence more variation is explained by the spatio-temporal intercept process $\eta(\mathbf{s}, t)$ than the pure error process $\epsilon(\mathbf{s}, t)$.

Maps of annual depositions are provided in the top panel of Figures 10 and 11. For sulfate deposition the validation mean square error for the annual totals from the 8 reserved sites for the IDW method is 20.4 while the same for our model is 8.1. The corresponding statistics for the nitrate deposition are 3.5 and 1.3. These show better performance by our model both for the sulfate and nitrate depositions. The highest wet sulfate deposition occurs near major emissions sources such as fossil fueled power plants (concentrated in the Ohio River Valley) and mobile sources in major populations centers. Lower values occur

near background monitoring sites. These 2001 patterns are similar to those reported by Brook *et al.* (1995) for eastern North America.

The lengths of the prediction intervals are provided as maps in panel (d) of Figures 10 and 11. As expected, the lengths are smaller for the predictive sites near the modeling sites and also for sites in the regions of low depositions.

The quarterly prediction maps are provided in Figures 12 and 14. Increased depositions are seen during the spring and summer months April to September analogous to the summary Figure 3. The lengths of the predictions intervals, plotted in Figures 13 and 15 show that the uncertainties in quarterly maps are reasonably consistent over the seasons.

6 Discussion

The paper has developed a data fusion approach using a measurement error specification to combine gridded CMAQ output and point level monitoring data. Model components have been linked using latent processes in a Bayesian hierarchical framework. We use this approach to investigate space-time wet deposition patterns over the eastern U.S. Compared to the current practice of predicting wet deposition from the monitoring data alone using IDW, a significant reduction in MSE, calculated over a set of validation sites, has been achieved. Inclusion of the significant covariate precipitation improves the predictive capability of our model, and these predictions can be expected to be better than the predictions based on the IDW method since that ignores the significant covariate.

The model was initially developed for sulfate deposition, but its success led us to consider nitrate deposition as well. The performance of the model for both sulfate and nitrate deposition encourages its application to other constituents of wet deposition.

It is also of interest to estimate dry deposition which is defined as the exchange of gases, aerosols, and particles between the atmosphere and earth's surface. Future analyses will focus on predicting total (wet plus dry) sulfur and nitrogen deposition. Using the total predictive surface it will be possible to estimate deposition 'loadings' as the integrated volume of total deposition over ecological regions of interest. For this, a new model for dry deposition has to be developed. If successful, this effort will lead to the first ever estimation of total deposition loadings, perhaps the most critical quantity for ecological assessments. Future work will also address trends in deposition to assess whether regulation has been

successful.

Appendix: Distributions for Gibbs sampling

Handling of the missing values:

Note that the transformation Equation (3) does not define a unique value of $Y(\mathbf{s}_i, t)$ and in addition, there will be missing values corresponding to the missing values in $Z(\mathbf{s}_i, t)$. Any missing value of $Y^*(\mathbf{s}_i, t)$ is sampled from the model value $N(\vartheta(\mathbf{s}_i, t), \sigma_\epsilon^2)$ for $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

The sampling of the missing $U^*(\mathbf{s}_i, t)$ for the precipitation process is a bit more involved. The sampling of the missing values must be done using the model (6) conditional on all the parameters. Since this model is a spatial model we must use the conditional distribution of $U^*(\mathbf{s}_i, t)$ given all the $U(\mathbf{s}_j, t)$ values for $j = 1, \ldots, n$ and $j \neq i$. This conditional distribution is obtained using the covariance matrix Σ_δ of $\boldsymbol{\delta}_t$ and is omitted for brevity.

Similarly, Equation (4) does not define unique values of $X(A_j, t)$ when $Q(A_j, t) = 0$. Those values, denoted by $X^*(A_j, t)$, are sampled using the model Equation (8), $X^*(A_j, t)$ is sampled from $N(\gamma_0 + \gamma_1 \tilde{v}(A_j, t), \sigma_\psi^2)$.

Conditional posterior distributions of $\boldsymbol{\theta}$

Straightforward calculation yields the following complete conditional distributions:

$$
\begin{aligned}
&\tfrac{1}{\sigma_\epsilon^2} \sim G\left(\tfrac{nT}{2} + \nu, \lambda + \tfrac{1}{2}\textstyle\sum_{t=1}^{T}(\mathbf{y}_t - \boldsymbol{\vartheta}_t)'(\mathbf{y}_t - \boldsymbol{\vartheta}_t)\right),\\
&\tfrac{1}{\sigma_b^2} \sim G\left(\tfrac{n}{2} + \nu, \lambda + \tfrac{1}{2}\mathbf{b}' S_b^{-1}\mathbf{b}\right),\\
&\tfrac{1}{\sigma_\eta^2} \sim G\left(\tfrac{nT}{2} + \nu, \lambda + \tfrac{1}{2}\textstyle\sum_{t=1}^{T}\boldsymbol{\eta}_t' S_\eta^{-1}\boldsymbol{\eta}_t\right)\\
&\tfrac{1}{\sigma_\delta^2} \sim G\left(\tfrac{nT}{2} + \nu, \lambda + \tfrac{1}{2}\textstyle\sum_{t=1}^{T}(\mathbf{u}_t - \alpha_0\mathbf{1} - \alpha_1\mathbf{v}_t)' S_\delta^{-1}(\mathbf{u}_t - \alpha_0\mathbf{1} - \alpha_1\mathbf{v}_t)\right),\\
&\tfrac{1}{\sigma_\psi^2} \sim G\left(\tfrac{JT}{2} + \nu, \lambda + \tfrac{1}{2}\textstyle\sum_{t=1}^{T}(\mathbf{x}_t - \gamma_0\mathbf{1} - \gamma_1\tilde{\mathbf{v}}_t)'(\mathbf{x}_t - \gamma_0\mathbf{1} - \gamma_1\tilde{\mathbf{v}}_t)\right)\\
&\tfrac{1}{\sigma_v^2} \sim G\left(\tfrac{nT}{2} + \nu, \lambda + \tfrac{1}{2}\textstyle\sum_{t=1}^{T}(\mathbf{v}_t - \tilde{\mathbf{v}}_t^{(1)})'(\mathbf{v}_t - \tilde{\mathbf{v}}_t^{(1)})\right),\\
&\tfrac{1}{\sigma_\zeta^2} \sim G\left(\tfrac{JT}{2} + \nu, \lambda + \tfrac{1}{2}\textstyle\sum_{t=1}^{T}\sum_{j=1}^{J}\left\{m_j(\zeta(A_j, t) - \bar{\zeta}(A_j, t))^2\right\}\right).
\end{aligned}
$$

where $\bar{\zeta}(A_j, t) = \sum_{i=1}^{J} h_{ji}\zeta(A_i, t)$.

Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ and $G_t = (\mathbf{1}, \mathbf{u}_t, \mathbf{v}_t)$ so that G_t is an $n \times 3$ matrix. The full conditional distribution of $\boldsymbol{\beta}$ is $N(\Lambda\boldsymbol{\chi}, \Lambda)$ where

$$
\Lambda^{-1} = \frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{T} G_t' G_t + 10^{-3} I_3, \quad \boldsymbol{\chi} = \frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{T} G_t'(\mathbf{y}_t - b_0\mathbf{x}_t + X_t\mathbf{b} + \boldsymbol{\eta}_t).
$$

The full conditional distribution of b_0 is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1} = \frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{T}\mathbf{x}_t'\mathbf{x}_t + 10^{-3},\quad \chi = \frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{T}\mathbf{x}_t'(\mathbf{y}_t - \beta_0\mathbf{1} - \beta_1\mathbf{u}_t - \beta_2\mathbf{v}_t - X_t\mathbf{b} - \boldsymbol{\eta}_t).$$

The full conditional distribution of $\mathbf{b}$ is $N(\Lambda\boldsymbol{\chi},\ \Lambda)$ where

$$\Lambda^{-1} = \frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{T}X_t'X_t + \Sigma_b^{-1},\quad \boldsymbol{\chi} = \frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{T}X_t'(\mathbf{y}_t - \beta_0\mathbf{1} - \beta_1\mathbf{u}_t - \beta_2\mathbf{v}_t - b_0\mathbf{x}_t - \boldsymbol{\eta}_t).$$

The full conditional distribution of $\boldsymbol{\eta}_t$ for $t = 1, \ldots, T$ is $N(\Lambda_t\boldsymbol{\chi}_t,\ \Lambda_t)$ where

$$\Lambda_t^{-1} = \frac{I_n}{\sigma_\epsilon^2} + \Sigma_\eta^{-1},\quad \boldsymbol{\chi}_t = \frac{1}{\sigma_\epsilon^2}(\mathbf{y}_t - \beta_0\mathbf{1} - \beta_1\mathbf{u}_t - \beta_2\mathbf{v}_t - b_0\mathbf{x}_t - X_t\mathbf{b}).$$

Let $G_t = (\mathbf{1}, \mathbf{v}_t)$ so that G_t is an $n \times 2$ matrix. The full conditional distribution of $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ is $N(\Lambda\boldsymbol{\chi},\ \Lambda)$ where

$$\Lambda^{-1} = \sum_{t=1}^{T}G_t'\Sigma_\delta^{-1}G_t + 10^{-3}I_2,\quad \boldsymbol{\chi} = \sum_{t=1}^{T}G_t'\Sigma_\delta^{-1}\mathbf{u}_t.$$

Let $G_t = (\mathbf{1}, \tilde{\mathbf{v}}_t)$ so that G_t is an $J \times 2$ matrix. The full conditional distribution of $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ is $N(\Lambda\boldsymbol{\chi},\ \Lambda)$ where

$$\Lambda^{-1} = \frac{1}{\sigma_\psi^2}\sum_{t=1}^{T}G_t'G_t + 10^{-3}I_2,\quad \boldsymbol{\chi} = \sum_{t=1}^{T}G_t'\mathbf{x}_t.$$

The full conditional distribution of ρ is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1} = \frac{1}{\sigma_\zeta^2}\sum_{t=1}^{T}\sum_{j=1}^{J}m_j e_{j,t-1}^2 + 10^{-3},\quad \chi = \frac{1}{\sigma_\zeta^2}\sum_{t=1}^{T}\sum_{j=1}^{J}m_j e_{jt}e_{j,t-1}$$

where $e_{jt} = \tilde{v}(A_j, t) - \bar{v}(A_j, t)$ and $\bar{v}(A_j, t) = \sum_{i=1}^{J}h_{ji}\tilde{v}(A_i, t)$.

Conditional posterior distributions of $\mathbf{V}_t$

Note that due to the missing and zero precipitation values the full conditional distribution of $\mathbf{V}_t$ will be in a restricted space. First, the unrestricted full conditional distribution of $\mathbf{v}_t$ is $N(\Lambda_t\boldsymbol{\chi}_t,\ \Lambda_t)$ where

$$\Lambda_t^{-1} = \beta_2^2\frac{I_n}{\sigma_\epsilon^2} + \alpha_1^2\Sigma_\delta^{-1} + \frac{I_n}{\sigma_v^2},\quad \text{and}\quad \boldsymbol{\chi}_t = \frac{\beta_2}{\sigma_\epsilon^2}\mathbf{a}_t + \alpha_1\Sigma_\delta^{-1}(\mathbf{u}_t - \alpha_0\mathbf{1}) + \frac{1}{\sigma_v^2}\tilde{\mathbf{v}}_t^{(1)},$$

where $\mathbf{a}_t = \mathbf{y}_t - \beta_0\mathbf{1} - \beta_1\mathbf{u}_t - b_0\mathbf{x}_t - X_t\mathbf{b} - \boldsymbol{\eta}_t$. From this n-dimensional joint distribution we obtain the conditional distribution $V(\mathbf{s}_i, t) \sim N(\mu_{it}, \Xi_{it})$, say. If the precipitation value, $p(\mathbf{s}_i, t)$, is missing then there will be no constraint on $V(\mathbf{s}_i, t)$ and we sample $V(\mathbf{s}_i, t)$ unrestricted from $N(\mu_{it}, \Xi_{it})$. If on the other hand the observed precipitation

value is zero, $p(\mathbf{s}_i, t) = 0$, we must sample $V(\mathbf{s}_i, t)$ to be negative, i.e we sample from $N(\mu_{it}, \Xi_{it})I(V(\mathbf{s}_i, t) < 0)$. Corresponding to non-zero precipitation value $p(\mathbf{s}_i, t) > 0$ we sample $V(\mathbf{s}_i, t)$ from $N(\mu_{it}, \Xi_{it})I(V(\mathbf{s}_i, t) > 0)$.

Conditional posterior distributions of $\tilde{\mathbf{V}}_t$

The full conditional distribution of $\tilde{\mathbf{V}}_t = (\tilde{\mathbf{V}}_t^{(1)}, \tilde{\mathbf{V}}_t^{(2)})$ for any t is $N(\Lambda_t \boldsymbol{\chi}_t, \Lambda_t)$ where

$$\Lambda_t^{-1} = \begin{pmatrix} \frac{I_n}{\sigma_v^2} & 0 \\ 0 & 0 \end{pmatrix} + \gamma_1^2 \frac{I_J}{\sigma_\psi^2} + (1 + I(t < T)\rho^2) D^{-1}(I - H),$$

$$\boldsymbol{\chi}_t = \begin{pmatrix} \frac{1}{\sigma_v^2}\mathbf{v}_t \\ 0 \end{pmatrix} + \frac{\gamma_1}{\sigma_\psi^2}(\mathbf{x}_t - \gamma_0 \mathbf{1}) + \rho D^{-1}(I - H)(\tilde{\mathbf{v}}_{t-1} + I(t < T)\tilde{\mathbf{v}}_{t+1})$$

where $I(t < T) = 1$ if $t = 1, \ldots, T-1$ and 0 otherwise.

Note that this full conditional distribution is a J-variate normal distribution where J is possibly very high (33,390 in our example) and simultaneous update is computationally prohibitive. In addition, we need to incorporate the constraints implied by the first stage likelihood specification (5). The partition of $\tilde{\mathbf{V}}_t$, however, suggests an immediate univariate sampling scheme as follows.

The conditional prior distribution for $\tilde{V}(A_j, t)$, for each j and t, from the vectorized specification (12), as calculated above, is given by $N\left(\xi_{jt}, \omega_{jt}^2\right)$ where:

$$\omega_{jt}^2 = \sigma_\zeta^2 \frac{1}{m_j(1 + I(t < T)\rho^2)} \quad \text{and} \quad \xi_{jt} = r_{jt} + \sum_{i=1}^{J} h_{ji}(\tilde{v}(A_i, t) - r_{it})$$

where r_{jt} is the jth element of

$$\mathbf{r}_t = \frac{\rho}{1 + I(t < T)\rho^2)}(\tilde{\mathbf{v}}_{t-1} + I(t < T)\tilde{\mathbf{v}}_{t+1}).$$

The form of the likelihood contribution for $\tilde{V}(A_j, t)$ will depend on whether $\tilde{V}(A_j, t)$ is one of $\tilde{\mathbf{V}}_t^{(1)}$ or one of $\tilde{\mathbf{V}}_t^{(2)}$. For each component $\tilde{V}(A_j, t)$ of $\tilde{\mathbf{V}}_t^{(1)}$ we extract the full conditional distribution to be viewed as the likelihood contribution from the joint distribution $N(\Lambda_{(1),t}\boldsymbol{\chi}_{(1),t}, \Lambda_{(1),t})$ where

$$\Lambda_{(1),t}^{-1} = \frac{I_n}{\sigma_v^2} + \gamma_1^2 \frac{I_n}{\sigma_\psi^2} \quad \text{and} \quad \boldsymbol{\chi}_{(1),t} = \frac{1}{\sigma_v^2}\mathbf{v}_t + \frac{\gamma_1}{\sigma_\psi^2}(\mathbf{x}_t - \gamma_0 \mathbf{1}).$$

This conditional likelihood contribution is given by $N\left(\mu_{jt}, \Xi^2\right)$ where

$$\mu_{jt} = \Xi^2 \left\{ \tilde{v}(A_j, t)/\sigma_v^2 + \gamma_1(x(A_j, t) - \gamma_0)/\sigma_\psi^2 \right\}, \quad \Xi^2 = 1/(1/\sigma_v^2 + \gamma_1^2/\sigma_\psi^2).$$

For each component $\tilde{V}(A_j,t)$ of $\tilde{\mathbf{V}}_t^{(2)}$ the likelihood contribution is also denoted by the normal distribution $N\left(\mu_{jt},\Xi^2\right)$ where

$$\mu_{jt}=\frac{x(A_j,t)-\gamma_0}{\gamma_1} \text{ and } \Xi^2=\frac{\sigma_\psi^2}{\gamma_1^2}.$$

Now the un-constrained full conditional distribution of $\tilde{V}(A_j,t)$, according to the second stage likelihood and prior specification, is obtained by combining the likelihood contribution $N\left(\mu_{jt},\Xi^2\right)$ and the prior conditional distribution $N\left(\xi_{jt},\omega_{jt}^2\right)$ and is given by $N\left(\Lambda_{jt}\chi_{jt},\Lambda_{jt}\right)$ where

$$\Lambda_{jt}^{-1}=\Xi^{-2}+\omega_{jt}^{-2},\quad \chi_{jt}=\Xi^{-2}\mu_{jt}+\omega_{jt}^{-2}\xi_{jt}.$$

In order to respect the constraints implied by the first stage specification we simulate the $\tilde{V}(A_j,t)$ to be positive if $X(A_j,t)>0$ and negative otherwise.

REFERENCES

Banerjee, S., Carlin, B.P. and Gelfand, A. E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC.

Bilonick, R. A. (1985) The space-time distribution of sulfate deposition in the northeastern United States. *Atmos. Env.* **19**, 1829–1845.

Brown, P. J., Le, N. D., Zidek, J. V. (1994) Multivariate spatial interpolation and exposure to air pollutants. *Canad. J. Stat.*, **22**, 489–510.

Brook, J. R., P. J. Samson, and S. Sillman (1995) Aggregation of selected three-day periods to estimate annual and seasonal wet deposition totals for sulfate, nitrate, and acidity. Part I: A synoptic and chemical climatology for eastern North America. *J. Appl. Meteorol.*, 34, 297-325.

Carroll, R. J., Chen, R., George, E. I., Li, T.H., Newton, H.J., Schmiediche, H. and Wang, N. (1997) Ozone exposure and population density in Harris County, Texas. *J. Am. Statist. Ass.*, **92**, 392–404.

Fuentes, M. and Raftery, A. (2005). Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations with outputs from Numerical Models. *Biometrics*, **61**, 36–45.

Gelfand, A. E. and Ghosh, S. K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika*, **85**, 1–11.

Goulard, M.and Voltz, M. (1992). Linear coregionalization model:tools for estimation and choice of multivariate variograms *Mathematical Geology*, **24**, 269–286.

Gotway, C.A. and Young, L.J. (2002). Combining incompatible spatial data. *J. Am. Statist. Ass.*, **97**, 632–648.).

Grimm, J. W. and Lynch, J. A. (2004). Enhanced wet deposition estimates using modeled precipitation inputs. *Environmental Monitoring and Assessment* **90**, 243–268.

Haas, T. C. (1990a). Lognormal and moving window methods of estimating acid deposition. *J. Am. Statist. Ass.*, **85**, 950-963.

Haas, T. C. (1990b). Kriging and automated variogram modeling within a moving window. *Atmos. Envir.* **24A**, 1759–1769.

Haas, T. C. (1995). Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *J. Am. Statist. Ass.*, **90**, 1189–1199.

Haas, T. C. (1996). Multivariate spatial prediction in the presence of non-linear trend and covariance non-stationarity. *Environmetrics*, **7**, 145–165.

Huerta, G., Sanso, B., and Stroud, J. R. (2004) A spatiotemporal model for Mexico City ozone levels. *J. R. Statist. Soc.*, Series C, **53**, 231–248.

Le, N. and Zidek, J. (1992) Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *J. Mult. Ana.*, **43**, 351–374.

Oehlert, G. W. (1993) Regional trends in sulfate wet deposition. *J. Am. Statist. Ass.*, **88**, 390–399.

Rappold, A. G., Gelfand, A. E. and Holland, D. M. (2008) Modeling mercury deposition through latent space-time processes. *J. R. Statist. Soc.*, Series C, to appear.

Sahu, S. K. and Mardia, K. V. (2005) A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels *J. R. Statist. Soc.*, Series C, **54**, 223–244.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2006) Spatio-temporal modeling of fine particulate matter, *J. Agr., Bio., and Env. Stat.,* **11**, 61–86.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2007) High Resolution Space-Time Ozone Modeling for Assessing Trends. *J. Am. Statist. Assoc.*, **102**, 1221–1234.

Sampson, P. and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Statist. Ass.*, **87**, 108–119.

Stein, M. L. (2005) Space time covariance functions. *J. Am. Statist. Ass.*, **100**, 310-321.

Wikle, C. K. (2003) Hierarchical models in environmental science. *Int. Stat. Rev.,* **71**, 181–199.

Zhang, H. (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Am. Statist. Ass.*, **99**, 250–261.

Disclaimer:

The U.S. Environmental Protection Agency's Office of research and Development partially collaborated in the research described here. Although it has been reviewed by EPA and approved for publication, it does not necessarily reflect the Agency's policies or views. The authors thank Gary Lear and Norm Possiel of the U.S. EPA for providing the monitoring data and CMAQ output.

Table 1: Estimation of the parameters for the sulfate and nitrate models. CI stands for equal tailed credible intervals.

	Sulfate			Nitrate		
	mean	sd	95%CI	mean	sd	95%CI
α_0	–0.4497	0.0871	(–0.6189, –0.2733)	–0.3548	0.0596	(–0.4695, -0.2369)
α_1	0.1787	0.0379	(0.1017, 0.2499)	0.1522	0.0336	(0.0843, 0.2161)
β_0	–1.9414	0.0196	(–1.9784, –1.9012)	–1.9976	0.0192	(–2.0344, –1.9605)
β_1	0.9103	0.0067	(0.8972, 0.9240)	0.8412	0.0070	(0.8274, 0.8553)
β_2	0.0029	0.0062	(–0.0091, 0.0151)	0.0040	0.0060	(–0.0078, 0.0159)
b_0	0.0490	0.0053	(0.0386, 0.0599)	0.0535	0.0062	(0.0409, 0.0652)
γ_0	–3.0768	0.0035	(–3.0836, -3.0700)	–3.2177	0.0033	(–3.2242, –3.2112)
γ_1	0.8957	0.0034	(0.8891, 0.9025)	0.7368	0.0033	(0.7303, 0.7433)
ρ	0.7688	0.0012	(0.7664, 0.7712)	0.7492	0.0013	(0.7468, 0.7517)
σ^2_δ	2.6438	0.0602	(2.5254, 2.7631)	1.8694	0.0387	(1.7942, 1.9476)
σ^2_η	0.2812	0.0101	(0.2616, 0.3010)	0.3354	0.0105	(0.3149, 0.3564)
σ^2_ϵ	0.0718	0.0057	(0.0607, 0.0832)	0.0727	0.0074	(0.0588, 0.0878)
σ^2_ψ	2.5062	0.0033	(2.4997, 2.5127)	2.2148	0.0028	(2.2092, 2.2203)
σ^2_v	0.8087	0.0259	(0.7601, 0.8620)	0.7821	0.0237	(0.7366, 0.8290)
σ^2_ζ	0.4345	0.0011	(0.4322, 0.4367)	0.4340	0.0012	(0.4316, 0.4363)

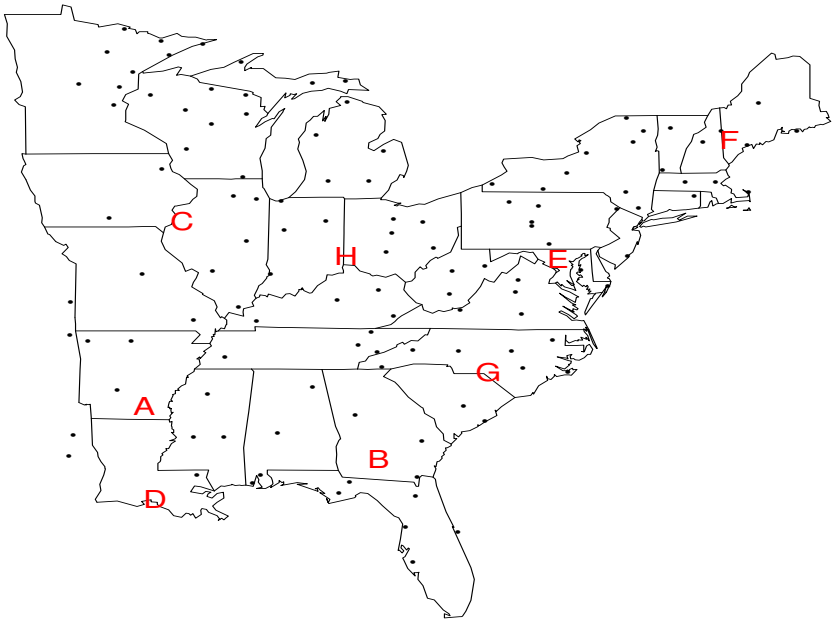

Figure 1: A map of the study region; points indicate the modeled NADP sites and the letters A-H denote the eight validation sites.

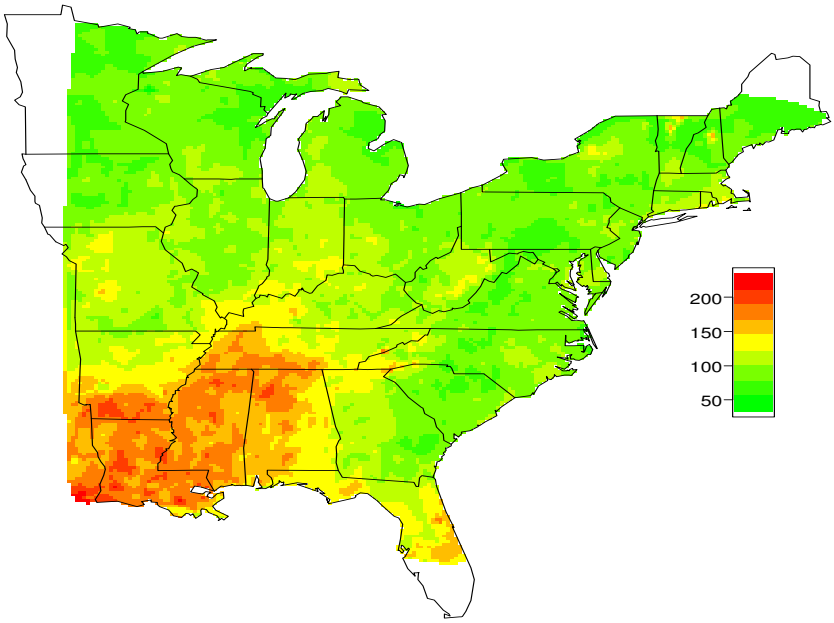

Figure 2: Map of annual total precipitation in 2001.

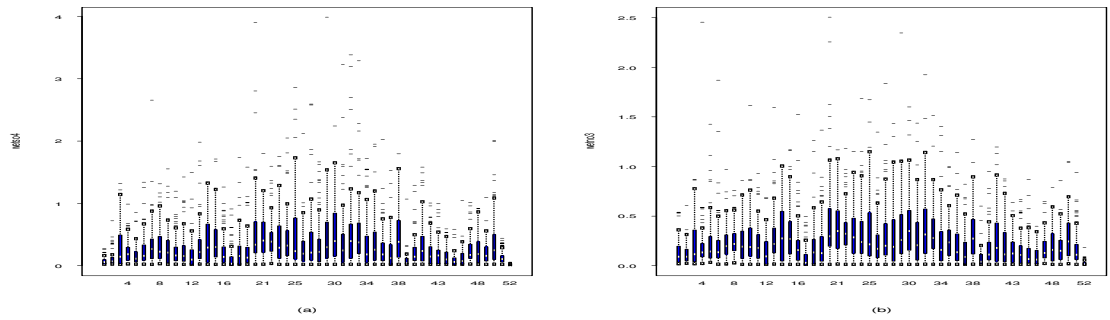

Figure 3: Boxplot of weekly depositions: (a) wet sulfate and (b) wet nitrate.

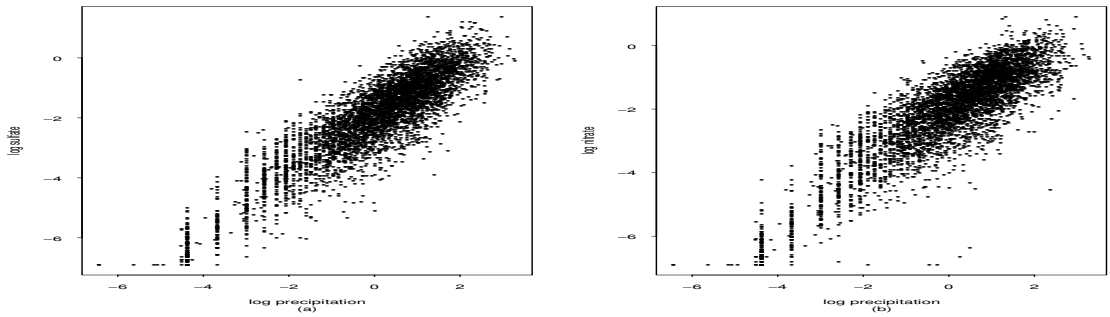

Figure 4: Deposition against precipitation on the log scale: (a) wet sulfate and (b) wet nitrate.

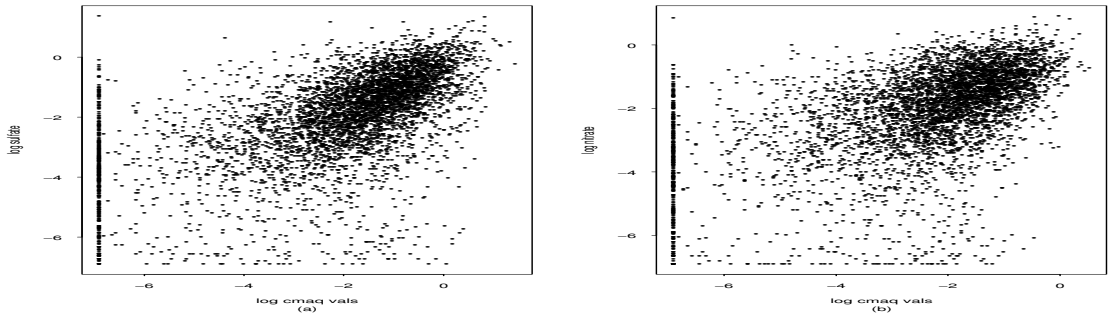

Figure 5: Deposition at the NADP sites against the CMAQ values in the grid cell covering the corresponding NADP site on the log scale: (a) wet sulfate and (b) wet nitrate.

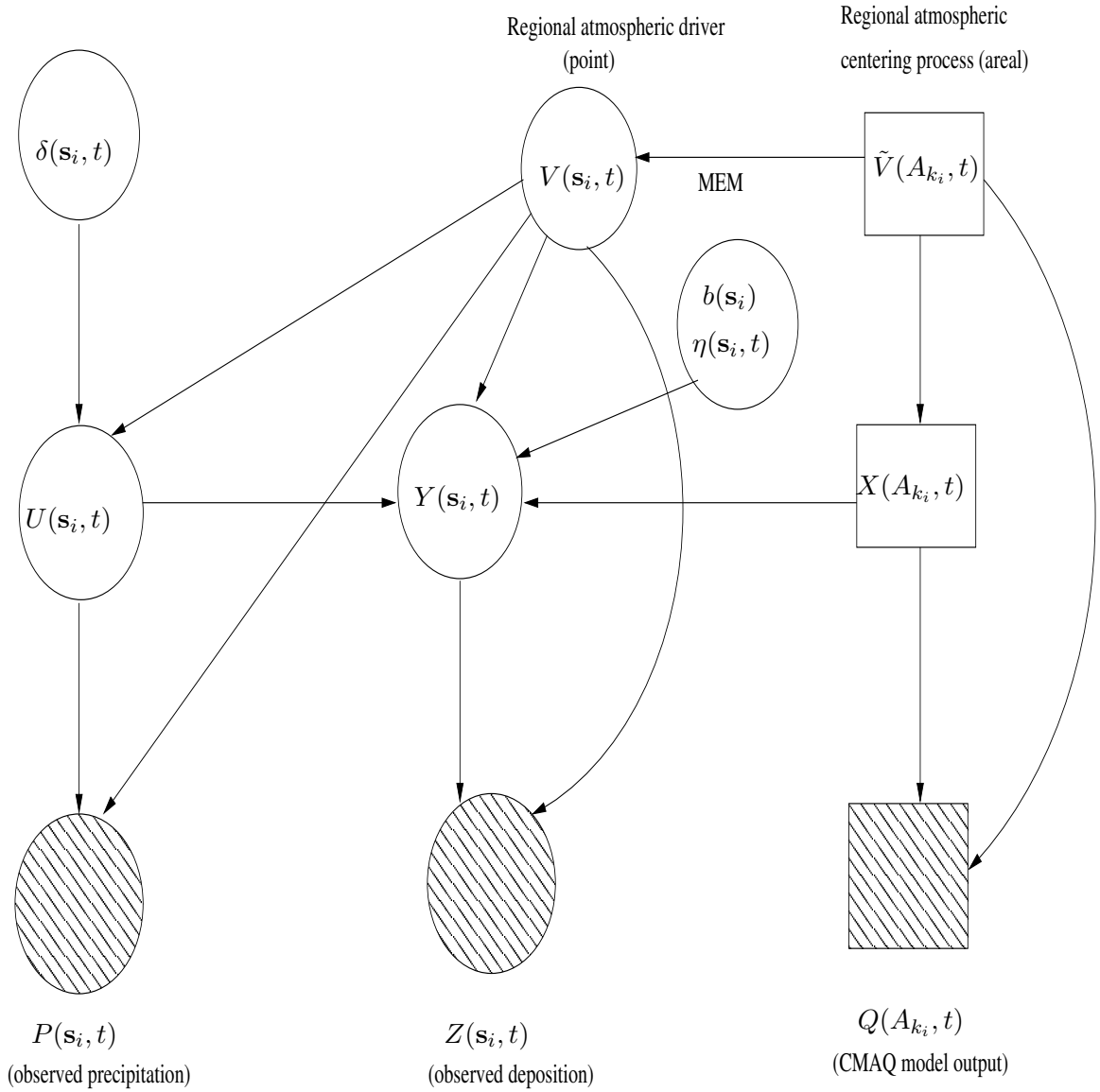

Figure 6: Graphical representation of the model. MEM stands for "measurement error model" to handle $Q(A_{k_i}, t) > 0$, $Z(\mathbf{s}_i, t) = 0$ or $Q(A_{k_i}, t) = 0$, $Z(\mathbf{s}_i, t) > 0$. An oval represents a point level random variable while a rectangular shape denotes an areal level one. Shaded shapes denote the observed random variables while the clear ones denote the latent un-observed ones.

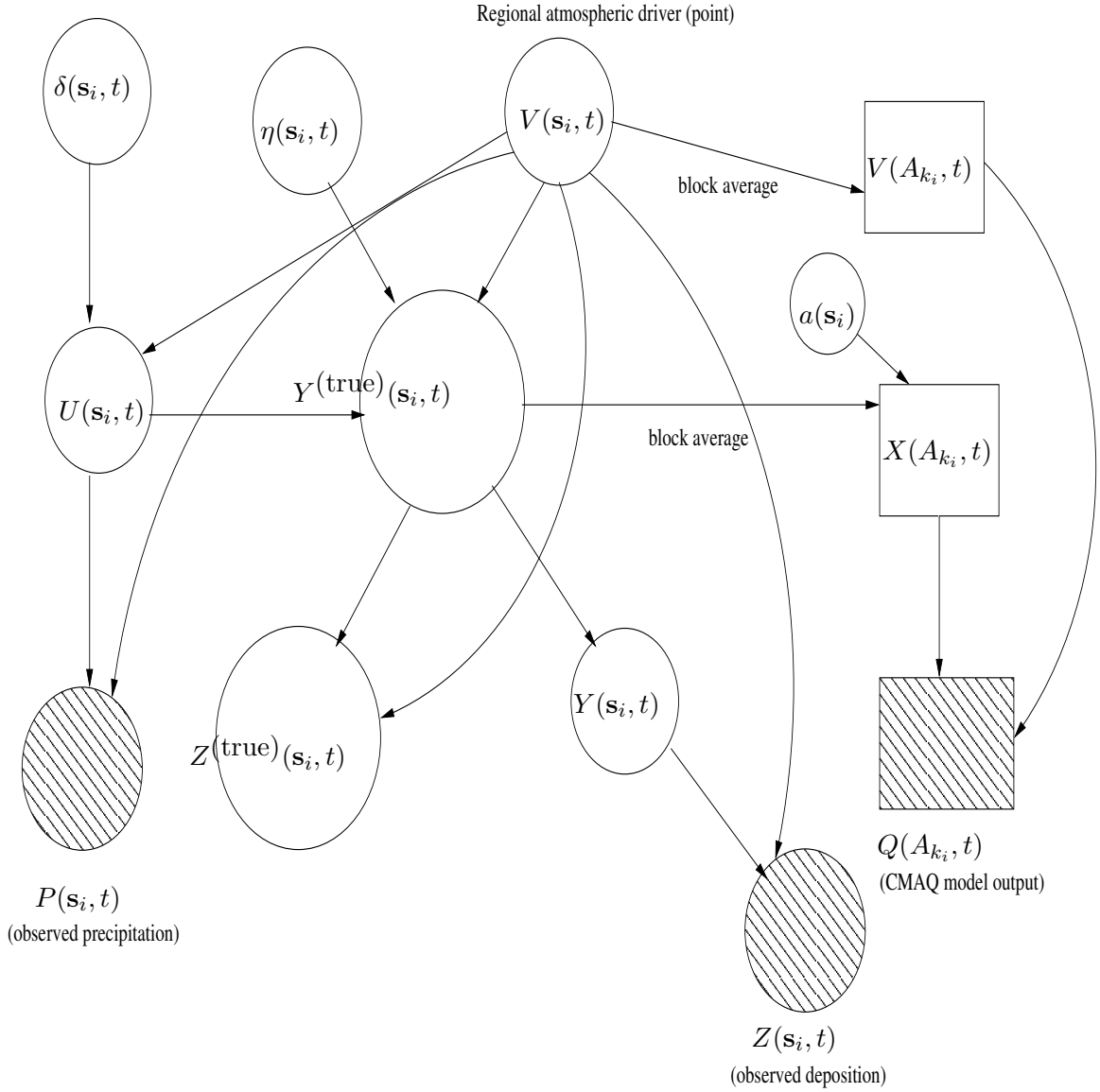

Figure 7: Graphical representation of a fusion model. Block averaging avoids the need for MEM; there is no $\tilde{V}(A_{k_i}, t)$ process.

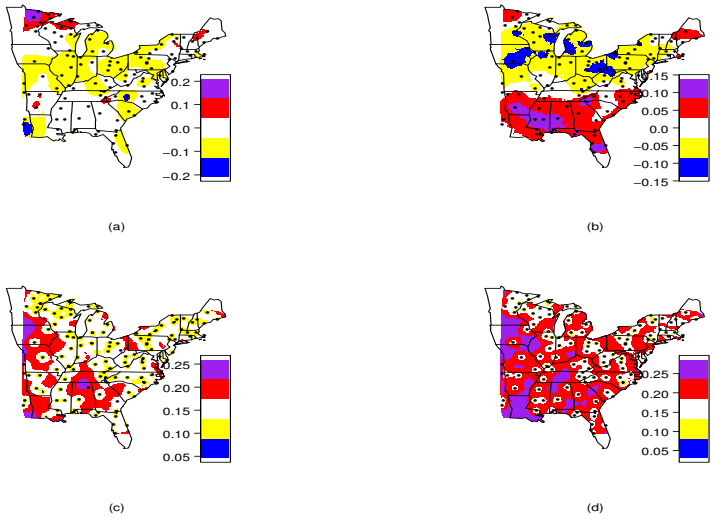

Figure 8: (a) The $b(\mathbf{s})$ surface for sulfate. (b) The $b(\mathbf{s})$ surface for nitrate. (c) The standard deviations of the $b(\mathbf{s})$ surface for sulfate. (d) The standard deviations of the $b(\mathbf{s})$ surface for nitrate.

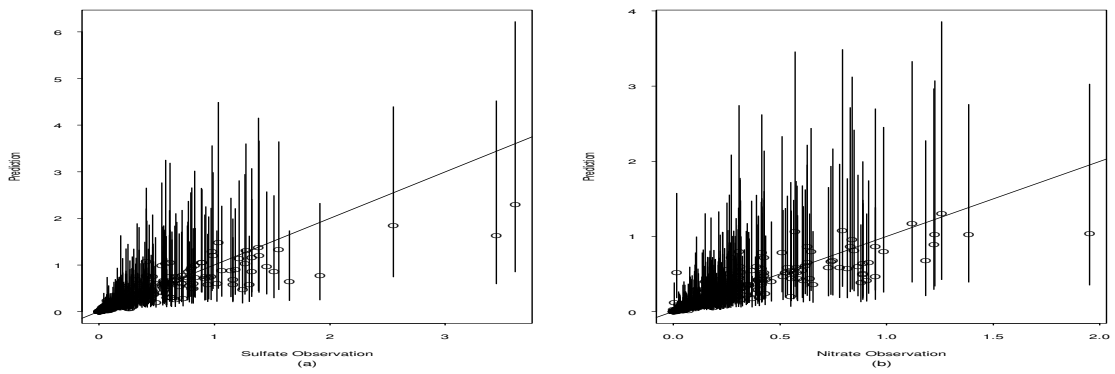

Figure 9: Validations versus the observed values at the 8 reserved sites. Validation prediction intervals are plotted as vertical lines. (a) wet sulfate and (b) wet nitrate.

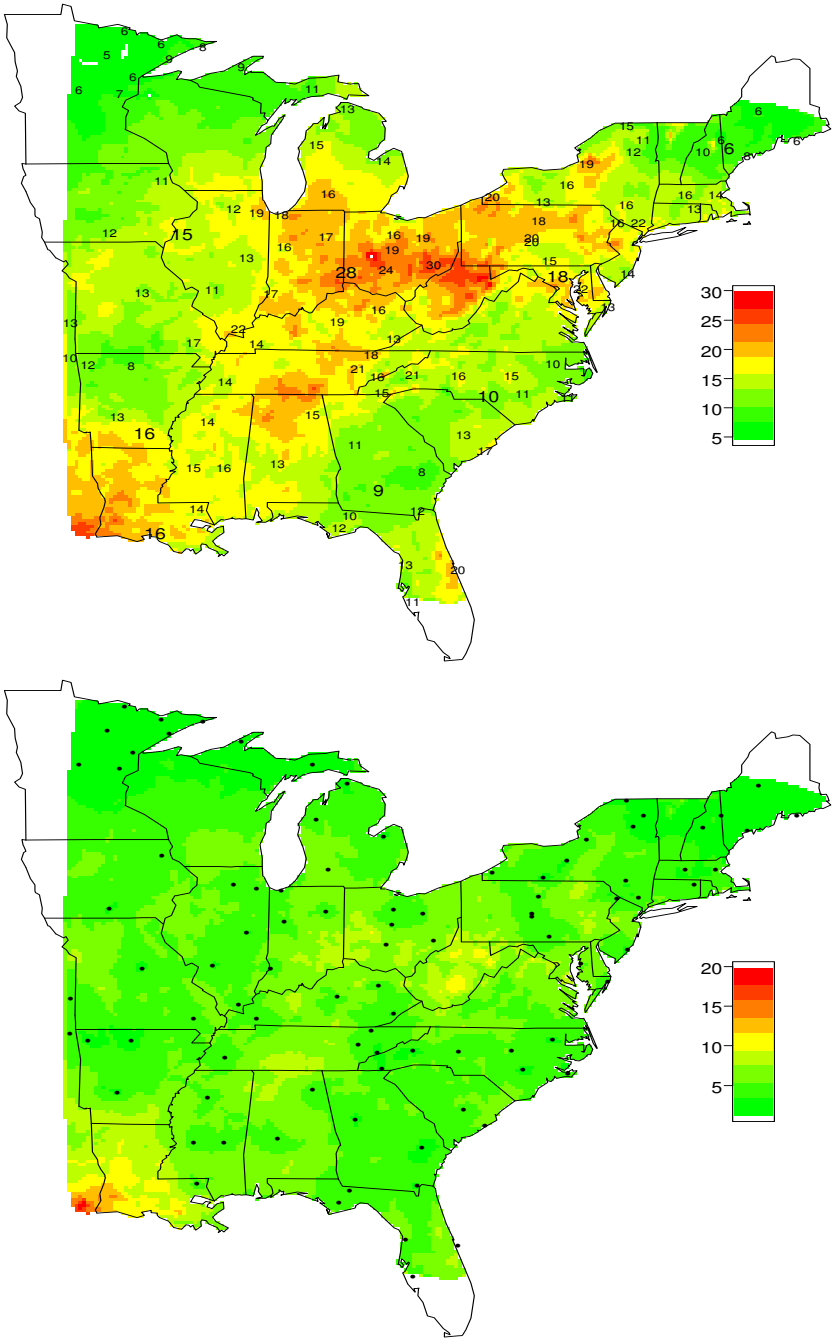

Figure 10: Analyses for sulfate. Top panel: The annual model predicted map. The observed annual totals are labeled; a larger font size is used for the validation sites. Bottom panel: Lengths of the prediction intervals.

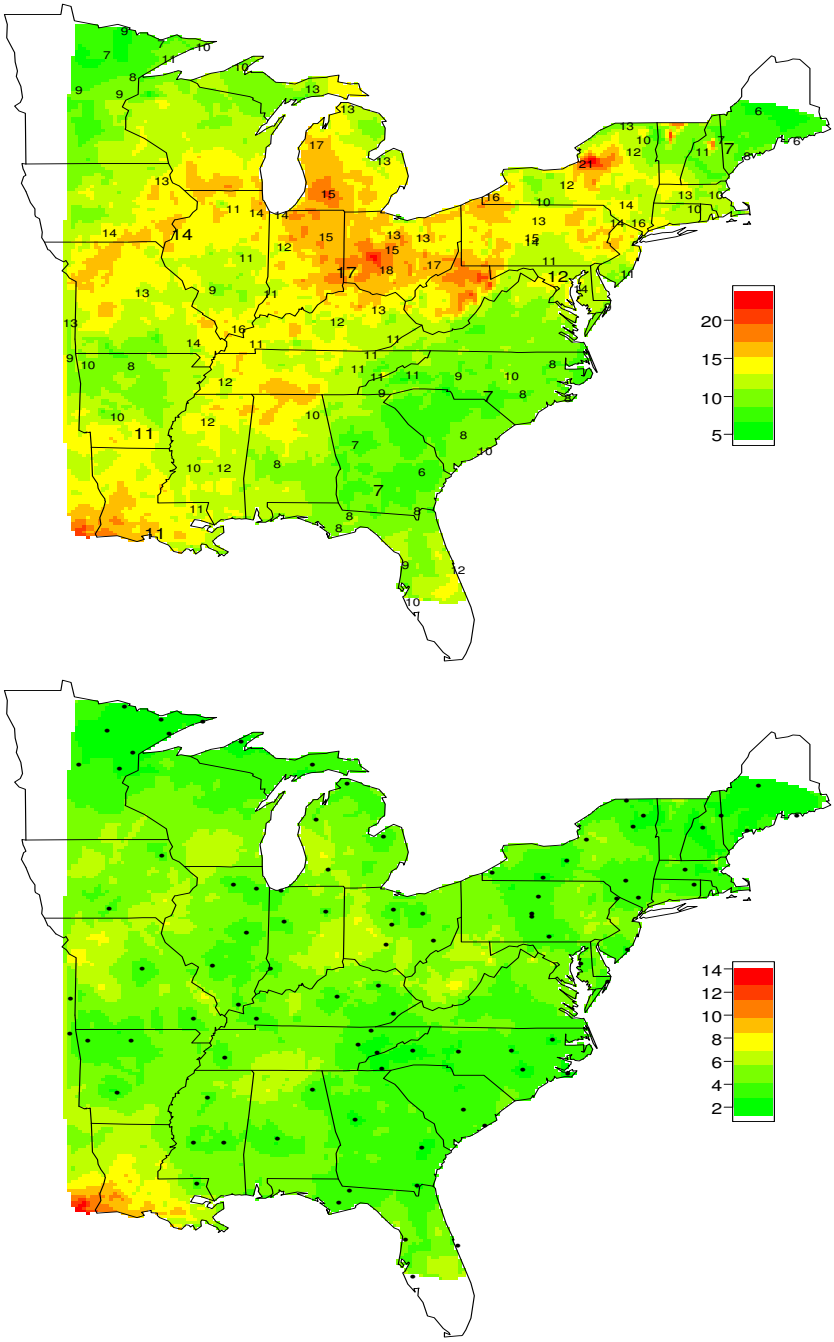

Figure 11: Analyses for nitrate. Top panel: The annual model predicted map. The observed annual totals are labeled; a larger font size is used for the validation sites. Bottom panel: Lengths of the prediction intervals.

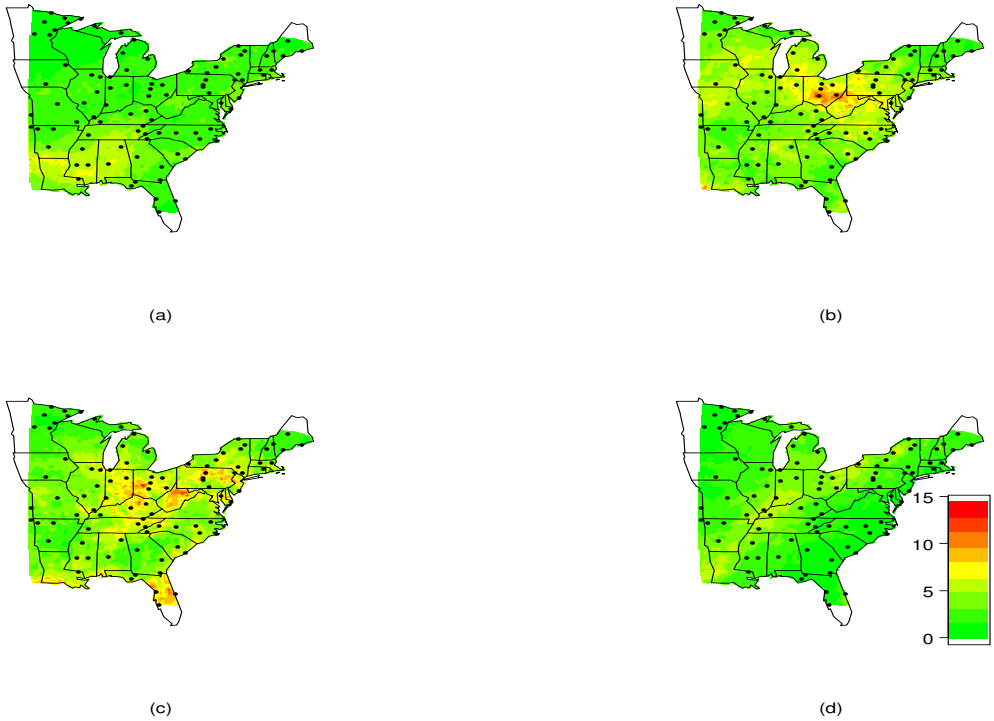

Figure 12: Sulfate prediction maps on 4 quarters. (a) Jan–Mar, (b) Apr-Jun, (c) Jul–Sep, (d) Oct-Dec.

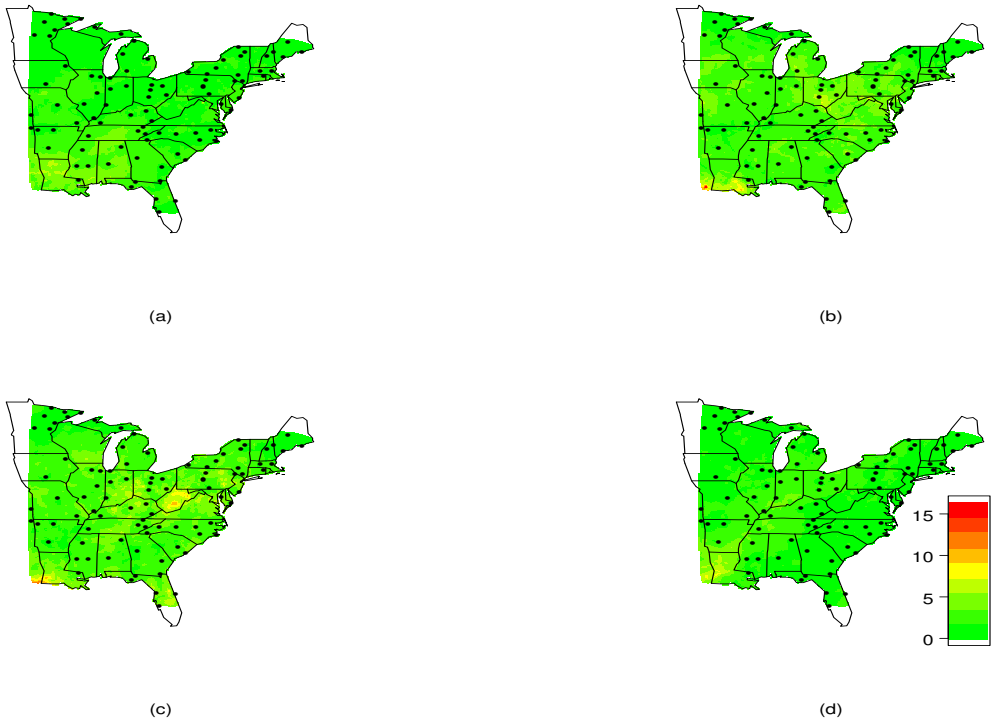

Figure 13: Maps showing the lengths of the 95% credible intervals for the sulfate predictions on 4 quarters. (a) Jan–Mar, (b) Apr-Jun, (c) Jul–Sep, (d) Oct-Dec.

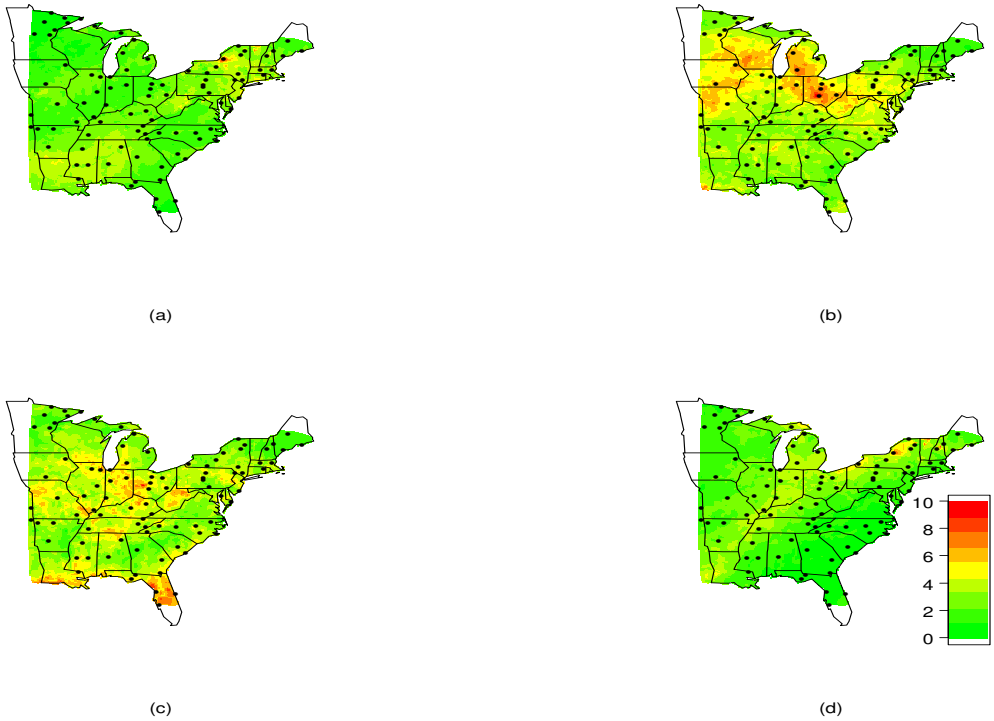

Figure 14: Nitrate prediction maps on 4 quarters. (a) Jan–Mar, (b) Apr-Jun, (c) Jul–Sep, (d) Oct-Dec.

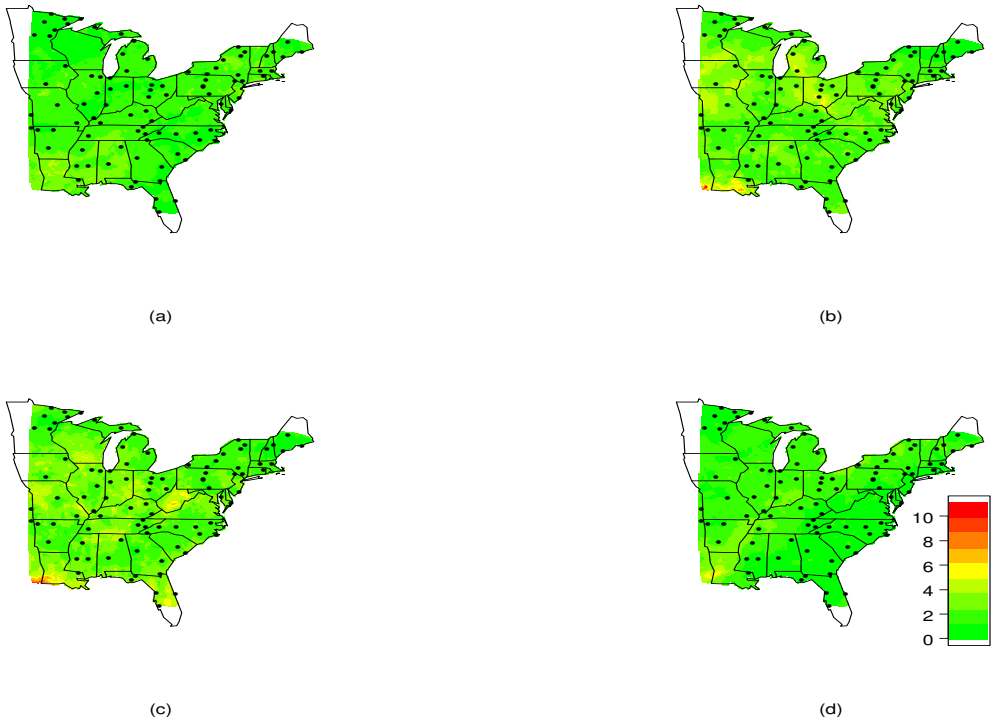

Figure 15: Maps showing the lengths of the 95% credible intervals for the nitrate predictions on 4 quarters. (a) Jan–Mar, (b) Apr-Jun, (c) Jul–Sep, (d) Oct-Dec.