Combining Monitoring Data and Computer model Output in Assessing Environmental Exposure

Alan E. GELFAND and Sujit K. SAHU *

December 21, 2008

Abstract

In the environmental exposure community numerical models are now widely available for a number of air pollutants. These models provide output exposures at various spatial and temporal resolutions. They are based on inputs from a number of factors such as meteorological conditions, land usage, and power station emission volumes, all of which are responsible for producing air pollution. Prediction of spatial surfaces may be available for current, past, and future time periods. For large spatial regions such as the entire United States, the spatial coverage of the available network of fixed monitoring stations can never match the coverage at which the computer models produce their output. However, the monitoring data will be more accurate than the computer model output since, up to measurement error, they provide the actual true levels; at each time point, they are a finite set of observations from the realization of the pollution process surface at that time. The latter typically require calibration. In any event, it seems natural to attempt to combine these two sets of information to make inference regarding the pollution exposure. In this chapter we review the currently available fully model-based methods in the literature and illustrate a novel, very recent, method with a real data example.

*Alan E. Gelfand is Professor, Department of Statistical Science, Duke University, Durham, NC, USA (Email: alan@stat.duke.edu) and 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).

1 Introduction

The demand for spatial models to assess regional progress in air quality has grown rapidly over the past decade. For improved environmental decision-making, it is imperative that such models enable spatial prediction to reveal important gradients in air pollution, offer guidance for determining areas in non-attainment with air standards, and provide air quality input to models for determining individual exposure to air pollution. Spatial prediction has the potential to suggest new perspectives in the development of emission control strategies and to provide a credible basis for resource allocation decisions, particularly with regard to network design.

Space-time modeling of pollutants has some history including, e.g., Guttorp et al. (1994), Haas (1995) and Carroll *et al.* (1997). In recent years, hierarchical Bayesian approaches for spatial prediction of air pollution have been developed (Brown et al., 1994; Le et al., 1997; Sun et al., 2000). Cressie et al. (1999) compared kriging and Markov-random field models in the prediction of PM_{10} (particles with diameters less than 10μm) concentrations around the city of Pittsburgh. Zidek et al. (2002) developed predictive distributions on non-monitored PM_{10} concentrations in Vancouver, Canada. They noted the under-prediction of extreme values in the pollution field, but their methodology provided useful estimates of uncertainties for large values. Sun et al. (2000) developed a spatial predictive distribution for the space-time response of daily ambient PM_{10} in Vancouver. They exploit the temporal correlation structure present in the observed data from several sites to develop a model with two components, a common deterministic trend across all sites plus a stochastic residual. They illustrate the methods by imputing daily PM_{10} fields in Vancouver. Kibria et al. (2000) developed a multivariate spatial prediction methodology in a Bayesian context for the prediction of $PM_{2.5}$ in the city of Philadelphia. This approach used both $PM_{2.5}$ and PM_{10} data at monitoring sites with different start-up times. Shaddick and Wakefield (2002) proposeed short term space-time modeling for PM_{10}.

Smith et al. (2003) proposed a spatio-temporal model for predicting weekly averages of $PM_{2.5}$ and other derived quantities such as annual averages within three southeastern states in the United States. The $PM_{2.5}$ field is represented as the sum of semi-parametric spatial and temporal trends, with a random component that is spatially correlated, but not temporally. These authors apply a variant of the expectation-maximization (EM) algorithm

to account for high percentages of missing data. Sahu and Mardia (2005) present a short-term forecasting analysis of $PM_{2.5}$ data in New York City during 2002. Within a Bayesian hierarchical structure, they use principal kriging functions to model the spatial structure and a vector random-walk process to model temporal dependence. Sahu et al. (2006) consider modeling of $PM_{2.5}$ through the use of rural and urban process models while Sahu et al. (2007) deal with misalignment between ozone data and meteorological information. Sahu et al. (2008b) develop a hierarchical space-time model for daily 8-hour maximum ozone concentration data covering much of the eastern United States. The model combines observed data and forecast output from a computer model known as the Community Multi-scale Air Quality (CMAQ) Eta forecast model (see below for references) so that the next day forecasts can be computed in real time. They validate the model with a large amount of set-aside data and obtain much improved forecasts of daily O_3 patterns. Berrocal et al. (2008) propose a downscaling approach by regressing the observed point level ozone concentration data on grid cell level computer model output with spatially varying regression co-efficients specified through a Gaussian process. Rappold et al. (2008) study wet mercury deposition over space and time. Finally, Wikle (2003) provides a nice overview of the role of hierarchical modeling in environmental science. With so much interest in space-time exposure prediction, attention to data fusion models to improve such prediction is not surprising.

1.1 Environmental computer models

Computer models are playing an increasing role in our quest to understand complex systems. In this regard, the discussion paper of Kennedy and O'Hagan (2001) reviews prediction and uncertainty analysis for systems which are approximated by complex mathematical models. These models are often implemented as computer codes and typically depend on a number of input parameters which determine the nature of the output. The input parameters are often unknown and are customarily estimated by ad hoc methods such as very crude fitting of the computer model to the observed data. Kennedy and O'Hagan present a Bayesian calibration technique which improves on this usual approach in two respects. First, Bayesian prediction methods allow one to account for all sources of uncertainty including the ones from the estimation of the parameters. Second, any inadequacy in the model specification, even under the best-fitting parameter values, is revealed by discrepancies between the observed

data and the model predictions. Illustration is provided using data from a nuclear radiation release at Tomsk and also from a more complex simulated nuclear accident exercise.

Cox et al. (2001) describe a statistical procedure for estimation of unknown parameters in a complex computer model from an observational or experimental data base. They develop methods for accuracy assessments of the estimates and illustrate their results in the setting of computer code which models nuclear fusion reactors. Fuentes et al. (2003) develop a formal method for evaluation of the performance of numerical models. They apply the method to an air quality model (essentially the CMAQ model) and discuss related issues in the estimation of nonstationary spatial covariance structures.

Turning to environmental computer models, high spatial resolution numerical model output is now widely available for various air pollutants. Our focus here is on the CMAQ forecast model. CMAQ is a modeling system which has been designed to approach air quality as a whole by including capabilities for modeling multiple air quality issues, including tropospheric ozone, fine particles, toxics, acid deposition, and visibility degradation. CMAQ was also designed to have multi-scale capabilities so that separate models are not needed for urban and regional scale air quality modeling. The target grid resolutions and domain sizes for CMAQ range spatially and temporally over several orders of magnitude. With the temporal flexibility of the model, simulations can be performed to evaluate longer term (annual to multi-year) pollutant climatologies as well as short term (weeks to months) transport from localized sources. The ability to handle a large range of spatial scales enables CMAQ to be used for urban and regional scale model simulations. See, e.g., http://www.epa.gov/asmdnerl/CMAQ/.

It is worth distinguishing the goal of CMAQ which is to provide ambient exposure at high spatial and temporal resolution from computer models that provide individual level exposure. In particular, the former, assimilated with station data, provide the ambient exposures which drive the latter. Again, the contribution of this chapter is to discuss fully model-based implementations of this fusion.

With regard to the latter, Zidek and his co-authors have written a series of papers considering prediction of human exposure to air pollution. In particular, Zidek et al. (2007) present a general framework for constructing a predictive distribution for the exposure to an environmental hazard sustained by a randomly selected member of a designated population. The individual's exposure is assumed to arise from random movement through

the environment, resulting in a distribution of exposure that can be used for environmental risk analysis. Zidek et al. (2005) develop a computing platform, referred to as pCNEM, to produce such distributions. This software is intended for simulating exposures to airborne pollutants. In the paper they illustrate with a model for predicting human exposure to PM_{10}.

Further work along these lines has been an objective of US Environmental Protection Agency (EPA) initiatives. The EPA's National Exposure Research Laboratory (NERL) has developed a population exposure and dose model, particularly for particulate matter (PM), called the Stochastic Human Exposure and Dose Simulation (SHEDS) model (Burke et al., 2003). SHEDS-PM uses a probabilistic approach that incorporates both variability and uncertainty to predict distributions of PM exposure, inhaled dose, and deposited dose for a specified population. SHEDS-PM estimates the contribution of PM from both outdoor and indoor sources (e.g., cigarette smoking, cooking) to total personal PM exposure and dose. In particular, SHEDS-PM generates a simulation population using US Census demographic data for the user-specified population with randomly assigned activity diaries of individuals. Output from the SHEDS-PM model includes distributions of exposure and dose for the specified population, as well as exposure and dose profiles for each simulated individual. It is Bayesian in its conception in the sense that the input parameters (e.g., air exchange rates, penetration rates, cooking and smoking emission rates) are drawn at random from suitable priors.

A similar EPA product, the Air Pollutants Exposure Model (APEX) was developed by the Office of Air Quality and Planning (Richmond et al., 2002). It is derived from the probabilistic National Ambient Air Quality Standards (NAAQS) Exposure Model for carbon monoxide (pNEM/CO). APEX serves as the human inhalation exposure model within the Total Risk Integrated Methodology (TRIM) model framework. APEX is intended to be applied at the local, urban, or consolidated metropolitan area scale and currently only addresses inhalation exposures. The model simulates the movement of individuals through time and space and their exposure to the given pollutant in various micro-environments (e.g., outdoors, indoors residence, in-vehicle). Results of the APEX simulations are provided as hourly and summary exposure and/or dose estimates, depending on the application, for each individual included in the simulation as well as summary statistics for the population modeled.

The format of the remainder of this Chapter is as follows. In Section 2 we review some algorithmic and pseudo-statistical approaches in weather prediction. Section 3 provides a review of current state of the art fusion methods for environmental data. We develop a non-dynamic downscaling approach based on our recent work (Sahu, Gelfand, and Holland, 2008a) in Section 4. A few summary remarks are provided in Section 5. Appendix A contains an introduction to the Gaussian Processes (GP) and Appendix B outlines the full conditional distributions for the downscaler approach proposed in Section 4.

2 Algorithmic and pseudo-statistical approaches in weather prediction

A convenient framework within to review algorithmic and pseudo-statistical approaches to data assimilation is in the context of numerical weather prediction. Kalnay (2003) provides a recent development of this material. Such assimilation has a long history dating at least to Charney (1951) who recognized that hand interpolation of available weather observations to a regular grid was too time consuming and that numerical interpolation methods were needed. Earliest work created local polynomial interpolations using quadratic trend surfaces in locations in order to interpolate observed values to grid values. Of course, in the past half century, such polynomial interpolation has come a long way to become a standard device in the statistician's toolkit; we do not detail this literature here.

Instead, we note that what emerged in the meteorology community was the recognition that a first guess (or background field or prior information) was needed (Bergthorsson and Döös, 1955), supplying the *initial conditions*. As short-range forecasts became better and better, their use as a first guess became universal. The climatological intuition here is worth articulating. Over "data-rich" areas the observational data dominates while in "data-poor" regions the forecast facilitates transport of information from the data-rich areas. Of course, in the setting of fully-specified models and fully model-based inference we can quantify this adaptation and the associated uncertainty. Indeed, this is the contribution of the following sections of this Chapter.

We illustrate several numerical approaches using, illustratively, temperature as the variable of interest. At time t, we let $T_{obs}(t)$ be an observed measurement, $T_b(t)$ a background level, $T_a(t)$ an assimilated value, and $T_{true}(t)$ the true value. An early scheme is

known as the successive corrections method (SCM) which obtains $T_{i,a}(t)$ iteratively through $T_{i,a}^{(r+1)}(t) = T_{i,a}^{(r)}(t) + \left[\sum_k w_{ik} \left\{T_{k,obs}(t) - T_{k,a}^{(r)}(t)\right\}\right] / \left(\sum_k w_{ik} + \epsilon^2\right)$. Here, i indexes the grid cells for the interpolation while k indexes the observed data locations. $T_{k,a}^{(r)}(t)$ is the value of the assimilator at the r-th iteration at the observation point k (obtained from interpolating the surrounding grid points). The weights, w_{ik}, can be defined in various ways but usually as a decreasing function of the distance between the grid point and the observation point. In fact, they can vary with iteration, perhaps becoming increasingly local. See, e.g., Cressman (1959) and Bratseth (1986).

Another empirical approach is called *nudging* or Newtonian relaxation. Suppose, suppressing location, we think about a differential equation driving temperature, e.g., $\frac{dT(t)}{dt} = a\left(T(t), t, \theta(t)\right)$. If we write $a(\cdot)$ as an additive form say $a(T(t), t) + \theta(t)$ and let $\theta(t) = (T_{obs}(t) - T(t))/\tau$ then τ controls the relaxation. Small τ implies that the $\theta(t)$ term dominates while large τ implies that the nudging effect will be negligible.

We next turn to a least squares approach. Again, suppressing location, suppose we assume that $T_{obs}^{(1)}(t) = T_{true}(t) + \epsilon_1(t)$ and $T_{obs}^{(2)}(t) = T_{true}(t) + \epsilon_2(t)$ where we envision two sources of observational data on the true temperature at t. The ϵ_l have mean 0 and variance $\sigma_l^2, l = 1, 2$. Then, with the variances known, it is a familiar exercise to obtain the best unbiased estimator of $T_{true}(t)$ based upon these two pieces of information. That is, $T_a(t) = a_1 T_{obs}^{(1)}(t) + a_2 T_{obs}^{(2)}(t)$ where $a_1 = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$ and $a_2 = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$. Of course, we obtain the same solution as the maximum likelihood estimates (MLE) if we use independent normal likelihoods for the $T_{obs}^{(l)}(t)$s.

A last idea here is simple sequential assimilation and its connection to the Kalman filter. In the univariate case suppose we write $T_a(t) = T_b(t) + \gamma(T_{obs}(t) - T_b(t))$. Here, $T_{obs}(t) - T_b(t)$ is referred to as the observational innovation or observational increment relative to the background. The optimal weight $\gamma = \sigma_{obs}^2/(\sigma_{obs}^2 + \sigma_b^2)$, analogous to the previous paragraph. Hence, we only need a prior estimate of the ratio of the observational variance to the background variance in order to obtain $T_a(t)$. To make this scheme dynamic, suppose the background is updated through the assimilation, i.e., $T_b(t+1) = h(T_a(t))$ where $h(\cdot)$ denotes some choice of forecast model. Then we will also need to create a revised background variance; this is usually taken to be a scalar (> 1) multiple of the variance of $T_a(t)$.

Finally, the multivariate assimilation idea is now clear. Now we collect the grid cell

variables to vector variables and write $T_a(t) = T_b(t) + W(Y_{obs}(t) - Y_b(t))$. Here, the vector Y_{obs} denotes variables that are different from the ones we seek to interpolate. For temperature, these might be Doppler shifts, radar reflectivities, or satellite radiances. Then, g is the nonlinear operator that converts background temperatures into guesses for these new variables. The dimension of Y is not necessarily the same as that of T. W is the gain matrix that usually appears in the Kalman filter. Finally, we introduce errors in the transitional stage, $T_b(t) = T_{true}(t) + \epsilon_b(t)$ and $T_a(t) = T_{true}(t) + \epsilon_a(t)$, as well as errors in the observational stage, i.e., for the $Y_{obs}(t)$. Assuming all errors are Gaussian, the dynamic model is specified and the Kalman filter can be implemented to fit the model.

3 Review of environmental exposure data fusion methods

Recall that our objective is to combine model output and station data to improve assessment of environmental exposure. Such synthesis is referred to as assimilation or fusion. Here we move from the more algorithmic strategies of the previous section to fully model-based approaches. In the next two subsections we review the work of Fuentes and Raftery (2005) which has received considerable attention and the very recent work of McMillan et al. (2008). A full development of the approach of Sahu et al. (2008a) with an example is deferred to the following sections.

3.1 Fusion modeling using stochastic integration

The fusion approach proposed by Fuentes and Raftery (2005) builds upon earlier Bayesian melding work in Poole and Raftery (2000). It conceptualizes a true exposure surface and views the monitoring station data as well as the model output data as varying in a suitable way around the true surface. In particular, the average exposure in a grid cell A, denoted by $Z(A)$, differs from the exposure at any particular location s, $Z(s)$. The so called change of support problem in this context addresses converting the point level $Z(s)$ to the grid level $Z(A)$ through the stochastic integral,

$$Z(A) = \frac{1}{|A_j|} \int_{A_j} Z(s)\, ds, \tag{1}$$

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

Let $Y(s)$ denote the true exposure corresponding to $Z(s)$ at a station s. The first model assumption is:

$$Z(s) = Y(s) + \epsilon(s) \tag{2}$$

where $\epsilon(s) \sim N(0, \sigma_\epsilon^2)$ represents the measurement error at location s. The true exposure process is assumed to be:

$$Y(s) = \mu(s) + \eta(s) \tag{3}$$

where $\mu(s)$ provides the spatial trend often characterized by known functions of the site characteristics such as the components of s, elevation etc. The error term $\eta(s)$ is a spatially colored process assumed to be the zero mean GP with a specified covariance function. (Appendix A provides an introduction to GP's.) The output of the computer model denoted by $Q(s)$ is often known to be biased and hence this is modeled as:

$$Q(s) = a(s) + b(s)Y(s) + \delta(s) \tag{4}$$

where $a(s)$ denotes the additive bias and $b(s)$ denotes the multiplicative bias. The error term, $\delta(s)$, is assumed to be a white noise process given by $N(0, \sigma_\delta^2)$. However, the computer model output is provided in a grid, $A_1, \ldots, A_J$ so the point level process is converted to a grid level one by the stochastic integral (1) for the model (4), i.e.

$$Q(A_j) = \int_{A_j} a(s)\, ds + \int_{A_j} b(s)Y(s)\, ds + \int_{A_j} \delta(s)\, ds.$$

It is acknowledged that unstable model fitting accrues to the case where we have spatially varying $b(s)$ so $b(s) = b$ is adopted. Spatial prediction at a new location s' is done through the posterior predictive distribution $p(Y(s')|Z, Q)$ where Z denote all the station data and Q denote all the grid-level computer output $Q(A_1), \ldots, Q(A_J)$.

This fusion strategy becomes computationally infeasible in the setting of fusing say, CMAQ data at $12km^2$ grid cells for the eastern United States with station data for this region. We have a very large number of grid cells with a relatively sparse number of monitoring sites. An enormous amount of stochastic integration is required. In this regard, a dynamic implementation over many time periods becomes even more infeasible. Recently Berrocal et al. (2008) have shown that the fusion strategy can be outperformed by their proposed downscaling approach both in terms of computing speed and out-of-sample validation.

3.2 Fusion modeling by upscaling

While the Fuentes and Raftery (2005) approach models at the point level, the strategy in McMillan et al. (2008) scales up to, models at, the grid cell level. In this fashion, computation is simplified and fusion with space-time data is manageable.

In particular, suppose that we have, say, n monitoring stations. As before, let $Q(A_j)$ denote the CMAQ output value for cell A_j while $Z_{A_j}(s_i)$ denotes the station data for site s_i within cell A_j, $i = 1, \ldots, k_j$. Of course, for most of the j's, k_j will be 0 since $n << J$. Let $Y(A_j)$ denote the true value for cell A_j.

Then, paralleling (2) and (4), for each $j = 1, \ldots, J$,

$$Z_{A_j}(s_i) = Y(A_j) + \epsilon_{A_j}(s_i), \quad i = 1, \ldots, k_j \tag{5}$$

and

$$Q(A_j) = Y(A_j) + b(A_j) + \gamma(A_j). \tag{6}$$

In (6), the CMAQ output is modeled as varying around the true value with a bias term, denoted by $b(A_j)$, specified using a B-spline model. Also, the ϵ's are assumed to be independently and identically distributed and so are the γ's, each with a respective variance component. So, the station data and the CMAQ data are conditionally independent given the true surface. Finally, the true surface is modeled analogously to (3). But now, the η's are given a CAR specification (see, e.g., Banerjee et al., 2004). For space-time data, McMillan et al. (2008) offer a dynamic version of this approach, formulated by assuming a dynamic CAR specification for the η's. They illustrate with a fusion for the year 2001.

4 A downscaling approach

Very recently, Sahu et al. (2008a) propose a modeling approach that avoids the computationally demanding stochastic integrations required in Fuentes and Raftery (2005) but models at the point rather than the grid cell level as in McMillan et al. (2008). In particular, they formalize a latent atmospheric process which is modeled at two different scales, at the point level to align with the station data and at the grid cell level to align with the resolution for the computer model output. The models at these two scales are connected through a measurement error model (MEM). The latent processes are introduced to capture point masses at 0 with regard to chemical deposition while the MEM circumvents the

stochastic integration in (1). 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 $Q(A_j)$ to point level ones, $Q(s_i)$ is not required. The only assumption is that $Q(A_j)$ is a reasonable surrogate for $Z(s_i)$ if the site s_i is within the grid cell A_j. In this sense, the approach is a downscaler, scaling the grid cell level CMAQ data to the point-level station data.

Sahu et al. (2008a) model the above fusion approach in a dynamic setting modeling weekly chemical deposition data over a year. They utilize precipitation information to model wet deposition since there can be no deposition without precipitation. They also handle occurrences of zero values in both precipitation and deposition. They introduce a 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. do 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. Sahu et al. illustrate their methods separately for both wet sulfate and wet nitrate deposition in the eastern United States.

4.1 The modeling detail

Here we present detail for the static version of the dynamic spatial model developed in Sahu et al. (2008). Let $R(s_i)$ and $Z(s_i)$ denote the observed precipitation and deposition respectively at a site $s_i, i = 1, \ldots, n$. We suppose that $R(s_i)$ and $Z(s_i)$ are driven by a point level latent atmospheric process, denoted by $V(s_i)$, and both take the value zero if $V(s_i) < 0$ to reflect that there is no deposition without precipitation. That is,

$$R(s_i) = \begin{cases} \exp\{U(s_i)\} & \text{if } V(s_i) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

and

$$Z(s_i) = \begin{cases} \exp\{Y(s_i)\} & \text{if } V(s_i) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The random variables $U(s_i)$ and $Y(s_i)$ are thus taken as log observed precipitation and deposition respectively when $V(s_i) > 0$. The models described below will specify their values when $V(s_i) \leq 0$ and/or the corresponding $R(s_i)$ or $Z(s_i)$ are missing.

Similar to (8) we suppose that the CMAQ model output at grid cell A_j, $Q(A_j)$, is positive if an areal level latent atmospheric process, denoted by $\tilde{V}(A_j)$, is positive,

$$Q(A_j) = \begin{cases} \exp\{X(A_j)\} & \text{if } \tilde{V}(A_j) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

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

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

The first stage likelihood implied by the definitions (7), (8) and (9) is given by:

$$p(R, Z, Q|U, Y, X, V, \tilde{V}) = p(R|U, V) \times p(Z|Y, V) \times p(Q|X, \tilde{V}) \tag{10}$$

which takes the form

$$\prod_{i=1}^{n} \left\{ 1_{\exp(u(s_i))} 1_{\exp(y(s_i))} I(v(s_i) > 0) \right\} \prod_{j=1}^{J} \left\{ 1_{\exp(x(A_j))} I(\tilde{v}(A_j) > 0) \right\}$$

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

4.2 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 $V(s_i)$. In particular, we assume the model:

$$U(s_i) = \alpha_0 + \alpha_1 V(s_i) + \delta(s_i), \quad i = 1, \ldots, n \tag{11}$$

where $\delta = (\delta(s_1), \ldots, \delta(s_n))'$ is an independent GP following the $N(0, \Sigma_\delta)$ 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 s_i and s_j. Using vector notation, the above specification is equivalently written as:

$$U \sim N\left(\alpha_0 + \alpha_1 V, \Sigma_\delta\right).$$

To model $Y(s_i)$, we assume that:

$$Y(s_i) = \beta_0 + \beta_1 U(s_i) + \beta_2 V(s_i) + \{b_0 + b(s_i)\} X(A_{k_i}) + \eta(s_i) + \epsilon(s_i), \tag{12}$$

for $i = 1, \ldots, n$ where, unless otherwise mentioned, A_{k_i} is the grid cell which contains the site s_i.

The error terms $\epsilon(s_i)$ 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 (12) is as follows. The term $\beta_1 U(s_i)$ is included because of the strong linear relationships between log-deposition and log-precipitation, see Figure 3. The term $\beta_2 V(s_i)$ captures any direct influence of the atmospheric process $V(s_i)$ on $Y(s_i)$ in the presence of precipitation.

It is anticipated that the relationship between the station data and the CMAQ model output will be roughly linear but that this relationship may vary locally. To specify a rich class of *locally* linear models we may think of a spatially varying slope for the regression of $Y(s_i)$ on log-CMAQ values $X(A_j)$, specified as $\{b_0 + b(s_i)\} X(A_{k_i})$ in (12). Writing $b = (b(s_1), \ldots, b(s_n))'$ we propose a mean 0 GP for b, i.e.

$$b \sim N(0, \Sigma_b)$$

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

The term $\eta(s_i)$ 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

$$\eta \sim N(0, \Sigma_\eta),$$

where $\eta = (\eta(s_1), \ldots, \eta(s_n))'$ and Σ_η has elements $\sigma_\eta(i,j) = \sigma_\eta^2 \exp(-\phi_\eta d_{ij})$. The regression model (12) is now equivalently written as:

$$Y \sim N(\vartheta, \sigma_\epsilon^2 I_n)$$

where $Y = (Y(s_1), \ldots, Y(s_n))$ and $\vartheta = \beta_0 + \beta_1 u + \beta_2 v + b_0 x + Xb + \eta$ where x is the n-dimensional vector with the ith element given by $x(A_{k_i})$ and X is a diagonal matrix whose ith diagonal entry is $X(A_{k_i})$, $i = 1, \ldots, n$ and I_n is the identity matrix of order n.

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

$$X(A_j) = \gamma_0 + \gamma_1 \tilde{V}(A_j) + \psi(A_j), \quad j = 1, \ldots, J. \tag{13}$$

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

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

where as before, $X = (X(A_1), \ldots X(A_J))'$ and $\tilde{V} = (\tilde{V}(A_1), \ldots \tilde{V}(A_J))'$, see the partitioning of $\tilde{V}$ below Equation (14) regarding the order of the grid cell indices $1, \ldots, J$.

We now turn to specification of the latent processes $V(s_i)$ and $\tilde{V}(A_j)$. Note that it is possible to have $Z(s_i) > 0$ and $Q(A_{k_i}) = 0$ and vice versa since $Q(A_{k_i})$ is the output of a computer model which has not used the actual observation $Z(s_i)$. This implies that $V(s_i)$ and $\tilde{V}(A_{k_i})$ 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:

$$V(s_i) \sim N(\tilde{V}(A_{k_i}), \sigma_v^2), i = 1, \ldots, n \tag{14}$$

where σ_v^2 is unknown. Without loss of generality we write $\tilde{V} = (\tilde{V}^{(1)}, \tilde{V}^{(2)})$ where the n-dimensional vector $\tilde{V}^{(1)}$ contains the values for the grid cells where the n observation sites are located and $\tilde{V}^{(2)}$ contains the values for the remaining $J - n$ grid cells. The specification (14) can now be written equivalently as

$$V \sim N(\tilde{V}^{(1)}, \sigma_v^2 I_n).$$

The latent process $\tilde{V}(A_j)$ is assumed to follow a conditionally auto-regressive (CAR) process in space (see e.g. Banerjee et al., 2004). That is,

$$\tilde{V}(A_j) \sim N\left(\sum_{i=1}^{J} h_{ji}\tilde{V}(A_i), \frac{\sigma_\zeta^2}{m_j}\right) \tag{15}$$

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. The above improper CAR specification can be written as:

$$p(\tilde{V}|\sigma_\zeta^2) \propto \exp\left\{-\frac{1}{2}\tilde{V}'D^{-1}(I - H)\tilde{V}\right\} \tag{16}$$

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:

$$p(Y|U, V, X, \eta, b, \theta) \times p(\eta|\theta) \times p(U|V, \theta) \times p(V|\theta) \times p(X|\tilde{V}, \theta) \times p(\tilde{V}|\theta) \times p(b|\theta)$$

where θ denote the parameters α_0, α_1, β_0, β_1, β_2, b_0, γ_0, γ_1, ρ, σ_δ^2, σ_b^2, σ_η^2, σ_ϵ^2, σ_ψ^2, σ_v^2 and σ_ζ^2. See Appendix B for the prior distributions, the form of the joint posterior distribution and the full conditional distributions needed for Gibbs sampling.

4.3 Spatial interpolation at a new location

We can interpolate the deposition surface using the above models as follows. Consider the problem of predicting $Z(s')$ at any new location s' falling on the grid cell A'. The prediction is performed by constructing the posterior predictive distribution of $Z(s')$ which in turn depends on the distribution of $Y(s')$ as specified by Equation (12) along with the associated $V(s')$. 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 s'. If precipitation information is available and there is no positive precipitation, i.e. $r(s') = 0$, then we have $Z(s') = 0$ and no further sampling is needed, since there can be no deposition without precipitation. Now suppose that there is positive precipitation, i.e. $r(s') > 0$, then set $u(s') = \log(r(s'))$. We need to generate a sample $Y(s')$. We first generate $V(s') \sim N(\tilde{V}(A'), \sigma_v^2)$ following the measurement error model (14). Note that $\tilde{V}(A')$ is already available for any grid cell A' (within the study region) from model fitting, see Equation (15). Similarly, $X(A')$ is also available either as the log of the CMAQ output, $\log(Q(A'))$, if $Q(A') > 0$ or from the MCMC imputation when $Q(A') = 0$, see Appendix B. To sample $\eta(s')$ we note that:

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

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

$$\eta(s')|\eta, \theta \sim N\left[S_{\eta,12}S_\eta^{-1}\eta, \ \ \sigma_\eta^2\left(1 - S_{\eta,12}S_\eta^{-1}S_{\eta,21}\right)\right]. \tag{17}$$

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

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

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

$$b(s')|b, \theta \sim N\left[S_{b,12}S_b^{-1}b, \ \sigma_b^2\left(1 - S_{b,12}S_b^{-1}S_{b,21}\right)\right]. \tag{18}$$

If it is desired to predict $Z(s')$ where $R(s')$ is not available, we proceed as follows. We generate $V(s') \sim N(\tilde{V}(A'), \sigma_v^2)$ following the measurement error model (14). If this

$V(s') < 0$, then we set both $r(s')$ and $Z(s')$ to zero. If, however, $V(s') > 0$ we need to additionally draw $U(s')$ using the precipitation model (11). For this we note that,

$$\begin{pmatrix} U(s') \\ U \end{pmatrix} \sim N\left[\begin{pmatrix} \alpha_0 + \alpha_1 V(s') \\ \alpha_0 + \alpha_1 V \end{pmatrix}, \sigma_\delta^2 \begin{pmatrix} 1 & S_{\delta,12} \\ S_{\delta,21} & S_\delta \end{pmatrix}\right],$$

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

$$U(s')|U, \theta \sim N\left[\mu(s'), \quad \sigma_\delta^2\left(1 - S_{\delta,12} S_\delta^{-1} S_{\delta,21}\right)\right], \tag{19}$$

where

$$\mu(s') = \alpha_0 + \alpha_1 V(s') + S_{\delta,12} S_\delta^{-1}(U - \alpha_0 - \alpha_1 v).$$

If $Z(s')$ is not inferred to be zero then we set it to be $\exp\{Y(s')\}$. If we want the predictions of the smooth deposition surface without the nugget term we simply ignore the nugget term $\epsilon(s')$ in generating $Y(s')$.

4.4 An illustration

We illustrate with weekly wet deposition data for 2001 from 120 sites monitored by the National Atmospheric Deposition Program (NADP, `nadp.sws.uiuc.edu`) in the eastern United States, see Figure 1. We analyze data from the year 2001 since this is the year for which the most recent outputs from the CMAQ model for wet chemical deposition are currently available. These outputs are available for $J = 33,390$ grid cells covering the study region. Our approach is applied separately to the wet sulfate and wet nitrate data. Since there is no need to make any simultaneous inference, a joint model is not required. There is high correlation between the two types of deposition but this is expected since both are driven by precipitation. To facilitate spatial interpolation, we also use weekly precipitation data obtained from a network of 2827 sites located inside the study region.

We model the data separately for the week of January 16-22 and May 22–28 in 2001 to make a comparison between a week in the winter and another one in the summer. Deposition data for these two weeks show significant differences according to classical t-tests, see Figure 2. This confirms the fact that the deposition levels are generally higher during the wet summer months and lower during the drier winter months, see e.g. Brook et al. (1995). However, in both the weeks there is strong linear relationship between deposition and precipitation (on the log-scale), see Figure 3. There is also some, although

not very strong, linear relationship between observed NADP data and the CMAQ output for the corresponding grid cell containing the NADP site, see Figure 4. Deposition and precipitation values that are 0 are ignored in obtaining the above Figures 3 and 4.

The spatial interpolation maps are provided in Figure 5 for sulfate and Figure 7 for nitrate. As seen in Figure 2 the model has reconstructed higher levels of both deposition types for May 22-28 than that for January 16-22. Observe that the grey scales are different for the two weeks in each of the Figures 5 and 7. In the corresponding week similar spatial patterns are seen for the sulfate and nitrate deposition values, as expected. Figures 6 and 8 provide the standard deviation maps for the predictions in Figures 5 and 7. From these figures we conclude that higher levels of deposition values are predicted with higher levels of uncertainty which is common in this sort of data analysis.

Parameter estimates, model choice analysis, and full spatio-temporal analysis of all the 52 week's dataset with a dynamic version of the foregoing model is presented in the paper of Sahu et al. (2008a), and hence are not repeated here. They also discuss methods for choosing the decay parameter values ϕ_δ, ϕ_b and ϕ_η. In addition, they validate the models with set aside data and, by suitable aggregation, obtain total annual deposition maps along with their uncertainties.

5 Further discussion

A related question of interest is to estimate dry deposition which is defined as the exchange of gases, aerosols, and particles between the atmosphere and earth's surface. Such analysis will enable prediction of 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. If successful, this effort will lead to the first ever estimation of total deposition loadings, perhaps the most critical quantity for making ecological assessments. Future work will also address trends in deposition to assess whether regulation has been successful.

Acknowledgment

The authors thank David Holland, Gary Lear and Norm Possiel of the U.S. EPA for many helpful comments and also for providing the monitoring data and CMAQ model output

used in this chapter.

Appendix A: Gaussian Processes

Gaussian processes play a key role in modeling for spatial and spatio-temporal data. By now, there is an extensive literature on formalizing and characterizing stochastic processes along with analysis their behavior. However, in a practical setting, ensuring that a stochastic process has been properly defined when the index of the process is over a continuum, say a spatial region, requires care. The primary issue is to guarantee that the joint distribution associated with the entire collection of random variables is consistently defined. The usual strategy is to define the process through its finite dimensional distributions and verify that these distributions satisfy a *consistency condition.* In this regard, the Gaussian process becomes very attractive since its finite dimensional distributions are all multivariate normals. Specification over the set D only requires a mean function, $\mu(s), s \in D$ and a valid covariance function, $C(s, s') = \text{Cov}(Y(s), Y(s'))$, the latter supplying the covariance matrix associated with any finite set of locations.

The convenient conditional distribution theory associated with the multivariate normal distribution is at the heart of kriging (spatial prediction); the convenient marginal distributions facilitate local model specification. Moreover, the only finite dimensional distributions within the class of jointly elliptical distributions that are able to support a stochastic process over a continuum are normals (or scale mixtures of normals).

Additionally, spatial dependence is typically introduced into the modeling in the form of spatial random effects. In general, random effects are modeled as normal variables so a multivariate normal specification for such effects in a spatial setting seems appropriate. In this regard, hierarchical modeling naturally emerges. The spatial random effects are introduced at the second stage of modeling. They appear in the mean (perhaps on a transformed scale if the first stage specification is non-Gaussian as in a spatial generalized linear model such as a binary process where the observation at any location is a 0 or a 1). In this regard, there is practical interest in these random effects. Given their prior specification, the associated revised posterior distributions are of interest with regard to "seeing" spatial pattern, again emphasizing the role of Bayesian inference in spatial analysis. While the genesis of spatial modeling for data over a continuum was based primarily on simple least

squares theory, modern, fully model-based spatial analysis is almost entirely done within a Bayesian framework.

Specification of a valid covariance function is a separate issue. Such a function must be positive definite, i.e., for any number of and set of spatial locations, the resultant covariance matrix must be positive definite. By now there is a rich literature regarding the choice of such functions in space, enabling isotropy, stationarity, and non-stationarity, and in space time, enabling space-time dependence in association. See, e.g., the book of Banerjee et al. (2004) and the recent paper of Stein (2005) and references therein.

Finally, Bayesian computation for space and space-time data analysis is much more demanding than usual analysis. Of course, this is true in general, for fitting hierarchical models but the rewards of full inference will usually justify the effort. Bayesian software to fit spatial data models includes Winbugs (`http://www.mrc-bsu.cam.ac.uk/bugs/`), and two R-packages Geo-R (Ribeiro and Diggle, 1999) and and SpBayes (Finley, et al., 2008).

Appendix B: Distributions for Gibbs sampling

Prior and posterior distributions:

We 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 is normally distributed with mean 0 and variance 10^3, 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.

The log of the likelihood times prior in the second stage specification up to an additive

constant is given by:

$$
\begin{aligned}
&-\tfrac{n}{2}\log(\sigma_\epsilon^2) - \tfrac{1}{2\sigma_\epsilon^2}(y-\vartheta)'(y-\vartheta) - \tfrac{n}{2}\log(\sigma_\eta^2) - \tfrac{1}{2\sigma_\eta^2}\eta' S_\eta^{-1}\eta \\
&-\tfrac{n}{2}\log(\sigma_\delta^2) - \tfrac{1}{2\sigma_\delta^2}(u-\alpha_0-\alpha_1 v)' S_\delta^{-1}(u-\alpha_0-\alpha_1 v) \\
&-\tfrac{n}{2}\log(\sigma_v^2) - \tfrac{1}{2\sigma_v^2}(v-\tilde{v}^{(1)})'(v-\tilde{v}^{(1)})' \\
&-\tfrac{J}{2}\log(\sigma_\psi^2) - \tfrac{1}{2\sigma_\psi^2}\left(x-\gamma_0-\gamma_1\tilde{v}\right)'\left(x-\gamma_0-\gamma_1\tilde{v}\right) \\
&-\tfrac{J}{2}\log(\sigma_\zeta^2) - \tfrac{1}{2}\tilde{v}' D^{-1}(I-H)\tilde{v} \\
&-\tfrac{n}{2}\log(\sigma_b^2) - \tfrac{1}{2\sigma_b^2} b' S_b^{-1} b + \log(p(\theta))
\end{aligned}
$$

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

Handling of the missing values:

Note that the transformation Equation (8) does not define a unique value of $Y(s_i)$ and in addition, there will be missing values corresponding to the missing values in $Z(s_i)$. Any missing value of $Y^*(s_i)$ is sampled from the model (12).

The sampling of the missing $U^*(s_i)$ for the precipitation process is a bit more involved. The sampling of the missing values must be done using the model (11) conditional on all the parameters. Since this model is a spatial model we must use the conditional distribution of $U^*(s_i)$ given all the $U(s_j)$ values for $j = 1, \ldots, n$ and $j \neq i$. This conditional distribution is obtained using the covariance matrix Σ_δ of δ and is omitted for brevity.

Similarly, Equation (9) does not define unique values of $X(A_j)$ when $Q(A_j) = 0$. Those values, denoted by $X^*(A_j)$, are sampled using the model Equation (13), $X^*(A_j)$ is sampled from $N(\gamma_0 + \gamma_1\tilde{v}(A_j), \sigma_\psi^2)$.

Conditional posterior distribution of θ

Straightforward calculation yields the following full conditional distributions:

$$
\begin{aligned}
&\tfrac{1}{\sigma_\epsilon^2} \sim G\left(\tfrac{n}{2}+\nu, \lambda + \tfrac{1}{2}(y-\vartheta)'(y-\vartheta)\right), \\
&\tfrac{1}{\sigma_b^2} \sim G\left(\tfrac{n}{2}+\nu, \lambda + \tfrac{1}{2} b' S_b^{-1} b\right), \\
&\tfrac{1}{\sigma_\eta^2} \sim G\left(\tfrac{n}{2}+\nu, \lambda + \tfrac{1}{2}\eta' S_\eta^{-1}\eta\right), \\
&\tfrac{1}{\sigma_\delta^2} \sim G\left(\tfrac{n}{2}+\nu, \lambda + \tfrac{1}{2}(u-\alpha_0-\alpha_1 v)' S_\delta^{-1}(u-\alpha_0-\alpha_1 v)\right), \\
&\tfrac{1}{\sigma_\psi^2} \sim G\left(\tfrac{J}{2}+\nu, \lambda + \tfrac{1}{2}\left(x-\gamma_0-\gamma_1\tilde{v}\right)'\left(x-\gamma_0-\gamma_1\tilde{v}\right)\right), \\
&\tfrac{1}{\sigma_v^2} \sim G\left(\tfrac{n}{2}+\nu, \lambda + \tfrac{1}{2}(v-\tilde{v}^{(1)})'(v-\tilde{v}^{(1)})\right), \\
&\tfrac{1}{\sigma_\zeta^2} \sim G\left(\tfrac{J}{2}+\nu, \lambda + \tfrac{1}{2}\sum_{j=1}^{J}\left\{m_j(\tilde{V}(A_j)-\mu_j)^2\right\}\right).
\end{aligned}
$$

where $\mu_j = \sum_{i=1}^{J} h_{ji}\tilde{V}(A_i)$.

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

$$\Lambda^{-1} = \frac{1}{\sigma_\epsilon^2}G'G + 10^{-3}I_3, \quad \chi = \frac{1}{\sigma_\epsilon^2}G'(y - b_0 x + Xb + \eta).$$

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

$$\Lambda^{-1} = \frac{1}{\sigma_\epsilon^2}x'x + 10^{-3}, \quad \chi = \frac{1}{\sigma_\epsilon^2}x'(y - \beta_0 - \beta_1 u - \beta_2 v - Xb - \eta).$$

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

$$\Lambda^{-1} = \frac{1}{\sigma_\epsilon^2}X'X + \Sigma_b^{-1}, \quad \chi = \frac{1}{\sigma_\epsilon^2}X'(y - \beta_0 - \beta_1 u - \beta_2 v - b_0 x - \eta).$$

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

$$\Lambda^{-1} = \frac{I_n}{\sigma_\epsilon^2} + \Sigma_\eta^{-1}, \quad \chi = \frac{1}{\sigma_\epsilon^2}(y - \beta_0 - \beta_1 u - \beta_2 v - b_0 x - Xb).$$

Let $G = (1, v)$ so that now G is a $n \times 2$ matrix. The full conditional distribution of $\alpha = (\alpha_0, \alpha_1)$ is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1} = G'\Sigma_\delta^{-1}G + 10^{-3}I_2, \quad \chi = G'\Sigma_\delta^{-1}u.$$

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

$$\Lambda^{-1} = \frac{1}{\sigma_\psi^2}G'G + 10^{-3}I_2, \quad \chi = G'x.$$

Conditional posterior distribution of V

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

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

where $a = y - \beta_0 - \beta_1 u - b_0 x - Xb - \eta$. From this n-dimensional joint distribution we obtain the conditional distribution $V(s_i) \sim N(\mu_i, \Xi_i)$, say. If the precipitation value, $r(s_i)$, is missing then there will be no constraint on $V(s_i)$ and we sample $V(s_i)$ unrestricted from $N(\mu_i, \Xi_i)$. If on the other hand the observed precipitation value is zero, $r(s_i) = 0$, we must

sample $V(s_i)$ to be negative, i.e we sample from $N(\mu_i, \Xi_i)I(V(s_i) < 0)$. Corresponding to non-zero precipitation value $r(s_i) > 0$ we sample $V(s_i)$ from $N(\mu_i, \Xi_i)I(V(s_i) > 0)$.

Conditional posterior distribution of $\tilde{V}$

The full conditional distribution of $\tilde{V} = (\tilde{V}^{(1)}, \tilde{V}^{(2)})$ is $N(\Lambda\chi, \Lambda)$ where

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

$$\chi = \begin{pmatrix} \frac{1}{\sigma_v^2} v \\ 0 \end{pmatrix} + \frac{\gamma_1}{\sigma_\psi^2}(x - \gamma_0).$$

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 (10).

The partitioning of $\tilde{V}$, however, suggests an immediate univariate sampling scheme detailed below. First, note that the conditional prior distribution for $\tilde{V}(A_j)$ from the vectorized specification (16), as calculated above is given by $N\left(\xi_j, \omega_j^2\right)$ where:

$$\omega_j^2 = \sigma_\zeta^2 \frac{1}{m_j} \quad \text{and} \quad \xi_j = \sum_{i=1}^{J} h_{ji} \tilde{v}(A_i).$$

Now for each component $\tilde{V}(A_j)$ of $\tilde{V}^{(1)}$ we extract the full conditional distribution to be viewed as the likelihood contribution from the joint distribution $N(\Lambda_{(1)}\chi_{(1)}, \Lambda_{(1)})$ where

$$\Lambda_{(1)}^{-1} = \frac{I_n}{\sigma_v^2} + \gamma_1^2 \frac{I_n}{\sigma_\psi^2} \quad \text{and} \quad \chi_{(1)} = \frac{1}{\sigma_v^2} v + \frac{\gamma_1}{\sigma_\psi^2}(x^{(1)} - \gamma_0),$$

where $x = (x^{(1)}, x^{(2)})$, partitioned analogously to $\tilde{V}$. This conditional likelihood contribution is given by $N\left(\mu_j, \Xi^2\right)$ where

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

The conditional likelihood contribution for each component of $\tilde{V}^{(2)}$ is the normal distribution $N\left(\mu_j, \Xi^2\right)$ where

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

Now the un-constrained full conditional distribution of $\tilde{V}(A_j)$, according to the second stage likelihood and prior specification, is obtained by combining the likelihood contribution

$N\left(\mu_j, \Xi^2\right)$ and the prior conditional distribution $N\left(\xi_j, \omega_j^2\right)$ and is given by $N\left(\Lambda_j \chi_j, \Lambda_j\right)$ where

$$\Lambda_j^{-1} = \Xi^{-2} + \omega_j^{-2}, \quad \chi_j = \Xi^{-2}\mu_j + \omega_j^{-2}\xi_j.$$

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

References

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

Bergthorsson, P. and Döös, B. (1955). Numerical weather map analysis *Tellus 7*, 329-340.

Berrocal, V. J. Gelfand, A. E. and Holland, D. M. (2008). A spatio-temporal downscaler for output from numerical models. *Submitted*

Bratseth, A. M. (1986). Statistical interpolation by means of successive corrections, *Tellus 38A*, 439-447.

Brook, J. R., Samson, P. J. and Sillman, S. (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. *Journal of Applied Meteorology 34*, 297-325.

Brown, P. J., Le, N. D., Zidek, J. V. (1994). Multivariate spatial interpolation and exposure to air pollutants. *The Canadian Journal of Statistics 22*, 489-510.

Burke, J. M., Vedantham,R., McCurdy, T.R., Xue, J.,and A. H. Ozkaynak. A.H. (2003). SHEDS-PM: A Population Exposure Model for Predicting Distributions of PM Exposure and Dose From Both Outdoor and Indoor Sources. *Presented at International Society of Exposure Analysis*, Stresa, Italy, September 21-25, 2003.

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. *Journal of the American Statistical Association 92*, 392–404.

Charney, J.G. (1951). Dynamic forecasting by numerical process, *Compendium of Meteorology* American Meteorological Society, Boston, MA

Cox, D. D., Park, J. S. and Singer C. E.(2001). A statistical method for tuning a computer code to a data base. *Computational Statistics and Data Analysis 37*, 77–92.

Cressie, N., Kaiser, M. S., Daniels, M. J., Aldworth, J., Lee, J., Lahiri, S. N., Cox, L. (1999). Spatial Analysis of Particulate Matter in an Urban Environment. In *GeoEnv II: Geostatistics for Environmental Applications*, (Eds. J. Gmez-Hernndez, A. Soares, R. Froidevaux)) Kluwer:Dordrecht, pp 41-52.

Cressman, G.P. (1959). An operational objective analysis system, *Monthly Weather Review 87*, 367-374.

Finley, A. O., Banerjee, S. and Carlin, B. P. (2008). Univariate and multivariate spatial modeling. Technical Report, Department of Bio-statistics, University of Minnesota, `http://blue.for.msu.edu/software`.

Fuentes M., Guttorp P., and Challenor, P. (2003). Statistical assessment of numerical models. *International Statistical Review 71*, 201-221.

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.

Guttorp, P., Meiring, W. and Sampson, P. D. (1994). A Space-time Analysis of Ground-level Ozone Data. *Environmetrics 5*, 241-254.

Haas, T. C. (1995). Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association 90*, 1189-1199.

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press: Cambridge.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B 63*, 425-464.

Kibria, B. M. G., Sun, L., Zidek, J. V., and Le, N. D. (2002). Bayesian spatial prediction of random space-time fields with application to mapping PM2.5 exposure. *Journal of the American Statistical Association 97*, 112-124.

Le, N. D., Sun, W., Zidek, J. V. (1997). Bayesian multivariate spatial interpolation with data missing by design. *Journal of the Royal Statistical Society, Series B 59*, 501-510.

McMillan, N., Holland, D., Morara, M. and Feng, J. (2008). Combining numerical model output and particulate data using Bayesian space-time modeling, *Environmetrics*, to appear.

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation models: The Bayesian melding approach *Journal of the American Statistical Association 95*, 1244-1255.

Rappold, A. G., Gelfand, A. E. and Holland, D. M. (2008). Modeling mercury deposition through latent space-time processes. *Journal of the Royal Statistical Society, Series C 57*, 187-205.

Ribeoro Jr, P.J. and Diggle, P.J. (1999). geoS: A geostatistical library for S-PLUS. Technical report ST-99-09, Dept of Maths and Stats, Lancaster University, `http://www.leg.ufpr.br/geoR/`.

Richmond, H.M., Palma T., Langstaff J., McCurdy T., Glenn G., and Smith L. (2002). Further Refinements and Testing of APEX (3.0): EPA's Population Exposure Model for Criteria and Air Toxic Inhalation Exposures. Joint Meeting of the Society of Exposure Analysis and International Society of Environmental Epidemiology, Vancouver, Canada.

Sahu, S. K. and Mardia, K. V. (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. K., Gelfand, A. E. and Holland, D. M. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological and Environmental Statistics 11*, 61-86.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2007). High Resolution Space-Time Ozone Modeling for Assessing Trends. *Journal of the American Statistical Association 102*, 1221-1234.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2008a). Fusing point and areal space-time data with application to wet deposition. *Submitted*

Sahu, S. K., Yip, S. and Holland, D. M. (2008b). Improved space-time forecasting of next day ozone concentrations in the eastern US. *Atmospheric Environment*, to appear.

Shaddick, G. and Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society, Series C 51*, 351-372.

Smith, R. L., Kolenikov, S. and Cox, L. H. (2003). Spatio-Temporal modelling of $PM_{2.5}$ data with missing values. *Journal of Geophysical Research-Atmospheres 108*, D249004, doi:10.1029/2002JD002914.

Stein, M.L. (2005). Space-time covariance functions. *Journal of the American Statistical Association 100*, 310-321.

Sun L., Zidek, J. V., Le, N. D. and Ozkaynak, H. (2000). Interpolating Vancouver's daily ambient PM10 field. *Environmetrics 11*, 651-663.

Wikle, C. K. (2003). Hierarchical models in environmental science. *International Statistical Review 71*, 181-199.

Zidek, J. V., Shaddick, G., White, R., Meloche, J., and Chatfield, C. (2005). Using a probabilistic model (pCNEM) to estimate personal exposure to air pollution Context Sensitive Links. *Environmetrics 16*, 481-493.

Zidek J. V., Shaddick, G., Meloche, J., Chatfield, C., and White, R. (2007). A framework for predicting personal exposures to environmental hazards. *Environmental and Ecological Statistics 14*, 411-431.

Zidek, J. V., Sun, L., Le, N., Ozkaynak, H.(2002). Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM_{10} field. *Environmetrics 13*, 595-613.

Figure 1: A map of the eastern U.S. with the 120 NADP sites plotted as points.

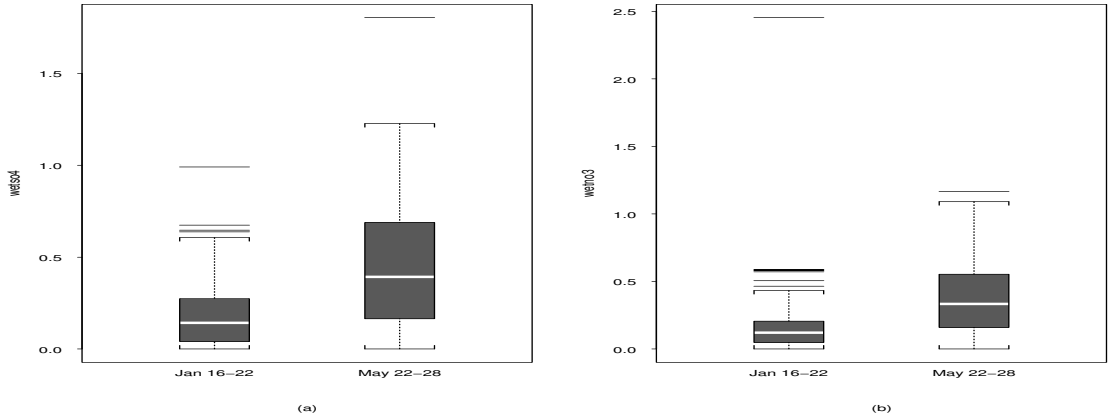

Figure 2: Boxplot of weekly depositions: (a) wet sulfate and (b) wet nitrate.

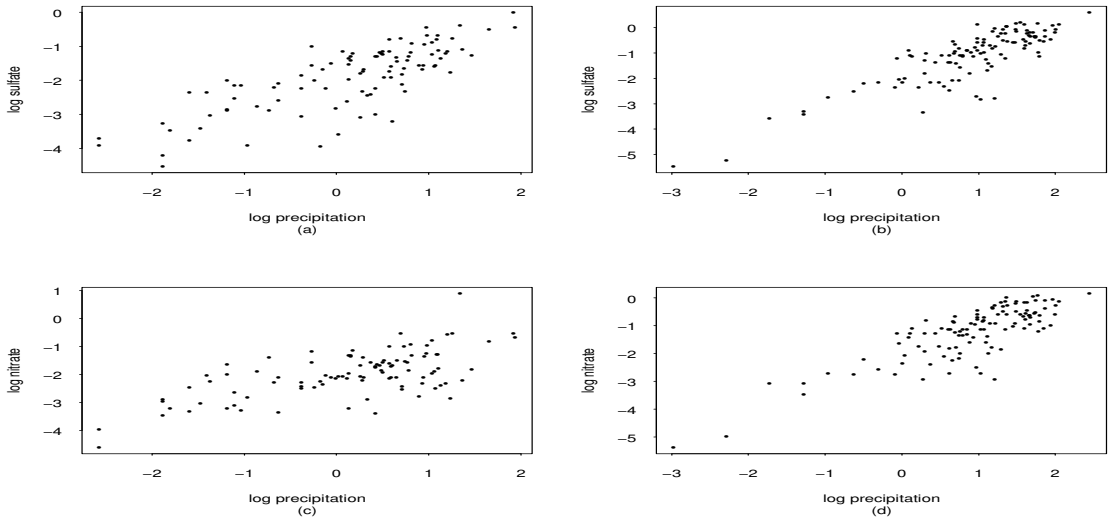

Figure 3: Plot of log deposition against log precipitation: (a) wet sulfate for January 16-22 (b) wet sulfate for May 22–28, (c) wet nitrate for January 16-22 (d) wet nitrate for May 22–28.

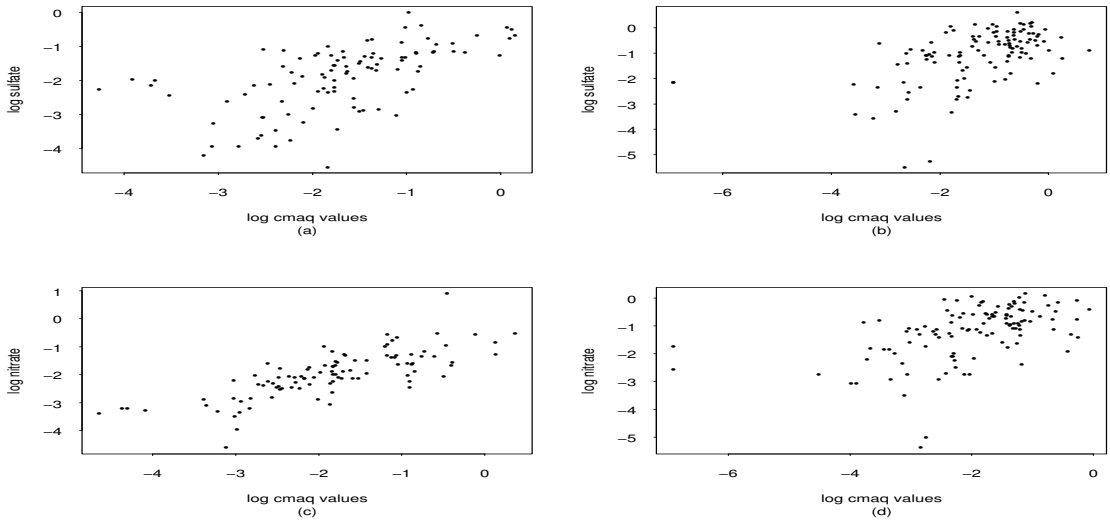

Figure 4: Plot of log deposition against log CMAQ value for the cell containing the corresponding NADP site: (a) wet sulfate for January 16-22 (b) wet sulfate for May 22–28, (c) wet nitrate for January 16-22 (d) wet nitrate for May 22–28.

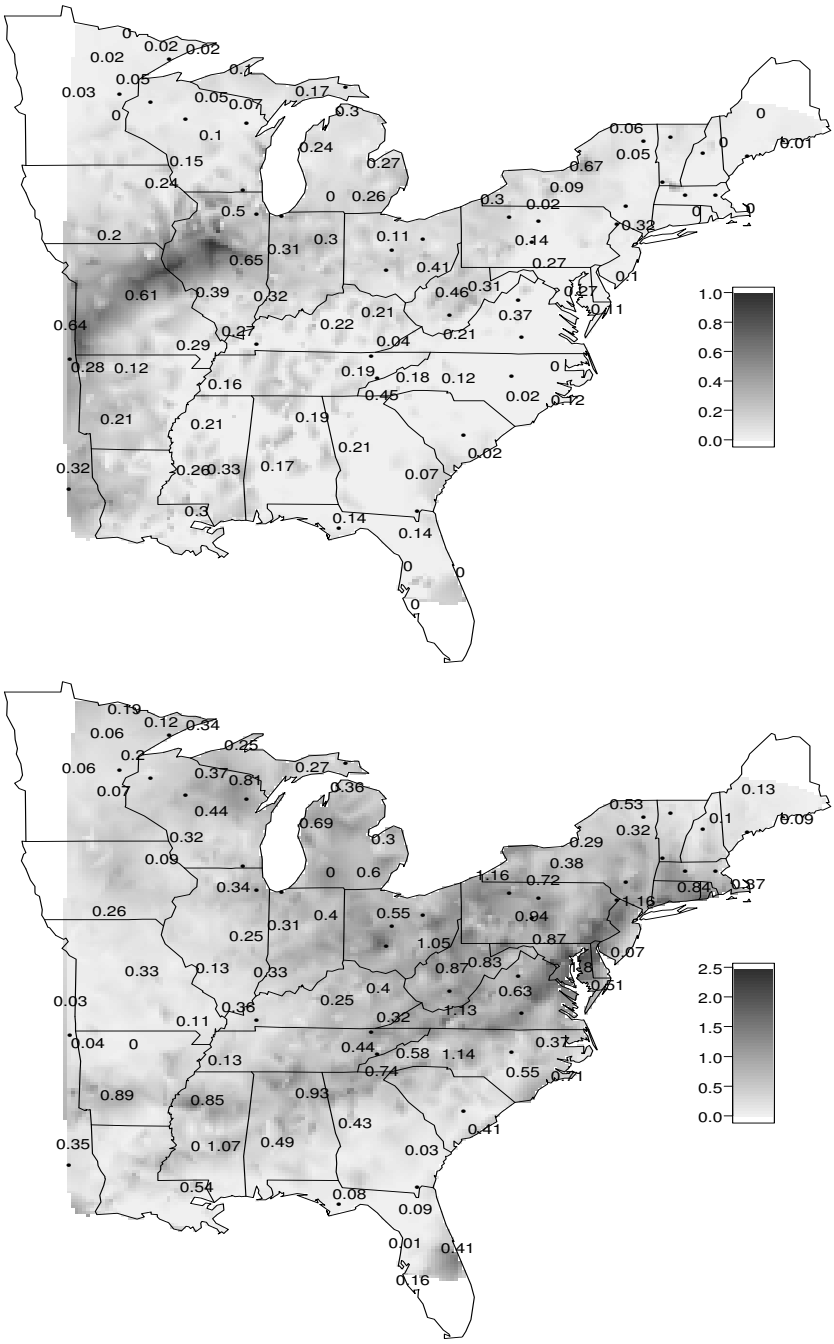

Figure 5: Analyses for sulfate. Top panel is for the model interpolated map for January 16-22 and the same map for May 22-28 is given in the bottom panel. Observed deposition values from some selected sites are superimposed. (For visual clarity we present only a subset of the monitoring data.)

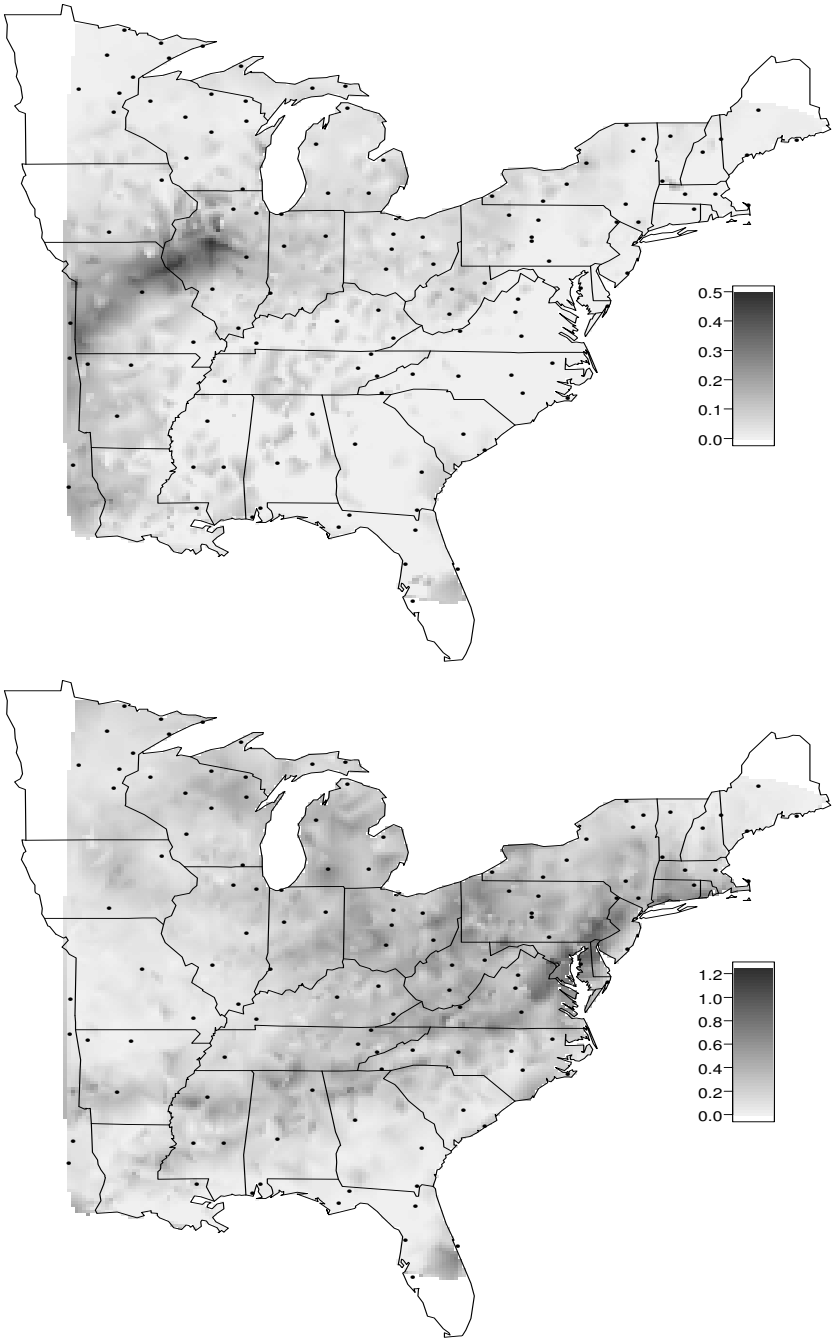

Figure 6: The standard deviation maps for the predictions in Figure 5.

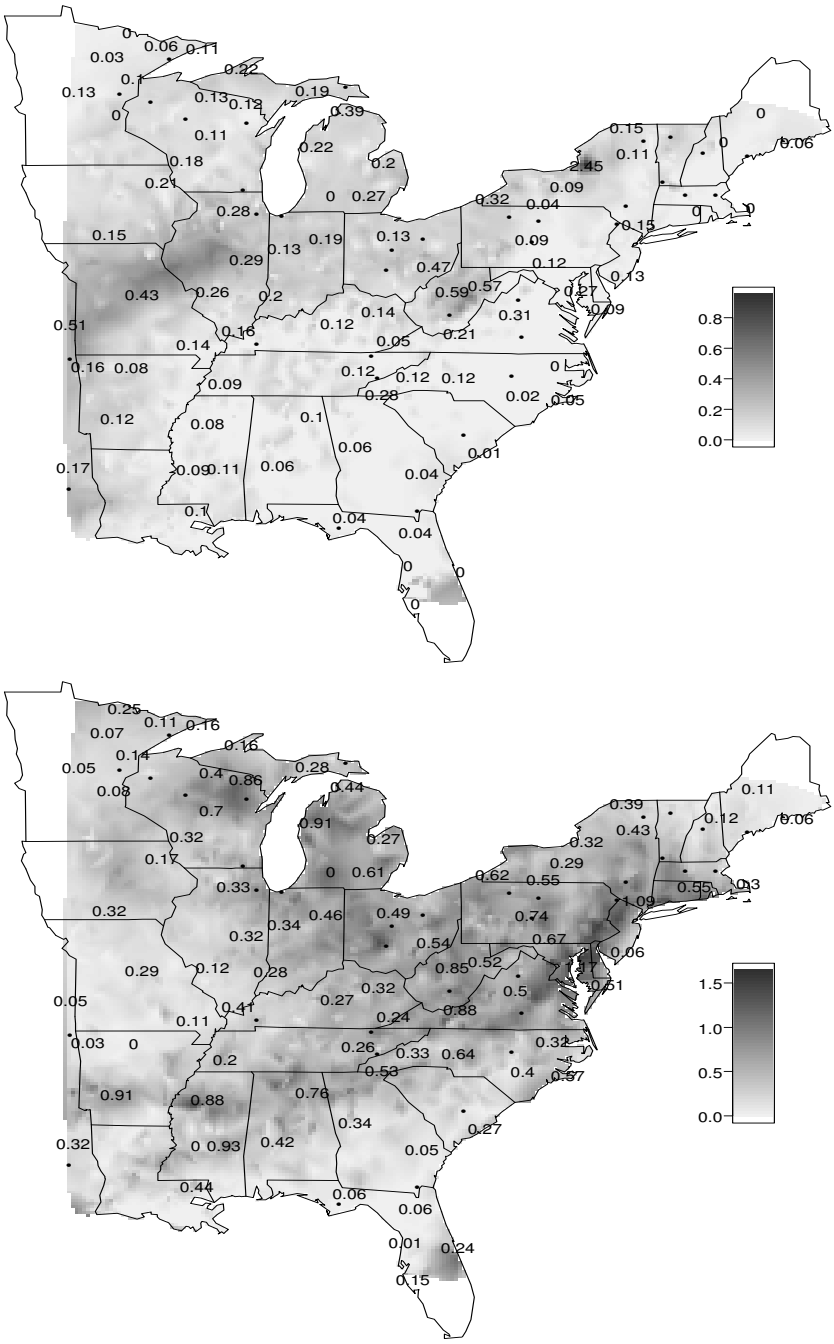

Figure 7: Analyses for nitrate. Top panel is for the model interpolated map for January 16-22 and the same map for May 22-28 is given in the bottom panel. Observed deposition values from some selected sites are superimposed. (For visual clarity we present only a subset of the monitoring data.)

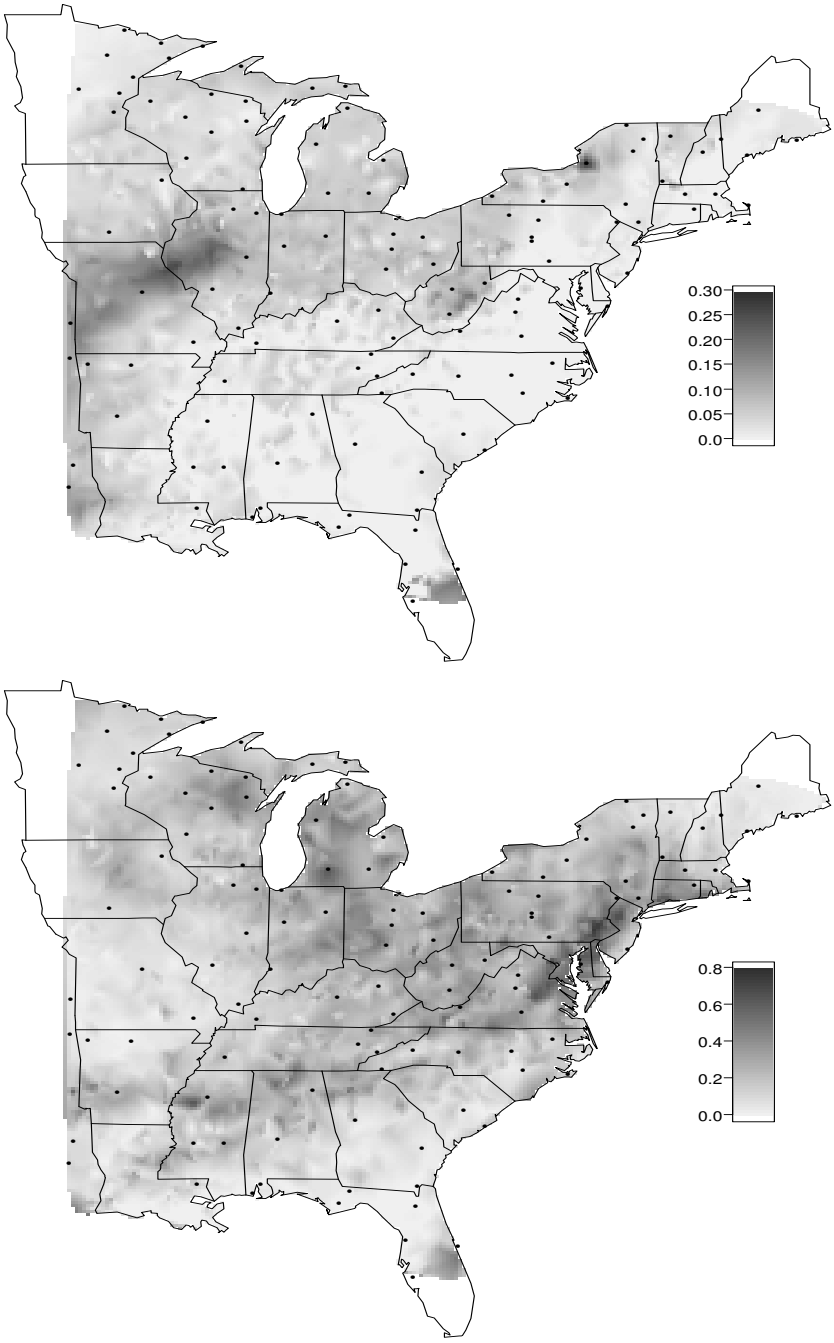

Figure 8: The standard deviation maps for the predictions in Figure 7.