Received: 19 April 2015, Revised: 8 September 2015, Accepted: 9 September 2015, Published online in Wiley Online Library: 26 October 2015

(wileyonlinelibrary.com) DOI: 10.1002/env.2369

# A data fusion approach for spatial analysis of speciated $PM_{2.5}$ across time

**Colin W. Rundel[a*], Erin M. Schliep[b], Alan E. Gelfand[a] and David M. Holland[c]**

**$PM_{2.5}$ exposure is linked to a number of adverse health effects such as lung cancer and cardiovascular disease. However, $PM_{2.5}$ is a complex mixture of different species whose composition varies substantially in both space and time. An open question is how these constituent species contribute to the overall negative health outcomes seen from $PM_{2.5}$ exposure. To this end, the Environmental Protection Agency as well as other federal, state, and local organization monitor total $PM_{2.5}$ along with its primary species on a national scale. From an epidemiological perspective, there is a need to develop effective methods that will allow for the spatially and temporally sparse observations to be used to predict exposures for locations across the entire United States.**

**Toward this objective, we have collected data from three separate monitoring station networks as well as output from a deterministic atmospheric computer model. We introduce a novel multi-level speciated $PM_{2.5}$ model, which captures the following features: (1) it fuses data from three monitoring networks; (2) it simultaneously models each of the five primary components of $PM_{2.5}$ from each network along with the computer model output; (3) it introduces species and network level measurement error models as well as total $PM_{2.5}$ measurement error models, all varying around the respective latent *true* levels; (4) it incorporates an unobserved "other" species component as well as a sum constraint such that the total is physically consistent (i.e., total must be equal to the sum of the primary species and "other"), which is not always the case with the observed data. Copyright © 2015 John Wiley & Sons, Ltd.**

**Keywords:** downscaling; latent process; Markov chain Monte Carlo; multi-level model; tobit (truncated) Gaussian process

## 1. INTRODUCTION

Epidemiologic and exposure studies pertaining to the health effects of total particulate matter ($PM_{2.5}$) have led the U.S. Environmental Protection Agency (US EPA) to establish a mass-based ambient air quality standard for this pollutant. $PM_{2.5}$ is a complex mixture of different species and that composition varies with season and location. Air managers and the broader public would like to understand which of the different species are most strongly connected to various adverse health effects. As a result, recent health analyses have focused on identifying strong associations between specific species and health outcomes. Thus, there is a need to develop reliable prediction of speciated $PM_{2.5}$ over space, particularly across the conterminous United State. This will facilitate understanding of linkages between speciated $PM_{2.5}$ and health effects as well as success of emission reduction programs designed to control those species that most contribute to public health issues.

Our contribution here is to develop a novel, process-motivated multi-level model for speciated $PM_{2.5}$. First, the model fuses data from three monitoring station networks. These networks, described in more detail below, are the large-scale $PM_{2.5}$ federal reference monitoring network (FRM), the smaller urban chemical speciation network (CSN), and the rural speciated $PM_{2.5}$ interagency monitoring of protected visual environments (IMPROVE) network. Also included in the model is gridded 12 km output from the community multi-scale air quality (CMAQ) numerical atmospheric model. Second, we model each of the five primary species of $PM_{2.5}$—sulfate, nitrate, total carbonaceous matter, ammonium, and fine soil—through fusion of CSN, IMPROVE, and CMAQ and model total $PM_{2.5}$ through FRM, CSN, IMPROVE, and CMAQ. Third, the model introduces species level measurement error models, as well as total $PM_{2.5}$ measurement error models, all varying around the respective latent true levels. Lastly, a latent "other" species component is included to ensure that the totals are physically consistent. That is, that total $PM_{2.5}$ is always at least the sum of the primary species and is non-negative. This sum constraint on the true levels is particularly important as the observed data frequently violate this condition. Work of Calder (2008) is also in this spirit. To improve interpolation of $PM_{2.5}$, she incorporates $PM_{10}$ data into the model. She creates a model for the fine PM (particle size at most 2.5 $\mu$m) and the coarse PM (particle size between 2.5 and 10 $\mu$m) with a sum constraint to $PM_{10}$.

* *Correspondence to: Colin W. Rundel, Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708, U.S.A. E-mail: rundel@gmail.com*

a *Department Of Statistical Science, Duke University, Durham, NC 27708, U.S.A.*

b *Department of Statistics, University of Missouri, Columbia, MO 65211, U.S.A.*

c *U.S. Environmental Protection Agency, National Exposure Research Laboratory, Research Triangle Park, NC 27711, U.S.A.*

All models for the true levels are spatial, anticipating spatial dependence in speciated $PM_{2.5}$. The computer model output is incorporated through downscaling, which provides local calibration of the CMAQ output. We model at the weekly scale (higher temporal resolution is not possible because of the sampling frequency limitations of the monitoring networks) enabling convenient aggregation to seasonal or annual scale. These models are fitted to spatial data at the continental scale, allowing efficient prediction/interpolation at any location as well as week. The resulting model is hierarchical and is fitted within a Bayesian framework using a challenging Markov chain Monte Carlo (MCMC) algorithm. Altogether, this is the first effort that employs all four data sources to provide a coherent fusion to infer about these primary species as well as the total $PM_{2.5}$.

Integrated 24 h measurements of speciated $PM_{2.5}$ are routinely provided by the mostly urban CSN and the mostly rural IMPROVE network. The CSN monitors are operated by state and local agencies, while IMPROVE sites are operated by the National Park Service. The mixture of different chemical species, which make up $PM_{2.5}$, varies by season and location because of differences in emissions and weather conditions that drive the formation and transport of $PM_{2.5}$. In addition to the primary species listed above, $PM_{2.5}$ also contains additional ions and elements which we will refer to as "other".

Time-series analyses have been used to investigate the influence of speciated $PM_{2.5}$ components on daily mortality (Lippmann *et al.*, 2006; Dominici *et al.*, 2007; Franklin *et al.*, 2008) and on hospital admissions (Bell *et al.*, 2013). More recently, Ito *et al.* (2013) investigated time series analyses to overcome the potential limitations of the time-varying sampling frequency of CSN data. Rich *et al.* (2013) used a statistical space-time bias-correction model for CMAQ to evaluate whether the relative odds of transmural myocardial infarction associated with increased $PM_{2.5}$ concentration is modified by fractions of the major $PM_{2.5}$ species by requiring predicted concentrations to be positive and the sum of the species to equal total $PM_{2.5}$. Kim *et al.* (2013) investigated the utility of using CSN and IMPROVE monitoring data for epidemiologic applications. They note the sparse spatial coverage of each network, and that the differences in analysis and sampling protocols may affect their utility in developing exposure predictions and resulting estimates of health effects. Choi *et al.* (2009) presented a multivariate spatial-temporal model for speciated $PM_{2.5}$ using a Bayesian hierarchical framework with spatio-temporal varying coefficients. They use speciated monitoring data from CSN and IMPROVE and, adopting a compositional data perspective, express the mean of each $PM_{2.5}$ component as a proportion of the total $PM_{2.5}$, ensuring that the sum of the components or species equals the total but losing the actual levels of exposure. A coregionalized spatio-temporal model is used to account for the dependency structures of all species, covariance parameters within a complex hierarchical model with many variance components.

The statistical literature on data fusion can be grouped into two distinct approaches. One is Bayesian melding (Fuentes and Raftery, 2005) where observed data are "fused" with gridded atmospheric model output by using latent point-level processes for both sources of spatial information. The numerical model output is then expressed as a linearly calibrated integral over a grid cell of the latent point-level process while the monitoring data is related to the latent process via a measurement error model. McMillan *et al.* (2010) describe a grid cell level upscaling spatio-temporal fusion model that provides predictions at the grid cell level.

The second approach uses a two-stage regression, originating with Guillas *et al.* (2008). Liu *et al.* (2008) present an ad hoc method to allow the coefficients of the linear regression to be spatially interpolated. Berrocal *et al.* (2010b,2010a) propose univariate and bivariate downscaler models that use spatial regression to relate the point monitoring data to the gridded numerical model output with spatially varying coefficients modeled as Gaussian processes. Berrocal *et al.* (2012) extended this model to allow adaptive smoothing of the numerical model output to achieve stronger association with the monitoring data.

The paper is organized as follows. In Section 2, we discuss the data sources used in this paper. Section 3 focuses on the model specification, model fitting, and prediction under the model. Details of the MCMC model fitting are supplied in an appendix. Section 4 provides the data analysis results under our models. We conclude with Section 5 giving a brief summary and a view toward future work.

## 2. THE DATA

We fuse ambient pollution data from four distinct sources: monitoring data from the CSN, IMPROVE, FRM monitoring networks, and gridded numerical model output from the CMAQ model. The monitoring locations of the three networks are shown in Figure 1. We focus on modeling total $PM_{2.5}$ (in $\mu g/m^3$) and the five major $PM_{2.5}$ species: sulfate, nitrate, total carbonaceous matter (defined as 1.4[Organic C] + [Elemental C]), ammonium, and fine soil or crustal material (defined as 2.2[Al] + 2.49[Si] + 1.63[Ca] + 2.42[Fe] + 1.94[Ti]). For the IMPROVE network, ammonium concentration is *inferred* from nitrate and sulfate concentrations (defined as $18/62[NO_3] + 36/96[SO_4]$).

The speciated particulate CSN (http://epa.gov/ttnamti1/speciepg.html) is a companion network to the total particulate FRM network that was implemented to support the $PM_{2.5}$ national air quality standard. The US EPA established CSN to provide national measurements of speciated $PM_{2.5}$ data at mostly urban locations. For as part of the routine monitoring operation, the CSN quantifies daily total fine particulate mass concentrations and most of the $PM_{2.5}$ species, including numerous trace elements, ions (e.g. sulfate, nitrate, and ammonium), and elemental and organic carbon. Speciated data are collected on a 3-day (38% of stations) or 6-day (62%) cycle depending on location. In this analysis, we use data from 221 CSN monitoring locations.

The second source of data is the IMPROVE network (http://vista.cira.colostate.edu/improve/). This network is smaller compared with CSN, consisting of 172 sites at mostly rural locations. The network produces observations of total $PM_{2.5}$ and the speciated components every 3 days. It was established in 1987 among federal and state monitoring agencies to provide information for determining the types of pollutants and sources responsible for visibility impairment in US National Parks and US wilderness areas. So CSN has an urban focus, while IMPROVE addresses monitoring remote environments. Each network uses different physical samplers and analytical protocols.

To supplement the speciated and total fine particulate mass data from these two networks, we use total $PM_{2.5}$ from the large FRM network (http://epa.gov/ttnamti1/pmfrm.html) consisting of 949 sites reporting on a daily (19% of stations), 3-day (68%) or 6-day (13%) monitoring cycle. This network monitors a variety of urban, suburban, and rural environments (with the major focus on urban areas) in support of the national air quality standard for $PM_{2.5}$. Although FRM does not monitor speciated $PM_{2.5}$, the total particulate mass values provide

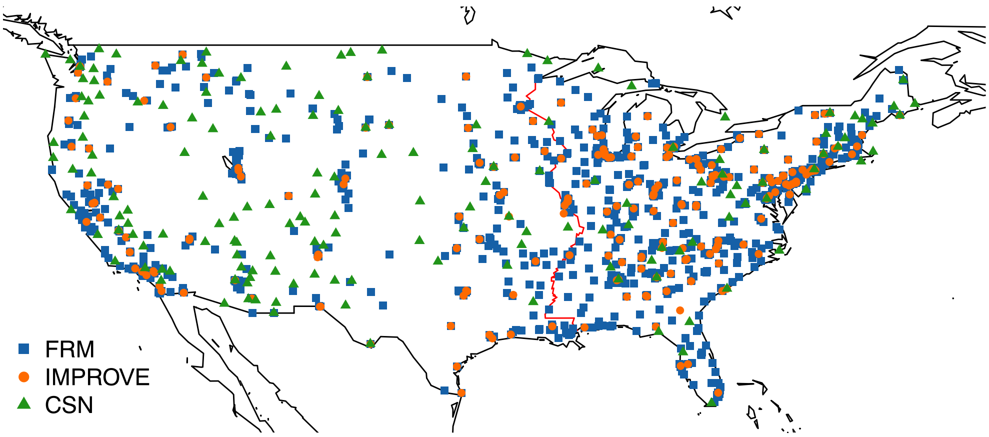

**Figure 1.** Location of chemical speciation network (CSN), interagency monitoring of protected visual environments (IMPROVE), and federal reference method (FRM) monitoring stations in 2007

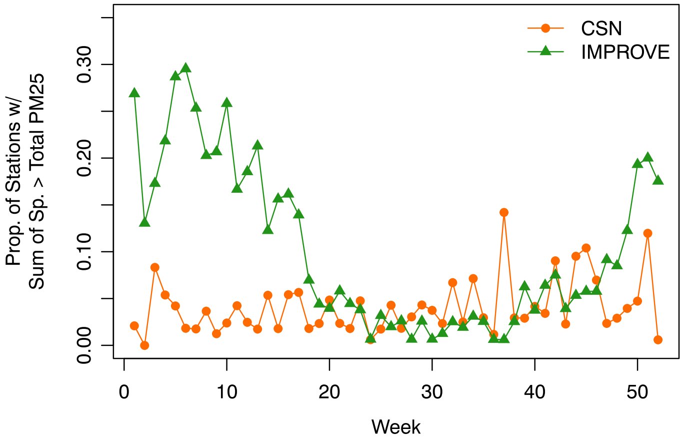

**Figure 2.** Comparison of the proportion of stations reporting a sum of $PM_{2.5}$ species that exceeds the observed total $PM_{2.5}$ for each week and network in 2007

additional information with regard to the constraint that the sum of the latent true levels for the five major species does not exceed the latent true total mass concentration.

The final source of data used in this analysis is the gridded numerical model output from the CMAQ model Version 4.7 (Foley *et al.*, 2009). This model employs Weather Research and Forecast model meteorology with gridded emissions of primary $PM_{2.5}$ and precursors to secondary $PM_{2.5}$. The National Emissions Inventory is the primary source for the emissions data. Aerosol transport, atmospheric chemistry, and secondary $PM_{2.5}$ formation are simulated to provide the CMAQ $PM_{2.5}$ species concentrations. Illustratively, we work with data from the year 2007 using 12 km grids, covering the conterminous United States. For each species, association between CMAQ and monitoring data is shown in Supporting information Figure 1.

Given the diversity of data used in this analysis, we immediately encounter several modeling challenges. Although CSN, IMPROVE, and FRM mostly report on the same sampling schedule, there is considerable temporal misalignment among monitoring sites (Supporting information Figure 2). Speciated data is not available from either CSN or IMPROVE on a daily basis, and the majority of CSN stations only report every 6 days. To avoid these issues, we model $PM_{2.5}$ total and species on a weekly scale whereby we average all CMAQ output and monitoring observations within each week. An additional challenge is that total recorded particulate mass can be less than the sum of the five major species defined previously. For 2007, the incidence of this issue is presented over all weeks by each monitoring network in Figure 2; we see that incidence is frequent. However, the amount of the disparity is generally small, with 80% of occurrences having less than 1 $\mu g/m^3$ of excess mass, and 98% having less than 5 $\mu g/m^3$.

## 3. STATISTICAL MODELS

In Section 3.1, we present the details of the complex and notationally demanding statistical model. In Section 3.2, we discuss priors and model fitting. In Section 3.3, we consider kriging (spatial prediction) of “true” and observed surfaces under this model, commenting briefly on assessment of model validation and model selection.

### 3.1. Specifying the model

We specify the data stage through measurement error models. Let $C_t^i(\boldsymbol{s})$ and $I_t^i(\boldsymbol{s})$ be the observed average monitoring value of PM$_{2.5}$ species $i$ at locations $\boldsymbol{s}$ for week $t$ for the CSN and IMPROVE networks, respectively. We denote the PM$_{2.5}$ species using superscripts: Sulfate = "1", Nitrate = "2", Ammonium = "3", Soil = "4", Carbon = "5", and Total = "tot". We use the superscript "o" to capture all of the "other" unmeasured species (e.g. trace elements (K, Mg, Ca), heavy metal (Cu, Fe), etc.) that, on average, account for 23% of total PM$_{2.5}$. Weekly average total PM$_{2.5}$ from the CSN and IMPROVE networks are denoted by $C_t^{tot}(\boldsymbol{s})$ and $I_t^{tot}(\boldsymbol{s})$, respectively, and by $F_t^{tot}(\boldsymbol{s})$ for the FRM network.

Under our joint model, we assume that all observed values are noisy observations of an underlying and unobserved true field. That is, let $Z_t^i(\boldsymbol{s})$ be the true level of the PM$_{2.5}$ species $i$ at locations $\boldsymbol{s}$ for week $t$. Additionally, at time $t$ and location $\boldsymbol{s}$, let $Z_t^o(\boldsymbol{s})$ denote the true level of the total of all of the "other" unmeasured species and let $Z_t^{tot}(\boldsymbol{s})$ denote the true total of PM$_{2.5}$. We have

$$Z_t^{tot}(\boldsymbol{s}) = \sum_{i=1}^{5} Z_t^i(\boldsymbol{s}) + Z_t^o(\boldsymbol{s}) \tag{1}$$

to link the latent species levels to the latent total level. Then, for the observations, with $i = 1, 2, ..., 5$,

$$C_t^i(\boldsymbol{s}) = Z_t^i(\boldsymbol{s}) + \epsilon_{C,t}^i(\boldsymbol{s}) \tag{2}$$

$$I_t^i(\boldsymbol{s}) = Z_t^i(\boldsymbol{s}) + \epsilon_{I,t}^i(\boldsymbol{s}) \tag{3}$$

and

$$C_t^{tot}(\boldsymbol{s}) = Z_t^{tot}(\boldsymbol{s}) + \epsilon_{C,t}^{tot}(\boldsymbol{s}) \tag{4}$$

$$I_t^{tot}(\boldsymbol{s}) = Z_t^{tot}(\boldsymbol{s}) + \epsilon_{I,t}^{tot}(\boldsymbol{s}) \tag{5}$$

$$F_t^{tot}(\boldsymbol{s}) = Z_t^{tot}(\boldsymbol{s}) + \epsilon_{F,t}^{tot}(\boldsymbol{s}) \tag{6}$$

The measurement error formulation earlier adopts weekly variance components. That is, all of the $\epsilon$'s are independent with distinct variance components. Altogether there are 13 variance components; $5 \times 2$ for the species plus 3 for the totals for each week. There is no observed "o", so no additional measurement error modeling is needed. This formulation does not incorporate dependence between any of the measurement error terms, even for species measurements made at the same monitoring location. This assumption is based on the fact that physical sampling and measurement processes for each species and for total PM are distinct and independent at all of the CSN and IMPROVE monitoring locations.

We model the $Z^i$'s and $Z^o$'s, which, under the sum constraint, yields implicit realizations for the $Z^{tot}$'s. We build regression models for the $Z^i$'s in the form of downscalers to incorporate CMAQ output in the model. We follow the approach of Berrocal *et al.* (2010b) to take advantage of the fact that there are 97,416 CMAQ grid cells for the continental United States, while we have only 221, 172, and 949 monitoring sites for CRM, IMPROVE, and FRM, respectively. Our model differs from Berrocal *et al.* (2010b) in that our downscaling is latent (i.e., the $Z$'s are not observed); we are downscaling CMAQ to the *true* species concentrations, not directly to the data.

The concentrations of each PM$_{2.5}$ species, that is, all $Z$'s must be non-negative. There are two customary ways to accomplish this. One is through a tobit transformation ($U = \mathtt{max}(0, \tilde{U})$ where $\tilde{U}$ is a normally distributed random variable); the second is through a log-normal. We have explored both in substantial detail, concluding that, for the species ($i$'s); the tobit model is a better choice. Specifically, we favor the tobit model over the log model because its density at 0 is not forced to be 0, which accords well with the observed species data (a large number of very small observed values, Figure 3). Additionally, the tobit model provides better behaved downscalers than those on the exponential scale (arising from a normal distribution for log exposures). Lastly, model comparison, using hold out stations, generally led to preference for the tobit model. Hence, in the interest of space, we present only the results for the tobit specification.

The tobit model is implemented by defining $Z_t^i(s) = \max\,(0, \tilde{Z}_t^i(s))$ where

$$\tilde{Z}_t^i(\boldsymbol{s}) = \beta_{0,t}^i + \beta_{0,t}^i(\boldsymbol{s}) + \beta_{1,t}^i\ Q_t^i(B_{\boldsymbol{s}}) \tag{7}$$

Here, $\beta_{0,t}^i$ and $\beta_{1,t}^i$ serve as additive and scale bias adjustments to the CMAQ prediction with $\beta_{0,t}^i(\boldsymbol{s})$ providing local spatial adjustment to the intercept. For the grid block $B_{\boldsymbol{s}}$, containing location $s$, the weekly average CMAQ output on week $t$ is denoted by $Q_t^i(B_{\boldsymbol{s}})$. We model $\beta_{0,t}^i(\boldsymbol{s})$ as a Gaussian process with exponential covariance, that is, $\beta_{0,t}^i(\boldsymbol{s}) = \sqrt{\sigma_t^{2\,i}}\ w_t^i(\boldsymbol{s})$ where $w_t^i(\boldsymbol{s})$ is a zero mean, unit variance, Gaussian process with $\mathrm{corr}(w_t^i(\boldsymbol{s}), w_t^i(\boldsymbol{s}')) = \exp(-\phi_t^i|\boldsymbol{s} - \boldsymbol{s}'|)$. Note that there is no pure error added in (7). The latent $\tilde{Z}$'s provide a smooth spatial surface. Measurement error is introduced at the first stage in (2)-(6).

While it would be attractive to introduce spatially varying slopes, $\beta_{1,t}^i(\boldsymbol{s})$ (centered around $\beta_{1,t}^i$) correlated with the $\beta_{0,t}^i(\boldsymbol{s})$, this would require the introduction of a bivariate process model (Wackernagel, 2003; Schmidt and Gelfand, 2003; Gelfand *et al.*, 2004). Previous work (Fuentes and Raftery, 2005; Berrocal *et al.*, 2010b) has shown that, in practice, the $\beta_{1,t}(\boldsymbol{s})$'s are difficult to identify and that the deviations, $\beta_{1,t}^i(\boldsymbol{s}) - \beta_{1,t}^i$ are rarely significantly different from 0. This is apart from the computational complexity they would introduce in fitting the foregoing model. Thus, we have elected not to consider them further here.

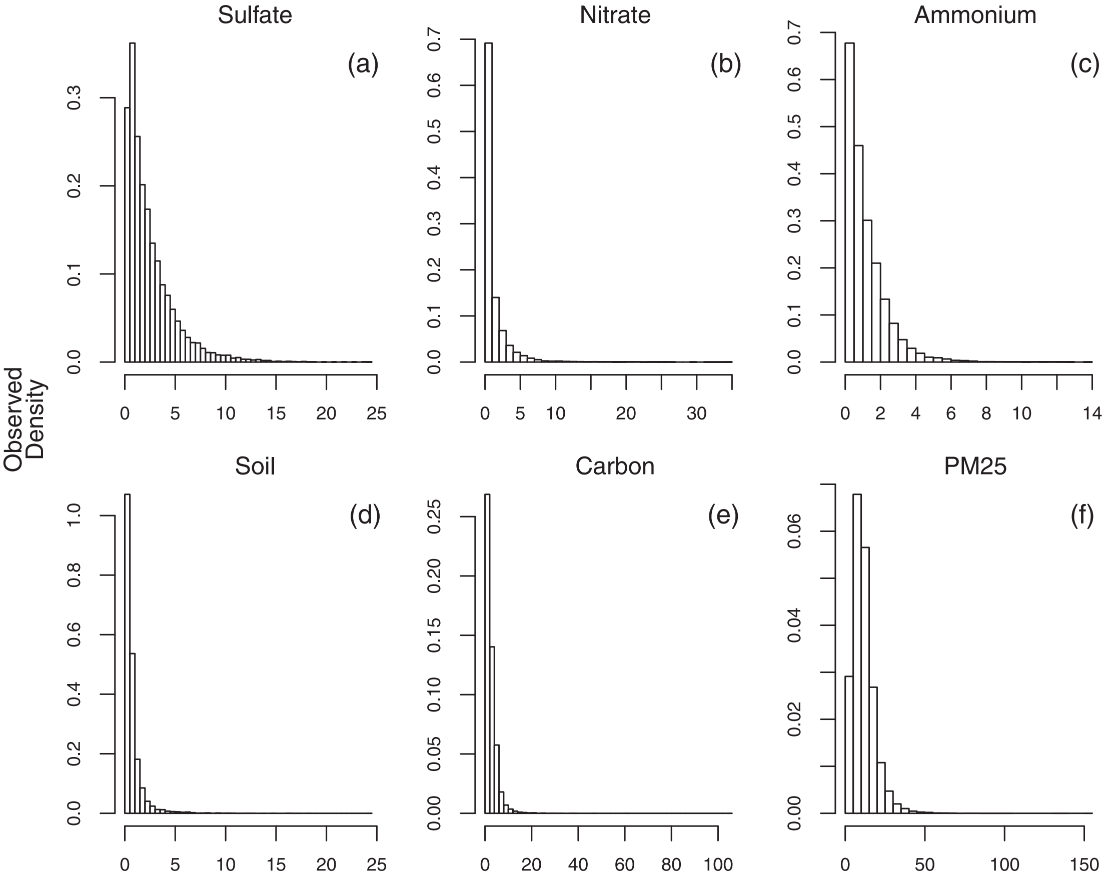

**Figure 3.** Histograms of the observed species and total PM$_{2.5}$ concentrations across the CSN and IMPROVE networks

We model the true value of "other" species, $Z^o_t(s)$, in the same way as the primary species, $Z^o_t(s) = \max(0, \tilde{Z}^o_t(s))$ where

$$\tilde{Z}^o_t(s) = \beta^o_{0,t} + \beta^o_{0,t}(s) + \beta^o_{1,t} Q^o_t(B_s) \tag{8}$$

Here, $\beta^o_{0,t}(s)$ is a zero mean Gaussian process with exponential covariance parameterized by $\phi^o_t$ and ${\sigma^2}^o_t$, decay and scale parameters, respectively. As "other" is not directly modeled by CMAQ, we calculate it by subtraction using

$$Q^o_t(B_s) = Q^{tot}_t(B_s) - \sum_{i=1}^{5} Q^i_t(B_s) \tag{9}$$

values of "other" in the model, with the tobit non-negative.

Returning to the temporal aspect, we work at a weekly scale. We fit the model to each week independently as it offers significant opportunities for improved computational efficiency, through parallelization of model fitting and prediction. Furthermore, we have introduced weekly variance and decay parameters for additional flexibility. Additionally, at weekly scale, there is small correlation between observations across weeks. To explore this further, we fit separate species models (identical to the models described previously but without the total sum constraint) with temporal dependence in the form of an autoregressive term added to $\beta^i_{0,t}(s)$ and without temporal dependence. The results of these models (not shown) indicate that the addition of temporal dependence does not consequentially improve predictive performance in terms of RMSPE, in support of the joint independence across time model.

Figure 4 presents a graphical view of the model, making clear the various levels. For a given $t$, the joint distribution of all of the variables, suppressing parameters, can be written concisely as the product of all species $i$ (including "other") and locations $s$

$$\prod_s \left[C^{tot}_t(s), I^{tot}_t(s), F^{tot}_t(s) \mid Z^{tot}_t(s)\right] \times \prod_i \left[C^i_t(s), I^i_t(s) \mid Z^i_t(s)\right] \left[Z^i_t(s) \mid \tilde{Z}^i_t(s)\right] \left[\tilde{Z}^i_t(s) \mid Q^i_t(B_s)\right] \tag{10}$$

Our model is, in essence, the simultaneous modeling of univariate downscaler models for each species subject to the summation constraint that their mass plus "other" must be equal to the true value of total PM$_{2.5}$ IMPROVE data.

To further enrich the model, we consider the introduction of dependence among the PM$_{2.5}$ species using the multivariate downscaling methods developed in Berrocal *et al.* (2010a). We explore this by introducing bivariate dependence between sulfate and ammonium, the two species with highest correlation (Supporting information Figure 3). We captured this dependence by jointly modeling the local spatial

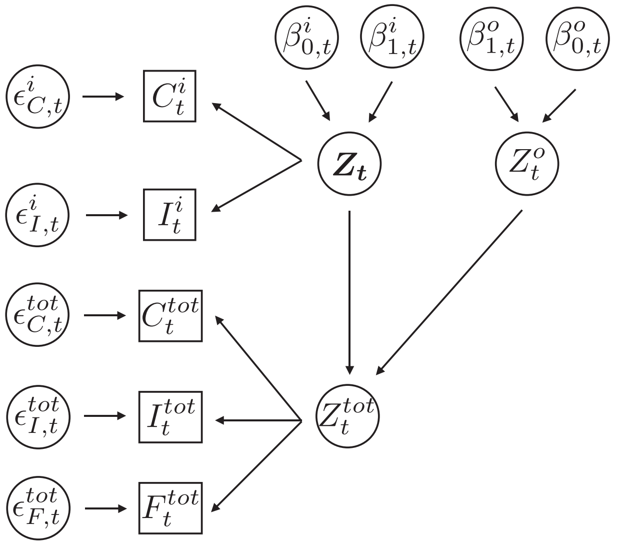

**Figure 4.** Diagram of model framework for total and speciated $PM_{2.5}$. Arrows indicate the dependence between model parameters

adjustment terms ($\beta_0^1(\boldsymbol{s})$ and $\beta_0^3(\boldsymbol{s})$) for these species using a coregionalization approach (Gelfand *et al.*, 2004). That is, $\begin{pmatrix}\beta_0^1(\boldsymbol{s})\\ \beta_0^3(\boldsymbol{s})\end{pmatrix} = \boldsymbol{A}\begin{pmatrix}w_0(\boldsymbol{s})\\ w_1(\boldsymbol{s})\end{pmatrix}$ where $w_0(\boldsymbol{s})$ and $w_1(\boldsymbol{s})$ are independent mean zero, unit variance Gaussian processes with exponential correlation functions and $\boldsymbol{A}$ is a lower triangular matrix with positive diagonal entries. We note that introducing dependence among all five species would require the unknown $\boldsymbol{A}$ to be 5 × 5 with five independent $w$ processes. Because this coregionalization is solely on the latent levels of the model, it is unlikely to be well identified, resulting in poorly behaved MCMC model fitting, apart from introducing a significant additional computational burden.

### 3.2. Prior distributions and model fitting

We complete the specification of the joint models by specifying priors and hyperpriors for the model parameters. We adopt noninformative and weakly informative priors that preserve conjugacy where possible. Recall that the models are fitted separately for each week $t$ and in parallel. The measurement error terms, $\epsilon_t^i(\boldsymbol{s})$, for each species and total $PM_{2.5}$ are assumed to be normal with mean zero and variance $\tau^2_{\epsilon_t^i}$; the $\tau^2$'s have inverse gamma(2,2) priors. Similarly, $\sigma_t^{2\,i}$, the variance parameters for the covariance functions for $\beta_0^i(\boldsymbol{s})$, are also given inverse gamma(2,2) prior distributions. These priors are vague with infinite variance and centering used in previous downscaling work (Berrocal *et al.* (2010a)); we found little sensitivity to this choice. The $\beta^i_{0,t}$ and $\beta^i_{1,t}$ are given normal priors with mean zero and large variance (500). Finally, the spatial decay parameters, $\phi_t^i$, are given uniform prior distributions with support such that the effective range is between 0 and half the maximum distance between monitoring stations (roughly 1400 miles). In the case of the tobit model with species dependence, we again follow the approach used in Berrocal *et al.* (2010a) and adopt log-normal priors for the diagonal entries of $\boldsymbol{A}$ with vague standard deviations (4) and a normal prior for the off-diagonal entry with mean zero and a large variance (500).

The models are fitted using a hybrid MCMC with a Metropolis-Hastings within Gibbs sampler. Parameters $\tau^2_{\epsilon_t^i}$ and $\sigma_t^{2\,i}$ are updated using Gibbs steps. The introduction of the tobit transformation breaks conjugacy for parameters $\beta_0^i$, $\beta_1^i$, and $\beta_0^i(\boldsymbol{s})$ requiring Metropolis-Hastings updates. All $\phi_t^i$'s require a Metropolis-Hastings step as well. Similarly, in the tobit model with species dependence, $\boldsymbol{A}$, $w_0(\boldsymbol{s})$, and $w_1(\boldsymbol{s})$ also require Metropolis Hastings updates. For the parameters updated via Metropolis-Hastings, we use a univariate random walk proposal distribution with proposal variances tuned using the methods described in Vihola (2011) during the burn-in period to achieve univariate acceptance rates close to 45%. See the Appendix A for the conditional distributions used to update each parameter and Supporting information Figure 4 for example parameter MCMC trajectories.

### 3.3. Prediction and validation

Prediction under the model in Section 3.1 is carried out through generation of posterior predictive samples. Such samples are developed through composition, using the posterior draws of the parameters along with the appropriate conditional distribution of the variables at the new locations given the variables at the fitting locations. Of interest for a new location $\boldsymbol{s}'$ are the posterior predictive distributions for the $Z$'s, the $C$'s, and the $F$'s. The $Z$'s are used for kriging, that is, for interpolation of species levels in a given week at a new location. The predictive distributions for the $C$'s and $F$'s can be used for model validation and model choice. That is, using hold out data, comparison can be made with the observed values using as follows: (i) root mean square predictive error (RMSPE); (ii) continuous rank probability scores (CRPS); see Gneiting and Raftery (2007); and (iii) empirical coverage versus nominal coverage for predictive intervals. Note that the $Z$'s can be used for RMSPE but not for empirical coverage or CRPS because they are not viewed as real data; they do not carry the uncertainty associated with observed data, and so we predict posterior observed values using the network level measurement error terms.

More precisely, suppressing $t$ in the notation, posterior predictive samples for each species are constructed for prediction locations $\boldsymbol{s}'$ by drawing $\beta_0^i(\boldsymbol{s}') \mid \{\beta_0^i(\boldsymbol{s})\}, \phi^i, \sigma^{2\,i}$. Along with posterior samples of $\beta_0^i$ and $\beta_1^i$, we construct predictions of the "true" concentration for each PM$_{2.5}$ species $Z^i(\boldsymbol{s}')$ (including "other") as well as total PM$_{2.5}$, $Z^{tot}(\boldsymbol{s}')$. This approach allows large spatial scale prediction for a fine grid of locations weekly across North America.

## 4. RESULTS

We present the results of fitting three variants of the joint tobit model. The first is the full joint tobit model as previously discussed; the second is an extension that employs bivariate coregionalization to allow for correlation between sulfate and ammonium. The third model has all $\beta_1^i$'s fixed at 0 thereby removing CMAQ from the model. This simplified model can be viewed as basic kriging subject to a sum constraint. Each model was fitted for data from each of the 52 weeks of 2007 for 50,000 MCMC iterations discarding the first 40,000 as burn-in. In Section 4.1, we report on our model comparison, and then in Section 4.2. we present our findings for the tobit models.

### 4.1. Model comparison

To assess the performance of our model variants, we randomly select 10% of CSN and IMPROVE stations as a validation set for out-of-sample prediction.[†] This validation set also includes all locations with incomplete observations (missing data) for the given week and is identical across the different model variants.

For each validation location $s$ and week $t$, we sample the posterior predictive distributions of $C_t^i(\boldsymbol{s})$ or $I_t^i(\boldsymbol{s})$, depending on the location's network membership, for all five species, "other", and total PM$_{2.5}$ ($i \in \{1, \ldots, 5, o, tot\}$). These posterior predictive samples are generated using the general prediction framework previously described for $Z^i(\boldsymbol{s})$ with the addition of draws of $\epsilon_C^i(\boldsymbol{s}') \mid \tau^2_{\epsilon^i_{C,t}}$ or $\epsilon_I^i(\boldsymbol{s}') \mid \tau^2_{\epsilon^{I,i}_t}$ to generate "observed" values from our predicted "true" values. These posterior predictive samples of $C_t^i(\boldsymbol{s})$ and $I_t^i(\boldsymbol{s})$ are then used to compute the root mean square predictive error, continuous ranked probability score, and the empirical coverage of the nominal 90% predictive interval.

We compute the root mean square predictive error of the predictions as

$$\text{RMSPE} = \sqrt{\frac{1}{\sum_t n_t} \sum_{t=1}^{52} \sum_{r=1}^{n_t} \left(\widehat{Y}_t(\boldsymbol{s}_r) - Y_t(\boldsymbol{s}_r)\right)^2 \cdot \mathbf{I}\left(Y_t(s_r)\right)} \tag{11}$$

where $n_t$ is the number of validation locations at time $t$, and $\boldsymbol{s}_r$ is the $r$th validation location. For $t$ and $\boldsymbol{s}_r$, $\widehat{Y}_t(\boldsymbol{s}_r)$ is the posterior predictive meanl; $Y_t(\boldsymbol{s}_r)$ is the observed value; and $\mathbf{I}(Y_t(s_r))$ is equal to 1 if $Y_t(s_r)$ is observed and 0 otherwise.

Additionally, to assess prrdictive performance, we employ the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007). The CRPS assesses both the calibration and sharpness of the posterior predictive distribution. It is defined as

$$\text{CRPS}(F, y) = \int \left(F(z) - \mathbf{1}_{\{z \geq y\}}\right)^2 dz \tag{12}$$

where $F$ is the empirical cumulative predictive distribution function, $y$ is the observed value; and $\mathbf{1}$ is the Heaviside function, that is equal to 1 if $z$ is greater than $y$ and 0 otherwise. A convenient computational form with posterior predictive samples can be created using Monte Carlo integrations for expression (21) in (Gneiting and Raftery, 2007). The reported CRPS values represent the average CRPS values over all validation locations for each time point. The CRPS has the desirable property that it compares the entire posterior predictive distribution to the observed hold out data. As defined, small CRPS is associated with a well-centered and concentrated predictive distribution.

Table 1 shows RMSPE, CRPS, 90% empirical coverage, and out-of-sample $R^2$ for all three tobit models discussed previously, tobit model with species dependence between sulfate and ammonium and the tobit model without CMAQ. Empirical coverage is used to assess model adequacy, while RMSPE and CRPS are used for model comparison. It is clear that the removal of CMAQ produces weaker model performance. For the other two models, at the species level, performance is mixed but, with regard to overall PM$_{2.5}$, the model with species dependence is clearly preferred[‡]. So, in the following sections, we focus on the results for this model. Predictive performance is elaborated in Supporting information Figure 6, which presents scatter plots of observed versus posterior predicted mean values of total PM$_{2.5}$ and its species. Also shown is the proportion of under-prediction for each species to assess bias. With the frequent occurrence of the species sum exceeding the total (recalling Figure 2), we might anticipate under-estimation at the species level and over-estimation for the total. However, the proportions do not suggest that there is consequential bias.

Lastly, with regard to the discrepancy in species and total predictive behavior between the tobit with independent downscaling and the tobit with species dependence, we note the as follows. In comparing the models, two aspects are in play. One is the frequently observed inconsistency between species sums and total PM$_{2.5}$. The other is the fact that, from the three datasets, we have far more total PM$_{2.5}$ measurements than individual species measurements. As a result, we expect the more flexible tobit model with species dependence to perform better in predicting total PM$_{2.5}$. However, both models downscale CMAQ with regard to the *true* levels (species and total) and both ensure that total PM$_{2.5}$ and its major species are consistent with regard to the sum constraint. So better prediction of the total for the more flexible model will require compensation in prediction at the species level, suggesting somewhat less fidelity to the observed species data.

[†] In fact, we have performed 10-fold cross-validation, holding out 10 distinct and randomly selected sets of 10%. Subsequently, we report the results for one set but results are consistent across the sets; see Supporting information Figure 5.

[‡] The mixed performance at the species level suggests that there will be no anticipated benefit in attempting to introduce dependence among all five species.

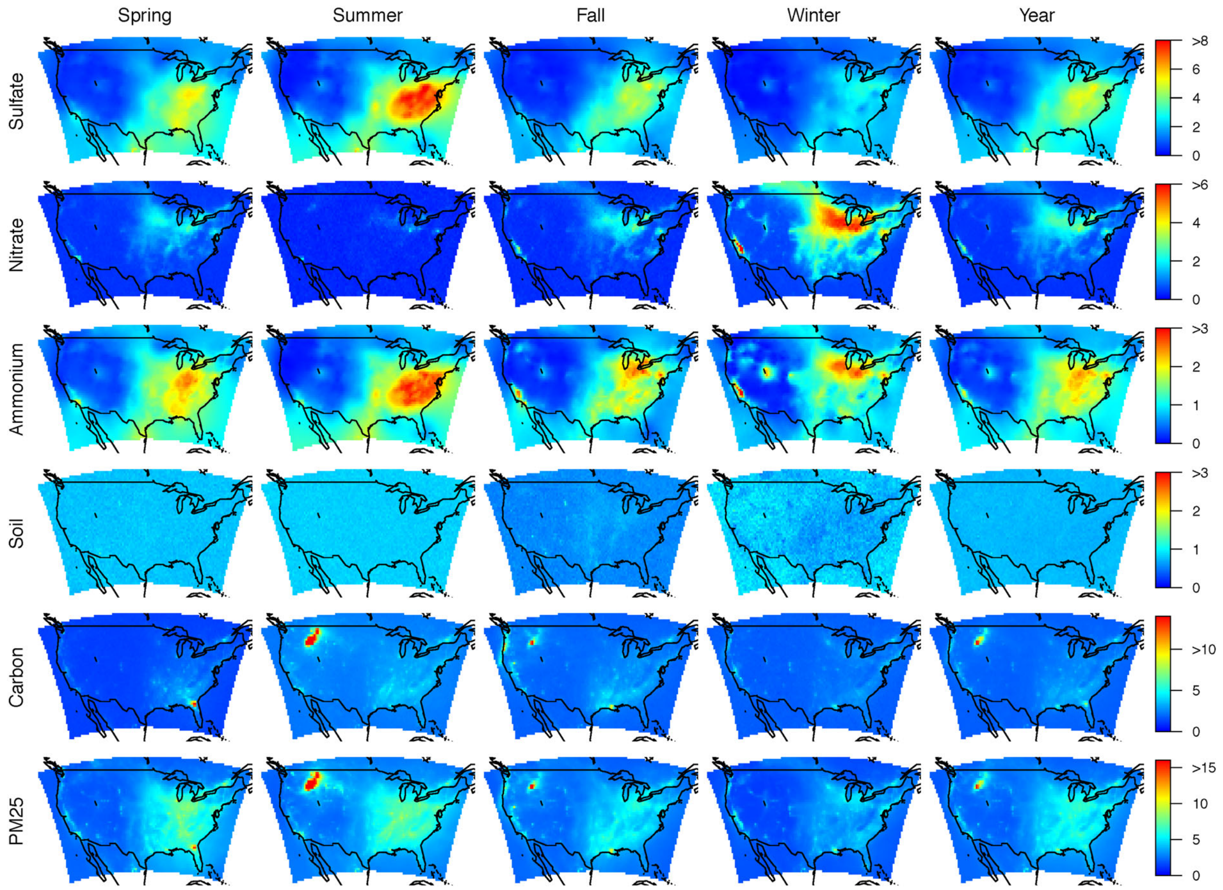

**Figure 5.** Maps of the seasonal and yearly average of posterior mean predicted PM$_{2.5}$ species and total PM$_{2.5}$ from the tobit with species dependence model

**Table 1.** Model validation results of posterior predictive means from joint downscaler model variants for randomly selected hold out locations

| | | Sulfate | Nitrate | Ammonium | Soil | Carbon | PM$_{2.5}$ |
|---|---|---|---|---|---|---|---|
| RMSPE | tobit | 1.16 | 1.77 | 0.72 | 1.38 | 3.99 | 5.51 |
| | tobit w/ Sp. Dep | 1.09 | 1.82 | 0.80 | 1.78 | 4.43 | 4.46 |
| | tobit w/o CMAQ | 1.35 | 2.37 | 0.86 | 1.43 | 4.20 | 6.33 |
| CRPS | tobit | 0.55 | 0.56 | 0.32 | 0.47 | 1.28 | 2.53 |
| | tobit w/ Sp. Dep | 0.50 | 0.62 | 0.34 | 0.61 | 1.40 | 2.10 |
| | tobit w/o CMAQ | 0.63 | 0.77 | 0.36 | 0.49 | 1.47 | 3.04 |
| EmpCov | tobit | 0.93 | 0.93 | 0.93 | 0.90 | 0.91 | 0.90 |
| | tobit w/ Sp. Dep | 0.91 | 0.94 | 0.91 | 0.91 | 0.93 | 0.90 |
| | tobit w/o CMAQ | 0.93 | 0.94 | 0.94 | 0.89 | 0.91 | 0.92 |
| OoS $R^2$ | tobit | 0.708 | 0.581 | 0.631 | 0.327 | 0.203 | 0.564 |
| | tobit w/ Sp. Dep | 0.757 | 0.543 | 0.592 | 0.131 | 0.182 | 0.726 |
| | tobit w/o CMAQ | 0.601 | 0.275 | 0.485 | 0.286 | 0.115 | 0.425 |

These include root mean square predictive error (RMSPE), average continuous rank probability score (CRPS), empirical 90% coverages (EmpCov), and out of sample $R^2$ (OoS $R^2$).

### 4.2. Results of the tobit analysis

Examples of seasonal and annual predictive surfaces for the United States for the tobit model with species dependence are presented in Figure 5. Equivalent maps for the original CMAQ data are presented in Supporting information Figure 7. These predictions are also used to construct the stacked time series plots in Figure 6; the prediction results are aggregated in space instead of time showing the relative contribution of each species to total $PM_{2.5}$ by region throughout the year. These regions match those used in Choi *et al.* (2009) and are given in Figure 7.

As seen in Figure 5, there are clear seasonal spatial patterns of speciated and total $PM_{2.5}$. Sulfate levels are generally higher in the warm summer months and account for the largest fraction of total $PM_{2.5}$ in the eastern United States. The largest sources of sulfate in the eastern

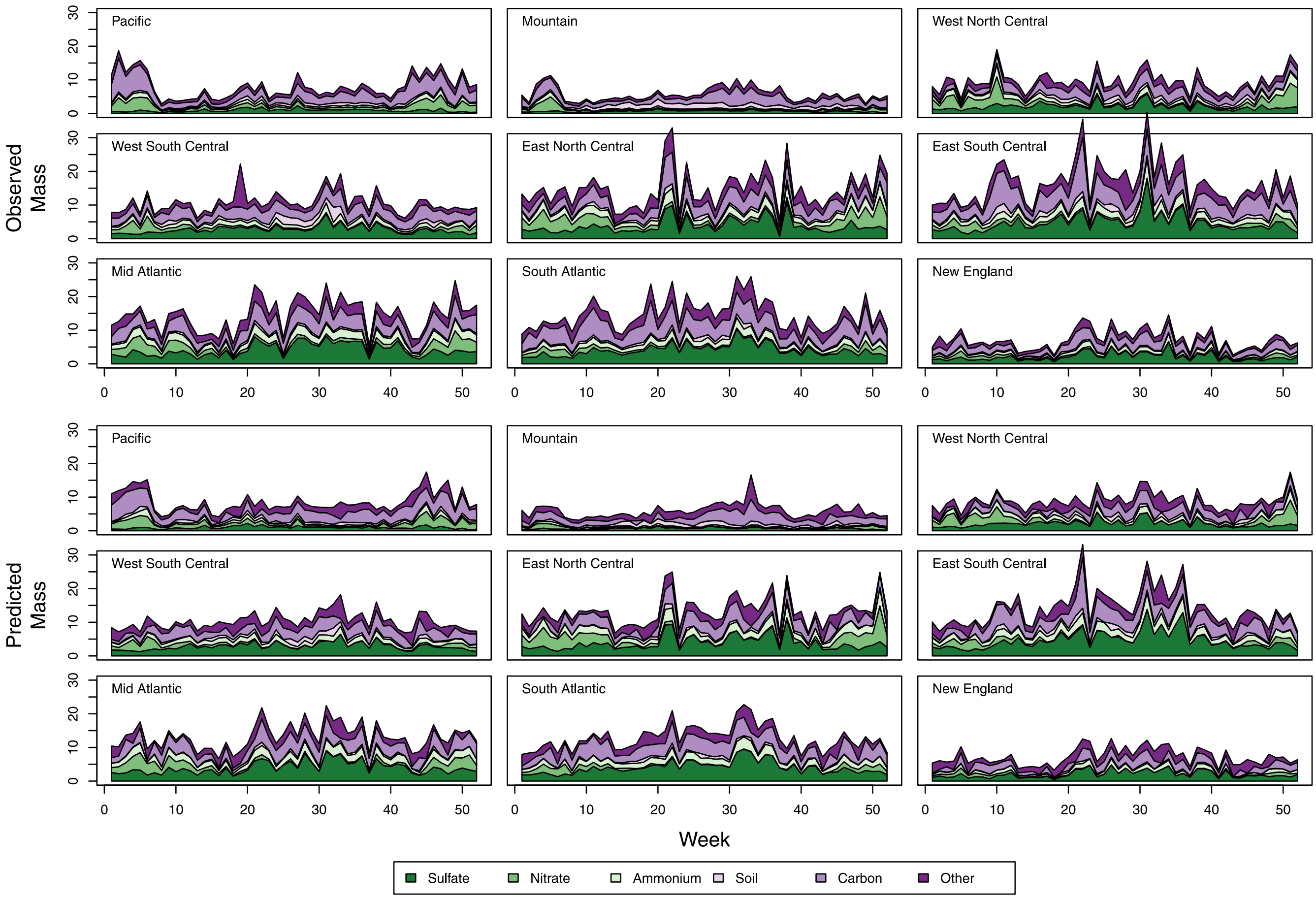

**Figure 6.** Stacked timeseries plots of the observed mass at CSN and IMPROVE monitoring locations aggregated to multistate regions defined in Figure 7 and the posterior mean predicted $PM_{2.5}$ species from the tobit with species dependence model aggregated to the same regions

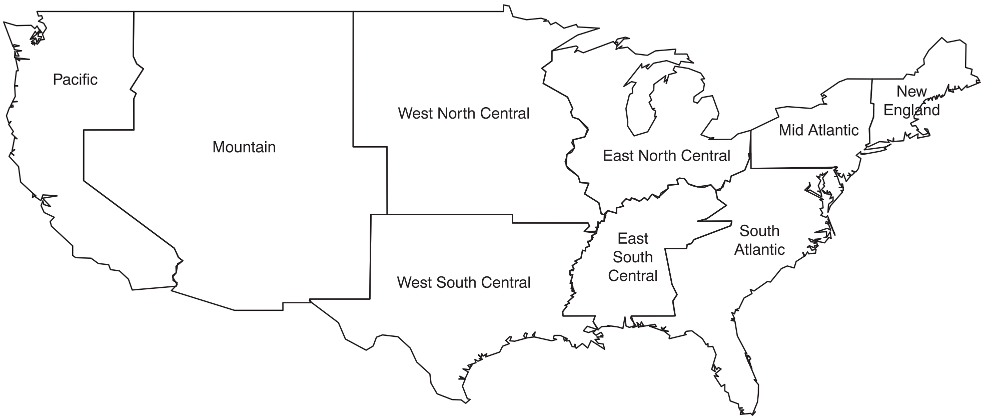

**Figure 7.** Map of the regions used to aggregate predictions based on regions used in Choi *et al.* (2009)

United States are $SO_2$ emissions from electric utilities and industrial boilers. Increased photochemical activity in the summer produces higher levels of sulfate. Nitrate concentrations are higher in the winter, particularly in the upper midwest because of increased ammonia availability in that region. The largest sources of nitrates are $NO_x$ emissions from highway vehicles and electric utilities. Ammonium levels exhibit strong seasonal variation where the highest concentrations occur during the summer in the central eastern United States. High predicted levels of soil are seen during the spring and summer seasons in the southwest United States because of low soil moisture and high wind speeds in this region. The highest levels of total carbon were found in the upper northwest during the summer months. For the rest of the year; the predicted carbon levels show slight seasonal variation with most predicted concentrations below 5 $\mu$g/m$^3$.

These seasonal spatial patterns are also clear from the stacked time series plots in Figure 6. These plots reflect predictions and observed data at CSN and IMPROVE monitoring locations enabling comparison of species compositions through time and by region. These plots largely agree in terms of major features and composition, with the most significant difference being that our model appears to smooth some of the larger spikes in the data (e.g. late spring in the East North Central region).

## 5. SUMMARY AND FUTURE WORK

We have proposed a downscaling approach for modeling speciated $PM_{2.5}$, which accomplishes a fusion of four data sources. The downscaling is carried out at the level of latent true levels with measurement error introduced to accommodate challenges with the observed monitoring data. We have modeled each species on its own scale and incorporated an “other” component in order to provide coherence between species and total. We fit independent weekly models and provide summaries for the entire continental United States. We show that downscaling CMAQ data substantially improves upon what is essentially ordinary kriging at the species level with a sum constraint. Our modeling could be aggregated to provide seasonal or annual prediction but, for such prediction, we recommend modeling at the desired temporal resolution to yield better predictions with less uncertainty.

More and more speciated data are being collected, enabling assessment of change in speciation across years as well as across space, for example, are some species levels increasing while others are perhaps stable or decreasing. Also, there are some cities which collect their own $PM_{2.5}$ component data. With this data, we can imagine further validation of our model as well as potential refinement. Lastly, an important long term goal is to use these data and models to aide the exploration of links between exposure to adverse health outcomes. With a better understanding of the nature of species-level exposure, we can better learn about which species are more influential with regard to particular outcomes.

## Acknowledgements

The work of the first two authors was supported in part by funding through the U.S. EPA’s Office of Research and Development using EPA contracts EP-13-D-000089 and EP-13-D-000257. The authors thank K. Wyat Appel of the National Exposure Research Laboratory/Atmospheric Modeling and Analysis Division (US EPA) for providing the CMAQ output and speciated monitoring data used in this analysis.

## DISCLAIMER

The U.S. Environmental Protection Agency through its Office of Research and Development partially collaborated in this research. Although it has been reviewed by the Agency and approved for publication, it does not necessarily reflect the Agency’s policies or views.

## REFERENCES

Bell ML, Ebisu K, Leaderer B. 2013. Associations of PM2. 5 constituents and sources with hospital admissions: analysis of four counties in connecticut and massachusetts (USA) for persons. *Envir. Health Perspectives* **12**(2):138–144.

Berrocal VJ, Gelfand AE, Holland DM. 2012. SpaceŘtime data fusion under error in computer model output: an application to modeling air quality. *Biometrics* **68**(3):837–848.

Berrocal V, Gelfand A, Holland D. 2010a. A bivariate space time downscaler under space and time misalignment. *The Annals of Applied Statistics* **4**(4): 1942–1975.

Berrocal VJ, Gelfand AE, Holland DM. 2010b. A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics* **15**(2):176–197.

Calder C. 2008. A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics* **19**(1):39–48.

Choi J, Fuentes M, Reich BJ, Davis JM. 2009. Multivariate spatial-temporal modeling and prediction of speciated fine particles. *Journal of Statistical Theory and Practice* **3**(2):407–418.

Dominici F, Peng R, Ebisu K. 2007. Does the effect of PM10 on mortality depend on PM nickel and vanadium content? A reanalysis of the NMMAPS data. *Environ. Health Perspectives* **115**:1701–1703.

Foley KM, Roselle SJ, Appel KW, Bhave PV, Pleim JE, Otte TL, Mathur R, Sarwar G, Young JO, Gilliam RC, Nolte CG, Kelly JT, Gilliland AB, Bash JO. 2009. Incremental testing of the community multiscale air quality (CMAQ) modeling system version 4.7. *Geoscientific Model Development Discussions* **2**(2):1245–1297.

Franklin M, Koutrakis P, Schwartz J. 2008. The role of particle composition on the association between PM2. 5 and mortality. *Epidemiology* **19**:680–689.

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

Gelfand A, Schmidt A, Banerjee S, Sirmans C. 2004. Nonstationary multivariate process modeling through spatially varying coregionalization. *Test* **13**(2):263–312.

Gneiting T, Raftery A. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477):359–378.

Guillas S, Bao J, Choi Y, Wang Y. 2008. Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. *Atmospheric Environment* **42**(5):1338–1348.

Ito K, Ross Z, Zhou J, Nádas A, Lippmann M, Thurston GD. 2013. NPACT Study 3. timeseries analysis of mortality, hospitalizations, and ambient PM2. 5 and its components. *National Particle Component Toxicity (NPACT) Initiative: Integrated Epidemiologic and Toxicologic Studies of the Health Effects of Particulate Matter Components,* 95–126.

Kim S, Sheppard L, Larson TV, *et al.* 2013. Issues related to combining multiple speciated PM2.5 data sources in spatio-temporal exposure models for epidemiology: the NPACT case study. (UW Biostatistics Working Paper Series, working paper 397). http://biostats.bepress.com/uwbiostat/paper397.

Lippmann M, Ito K, Hwang J. 2006. Cardiovascular effects of nickel in ambient air. *Environmental Health Perspectives* **144**:1662–1669.

Liu Z, Le N, Zidek J. 2008. Combining measurements and physical model outputs for the spatial prediction of hourly ozone space-time fields. *Technical Report*, The University of British Columbia, Vancouver, BC, Canada.

McMillan NJ, Holland DM, Morara M, Feng J. 2010. Combining numerical model output and particulate data using Bayesian spacetime modeling. *Environmetrics* **21**(1):48–65.

Rich D, Ozkaynak H, Crooks J. 2013. The triggering of myocardial infarction by fine particles is enhanced when particles are enriched in secondary species. *Environmental Science and Technology* **47**:9414–9423.

Schmidt A, Gelfand A. 2003. A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research: Atmospheres* **108**(D24):8783.

Vihola M. 2011. Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and Computing* **22**(5):997–1008.

Wackernagel H. 2003. *Multivariate Geostatistics* 3rd ed. Springer: New York.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.