

Multivariate spatial-temporal modeling and prediction of speciated fine particles ¹

Jungsoon Choi, Brian J. Reich, Montserrat Fuentes, and Jerry
M. Davis

Abstract

Fine particulate matter (PM_{2.5}) is an atmospheric pollutant that has been linked to serious health problems, including mortality. PM_{2.5} has five main components: sulfate, nitrate, total carbonaceous mass, ammonium, and crustal material. These components have complex spatial-temporal dependency and cross dependency structures. It is important to gain better understanding about the spatial-temporal distribution of each component of the total PM_{2.5} mass, and also to estimate how the composition of PM_{2.5} changes with space and time to conduct spatial-temporal epidemiological studies of the association of these pollutants and adverse health effects. We introduce a multivariate spatial-temporal model for speciated PM_{2.5}. Our hierarchical framework combines different sources of data and accounts for bias and measurement error in each data source. In addition, a spatiotemporal extension of the linear model of coregionalization is developed to account for spatial and temporal dependency structures for each component as well as the associations among the components. We apply our framework to speciated PM_{2.5} data in the United States for the year 2004.

Key words: Air pollution; Bayesian inference; linear coregionalization model; multivariate spatiotemporal processes; speciated particulate matter.

¹J. Choi is a Graduate Student at the Department of Statistics, North Carolina State University. B. J. Reich is a Postdoctoral Fellow at the Department of Statistics, North Carolina State University. M. Fuentes is an Associate Professor at the Department of Statistics, North Carolina State University. (Email: fuentes@ncsu.edu). J. M. Davis is a Professor in the Department of Marine Earth and Atmospheric Sciences, North Carolina State University.

1 Introduction

The study of the association between ambient particulate matter (PM) and human health has received much attention in epidemiological studies over the past few years. Özkaynak and Thurston (1987) conducted an analysis of the association between several particle measures and mortality. Their results showed the importance of considering particle size, composition, and source information when modeling particle pollution health effects. In particular, fine particle matter, $PM_{2.5}$ ($< 2.5\mu m$ in diameter), is an atmospheric pollutant that has been linked to numerous adverse health effects (e.g., respiratory and cardiovascular diseases). $PM_{2.5}$ is a mixture of pollutants which the U.S. EPA (2003) classified into five main components: sulfate, nitrate, total carbonaceous mass (TCM), ammonium, and crustal material (including calcium, iron, silicon, aluminum, and titanium).

Rao et al. (2003) and Malm et al. (2004) investigated the spatial and temporal patterns of speciated $PM_{2.5}$, but they only conducted an exploratory analysis of speciated $PM_{2.5}$. The research presented here is part of a larger project to study the association between speciated $PM_{2.5}$ and adverse health outcomes across the entire U.S. In order to investigate the health effects associated to speciated fine PM across space and time, we need to interpolate speciated $PM_{2.5}$ at the locations and times of interest. Our goal is to develop a statistical framework using all available sources of data about speciated $PM_{2.5}$ to investigate the spatial-temporal patterns of speciated $PM_{2.5}$ and then predict speciated $PM_{2.5}$ at all locations and times of interest.

In this article we introduce a new statistical framework to combine information for spe-

ciated $\text{PM}_{2.5}$ from two monitoring networks while accounting for potential bias and measurement error in each network. Daily speciated $\text{PM}_{2.5}$ measurements are available at a limited number of monitoring sites and missing values are common. Therefore, we supplement these observations with measurements of the total $\text{PM}_{2.5}$. These observations are indirectly informative about the individual components and greatly expand our spatial and temporal coverage. Incorporating total $\text{PM}_{2.5}$ measurements poses a challenging data fusion problem. In our Bayesian approach, all of the data sources are simultaneously represented in terms of the underlying true sum of the five main components and the true proportions of each speciated component relative to the total. We develop a spatiotemporal multinomial logistic model which allows the proportions to vary smoothly across space and time. Also, we extend the linear model of coregionalization to the spatiotemporal setting to account for the complex dependency structures of the speciated $\text{PM}_{2.5}$.

We use a speciated $\text{PM}_{2.5}$ data set that has not previously been analyzed. To our knowledge, this is the first time that a statistical framework has been used to analyze speciated $\text{PM}_{2.5}$ across the entire United States. We show that the total $\text{PM}_{2.5}$ measurements are generally positively biased relative to the sum of the speciated components and that magnitude of the bias varies across the U.S. We also show that the proportions of each component vary considerable across space and time and that accounting for the cross-dependency in the speciated components dramatically improves prediction.

The article is organized as follows. In Section 2 we describe the data used in this study. In Section 3, we present a Bayesian hierarchical multivariate spatial-temporal model for

speciated $\text{PM}_{2.5}$ along with computational details. In Section 4 we present the results, and in Section 5 we offer a general discussion.

2 Data Description

$\text{PM}_{2.5}$ data from two monitoring networks and meteorological data in the conterminous United States for year 2004 were used in this study. The first source of $\text{PM}_{2.5}$ data is the Speciated Trends Network (STN) established by the U.S. Environmental Protection Agency (EPA) in 1999. The STN measures speciated $\text{PM}_{2.5}$ either every day, every third day or every sixth day. It included about 200 monitoring stations in 2004, which were mostly in urban areas. Even though the STN collects numerous trace elements, elemental carbon, organic carbon, and ions (sulfate, nitrate, sodium, potassium, ammonium), we only consider the five main $\text{PM}_{2.5}$ components described in the previous section. In the STN, sulfate, nitrate, and ammonium are measured independently. Total carbonaceous mass is sum of elemental carbon mass and estimated organic carbon mass which is $1.4 \times ([OC] - 1.53)$, where $[OC]$ is the measured organic carbon value, 1.4 is the factor to correct organic carbon mass for other elements (Rao et al., 2003), and 1.53 is the blank correction factor to adjust for sampling artifacts (Flanagan et al., 2003). Elemental carbon mass is also measured at STN monitoring stations. Crustal material is computed using the IMPROVE equation (Malm et al., 2004) for the five most prevalent trace elements.

Since the $\text{PM}_{2.5}$ data at the STN monitoring stations provide sparse spatial coverage, using only the STN monitoring data might be insufficient for modeling the speciated fine

PM over the entire United States. Therefore, we use the total $\text{PM}_{2.5}$ data from the Federal Reference Method (FRM) monitoring network which includes rural and urban sites, and measures $\text{PM}_{2.5}$ either every day, every third day, or every sixth day. While the STN is a smaller network, the FRM network is a large national network which consisted of about 1000 monitoring stations in 2004.

Meteorological data for 2004 are provided by the U.S. National Climate Data Center. We use five daily meteorological variables: minimum temperature, maximum temperature, dew point temperature, wind speed, and pressure.

3 Statistical Model

While speciated $\text{PM}_{2.5}$ is only available at STN stations, information about the sum of the five components comes from both STN and FRM networks. Therefore we expect the sum of the five components to be better identified than the individual components. Also, exploratory analysis suggests that while the values of the speciated components vary considerable across space and time, the proportions of each component to the total are more stable, i.e., the proportions have larger spatial and temporal autocorrelations. Therefore we parameterize our statistical model in terms of the sum of the five components and the relative proportion of the each component to the sum.

Let $\hat{\mathbf{V}}(\mathbf{s}, t) = (\hat{V}_1(\mathbf{s}, t), \dots, \hat{V}_5(\mathbf{s}, t))^T$ be a vector of the observed speciated $\text{PM}_{2.5}$ at location \mathbf{s} and at time t from the STN network and $\hat{Z}_F(\mathbf{s}, t)$ be the observed total $\text{PM}_{2.5}$ from

the FRM network. Our model for these data is

$$\hat{Z}_F(\mathbf{s}, t) = a(\mathbf{s}, t) + Z(\mathbf{s}, t) + \epsilon_z(\mathbf{s}, t) \quad (1)$$

$$\hat{V}_k(\mathbf{s}, t) = \theta_k(\mathbf{s}, t)Z(\mathbf{s}, t) + \epsilon_k(\mathbf{s}, t), \quad (2)$$

where $Z(\mathbf{s}, t)$ is the true sum of the five components, $a(\mathbf{s}, t)$ is a bias term to account for systematic differences between the two networks (e.g., FRM measures more than the five main components), $\theta_k(\mathbf{s}, t)$ is the proportion of the sum attributed to component k , and $\epsilon_z(\mathbf{s}, t)$ and $\epsilon_k(\mathbf{s}, t)$ are errors. A possible extension of this model is to assume the FRM data has both additive and multiplicative bias. However, given that the STN data are modeled as multiples of $Z(\mathbf{s}, t)$, it would be difficult for the data to identify a multiplicative bias for FRM, so it is omitted.

The remainder of this section describes the statistical models for each component of the hierarchical framework. Section 3.1 describes the spatiotemporal model for $Z(\mathbf{s}, t)$, Section 3.2 gives the multinomial logistic model for the proportions $\theta_k(\mathbf{s}, t)$, and Section 3.3 describes model for the errors $\epsilon_k(\mathbf{s}, t)$. Section 3.4 provides computational details.

3.1 Model for the sum of the five main components

The true sum of the five components, $Z(\mathbf{s}, t)$, is modeled using a dynamic spatiotemporal linear model (Gelfand et al., 2005). We assume

$$Z(\mathbf{s}, t) = \mathbf{M}^T(\mathbf{s}, t)\boldsymbol{\beta}(\mathbf{s}, t) + \delta_z(\mathbf{s}, t), \quad (3)$$

where $\mathbf{M}^T(\mathbf{s}, t)$ is a vector of meteorological variables (described in Section 2) and sine and cosine functions with one-year periods to capture temporal trends and $\delta_z(\cdot, t) = (\delta_z(s_1, t), \dots, \delta_z(s_n, t))$ is normal with mean $\psi_z \delta_z(\cdot, t-1)$ and, based on exploratory analysis, exponential covariance $\sigma_\delta^2 \exp(-\|\mathbf{s} - \mathbf{s}'\|/\phi_\delta)$. In general, the regression coefficients $\beta(\mathbf{s}, t)$ may vary across space and time. In Section 4's application $\beta(\mathbf{s}, t)$ is constant over time and constant within nine geographic regions. We use priors $\psi_z \sim U(0, 1)$, $\phi_\delta \sim U(1, 1000)$, and $\sigma_\delta \sim U(0, 100)$.

Exploratory analysis suggests the additive bias varies over space and time, and we model $a(\mathbf{s}, t)$ using a hierarchical framework,

$$a(\mathbf{s}, t) = a_1(t) + a_2(\mathbf{s}, t), \quad (4)$$

$$a_1(t) = h(t) + e_1(t), \quad (5)$$

$$a_2(\mathbf{s}, t) = \delta_{a_2} a_2(\mathbf{s}, t-1) + e_2(\mathbf{s}, t), \quad (6)$$

where $a_1(t)$ represents the overall temporal trend in the bias of the RCFM data, and $h(t)$ is a smoothing function of time to explain seasonality in the additive bias term. The process $a_2(\mathbf{s}, t)$ accounts for the spatial-temporal structure which is not captured by the overall temporal trend. We assume the process $a_2(\mathbf{s}, t)$ is an AR(1) with coefficient δ_{a_2} , and e_1 and e_2 are independent white noise processes and are independent of the process Z .

The errors $\epsilon_z(\mathbf{s}, t)$ are modelled as $\epsilon_z(\mathbf{s}, t) \stackrel{iid}{\sim} N(0, \sigma_z^2)$. The results are somewhat sensitive to the prior of the standard deviation σ_z . Therefore σ_z is fixed based on prior information regarding the precision of the FRM monitoring devices. U.S. EPA (1997, 2000) suggests that the coefficient of variation (CV) is 15%. From the CV, the standard deviation can be calculated as $\text{sd} = \text{CV} * \text{mean}/100$, giving $\sigma_z = 1.796$.

3.2 Model for the speciated proportions

The proportion of each component to the total $\text{PM}_{2.5}$ mass varies over space and time, and we use a hierarchical framework to account for the spatial-temporal associations of the proportions. To ensure that the proportions add to one at each site and time we extend the multinomial logit model (McFadden, 1974) to the spatiotemporal setting. Let

$$\theta_k(\mathbf{s}, t) = \frac{\exp(\delta_k(\mathbf{s}, t))}{\sum_{j=1}^5 \exp(\delta_j(\mathbf{s}, t))} \quad (7)$$

where $\delta_j(\mathbf{s}, t)$ are independent across j and have the same dynamic spatiotemporal priors as $\delta_z(\mathbf{s}, t)$ in (3). For identification purposes, we fix $\delta_5(\mathbf{s}, t) = 0$ for all \mathbf{s} and t . In our study, crustal material is taken to be the 5th component because it is the most stable component (largest autocorrelations). Figure 1 shows the framework of the speciated fine PM.

3.3 Spatiotemporal linear coregionalization model

Even though the spatial-temporal dependency structures of the proportions are considered, it could be insufficient to capture the spatial-temporal dependency and the cross-dependency structures of speciated $\text{PM}_{2.5}$. We account for cross-dependency in the errors $\boldsymbol{\epsilon}(\mathbf{s}, t) = (\epsilon_1(\mathbf{s}, t), \dots, \epsilon_5(\mathbf{s}, t))^T$ using a spatiotemporal linear model of coregionalization (STLMC). The STLMC is an extension of the linear model of coregionalization (LMC) used in multivariate spatial analysis (Grzebyk and Wackernagel, 1994; Wackernagel, 1998; Gelfand et al., 2004). The basic idea of the STLMC is that dependent spatial-temporal

processes are expressed as linear combinations uncorrelated spatial-temporal processes, i.e.,

$$\boldsymbol{\epsilon}(\mathbf{s}, t) = \mathbf{A}\mathbf{w}(\mathbf{s}, t) + \boldsymbol{\epsilon}^w(\mathbf{s}, t), \quad (8)$$

where $\mathbf{w}^T(\mathbf{s}, t) = (w_1(\mathbf{s}, t), \dots, w_5(\mathbf{s}, t))$, \mathbf{A} is a 5×5 weight matrix explaining the association among the five variables, and $\boldsymbol{\epsilon}^w(\mathbf{s}, t)$ is Gaussian white noise. Without loss of generality, we assume \mathbf{A} is a lower triangular matrix. For computational convenience, we adopt a simple approach to model the spatial-temporal process $\mathbf{w}(\mathbf{s}, t)$. We assume that $w_i(\mathbf{s}, t)$, $i = 1, \dots, 5$, are independent Gaussian spatial-temporal processes with mean zero and separable spatial-temporal covariance, $Cov(w_i(\mathbf{s}_l, t_j), w_i(\mathbf{s}_{l'}, t_{j'})) = C_i^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \boldsymbol{\phi}_i)C_i^{(2)}(t_j, t_{j'}; \boldsymbol{\psi}_i)$, where $C_i^{(1)}$ is a spatial covariance with the parameter vector $\boldsymbol{\phi}_i$, and $C_i^{(2)}$ is a temporal autocovariance with the parameter vector $\boldsymbol{\psi}_i$. The STLMC in (8) implies $E(\boldsymbol{\epsilon}(\mathbf{s}, t)) = 0$ and

$$Cov(\boldsymbol{\epsilon}(\mathbf{s}_l, t_j), \boldsymbol{\epsilon}(\mathbf{s}_{l'}, t_{j'})) = \sum_{i=1}^5 C_i^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \boldsymbol{\phi}_i)C_i^{(2)}(t_j, t_{j'}; \boldsymbol{\psi}_i)\mathbf{T}_i, \quad (9)$$

where $\mathbf{T}_i = \mathbf{a}_i \mathbf{a}_i^T$ and \mathbf{a}_i is the i^{th} column vector of \mathbf{A} . Under this model, the covariance matrix of $\boldsymbol{\epsilon}$ at any site \mathbf{s} and time t is $\mathbf{T} = \sum_{i=1}^5 \mathbf{T}_i$.

We form $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \dots, \boldsymbol{\epsilon}_5^T)^T$ and $\boldsymbol{\epsilon}_i^T = (\boldsymbol{\epsilon}_i^T(t_1), \dots, \boldsymbol{\epsilon}_i^T(t_{N_t}))$ for $i = 1, \dots, 5$, where $\boldsymbol{\epsilon}_i^T(t_j) = (\epsilon_i(s_1, t_j), \dots, \epsilon_i(s_{N_s}, t_j))$ for $j = 1, \dots, N_t$. Then, the covariance matrix of $\boldsymbol{\epsilon}$ is

$$\boldsymbol{\Sigma}^\epsilon = \sum_{i=1}^5 \mathbf{T}_i \otimes \mathbf{U}_i \otimes \mathbf{R}_i, \quad (10)$$

where \otimes denotes the Kronecker product. Each \mathbf{R}_i is a $N_s \times N_s$ matrix with $(R_i)_{ll'} = C_i^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \boldsymbol{\phi}_i)$, which accounts for spatial associations. Each \mathbf{U}_i is a $N_t \times N_t$ matrix with $(U_i)_{jj'} = C_i^{(2)}(t_j, t_{j'}; \boldsymbol{\psi}_i)$, which explains temporal associations. This covariance matrix, $\boldsymbol{\Sigma}^\epsilon$,

is nonseparable, except in the special case of the STLMC where $C_i^{(1)} = C_{i'}^{(1)} = C^{(1)}$ and $C_i^{(2)} = C_{i'}^{(2)} = C^{(2)}$ for all $i, i' = 1, \dots, 5$. In this case, $\Sigma^\epsilon = \mathbf{T} \otimes \mathbf{U} \otimes \mathbf{R}$ for $(R)_{ll'} = C^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \phi)$ and $(U)_{jj'} = C^{(2)}(t_j, t_{j'}; \psi)$.

3.4 Computing details

All MCMC sampling is carried out in the freely-available software WinBUGS (<http://www.mrc-bsu.cam.ac.uk/bugs/>). For all MCMC sequences, we conducted MCMC convergence diagnosis using the Gelman and Rubin convergence diagnostics, autocorrelation functions, and trace plots. All the hyperpriors chosen here ensured to acceptable MCMC convergence. Further details are found in the Appendix.

For computational convenience, we carry out our analysis using a two-stage, empirical Bayes approach. We first estimate $Z(\mathbf{s}, t)$ using STN and FRM data and then analyze the STN data given $Z(\mathbf{s}, t)$. Estimates of $Z(\mathbf{s}, t)$ are taken to be the posterior means from the model

$$\hat{Z}_F(\mathbf{s}, t) = a(\mathbf{s}, t) + Z(\mathbf{s}, t) + e_z(\mathbf{s}, t) \quad (11)$$

$$\hat{Z}_R(\mathbf{s}, t) = Z(\mathbf{s}, t) + e_R(\mathbf{s}, t)$$

where $\hat{Z}_R(\mathbf{s}, t) = \sum_{k=1}^5 \hat{V}(\mathbf{s}, t)$ is the reconstructed total $\text{PM}_{2.5}$ from the STN data, $a(\mathbf{s}, t)$, $Z(\mathbf{s}, t)$, and $e_z(\mathbf{s}, t)$ are modelled as in Section 3.1, and $e_R(\mathbf{s}, t)$ is Gaussian white noise.

As a consequence on this two-stage approach, our posterior estimates of variability for the speciated components will reflect our uncertainty in the proportions $\theta(s, t)$, but not our

uncertainty in the total PM_{2.5} $Z(s, t)$. However, given the large amount of total PM_{2.5} data compared to speciated data, ignoring this second source of uncertainty may have little effect. Indeed, our calibration analysis in Section 4 shows that our prediction intervals maintain the proper coverage probability.

We use the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) to compare models. Deviance is defined as $D(V) = -2 \log(f(\hat{V}|V))$. The DIC statistic is $DIC = \bar{D} + p_D$ where $\bar{D} = E(D(\mathbf{V})|\hat{V})$ measures fit and $p_D = \bar{D} - D(E(\mathbf{V}|\hat{V}))$, the effective number of parameters, measures complexity. Models with smaller DIC are preferred. In addition to DIC , we also compare models using cross-validation. We randomly remove 10% of the observations and compute the root mean square prediction error (RMSPE) for the predictions for the deleted values.

4 Application

We apply our statistical framework to the daily speciated PM_{2.5} data in the United States in 2004. We begin by describing our analysis of the total PM_{2.5} mass using the hierarchical model in (11). Preliminary exploratory analysis suggests that the coefficients of the weather covariates are different in different regions. Therefore we implement our framework for the nine geographic regions as defined by the United States Census: New England; Middle Atlantic; East North Central; Midwest; South Atlantic; East South Central; West South Central; Mountain; Pacific. We assume that the regression parameters $\beta(\mathbf{s}, t)$ vary across regions but are constant (over space) within these regions and have $N(0, 100^2)$ priors.

In July 2004, the bias' posterior mean is positive for all nine geographic regions except for in the Pacific region. The mean is the largest in the South Atlantic region. The negative bias (the mean is -0.8) in the Pacific region is not surprising because in California during summer about 60 – 90% of the nitrate is lost due to evaporation in the FRM's total $\text{PM}_{2.5}$ mass measurement (Frank, 2006). In December 2004, the posterior mean bias is positive in all regions. Overall, the bias is higher in July than in December because the sulfate and ammonium concentrations are higher in the summer and the FRM total $\text{PM}_{2.5}$ mass includes a large amount of water during the summer (Frank, 2006).

Due to computational costs, we implement our entire spatial-temporal framework for speciated $\text{PM}_{2.5}$ using only California data, but we work with our framework over the entire United States at fixed times (June 14 and December 14) and at a fixed locations (Los Angeles, Phoenix, and New York City). Figure 2 maps of the estimated concentrations of sulfate and nitrate for the spatial analysis on June 14 and December 14. Overall, the sulfate concentration is highest in the Eastern U.S. and in the summer. In contrast, the nitrate concentration is the highest in the Western U.S. and in the winter.

The time series plots of the estimated concentrations of speciated $\text{PM}_{2.5}$ in Los Angeles, Phoenix, and New York City are presented in Figure 3. These three cities have markedly different temporal patterns, illustrating the difficulty in modeling speciated PM across the entire U.S. For all three cities, ammonium and crustal concentrations are relatively constant over time. In Los Angeles, the most abundant components are sulfate in the summer and nitrate and TCM in the winter. Sulfate is also high in the New York City in the summer, but

unlike Los Angeles the other components are fairly stable over time. TCM is the dominant component throughout the year in Phoenix.

Many of these patterns are also apparent in Figure 4's map of the estimated speciated $\text{PM}_{2.5}$ composition by region and by season in 2004. In this figure circle size corresponds to total $\text{PM}_{2.5}$ mass, and we can clearly see the spatial-temporal pattern of the total $\text{PM}_{2.5}$ mass. During the spring and summer (April - September) the total $\text{PM}_{2.5}$ mass is highest in the Eastern U.S. During the fall and winter (October - March) total $\text{PM}_{2.5}$ mass is fairly constant across space.

TCM has the highest proportion of the total $\text{PM}_{2.5}$ mass among the components over the entire U.S. Sulfate concentrations are highest during the summer in most of the Eastern U.S. and the Pacific region because increased photochemical reactions in the atmosphere increases sulfate formation (Baumgardner et al., 1999). Nitrate concentrations are highest during the winter (January-March) because high ammonia availability, low temperature, and high relative humidity favor ammonium nitrate condensation. We also see the seasonal pattern of ammonium concentration. On average, during the winter and spring (January-June), ammonium concentrations were about 3.2 times higher than during the summer and fall seasons. During the summer and fall, TCM concentrations are high because of secondary organic aerosol formation. Crustal material concentrations are higher in the eastern United States during the spring because of low soil moisture and high wind speeds. Also, these regions are impacted by North African dust during the spring (Malm et al., 2004).

Finally, to illustrate the need for our hierarchical framework, we present model diagnostics

using our entire spatiotemporal framework using only data from California in 2004. We compare three models. Model 1 is the statistical framework proposed in Section 3. Model 2 ignores the STLMC process $\mathbf{Aw}(\mathbf{s}, t)$, i.e., sets $\boldsymbol{\epsilon}(\mathbf{s}, t) = \boldsymbol{\epsilon}^w(\mathbf{s}, t)$ for all (\mathbf{s}, t) . Model 3 removes both the STLMC and the hierarchical framework of the proportion parameters, i.e., the proportion parameters are constant over space and time. The DIC is 950 ($p_D = 3444$) for Model 1, 8235 ($p_D = 1334$) for Model 2, and 15219 ($p_D = 10$) for Model 3. We also use the root mean squared prediction error (RMSPE). The RMSPE value for Model 1 is 0.075, for Model 2 it is 1.462, and for Model 3 it is 6.082. Thus, our framework has the lowest DIC and RMSPE values among the three models.

These results confirm the need for the multivariate spatiotemporal model. Table 1 summarizes the posterior correlations between components using the data from California in 2004. Several 95% posterior intervals do not cover zero. The strongest correlation (posterior mean 0.836) is between nitrate and ammonium. This relationship can also be seen Los Angeles in Figure 3, as both of these components have strong peaks in March and October.

In addition, we conduct a calibration analysis for the speciated $\text{PM}_{2.5}$ in Phoenix to test the performance of our framework. We select Phoenix because it has the fewest missing values. We randomly selected 30 observations in 2004, and we obtained 95% prediction intervals for the i^{th} time given the data, not using data from the i^{th} time we are predicting. Figure 5 plots the actual and predicted values. The percentages of the observed values that are outside the interval are 0% for sulfate and 3.3% for nitrate. We also did calibration analyses for the other three components in Phoenix. The percentages of the observed values

lying outside the interval are between 0% and 3.3%. It appears are model is well-calibrated.

5 Conclusion

In this article we present a flexible hierarchical framework to study speciated $\text{PM}_{2.5}$. The multivariate spatial-temporal model proposed here allows for spatial-temporal dependency for each component and cross dependency structures among the components. A hierarchical framework provides a natural way to investigate the spatiotemporally varying contribution of each component to the total $\text{PM}_{2.5}$ mass. Using our framework, we can estimate speciated $\text{PM}_{2.5}$ at unobserved locations of interest in the United States. We also introduce a new statistical framework to incorporate $\text{PM}_{2.5}$ data from different sources, which takes into account bias and measurement error over space and time. Diagnostics verify the performance of our model.

We found that the additive bias term of the FRM network is generally positive, that is the FRM's total $\text{PM}_{2.5}$ is higher than the sum of the speciated components measured by STN. However, in the Pacific region, we see different results during the summer season because of nitrate losses. In the eastern United States, the contribution of sulfate to the total $\text{PM}_{2.5}$ mass tends to be higher during the summer. In almost all regions, sulfate concentrations are higher during the summer. Also, the spatial differences in the sulfate concentrations are the largest during the summer. On average, the sulfate proportions and concentrations in the East South Central region are highest where sulfur is emitted from coal-fired sources (Malm et al., 2002). In general, nitrate concentrations are higher during the winter, and they are

also higher in urban areas because of high nitrogen oxide (NO_x) emissions from automobiles. During the summer, nitrate and ammonium concentrations in the western United States are low. TCM concentrations explain most of total $\text{PM}_{2.5}$ mass. It is found that TCM has high concentrations in the summer and fall seasons because of high fire-related activity. During the spring season, crustal material concentrations are high in the eastern United States. Our results for the speciated $\text{PM}_{2.5}$ are consistent with previous analyses (Malm et al., 2004).

Our approach has some limitations. The computational burden prohibits a fully-Bayesian analysis. We employ a two-stage algorithm that first fixes the sum of the five components and then estimates the individual components. It is not clear how much uncertainty is ignored by this approach. Also, the spatiotemporal process for total $\text{PM}_{2.5}$ ($Z(\mathbf{s}, t)$) and speciated $\text{PM}_{2.5}$ ($\mathbf{w}(\mathbf{s}, t)$) could be modeled as a nonstationary and/or nonseparable spatial-temporal process. However, the computational burden is exacerbated in these cases so we use simple spatial-temporal models.

The multivariate spatiotemporal model could also be applied in other areas, such as meteorology, ecological modeling, and exposure analysis. We are currently working on association between speciated $\text{PM}_{2.5}$ and daily mortality. The framework and results presented here will be essential for the health analysis.

Acknowledgements

The authors would like to thank Drs. Holland, Tesh, and Prakash at EPA for providing the data as well as helpful and insightful discussions.

References

- Baumgardner, R. E., Isil, S. S., Bowser, J. J., and Fitzgerald, K. M. (1999), “Measurements of rural sulfur dioxide and particle sulfate: Analysis of CASTNet data, 1987 through 1996,” *Journal of Air Waste Management Association*, 49, 1266-1279.
- Flanagan, J. B., Peterson, M. R., Jayanty, R. K. M., and Rickman, E. E. (2003), *Analysis of Speciation Network Carbon Blank Data*, RTP, NC: RTI International.
- Frank N. H. (2006), “Retained nitrate, hydrated sulfates, and carbonaceous mass in federal reference method fine particulate matter for six eastern U.S. cities,” *Journal of Air Waste Management Association*, 56, 500-511.
- Fuentes, M., Song, H., Ghosh, S. K., Holland, D. M., and Davis, J. M. (2006), “Spatial association between speciated fine particles and mortality,” *Biometrics*, 62, 855-863.
- Gelfand, A. E., and Smith, A. F. M. (1990), “Sampling-based approaches to calculating marginal densities,” *Journal of the American Statistical Association*, 85, 398-409.
- Gelfand, A. E., Banerjee, S., and Gamerman, D. (2005), “Spatial process modelling for univariate and multivariate dynamic spatial data,” *Environmetrics*, 16, 465-479.

- Gelfand, A.E., Schmidt, A.M., Banerjee, S., and Sirmans, C.F. (2004), Nonstationary Multivariate Process Modelling through Spatially Varying Coregionalization (with discussion), *Test* 13, 1-50.
- Grzebyk, M., and Wackernagel, H. (1994), "Multivariate analysis and spatial/temporal scales: real and complex models," in *Proceedings of the XVIIth International Biometrics Conference*, pp. 19-33.
- Malm, W. C., Schichtel, B. A., Ames, R. B., and Gebhart, K. A. (2002), "A 10-year spatial and temporal trend of sulfate across the United States," *Journal of Geophysical Research*, 107, D224627, doi:10.1029/2002JD002107.
- Malm, W. C., Schichtel, B. A., Pitchford, M. L., Ashbaugh, L. L., and Eldred, R. A. (2004), "Spatial and monthly trends in speciated fine particle concentration in the United States," *Journal of Geophysical Research*, 109, D03306, doi:10.1029/2003JD003739.
- McFadden, D. (1974), "Conditional logit analysis of qualitative choice behavior," in *Frontiers of Econometrics*, ed. P. Zarembka. New York: Academic Press, pp.105-142.
- Özkaynak, H., and Thurston, G. D. (1987), "Associations between 1980 U.S. mortality rates and alternative measures of airborne particle concentration," *Risk Analysis*, 7, 449-461.
- Rao, V., Frank, N., Rush, A., and Dimmick, F. (2003), "Chemical Speciation of PM_{2.5} in Urban and Rural Areas," *National Air Quality and Emissions Trends Report*, pp. 13-23.

- Spiegelhalter, D. J., Best, N. G., Carlin, B.P., and van der Linde, A. (2002), “Bayesian measures of model complexity and fit” (with discussion), *Journal of the Royal Statistical Society B*, 64, 583-639.
- U.S. Environmental Protection Agency (1997), “National Ambient Air Quality Standards for Particulate Matter; Final Rule, Part II,” *Federal Register*, 40, CFR Part 50.
- U.S. Environmental Protection Agency (2000), *Quality Assurance Guidance Document, Quality assurance Project Plan: PM_{2.5} Speciation Trends Network Field Sampling*, EPA 454/R-01-001, RTP, NC, Available at: <http://www.epa.gov/ttn/amtic/files/ambient/pm25/spec/1025sqap.pdf>.
- U.S. Environmental Protection Agency (2003), *National Air Quality and Emissions Trends Report, 2003*, Special Studies Edition, EPA 454/R-03-005, RTP, NC, Available at: <http://www.epa.gov/air/airtrends/aqtrnd03/>.
- Wackernagel, H. (1998), *Multivariate Geostatistics-An Introduction with applications* (2nd ed.), New York: Springer-Verlag.
- Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, M. (2001), “Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds,” *Journal of the American Statistical Association*, 95, 1076-1987.

Appendix: MCMC details

Here we describe separately MCMC algorithms for our models for total $\text{PM}_{2.5}$ (Section 3.4) and speciated $\text{PM}_{2.5}$ (Sections 3.2 and 3.3).

Algorithm for the sum of the five components

Missing values are updated at each MCMC iteration by drawing new values from their full-conditionals in (11). Given the complete data, blocked Gibbs sampling with day as the block unit is used to update δ (described below (3)) from its Gaussian conditional with

$$\text{Cov}(\delta(\cdot, t) | \text{rest}) = (\tau_F I + \tau_R I + (1 + \psi_z^2) * \Sigma^{-1})^{-1} \quad (12)$$

$$\text{E}(\delta(\cdot, t) | \text{rest}) = \text{Cov}(\delta(\cdot, t) | \text{rest}) (\tau_F r_F(\cdot, t) + \tau_R r_R(\cdot, t) + \psi_z \Sigma^{-1} [\delta(\cdot, t-1) + \delta(\cdot, t+1)]),$$

where τ_F and τ_R are the inverse variances of e_z and e_r , respectively, Σ is the spatial covariance matrix with elements $\sigma_\delta^2 \exp(-\|\mathbf{s} - \mathbf{s}'\|/\phi_\delta)$, $r_F(s, t) = \hat{Z}_F(s, t) - a(s, t) - M^T(s, t)\beta(s, t)$, and $r_R(s, t) = \hat{Z}_R(s, t) - M^T(s, t)\beta(s, t)$.

The regression parameters β for each of the nine regions in (3) have Gaussian conditionals and are updated individually using Gibbs sampling. The parameters that define a in (5) are updated using blocked Gibbs sampling similar to the algorithm for δ . The remaining variance/covariance parameters are updated using Metropolis-Hastings sampling.

Algorithm for speciated $\text{PM}_{2.5}$

As with total $\text{PM}_{2.5}$, missing speciated values are updated at each MCMC iteration by drawing new values from their full-conditionals in (1). However, unlike total $\text{PM}_{2.5}$, the proportions are not conjugate and must be updated using Metropolis sampling. $\delta_k(s, t)$ in (7) are updated individually using Gaussian candidate distributions, tuned to have acceptance ratios near 0.40. The STLMC parameters w in (8) are updated using blocked Gibbs sampling, similar to (12). The remaining variance/covariance parameters, including A in (8), are updated using Metropolis-Hastings sampling.

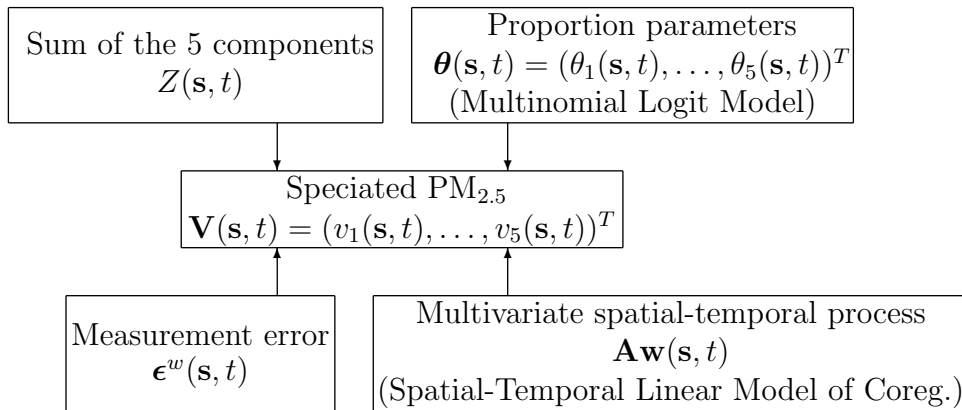
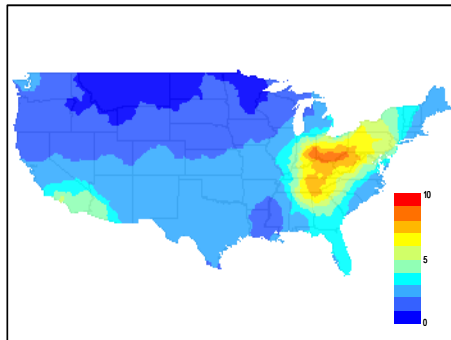
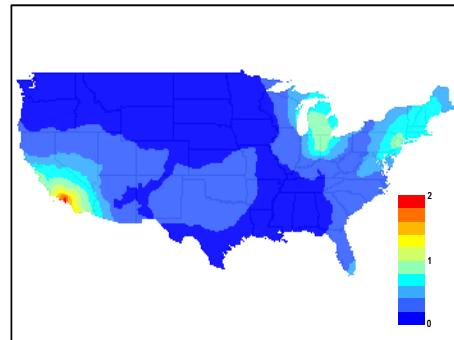


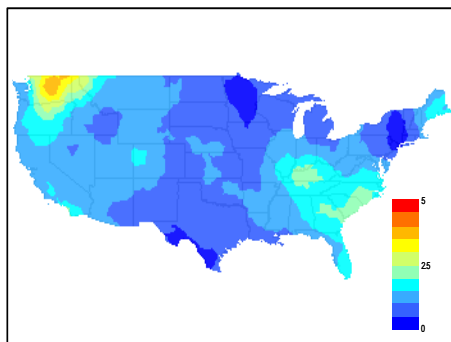
Figure 1: Model framework for speciated $\text{PM}_{2.5}$.



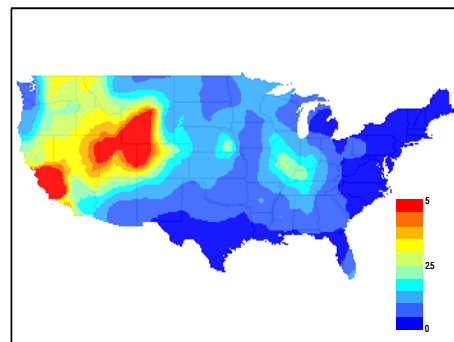
(a) Sulfate (June 14)



(b) Nitrate (June 14)



(c) Sulfate (December 14)



(d) Nitrate (December 14)

Figure 2: Maps of the posterior mean of sulfate concentration ($\mu g/m^3$) and nitrate concentration ($\mu g/m^3$) on June 14, 2004 and on December 14, 2004.

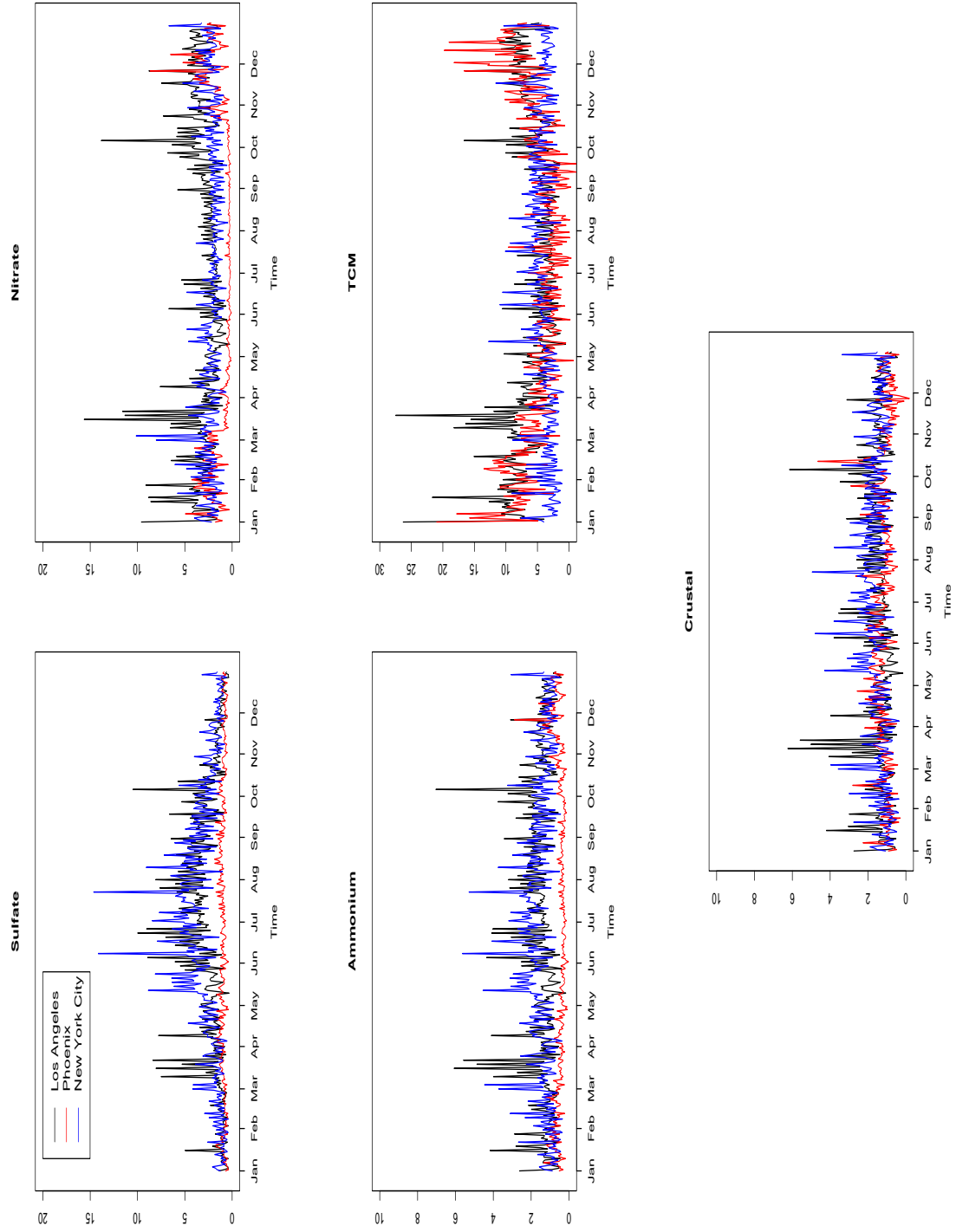
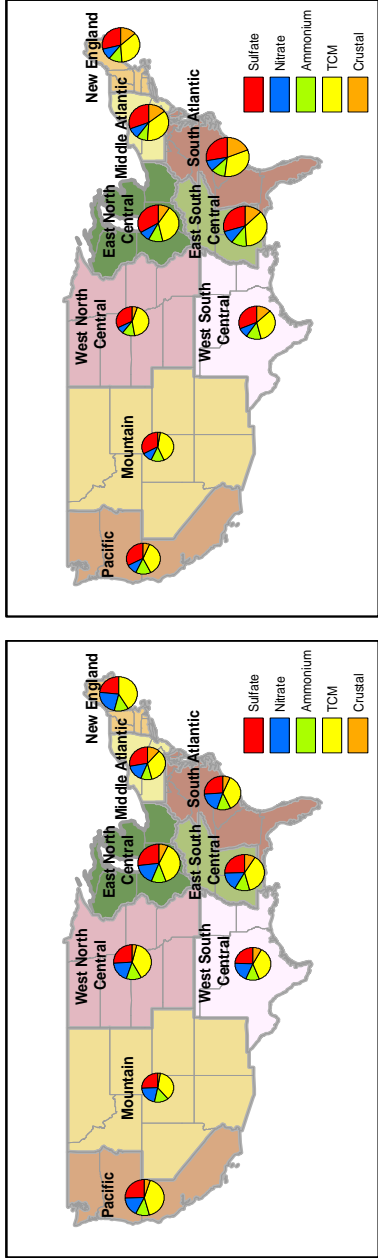
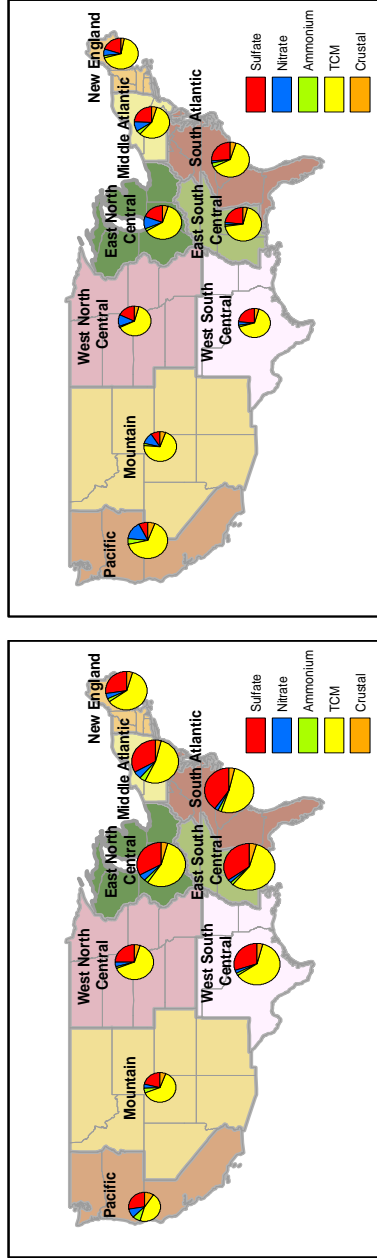


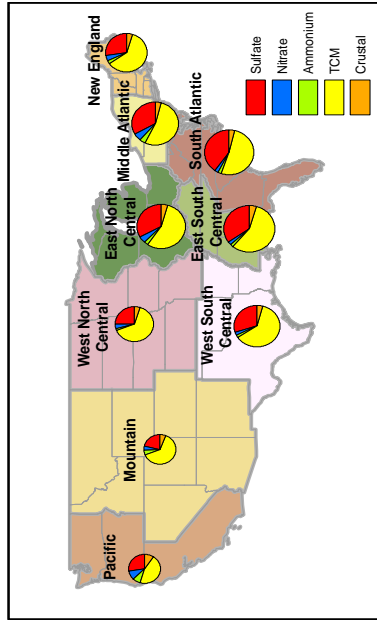
Figure 3: Time series plots of the estimated speciated $PM_{2.5}$ ($\mu g/m^3$) for three cities (Los Angeles, Phoenix, and New York City) in 2004.



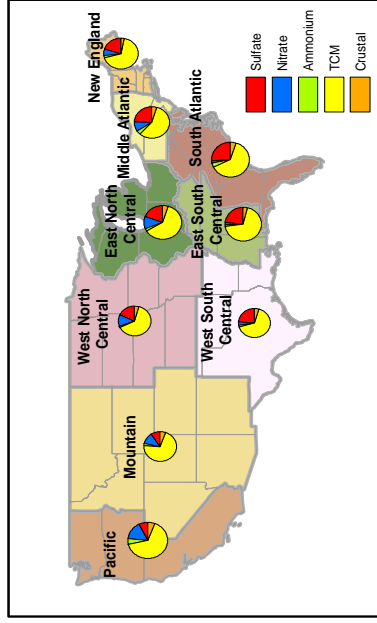
(a) January-March



(b) April-June



(c) July-September



(d) October-December

Figure 4: Maps of estimated speciated $PM_{2.5}$ composition by region and by season in 2004.

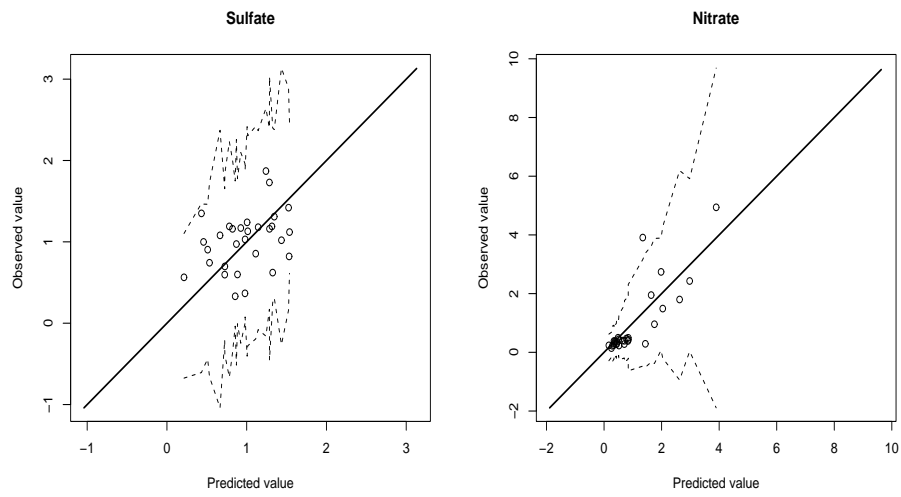


Figure 5: Model diagnostics for sulfate and nitrate in Phoenix: STN values of each component versus the mean of the predictive posterior distribution of the component values at time t eliminating the t^{th} observation. The dotted lines show the 95% prediction intervals.

Table 1: Posterior quantiles of the correlations between components, (i.e., $T_{ij}/\sqrt{T_{ii}T_{jj}}$) in California in 2004.

Parameters	2.5%	Mean	97.5%
Sulfate/Nitrate	-0.497	-0.151	-0.058
Sulfate/Ammonium	-0.131	0.448	0.607
Sulfate/TCM	-0.934	0.042	0.101
Sulfate/Crustal	-0.095	0.869	0.943
Nitrate/Ammonium	0.434	0.836	0.994
Nitrate/TCM	-0.095	0.038	0.286
Nitrate/Crustal	-0.741	-0.223	-0.104
Ammonium/TCM	-0.699	-0.558	-0.156
Ammonium/Crustal	-0.604	0.024	0.279
TCM/Crustal	-0.767	0.674	0.991