Multivariate spectral downscaling for PM$_{2.5}$ species

Yawen Guan*

Department of Statistics, University of Nebraska

Brian J Reich

Department of Statistics, North Carolina State University

James A Mulholland

School of Civil & Environmental Engineering, Georgia Tech

Howard H Chang

Department of Biostatistics and Bioinformatics, Emory University

*email: yguan12@unl.edu

SUMMARY: Fine particulate matter (PM$_{2.5}$) is a mixture of air pollutants that has adverse effects on human health. Understanding the health effects of PM$_{2.5}$ mixture and its individual species has been a research priority over the past two decades. However, the limited availability of speciated PM$_{2.5}$ measurements continues to be a major challenge in exposure assessment for conducting large-scale population-based epidemiology studies. The PM$_{2.5}$ species have complex spatial-temporal and cross dependence structures that should be accounted for in estimating the spatiotemporal distribution of each component. Two major sources of air quality data are commonly used for deriving exposure estimates: point-level monitoring data and gridded numerical computer model simulation, such as the Community Multiscale Air Quality (CMAQ) model. We propose a statistical method to combine these two data sources for estimating speciated PM$_{2.5}$ concentration. Our method models the complex relationships between monitoring measurements and the numerical model output at different spatial resolutions, and we model the spatial dependence and cross dependence among PM$_{2.5}$ species. We apply the method to combine CMAQ model output with major PM$_{2.5}$ species measurements in the contiguous United States in 2011.

KEY WORDS: Statistical downscaling, Multiresolution, Spectral analysis, Multivariate spatial, PM$_{2.5}$ species.

1. Introduction

Fine particulate matter (PM$_{2.5}$) is a mixture of air pollutants with aerodynamic diameters 2.5 micrometers and smaller. Exposure to PM$_{2.5}$ concentration has been linked to a variety of health problems, such as respiratory and cardiovascular diseases (Atkinson et al., 2014), as well as adverse birth outcomes (Li et al., 2017). However, while the health effects of the total PM$_{2.5}$ mass are well understood, there are limited and contradictory findings in the associations of different PM chemical components with the same health endpoints, potentially due to the differences in accuracy of exposure information (Grahame, 2009). The main challenges in predicting species exposure arise from limited observations and their complex correlation. The emission sources of PM$_{2.5}$, such as motor vehicles, coal-fired power plants and other manufacturing facilities are widespread and spatially heterogeneous. That, combined with the complex chemical reactions among pollutants result in highly temporally and spatially varying composition of PM$_{2.5}$. There is a need in developing computationally efficient statistical methods that can capture the complex structure and make reliable maps of species concentrations, which will facilitate epidemiologic research on identifying more reliable associations between specific species and health endpoints.

A common approach to produce accurate maps is to fuse disparate data sources such as deterministic computer model output and observations from monitoring stations (see, e.g., Gotway and Young, 2002; Cressie and Wikle, 2011, for reviews). For single pollutant predictions, numerous spatiotemporal methods have been developed to downscale gridded model output to point-level (e.g., Fuentes and Raftery, 2005; Berrocal et al., 2010b; McMillan et al., 2010). Extensions of these methods have also been proposed to handle ensemble forecasts (e.g., Berrocal et al., 2007; Feldmann et al., 2015; Schefzik, 2017), distribution matching (e.g., Clark et al., 2004; Gel et al., 2004; Huang et al., 2019), extreme value analysis (e.g., Mannshardt-Shamseldin et al., 2010; Reich et al., 2013; Bechler et al., 2015), and to

account for forecast errors (Berrocal et al., 2012). Less attention is paid for joint downscaling of multiple pollutants, where capturing the complex interactions and dependence among species are essential to improve accuracy. A few most relevant methods for our purposes are summarized here. Berrocal et al. (2010a) propose a bivariate spatial downscaler for PM$_{2.5}$ and ozone that both accounts for bias in the model output for each type of air pollution and also borrows strength across pollutants to improve precision. Choi et al. (2009), Crooks and Özkaynak (2014), and Rundel et al. (2015) jointly model speciated PM$_{2.5}$ and total PM$_{2.5}$ combining data from different monitoring networks. Huang et al. (2018) jointly model several pollutants and their health effects. These methods correct for the additive and multiplicative biases in the numerical model, but pay little attention to such biases in different spatial scales of the numerical model.

We propose a multivariate spectral downscaler for speciated PM$_{2.5}$. We model the associations between the numerical model output and observations at multiple resolutions using a spectral analysis. This provides insights on the performance of the numerical model at different spatial scales and can potentially improve predictions by using only the appropriate scales. We further exploit the cross and spatial dependence among observations by joint modeling multiple pollutants to improve prediction. This is particularly useful for PM$_{2.5}$ constituents, because their observations are more sparse than the total PM$_{2.5}$ due to the limited number of monitors and lower sampling frequencies. The proposed method is fit using parallel MCMC (Scott et al., 2016), where inference based on the full data is made by combining posterior distributions based on small batches of the data set. This work is an extension of the univariate spectral downscaler proposed by Reich et al. (2014), where their focus is prediction for a single pollutant.

The rest of this paper is organized as follows. Section 2 introduces the data sources used in the proposed downscaler. Section 3 provides details on the spectral analysis, the univariate

and multivariate spectral downscalers and model fitting. Section 4 presents the data analysis results for the contiguous US in 2011. We conclude with Section 5 giving a brief summary and some final remarks on the proposed approach. Details on exploratory data analysis to verify model assumptions and a simulation study are presented in the Supplementary Materials.

2. Data

We consider daily 24-hour average estimates and measurements of total PM$_{2.5}$ and its major constituents: elemental carbon (EC), organic carbon (OC), nitrate (NO$_3$), sulfate (SO$_4$) and ammonium (NH$_4$). Two sources of information were obtained from the United States Environmental Protection Agency (EPA): monitoring data for the Air Quality System and model output from the Community Multiscale Air Quality (CMAQ) model. The raw station data are right skewed, therefore, we analyze the log transformed station data and apply the same transformation on the CMAQ data (see Supplementary Materials S.1 for plots of the station data before and after transformation).

In our analysis, monitoring data come from 845 stations (including all PM2.5 monitors for regulatory purposes and monitors from SPECIATE network) located throughout the contiguous United States (a map of station locations is shown in the Supplementary Material S.1). Not all stations measure the same pollutants, among them, 734 measure only total PM$_{2.5}$, 15 measure only the constituents, 95 measure both total PM$_{2.5}$ and its constituents, and 1 measures only NO$_3$ and SO$_4$. For total PM$_{2.5}$, these measurements are taken daily, 1-in-3 or 1-in-6 days; for speciated PM$_{2.5}$, these measurements are mostly taken 1-in-3 or 1-in-6 days. Figure 1 (left) shows the monitoring data for both PM$_{2.5}$ and EC on Jan 1st and Jan 3rd, 2011. The PM$_{2.5}$ total and its species measurements are spatially sparse and not all are measured at the same locations (referred to as spatial misalignment), and the sampling locations of PM$_{2.5}$ species vary greatly across days.

CMAQ is a deterministic model that simulates daily pollutant concentrations accounting

for emission sources, and complex atmospheric chemistry and physics. It provides gridded maps of multipollutant concentrations, which can serve as proxy data for locations without a monitoring station. The version v5.02 used in this paper provides output at 12 km spatial resolution, on a grid of 299 by 459 cells for the contiguous US. As an example, Figure 1 (right) shows the CMAQ output for both PM$_{2.5}$ and EC on Jan 1st and Jan 3rd, 2011.

[Figure 1 about here.]

3. Modeling Framework

Here we present a multivariate spectral downscaler that models the complex relationship among pollutants and two data sources. We begin by reviewing the univariate spectral downscaler proposed by Reich et al. (2014), where they analyzed the correlation between CMAQ and monitoring data of ozone in the spectral domain. For simplicity, we first present purely spatial multivariate downscaling and then discuss the temporal aspect.

3.1 *Spectral Representation*

We provide a brief background on spectral analysis; a more comprehensive review is given by Fuentes and Reich (2010). A stationary spatial process $X(\boldsymbol{s})$ has a spectral representation,

$$X(\boldsymbol{s}) = \int \exp(i\boldsymbol{\omega}^T \boldsymbol{s})Z(\boldsymbol{\omega})d\boldsymbol{\omega}, \tag{1}$$

where $\boldsymbol{s} \in R^2$ is a spatial location, $\boldsymbol{\omega} \in R^2$ is a frequency and $Z(\boldsymbol{\omega})$ is the spectral process corresponding to $X(\boldsymbol{s})$. If we further assume that $X(\boldsymbol{s})$ is a mean-zero Gaussian process, then $Z(\boldsymbol{\omega})$ is also Gaussian with $E[Z(\boldsymbol{\omega})] = 0$, $V[Z(\boldsymbol{\omega})] = \sigma^2 f(\boldsymbol{\omega})$, and $Cov[Z(\boldsymbol{\omega}), Z(\boldsymbol{\omega}')] = 0$. The smoothness of the spatial process is determined by the spectral density $f(\boldsymbol{\omega})$, which weights the frequencies. If more weight is given to small frequencies then the spatial process will be smoother; if more weight is given to large frequencies then the spatial process will be relatively rougher. Spectral methods are particularly useful for decorrelating spatial processes and decomposing spatial variation over different scales (or resolutions).

To illustrate how the spectral representation decomposes the spatial signal into different scales, we extract and plot the signal corresponding to different $||\boldsymbol{\omega}||$. For spatial process observed on a grid $\mathcal{D}(m_1 \times m_2)$ with $M = m_1 m_2$ equally spaced locations, we can compute $Z(\boldsymbol{\omega}_l), l = 1, \ldots, M$ by taking inverse discrete Fourier transform of X. The available Fourier frequencies are $\boldsymbol{\omega}_l \in 2\pi \mathcal{J}_M$, for $\mathcal{J}_M = m_1^{-1}\{\lfloor -(m_1-1)/2 \rfloor, \ldots, m_1 - \lfloor m_1/2 \rfloor\} \times m_2^{-1}\{\lfloor -(m_2-1)/2 \rfloor, \ldots, m_2 - \lfloor m_2/2 \rfloor\}$. Then the spatial signal with frequencies $||\boldsymbol{\omega}|| \in [L, U)$ is approximated by the integral in (1) integrated over [L,U)

$$\tilde{X}_{[L,U)}(\boldsymbol{s}) = \sum_{||\boldsymbol{\omega}_l|| \in [L,U)} \exp(i\boldsymbol{\omega}_l^T \boldsymbol{s}) Z(\boldsymbol{\omega}_l), \tag{2}$$

For CMAQ gridded multipollutant concentrations, we explore the signals at different scales for each pollutant to see their contributions to the spatial process. An example of the extracted signals for CMAQ PM$_{2.5}$ on Sept 21, 2011 is shown in Figure 2; here, we divide the range of available frequency magnitude $[0, \sqrt{2}\pi]$ into 8 equal-width bins with width $\pi/5$, and filter the PM$_{2.5}$ concentration to retain the signals corresponding to each bin. We see that the signals at low frequencies $[0, \pi/5)$ (Figure 2a) resemble the large spatial-trend of CMAQ PM$_{2.5}$ (Figure 2e), while at higher frequencies (Figures 2b - 2c) resemble small-scale information (Figure 1e); finally, the signals corresponding to high frequencies (Figure 2d) resembles large-scale features, as a consequence of the aliasing effect (see Supplementary Materials S.2).

[Figure 2 about here.]

3.2 *Univariate Spectral Downscaler*

The univariate spectral downscaler of Reich et al. (2014) models the association between monitoring data and CMAQ for a single pollutant. Let $Y(\boldsymbol{s})$ be monitoring data and $X(\boldsymbol{s})$ be the CMAQ output at the grid cell containing the station location $\boldsymbol{s}$. A joint spectral model for two stationary mean-zero spatial processes is

$$X(\boldsymbol{s}) = \int \exp(i\boldsymbol{\omega}^T\boldsymbol{s})Z_1(\boldsymbol{\omega})d\boldsymbol{\omega}$$

$$Y(\boldsymbol{s}) = \int \exp(i\boldsymbol{\omega}^T\boldsymbol{s})Z_2(\boldsymbol{\omega})d\boldsymbol{\omega}, \qquad (3)$$

where $Z_i(\boldsymbol{\omega}), i = 1, 2$ are Gaussian spectral processes with $E[Z_i(\boldsymbol{\omega})] = 0$, $V[Z_i(\boldsymbol{\omega})] = \sigma_i^2 f_i(\boldsymbol{\omega})$, and $Cov[Z_i(\boldsymbol{\omega}), Z_j(\boldsymbol{\omega}')] = 0, i, j = 1, 2$ for $\boldsymbol{\omega} \neq \boldsymbol{\omega}'$. The dependence between the two processes is captured via their correlation at each frequency $Cor[Z_1(\boldsymbol{\omega}), Z_2(\boldsymbol{\omega})] = \phi(\boldsymbol{\omega}) = \phi(-\boldsymbol{\omega}) \in (-1, 1)$. For example, if the two processes are highly correlated at the low frequency (small $||\boldsymbol{\omega}||$), they should exhibit similar large-scale spatial trends; if high correlation exits at the high frequency (large $||\boldsymbol{\omega}||$), then the two process should exhibit similar small-scale spatial features.

Since CMAQ data X are known at all grid points, we consider the conditional model of monitoring data given CMAQ. In the spectral domain, we can write the conditional distribution of the spectral processes for each frequency as follows,

$$Z_2(\boldsymbol{\omega}) = \alpha(\boldsymbol{\omega})Z_1(\boldsymbol{\omega}) + Z^*(\boldsymbol{\omega}), \qquad (4)$$

where $\alpha(\boldsymbol{\omega}) = \phi(\boldsymbol{\omega})\frac{\sigma_2\sqrt{f_2(\boldsymbol{\omega})}}{\sigma_1\sqrt{f_1(\boldsymbol{\omega})}}$, $Z^*(\boldsymbol{\omega})$ is a Gaussian variable with $E[Z^*(\boldsymbol{\omega})] = 0$ and $Var[Z^*(\boldsymbol{\omega})] \leqslant Var[Z_2(\boldsymbol{\omega})]$, independent of $Z_1(\boldsymbol{\omega})$ and independent over $\boldsymbol{\omega}$. In the spatial domain, the conditional distribution of $Y(\boldsymbol{s})|X(\boldsymbol{s}')$ for all $\boldsymbol{s}'$ integrates over all frequencies,

$$Y(\boldsymbol{s})|X(\boldsymbol{s}') \text{ for all } \boldsymbol{s}' = \int \exp(i\boldsymbol{\omega}^T\boldsymbol{s})\alpha(\boldsymbol{\omega})Z_1(\boldsymbol{\omega})d\boldsymbol{\omega}$$

$$+ \int \exp(i\boldsymbol{\omega}^T\boldsymbol{s})Z^*(\boldsymbol{\omega})d\boldsymbol{\omega}, \qquad (5)$$

where the first term is the conditional mean of $Y(\boldsymbol{s})$ given X, and the second term is a mean-zero Gaussian process that represents the remaining variability not captured X.

3.3 *Multivariate Spectral Downscaler*

We extend the univariate spectral downscaler to the multivariate setting, such that the cross-species association between CMAQ and station data can be estimated from the data. Let

$\boldsymbol{X}(\boldsymbol{s}) \in \mathcal{R}^J$ and $\boldsymbol{Y}(\boldsymbol{s}) \in \mathcal{R}^K$ be vectors of the multiple pollutants from CMAQ and stations, respectively. A joint spectral model for $\boldsymbol{X}(\boldsymbol{s})$ and $\boldsymbol{Y}(\boldsymbol{s})$ is,

$$\begin{aligned}
\boldsymbol{X}(\boldsymbol{s}) &= \int \exp(i\boldsymbol{\omega}^T\boldsymbol{s})\boldsymbol{Z_1}(\boldsymbol{\omega})d\boldsymbol{\omega} \\
\boldsymbol{Y}(\boldsymbol{s}) &= \int \exp(i\boldsymbol{\omega}^T\boldsymbol{s})\boldsymbol{Z_2}(\boldsymbol{\omega})d\boldsymbol{\omega},
\end{aligned} \tag{6}$$

where $\boldsymbol{Z_1}(\boldsymbol{\omega}) \in \mathcal{R}^J$ and $\boldsymbol{Z_2}(\boldsymbol{\omega}) \in \mathcal{R}^K$ are mean-zero Gaussian vector with $J \times J$ and $K \times K$ covariance matrix $\Sigma_{ii}(\boldsymbol{\omega})$ for i=1 and 2 respectively, and $J \times K$ covariance matrix $\Sigma_{12}(\boldsymbol{\omega})$.

Similar to the univariate case, we model station data conditional on CMAQ. It can be shown that the conditional model at frequency $\boldsymbol{\omega}$ is,

$$\boldsymbol{Z}_2(\boldsymbol{\omega}) = A(\boldsymbol{\omega})\boldsymbol{Z}_1(\boldsymbol{\omega}) + \boldsymbol{Z}^*(\boldsymbol{\omega}), \tag{7}$$

where $A(\boldsymbol{\omega}) = \Sigma_{21}(\boldsymbol{\omega})\Sigma_{11}^{-1}(\boldsymbol{\omega})$ is a $K \times J$ matrix, and $\boldsymbol{Z}^*(\boldsymbol{\omega}) \in \mathcal{R}^K$ is a K-dimensional mean-zero Gaussian vector with $\mathrm{Cov}[\boldsymbol{Z}^*(\boldsymbol{\omega})] = \Sigma_{22}(\boldsymbol{\omega}) - \Sigma_{21}(\boldsymbol{\omega})\Sigma_{11}^{-1}(\boldsymbol{\omega})\Sigma_{12}(\boldsymbol{\omega})$ and independent of $\boldsymbol{Z}_1(\boldsymbol{\omega})$. The resulting conditional model in the spatial domain is

$$\begin{aligned}
\boldsymbol{Y}(\boldsymbol{s})|\boldsymbol{X}(\boldsymbol{s}') \text{ for all } \boldsymbol{s}' &= \int \exp(i\boldsymbol{\omega}^T\boldsymbol{s})A(\boldsymbol{\omega})\boldsymbol{Z}_1(\boldsymbol{\omega})d\boldsymbol{\omega} \\
&+ \int \exp(i\boldsymbol{\omega}^T\boldsymbol{s})\boldsymbol{Z}^*(\boldsymbol{\omega})d\boldsymbol{\omega},
\end{aligned} \tag{8}$$

where the first term is the conditional mean $\boldsymbol{\mu}(\boldsymbol{s}) = E[\boldsymbol{Y}(\boldsymbol{s})|\boldsymbol{X}(\boldsymbol{s}') \text{ for all } \boldsymbol{s}']$ and the second term is a mean-zero stationary multivariate Gaussian process.

Since the regression coefficient matrix $A(\boldsymbol{\omega})$ between the two spectral processes are unknown at every frequency, extending Reich et al. (2014) to the multivariate setting, we model each element of the $K \times J$ matrix as a linear combination of B B-spline basis functions $\mathcal{B}_1(\boldsymbol{\omega}), \ldots, \mathcal{B}_B(\boldsymbol{\omega})$,

$$A_{kj}(\boldsymbol{\omega}) = \sum_{b=1}^{B} \beta_{kjb}\mathcal{B}_b(\boldsymbol{\omega}). \tag{9}$$

Substituting (9) into (8), the conditional mean becomes $\mu_k(\boldsymbol{s}) = \sum_{j=1}^{J}\sum_{b=1}^{B}\tilde{X}_{jb}(\boldsymbol{s})\beta_{kjb}$, where $\tilde{X}_{jb}(\boldsymbol{s}) = \int \mathcal{B}_b(\boldsymbol{\omega})\exp(i\boldsymbol{\omega}^T\boldsymbol{s})Z_{1j}(\boldsymbol{\omega})d\boldsymbol{\omega}$ are the constructed spectral covariates. For data on a grid, the integration is approximated by $\tilde{X}_{jb} = \sum_{l=1}^{M}\mathcal{B}_b(\boldsymbol{\omega}_l)\exp(i\boldsymbol{\omega}_l^T\boldsymbol{s})Z_{1j}(\boldsymbol{\omega}_l),$

which can be computed efficiently by applying fast Fourier transform (FFT) twice. We first obtain $\hat{Z}_{1j}(\boldsymbol{\omega}_l), l = 1, \ldots, M$ by applying inverse FFT on X_j, then weight the signals by multiplying the bth basis function evaluated at $||\boldsymbol{\omega}_l||$ and finally applying FFT on the product $\hat{Z}_{1j}(\boldsymbol{\omega}_l)\mathcal{B}_b(\boldsymbol{\omega}_l)$.

Since CMAQ data are known, all the spectral covariates only need to be computed once prior to model fitting, then the multivariate spectral downscaler can be written as a multivariate spatial regression model. To simplify notation, let $\boldsymbol{Y}_k = [Y_k(\boldsymbol{s}_1), \ldots, Y_k(\boldsymbol{s}_{n_k})]^T$ collect all station data for the k-th pollutant, and let $\tilde{\boldsymbol{X}}_{jb} = \left[\tilde{X}_{jb}(\boldsymbol{s}_1), \ldots, \tilde{X}_{jb}(\boldsymbol{s}_{n_k})\right]^T$ be the collection of spectral covariates for the j-th pollutant and b-th basis at the grid cell containing the station locations. Then the multivariate spatial regression model for the observations is

$$\boldsymbol{Y}_k = \beta_{k_0} + \sum_{j=1}^{J}\sum_{b=1}^{B} \beta_{kjb}\tilde{\boldsymbol{X}}_{jb} + \boldsymbol{w}_k + \boldsymbol{\epsilon}_k, k = 1, \ldots, K, \tag{10}$$

where $\boldsymbol{w} = [\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_K^T]^T$ are multivariate spatial random effects to capture the remaining dependence (details provided in the next section), and $\boldsymbol{\epsilon}_k$ are independent measurement errors corresponding to the k-th pollutant, $\boldsymbol{\epsilon}_k \sim N(0, \tau_k^2 \boldsymbol{I})$.

3.4 *Multivariate Spatial Model*

Exploratory analysis of the residuals after accounting for the spectral covariates suggests they have negligible temporal dependence. We model $\boldsymbol{w}(\boldsymbol{s}) = [w_1(\boldsymbol{s}), \ldots, w_K(\boldsymbol{s})]^T$ as a purely-spatial multivariate process. Here we introduce the process as if there are no missing data at each location, but the covariance matrix for $\boldsymbol{w}$ the stacked random effects with missing data is formed explicitly during model fitting. A commonly-used and easy-to-interpret model for multivariate spatial process is the linear model of coregionalization (LMC) proposed by Gelfand et al., (2004), where the process is expressed as a linear combination of independent spatial processes $\boldsymbol{w}(\boldsymbol{s}) = L\boldsymbol{v}(\boldsymbol{s})$. Without loss of generality, L is taken as a $K \times K$ lower triangular matrix and $\boldsymbol{v}(\boldsymbol{s}) = [v_1(\boldsymbol{s}), \ldots, v_K(\boldsymbol{s})]^T$ with $v_k(\boldsymbol{s}) \overset{ind}{\sim} GP(0, \rho_k(\cdot, \phi_k))$. That is, $\text{Cov}(v_k(\boldsymbol{s}), v_{k'}(\boldsymbol{s}'))$ is zero for $k \neq k'$, and $\rho_k(||\boldsymbol{s} - \boldsymbol{s}'||, \phi_k)$ otherwise; the spatial dependence

of each process is determined by the covariance function ρ_k. For convenience, we use the exponential covariance function, $\rho_k(||\boldsymbol{s} - \boldsymbol{s}'||, \phi_k) = \exp(-\phi_k||\boldsymbol{s} - \boldsymbol{s}'||)$, which assumes a decreasing spatial dependence as the Euclidean distance $||\boldsymbol{s} - \boldsymbol{s}'||$ increases with the rate of decay controlled by the spatial decay parameter ϕ_k.

This model is easy to interpret as the cross dependence and spatial dependence for the multivariate spatial process is explicit. For any location $\boldsymbol{s}$, the variance-covariance matrix for the cross dependence is $C = LL^T$, i.e., $\text{Cov}\{w_i(\boldsymbol{s}), w_j(\boldsymbol{s})\} = C_{ij}$. The spatial dependence within the i-th pollutant is $\text{Cov}\{w_i(\boldsymbol{s}), w_i(\boldsymbol{s}')\} = \sum_{j=1}^{K} L_{ij}^2 \rho_j(||\boldsymbol{s} - \boldsymbol{s}'||, \phi_j)$, and between the i-th and j-th pollutants is $\text{Cov}\{w_i(\boldsymbol{s}), w_j(\boldsymbol{s}')\} = \sum_{k=1}^{K} L_{ik}L_{jk}\rho_k(||\boldsymbol{s} - \boldsymbol{s}'||, \phi_k)$. For a model with missing data, we can denote the observation for a pollutant by a vector and stack all vectors to form a larger vector as shown in Equation (10); then, the elements of covariance matrix for $\boldsymbol{w}$ are explicit. Furthermore, this LMC model provides a spectral representation for the remaining variability after accounting for $\boldsymbol{X}$. The second term $\boldsymbol{Z}^*(\boldsymbol{\omega})$ in Equation (7) is therefore a mean-zero stationary multivariate Gaussian process with $\text{Cov}[\boldsymbol{Z}^*(\boldsymbol{\omega})] = L\Lambda(\boldsymbol{\omega})L' + D(\boldsymbol{\omega})$, where $\Lambda(\boldsymbol{\omega})$ is a diagonal matrix with elements $f_k(\boldsymbol{\omega})$ which are the spectral densities corresponding to $v_k(\boldsymbol{s})$, and $D(\boldsymbol{\omega})$ is a diagonal matrix with elements τ_k^2.

To complete the Bayesian model, we have the following prior specification. Both the additive bias $\boldsymbol{\beta}_0 = (\beta_{10}, \dots, \beta_{K0})^T$ and the mean parameters $\boldsymbol{\beta} = \{\beta_i : i = 1, \dots, p\}$ with dimension $p = KJB$ have independent normal priors $N(0, 100^2)$. For the coregionalization matrix L, we have normal priors $N(0, 10^2)$ for the off-diagonal entries and lognormal priors $LN(0, 10^2)$ for the K diagonal entires. Lastly, we use inverse gamma priors for the nuggets $\tau_k \overset{iid}{\sim} \text{invGamma}(\text{shape} = 2, \text{scale} = 0.1)$ and a uniform prior for the spatial decay parameters $\phi_k = \phi \sim U\left(\frac{3}{0.75 \times d}, \frac{3}{0.1 \times d}\right)$ where d is the maximum distance in the spatial domain; therefore, the prior corresponds to the effective range (defined as the distance at which spatial correlation drops to 0.05) between $0.1 \times d$ and $0.75 \times d$.

3.5 *Computation*

The Bayesian hierarchical model is fit using a MCMC algorithm that involves Gibbs sampling for updating $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$, and $\tau_1^2, \ldots, \tau_K^2$, and Metropolis-Hastings sampling for updating the common range parameter ϕ and entries of L. We fit the model to 3-day batches in parallel, then combine the MCMC samples using the consensus Monte Carlo algorithm (Scott et al., 2016) provided in the R package *parallelMCMCcombine* (Miroshnikov and Conlon, 2014) for inference. Parallel MCMC for "divide and conquer" to leverage parallel computing have been proposed for data with independent subsets (Wang and Dunson, 2013; Neiswanger et al., 2014; Scott et al., 2016). It partitions the data into small independent batches and performs MCMC sampling independently to each batch, then combines the samples to give samples from the approximate full-data posterior for inference. Since CMAQ model has captured most temporal variability in the observations, and the observations are collected typically 1-in-3 and 1-in-6 days, it is reasonable to assume that the multivariate spatial residuals are independent in time and thus the 3-day batches are independent.

Let $\boldsymbol{x}^T = \{x_1, \ldots, x_T\}$ be the full data from a probability density function $p(x|\theta)$ with model parameter θ. If the T observations are conditionally independent and partitioned into M non-overlapping batches, $\{\boldsymbol{x}^{t_1}, \ldots, \boldsymbol{x}^{t_M}\}$, then the full-data posterior and the partition subposterior distributions have the following relationship,

$$p(\theta|\boldsymbol{x}^T) \propto p(\boldsymbol{x}^T|\theta)p(\theta) = \prod_{t=1}^{T} p(x_t|\theta)p(\theta) = \prod_{m=1}^{M} p(\boldsymbol{x}^{t_m}|\theta)p(\theta)^{1/M} \propto \prod_{m=1}^{M} p(\theta|\boldsymbol{x}^{t_m}).$$

We can obtain MCMC samples $\theta_m^{(i)}, i = 1, \ldots, I$ from the subposterior distributions $p(\theta|\boldsymbol{x}^{t_m})$, $m = 1, \ldots, M$, where each MCMC sample is obtained independently. Scott et al. (2016) proposed to combine the samples using weighted averages to form a set of draws $\theta^{(i)}$ from the full-data posterior by letting $\theta^{(i)} = \left(\sum_{m=1}^{M} \Sigma_m^{-1}\right)^{-1} \sum_{m=1}^{M} \Sigma_m^{-1} \theta_m^{(i)}$, where $\Sigma_m = Var(\theta|\boldsymbol{x}^{t_m})$ is estimated by computing the sample covariance of $\theta_m^{(1)}, \ldots, \theta_m^{(I)}$. The full posterior samples are obtained by approximating the subposterior densities with a multivariate normal density

function which is optimal when the subposteriors are near Gaussian or when the sample size is large. In our implementation we found that the subposteriors of the regression coefficients satisfy the Gaussian assumption while the spatial parameters tend to be right skewed. However, the inference for regression coefficients and the prediction performance are not affected by the skewness as illustrated by a simulation (see Supplementary Materials S.4).

4. Analysis of Speciated PM$_{2.5}$ Across the US

4.1 *Exploratory Analysis*

We begin by exploring the relationship between station and CMAQ data at different spatial scales by estimating their associations at different frequency bands. For each CMAQ pollutant output $X_j, j = 1, \ldots, J$, we obtain $\tilde{X}_{j,[\delta_l,\delta_{l+1})}$ using (2) for 8 equal-width bins in the interval $[0, 8\pi/5)$ with $\delta_1 = 0$ and $\delta_8 = 8\pi/5$, this is chosen to cover the range of available frequency with magnitude in $[0, \sqrt{2}\pi]$. Then, we estimate the conditional mean by least squares assuming residual independence,

$$E[Y_k(\boldsymbol{s})|X_j(\boldsymbol{s}') \text{ for all } \boldsymbol{s}' \text{and } j] = \beta_{k_0} + \sum_{j=1}^{J} \sum_{l=1}^{10} \beta_{kjl} \tilde{X}_{j,[\delta_l,\delta_{l+1})}(\boldsymbol{s}), \quad k = 1, \ldots, K. \quad (11)$$

All cross-species predictors are included in the regression mean, for example, EC CMAQ output is used to predict total PM$_{2.5}$ station data, allowing potential complex relationship among pollutants to be estimated. The coefficient estimates from the regression of station data on spectral covariates with cross-species (which have been standardized) are shown in Table 1 in the Supplementary Materials S.3; the coefficients measure the association between station and CMAQ data at multiple spatial scales. Cross-species predictors are significant (highlighted in blue), for instance, CMAQ sulfate at large scale (240km) is a significant predictor for PM$_{2.5}$ suggesting that regional average sulfate is predictive of local total PM$_{2.5}$.

We have used spatially constant slopes β_{kjl} in the conditional mean, which assumes the relationship between the station data and spectral covariates is fixed across contiguous

US. This assumption is verified by estimating Equation (11) separately for 5 US regions: northeast, southeast, midwest, southwest and west (see Supplementary Materials S.3 for region partitions). The differences in the estimated association are mostly insignificant (see Supplementary Materials S.3), therefore in the remaining analysis we combined all regions for mean estimation.

We also analyze the spatiotemporal as well as cross-species dependence of the residuals to verify the assumptions that are made implicitly in the model and model fitting. These include temporal independence after accounting for spectral covariates, isotropic and stationary assumptions implied by the exponential covariance function, and further more assuming the same spatial decay parameter for all pollutants. Details on data exploratory analysis are shown in Supplementary Materials S.3, while a summary is presented here. To assess time dependence, we compute the autocorrelation function (ACF) for each pollutant at each site. For PM$_{2.5}$, the proportions of significant lags at lag-1 and -2 are relatively large, however all constituent species lack daily data to estimate dependence at these lags; for lag-3 and beyond, the autocorrelations are mostly insignificant across all sites for all pollutants. Therefore, it is reasonable to assume residual temporal independence and this agrees with the assumption in Rundel et al (2015). To assess spatial dependence of each pollutant, the daily empirical variogram is computed and a variogram assuming an exponential model is fit to the daily empirical vaiograms combined; we see that spatial range is estimated to be over 100 miles for all pollutants, except NO$_3$ has a smaller range of 29 miles. To assess cross-dependence among pollutants, empirical cross-correlogram is plotted for each pair of pollutants, and it shows that the cross-dependence among pollutants within distance <50 miles could be exploited to improve prediction.

4.2 *Data Analysis*

We assess model performance on spatial interpolation and prediction using cross-validation. For this, we randomly split the monitors into five folds. Each fold has roughly 20% each from monitoring sites measuring PM$_{2.5}$ only, speciated PM$_{2.5}$ only, and both. Four different mean structures are compared: the CMAQ covariate of a single pollutant, the CMAQ covariates with cross-species, the spectral covariates of a single pollutant and the spectral covariates with cross-species. The first two models with mean structures that are linear in CMAQ are called linear downscaler (LD), while the latter two are linear in spectral covariates are called spectral downscaler (SD). We also compare the model with and without residual dependence (i.e. $w_k(\boldsymbol{s}) = 0$); hence, eight models are considered. Models 1-4 assume independent errors and are fit by least squares using the *lm* function in *R*, while models 5-8 take into account spatial and cross dependence and are fit using MCMC as described in Section 3.5. The eight models are:

$$\text{LD:} \qquad Y_k(\boldsymbol{s}) = \beta_{k_0} + \beta_k X_k(\boldsymbol{s}) + \epsilon_k(\boldsymbol{s})$$

$$\text{LD + Cross:} \qquad Y_k(\boldsymbol{s}) = \beta_{k_0} + \sum_{j=1}^{J} \beta_{kj} X_j(\boldsymbol{s}) + \epsilon_k(\boldsymbol{s})$$

$$\text{SD:} \qquad Y_k(\boldsymbol{s}) = \beta_{k_0} + \sum_{b=1}^{B} \beta_{kb} \tilde{X}_{kb}(\boldsymbol{s}) + \epsilon_k(\boldsymbol{s})$$

$$\text{SD + Cross:} \qquad Y_k(\boldsymbol{s}) = \beta_{k_0} + \sum_{j=1}^{J} \sum_{b=1}^{B} \beta_{kjb} \tilde{X}_{jb}(\boldsymbol{s}) + \epsilon_k(\boldsymbol{s})$$

$$\text{Spatial LD:} \qquad Y_k(\boldsymbol{s}) = \beta_{k_0} + \beta_k X_k(\boldsymbol{s}) + w_k(\boldsymbol{s}) + \epsilon_k(\boldsymbol{s})$$

$$\text{Spatial LD + Cross:} \qquad Y_k(\boldsymbol{s}) = \beta_{k_0} + \sum_{j=1}^{J} \beta_{kj} X_j(\boldsymbol{s}) + w_k(\boldsymbol{s}) + \epsilon_k(\boldsymbol{s})$$

$$\text{Spatial SD:} \qquad Y_k(\boldsymbol{s}) = \beta_{k_0} + \sum_{b=1}^{B} \beta_{kb} \tilde{X}_{kb}(\boldsymbol{s}) + w_k(\boldsymbol{s}) + \epsilon_k(\boldsymbol{s})$$

$$\text{Spatial SD + Cross:} \qquad Y_k(\boldsymbol{s}) = \beta_{k_0} + \sum_{j=1}^{J} \sum_{b=1}^{B} \beta_{kjb} \tilde{X}_{jb}(\boldsymbol{s}) + w_k(\boldsymbol{s}) + \epsilon_k(\boldsymbol{s})$$

We fit all eight models separately for each season (Jan-Mar, Apr-Jun, July-Sep and Oct-Dec), allowing the regression coefficients and dependence structure to vary. For each season, the first 78 days of the season is the training period and the last 12 days is the testing period. Two types of cross-validation are performed for comparing model performance: interpolation/spatial prediction and forecast/temporal prediction. The former is computed over the training period and evaluates how well a model predict at locations without an observation; this is of interest, for example, in epidemiology research to study the effects of long-term exposures to PM$_{2.5}$ on health outcomes (Adams et al., 2015, e.g.). The latter is computed over the testing period and evaluates model performance in making short-term forecasts for quantifying air pollution levels at local and national scales.

For interpolation, we predict at the hold-out sites in each fold using the fitted model

trained on the other 4 folds. Both hold-out and training data are observed in the training period. We compare the predicted values with the hold-out observations by averaging the root mean squared errors (RMSEs) and correlations across 5 folds. The performance of the spatial models are better than the independent models, due to the borrowing strength from neighboring observations; therefore, Table 1 shows only the interpolation result of the spatial models. The models are comparable across all species; this indicates that when spatial and cross dependence among species are accounted for, increasing the complexity of the mean structure may not improve interpolation. Among all pollutants, the interpolation for EC has the lowest correlation, this is likely because EC is a marker of traffic emissions which typically have high spatial heterogeneity (U.S. EPA, 1999).

For temporal prediction, the model is trained using data from each 4 folds in the training period, and the fitted model predicts at all sites for the testing period. In this case, the independent models perform better than the spatial models, and their results are presented in Table 1. Including cross-species predictors improves prediction, for instance, LD + Cross outperforms LD and SD + Cross outperforms SD for all species except SO$_4$. The SD+Cross model also gives the lowest RMSEs for all species except SO$_4$. This suggests that when spatial and cross-species dependence are not taken in to account, increasing mean complexity improves prediction.

[Table 1 about here.]

Figure 3 shows the PM$_{2.5}$ predictions for North Carolina (NC) on September 21, 2011 from the independent models, as well as CMAQ PM$_{2.5}$ and station data for comparison. The maps show a similar overall pattern, however the LD, LC + Cross and SD models overestimate PM$_{2.5}$ for central NC while SD + Cross seems to correct the overestimation. The predicted map from SD + Cross model is smoother comparing to LD and LD + Cross

models, suggesting that only the large spatial scale features from CMAQ output are used in prediction.

[Figure 3 about here.]

To gain better insight on the spectral downscaler performance, Figure 4 shows the multipollutant concentrations by US region and season for 2011. Daily multipollutant concentration maps for the entire US are interpolated from spatial SD + Cross for the training period and predicted from the SD + Cross for the testing period. Then, both map prediction and the CMAQ outputs are averaged within the region and season, while station data are averaged across available monitoring sites and days within the region and season for comparison. CMAQ predicted lower concentrations for the west for the entire year, and for all region for the months of Apr-Jun (AMJ) and Jul-Sep (JAS) for all pollutants, while the downscaler also underpredicts but less than CMAQ. The overall lower concentration when compared to station data is most likely due to the sampling bias of the monitoring system, as there are more sites located in urban areas with high air pollution due to car emission, coal combustion and industry activities, etc. A comparison of the concentration differences between CMAQ and station data across region and season reveals that CMAQ tends to underpredict in the west region, while the downscaler alleviates the underprediction issue.

[Figure 4 about here.]

Further evaluation of the CMAQ model at different spatial scales can be assessed. Figure 5 shows the coherence, $A_{kj}(\boldsymbol{\omega})$, between station data and CMAQ estimated from the Spatial SD + Cross model fitted to Jul-Sep, 2011. The estimated association is a function of the frequency $\boldsymbol{\omega}$ and is converted to spatial scale by using $12\text{km} \times 2\pi/||\boldsymbol{\omega}||$, where 12km is the CMAQ output grid size. All observed pollutants are significantly related to their corresponding CMAQ outputs at spatial scale greater than 120km, except NH$_4$. However, for most pollutants local variation in CMAQ (period less then 120km) is not associated

with local variation in the station data. That is, the model is filtering the CMAQ output to remove high-frequency terms. Some observed pollutants are significantly related to cross-species CMAQ outputs, for example, EC is found significantly associated with NO$_3$ and NH$_4$. It may be that these species share common sources that affect air quality at this spatial range, such as emissions from the combustion of various fuel (U.S. EPA, 1999).

[Figure 5 about here.]

5. Discussion

The proposed spectral downscaler exploits the complex relationship between the station and proxy data. As shown in our application, explicitly modeling the association between the two data sources at different spatial scales improves temporal prediction performance for total and speciated PM$_{2.5}$, while joint modeling the multiple pollutants with multivariate spatial dependence improves spatial prediction. The associations estimated at different spatial scales provide a valuable alternative tool for model evaluation, and this is used to evaluate daily CMAQ multipollutant output for the first time.

We have implicitly assumed stationarity in modeling the station observations. Nonstationarity can be introduced by allowing the mean parameters $\boldsymbol{\beta}$ to vary spatially. However this is not implemented in our application, since the mean parameters estimated from least squares are mostly not significantly different by region (see Supplementary Materials S.3). We have also assumed the multivariate spatial random effect $\boldsymbol{w}$ is isotropic and stationary, since the mean structure in our model allows for flexible relationships between the response and numerical model, which possibly captures the underlying process that manifest spatial heterogeneity. One way to introduce nonstationarity is to allow the cross dependence matrix to be spatially varying by letting $L(\boldsymbol{s})$ vary across location; however, considering the sparsity

of the station data, we have chosen a parsimonious model that seems adequate for our application.

In our model fitting, we have used parallel MCMC for divide and conquer large data set. Since standard R packages for fitting multivariate spatial process (without replicates), such as *spBayes* (Finley et al, 2015), and for combining subposterior MCMC samples, such as *parallelMCMCcombine* (Miroshnikov and Conlon, 2014), are available in *R*. It is convenient for practitioners to apply this model to their applications. The only additional programming is to construct the spectral covariates which only need to be computed once. The data likelihood can be parallelized, one could develop their own code to fit the model to the entire data set without partitioning. R Code to implement the spectral downscaler and Supplementary Materials are available on Github `https://github.com/yawenguan/multires`

References

Adams, K., Greenbaum, D. S., Shaikh, R., van Erp, A. M., and Russell, A. G. (2015). Particulate matter components, sources, and health: Systematic approaches to testing effects. *Journal of the Air & Waste Management Association* **65,** 544–558. PMID: 25947313.

Appel, K. W., Napelenok, S. L., Foley, K. M., Pye, H. O. T., Hogrefe, C., Luecken, D. J., Bash, J. O., Roselle, S. J., Pleim, J. E., Foroutan, H., Hutzell, W. T., Pouliot, G. A., Sarwar, G., Fahey, K. M., Gantt, B., Gilliam, R. C., Heath, N. K., Kang, D., Mathur, R., Schwede, D. B., Spero, T. L., Wong, D. C., and Young, J. O. (2017). Description and evaluation of the Community Multiscale Air Quality (CMAQ) modeling system version 5.1. *Geoscientific Model Development* **10,** 1703–1732.

Atkinson, R. W., Kang, S., Anderson, H. R., Mills, I. C., and Walton, H. A. (2014). Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax* **69,** 660–665.

Bechler, A., Vrac, M., and Bel, L. (2015). A spatial hybrid approach for downscaling of extreme precipitation fields. *Journal of Geophysical Research: Atmospheres* **120,** 4534–4550.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010a). A bivariate space-time downscaler under space and time misalignment. *The Annals of Applied Statistics* **4,** 1942.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010b). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics* **15,** 176–197.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2012). Space-time data fusion under error in computer model output: an application to modeling air quality. *Biometrics* **68,** 837–848.

Berrocal, V. J., Raftery, A. E., and Gneiting, T. (2007). Combining spatial statistical and ensemble information for probabilistic weather forecasting. *Monthly Weather Review* **135,** 1386–1402.

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

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R. (2004). The Schaake shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology* **5,** 243–262.

Cressie, N. A. C. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data.* Wiley series in probability and statistics. Hoboken, N.J. Wiley.

Crooks, J. L. and Özkaynak, H. (2014). Simultaneous statistical bias correction of multiple PM2.5 species from a regional photochemical grid model. *Atmospheric Environment* **95,** 126–141.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review* **143,** 955–971.

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

Fuentes, M. and Reich, B. (2010). Spectral analysis for spatial data. In *Handbook of Spatial Statistics*, chapter 5, pages 57–78. CRC Press.

Gel, Y., Raftery, A. E., and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (gop) method (with discussion). *Journal of the American Statistical Association* **99,** 575–587.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association* **97,** 632–648.

Grahame, T. J. (2009). Does improved exposure information for PM2.5 constituents explain differing results among epidemiological studies? *Inhalation Toxicology* **21,** 381–393.

Huang, G., Lee, D., and Scott, E. M. (2018). Multivariate space-time modelling of multiple air pollutants and their health effects accounting for exposure uncertainty. *Statistics in Medicine* **37,** 1134–1148.

Huang, Y.-N., Reich, B. J., and Fuentes, M. (2019+). Complete spatial model calibration.

Annals of Applied Statistics In press.

Li, X., Huang, S., Jiao, A., Yang, X., Yun, J., Wang, Y., Xue, X., Chu, Y., Liu, F., Liu, Y., Ren, M., Chen, X., Li, N., Lu, Y., Mao, Z., Tian, L., and Xiang, H. (2017). Association between ambient fine particulate matter and preterm birth or term low birth weight: An updated systematic review and meta-analysis. *Environmental Pollution* **227,** 596 – 605.

Mannshardt-Shamseldin, E. C., Smith, R. L., Sain, S. R., Mearns, L. O., Cooley, D., et al. (2010). Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. *The Annals of Applied Statistics* **4,** 484–502.

McMillan, N., Holland, D. M., Morara, M., and Feng, J. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics* **21,** 48–65.

Miroshnikov, A. and Conlon, E. (2014). parallelMCMCcombine: Methods for combining independent subset markov chain Monte Carlo (MCMC) posterior samples to estimate a posterior density given the full data setR package version 1.0.

Neiswanger, W., Wang, C., and Xing, E. P. (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 623–632, Arlington, Virginia, United States. AUAI Press.

Reich, B., Cooley, D., Foley, K., Napelenok, S., and Shaby, B. (2013). Extreme value analysis for evaluating ozone control strategies. *The Annals of Applied Statistics* **7,** 739.

Reich, B. J., Chang, H. H., and Foley, K. M. (2014). A spectral method for spatial downscaling. *Biometrics* **70,** 932–942.

Rundel, C. W., Schliep, E. M., Gelfand, A. E., and Holland, D. M. (2015). A data fusion approach for spatial analysis of speciated PM2.5 across time. *Environmetrics* **26,** 515–525.

Schefzik, R. (2017). Ensemble calibration with preserved correlations: unifying and comparing ensemble copula coupling and member-by-member postprocessing. *Quarterly Journal of the Royal Meteorological Society* **143,** 999–1008.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management* **11,** 78–88.

U.S. EPA (1999). Particulate Matter (PM2.5) Speciation Guidance, Final Draft, October 7, 1999.

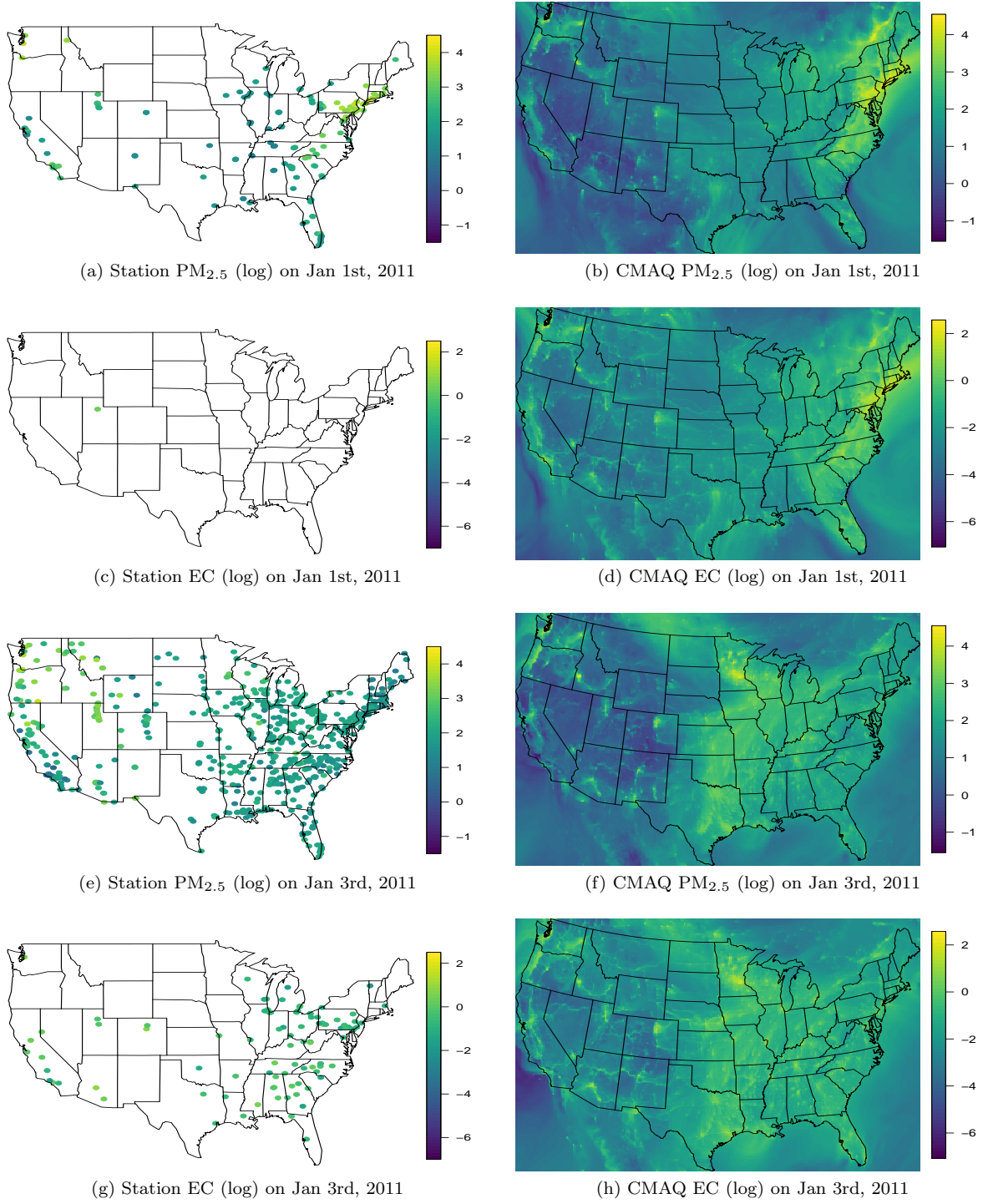Wang, X. and Dunson, D. B. (2013). Parallelizing MCMC via Weierstrass sampler. *CoRR* **abs/1312.4605,**.

(a) Station PM$_{2.5}$ (log) on Jan 1st, 2011

(b) CMAQ PM$_{2.5}$ (log) on Jan 1st, 2011

(c) Station EC (log) on Jan 1st, 2011

(d) CMAQ EC (log) on Jan 1st, 2011

(e) Station PM$_{2.5}$ (log) on Jan 3rd, 2011

(f) CMAQ PM$_{2.5}$ (log) on Jan 3rd, 2011

(g) Station EC (log) on Jan 3rd, 2011

(h) CMAQ EC (log) on Jan 3rd, 2011

Figure 1: Monitoring data of PM$_{2.5}$ and EC (left), and CMAQ output of PM$_{2.5}$ and EC concentrations (right) in unit log($\mu g/m^3$), for Jan 1 and 3, 2011.

(a) $\tilde{X}_{[0,\pi/5)}$

(b) $\tilde{X}_{[\pi/5,2\pi/5)}$

(c) $\tilde{X}_{[4\pi/5,\pi)}$

(d) $\tilde{X}_{[7\pi/5,8\pi/5)}$

(e) CMAQ PM$_{2.5}$ (log)
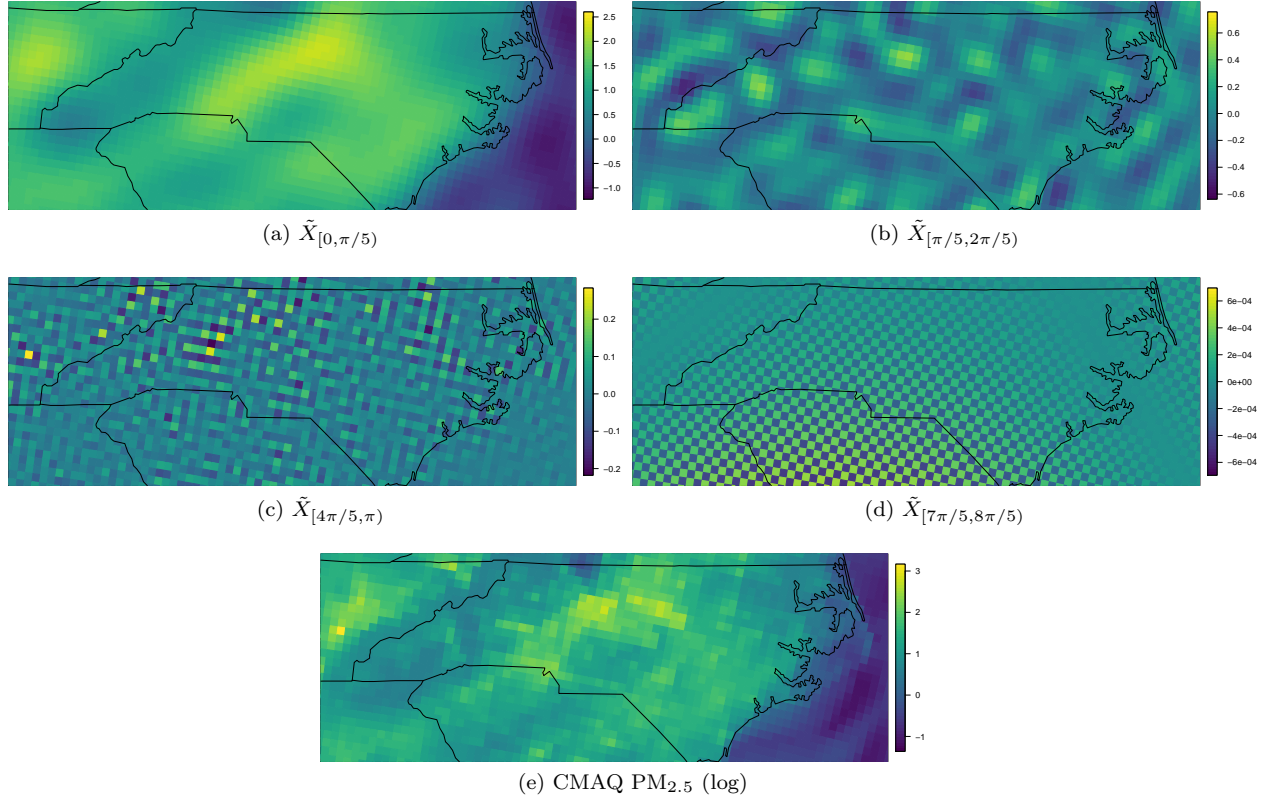
Figure 2: Filtered CMAQ PM$_{2.5}$ on Sept 21, 2011 by dividing the available frequency with magnitude of $[0,\sqrt{2}\pi]$ into 8 equal-width bins with width $\pi/5$; four of them are shown here. The signals at low frequencies $[0,\pi/5)$ (b) resemble the large spatial-trend ($>$120km), and at $[\pi/5,2\pi/5)$ (c) resemble spatial scale between 60 to 120km, while at higher frequencies $[4\pi/5,\pi)$ (d) resemble small scale (24$\sim$30km) information. Finally, the signals corresponding to the high frequencies (d) resembles large-scale features, as a consequence of aliasing. For comparison, CMAQ PM$_{2.5}$ is shown in (e).
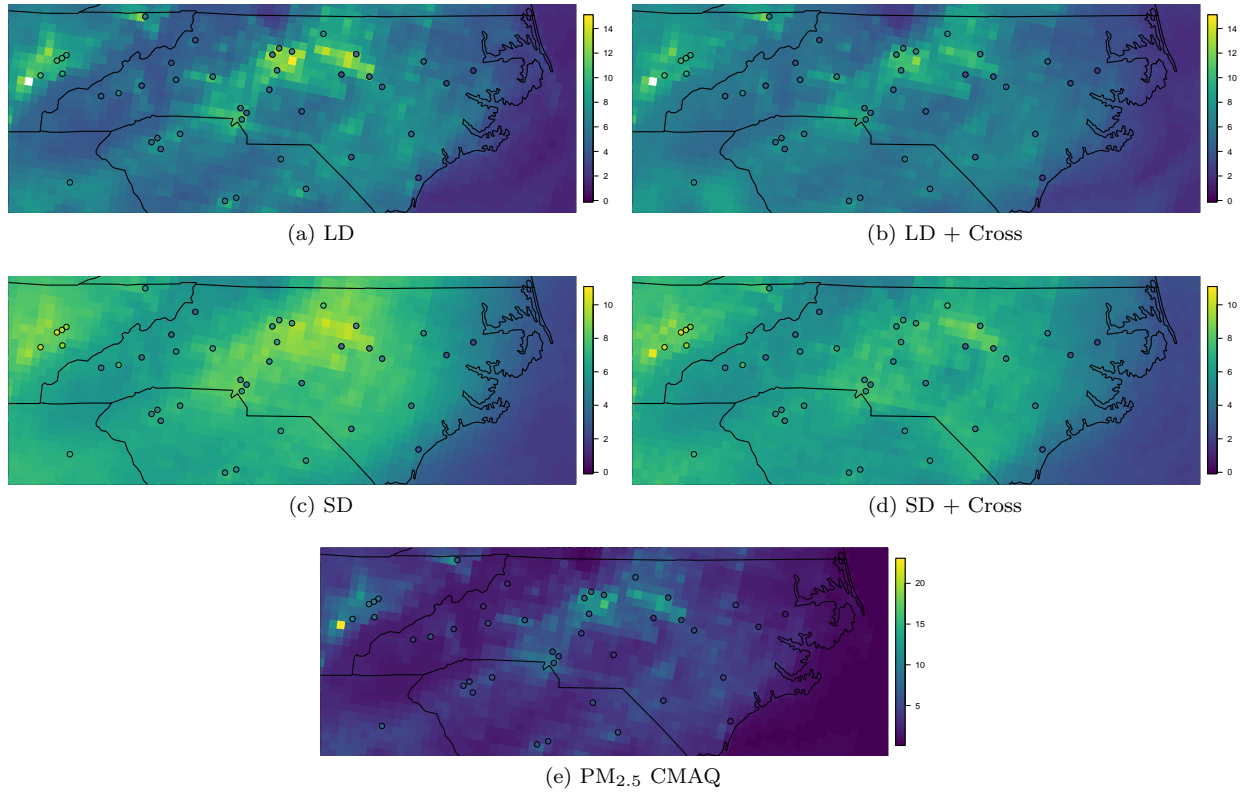
(a) LD

(b) LD + Cross

(c) SD

(d) SD + Cross

(e) PM$_{2.5}$ CMAQ

Figure 3: **Predicted PM**$_{2.5}$ **($\mu g/m^3$) for North Carolina** on Sept 21, 2011 from the linear downscaler ("LD") with and without CMAQ cross-species covariates, and the spectral downscaler ("SD") with and without cross-species spectral covariates. CMAQ PM$_{2.5}$ output and the station observation (colored circles) are plotted for comparison. The color scales for figures (a-b), (c-d) and (e) are different in order to show the pollution pattern.
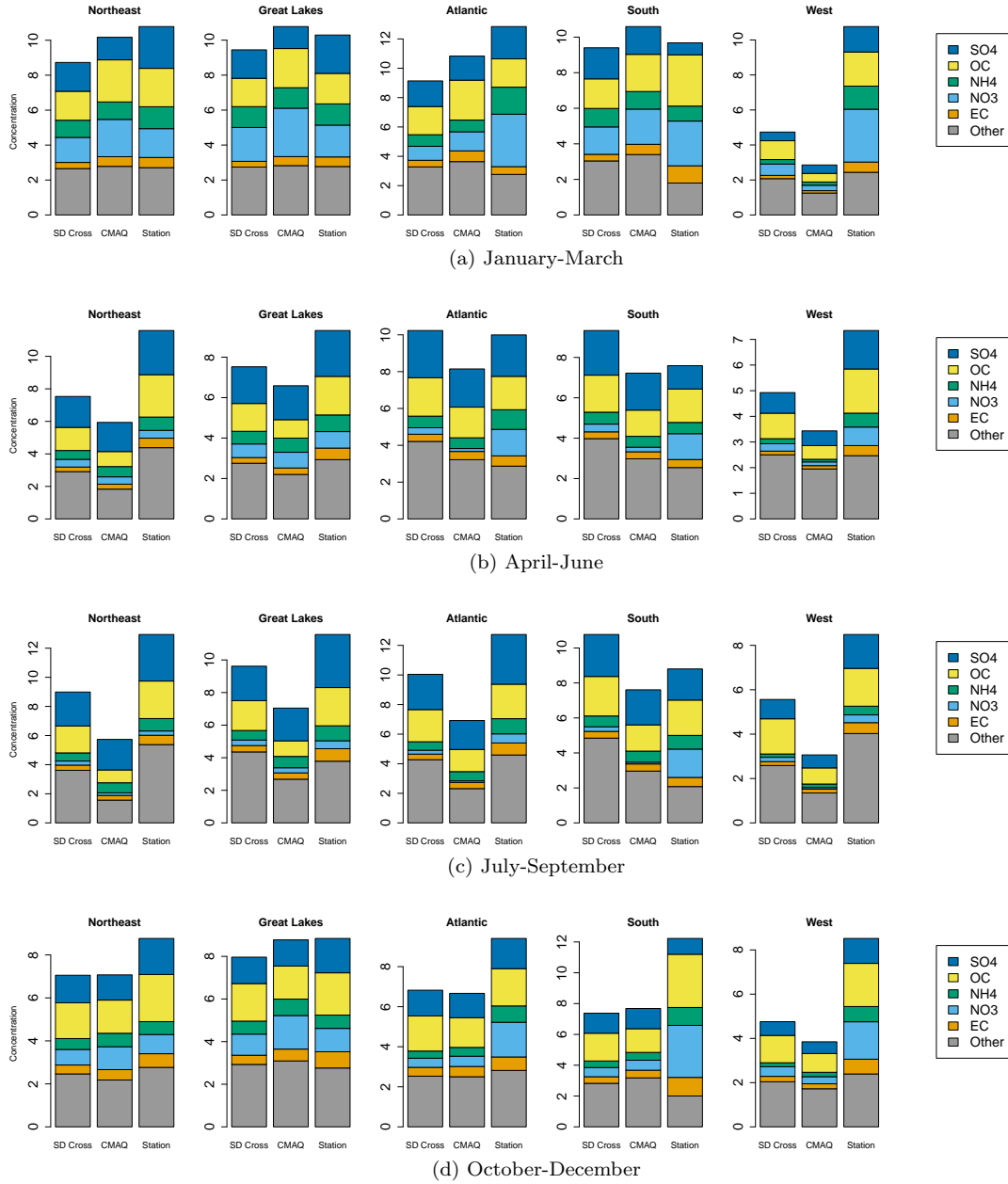
Figure 4: **Multipollutant concentrations ($\mu g/m^3$) by US region and season.** Results from spectral downscaler with cross-species predictors ("SD Cross") and CMAQ outputs are averaged within the region and season. Station data are averaged across available monitoring sites and days within the region and season. The "Other" pollutant is computed by taking the difference of total PM$_{2.5}$ and the sum of 5 constituent species. The five regions are grouped by states with similar PM$_{2.5}$ sources derived from principle component analysis according to the EPA (Appel et al., 2017). Northeast (Maine, New Hampshire, Vermont, Massachusetts,New York, New Jersey, Maryland, Delaware, Connecticut, Rhode Island, Pennsylvania, District of Columbia, Virginia and West Virginia), Great Lakes (Ohio, Michigan, Indiana, Illinois and Wisconsin), Atlantic (North Carolina, South Carolina, Georgia and Florida), south (Kentucky, Tennessee, Mississippi, Alabama, Louisiana, Missouri, Oklahoma and Arkansas) and west (California, Oregon, Washington, Arizona, Nevada and New Mexico).
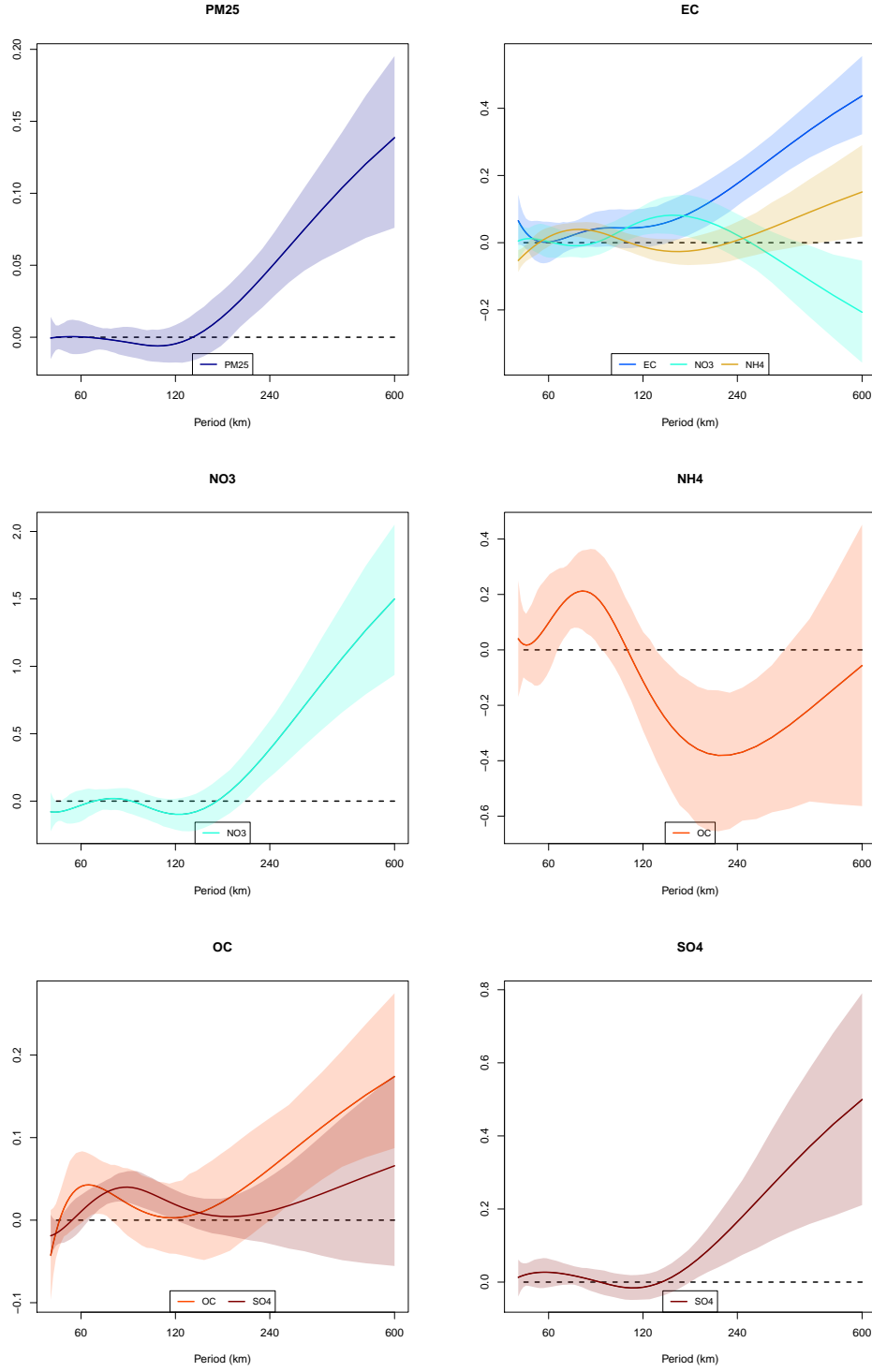
Figure 5: **Posterior mean (solid) and 95% interval (shade) of estimated association** plotted by period. The association function is declared significant if the coefficients are significant at a 5% significant level with Bonferroni correction.

Table 1: **Root mean squared error ($\mu g/m^3$) and correlation (parenthesis)** between observations and predicted values from multipollutant interpolation and prediction averaged over all seasons of 2011. Interpolation results of the spatial models are presented; they are the spatial linear downscaler ("SpLD") with and without CMAQ cross-species covariates, and the spatial spectral downscaler ("SpSD") with and without cross-species spectral covariates. Temporal prediction results of the independent models are presented; they are the linear downscaler ("LD") with and without CMAQ cross-species covariates, and the spectral downscaler ("SD") with and without cross-species spectral covariates. The results are based on 5-fold cross-validation.

Interpolation	PM25	EC	NO3	NH4	OC	SO4
SpLD	2.56(0.90)	0.44(0.58)	0.63(0.91)	0.41(0.87)	0.79(0.80)	0.75(0.88)
SpLD + Cross	2.56(0.90)	0.45(0.57)	0.66(0.91)	0.40(0.88)	0.79(0.80)	0.73(0.88)
SpSD	2.57(0.90)	0.44(0.59)	0.63(0.91)	0.41(0.87)	0.79(0.80)	0.72(0.89)
SpSD + Cross	2.56(0.90)	0.45(0.56)	0.74(0.88)	0.58(0.79)	0.83(0.77)	0.83(0.85)
Prediction	PM25	EC	NO3	NH4	OC	SO4
LD	5.11(0.40)	0.41(0.53)	2.18(0.57)	0.84(0.58)	1.26(0.46)	0.95(0.69)
LD + Cross	4.89(0.46)	0.41(0.56)	1.87(0.73)	0.80(0.60)	1.25(0.47)	0.98(0.66)
SD	5.06(0.41)	0.42(0.53)	2.14(0.60)	0.87(0.50)	1.28(0.45)	0.94(0.70)
SD + Cross	4.80(0.49)	0.40(0.58)	1.81(0.76)	0.79(0.63)	1.19(0.55)	0.94(0.69)