Supplementary Materials for Multivariate Spectral Downscaling for PM$_{2.5}$ Species

S.1 Monitoring Station Data

In our analysis, monitoring data come from 845 stations located throughout the contiguous United States. Figure 1 shows the locations of monitoring stations. Histograms of the raw station data are right skewed, therefore, log transformations are taken for each pollutant such that their marginal distributions are closer to normal (Figure 2).

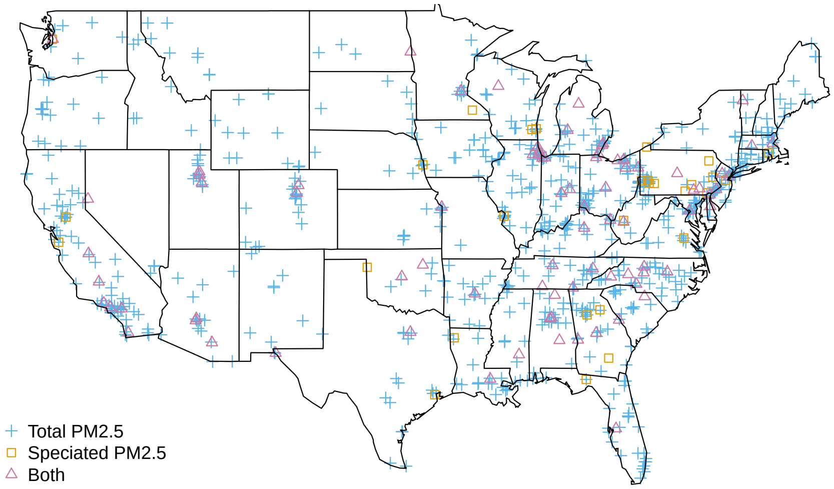


Figure 1: Monitoring station locations for total PM$_{2.5}$ only ("+"), speciated PM$_{2.5}$ only ("$\square$") and both ("$\triangle$").

S.2 Aliasing

For data observed at a uniformly grid with $\triangle$ units apart, the spectrum of the observations is concentrated in the principal interval $[-\pi/\triangle \leq \boldsymbol{\omega} < \pi/\triangle]^2$. Therefore, the signals corresponding to high frequencies outside the principal interval can not be distinguished, this is known as the aliasing phenomenon. In the construction of spectral covariates, we have divided the range of frequency based on magnitude $||\boldsymbol{\omega}||$ in Equation (2) and applied basis function smoothing also based on the magnitude in Equation (9). Hence the complex conjugates $Z(\boldsymbol{\omega})$ and $Z(-\boldsymbol{\omega})$ with the same frequency magnitude are combined to construct the spectral covariates.

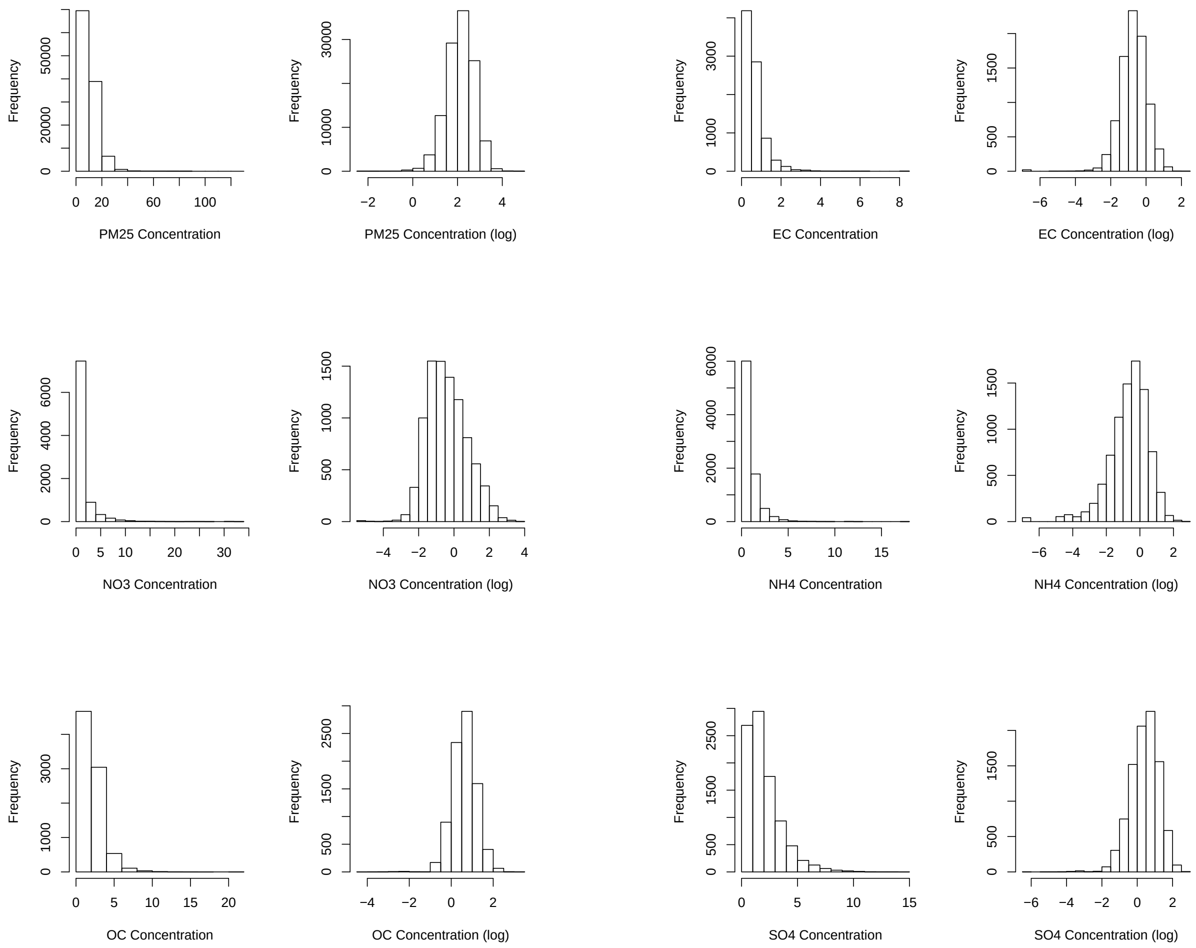
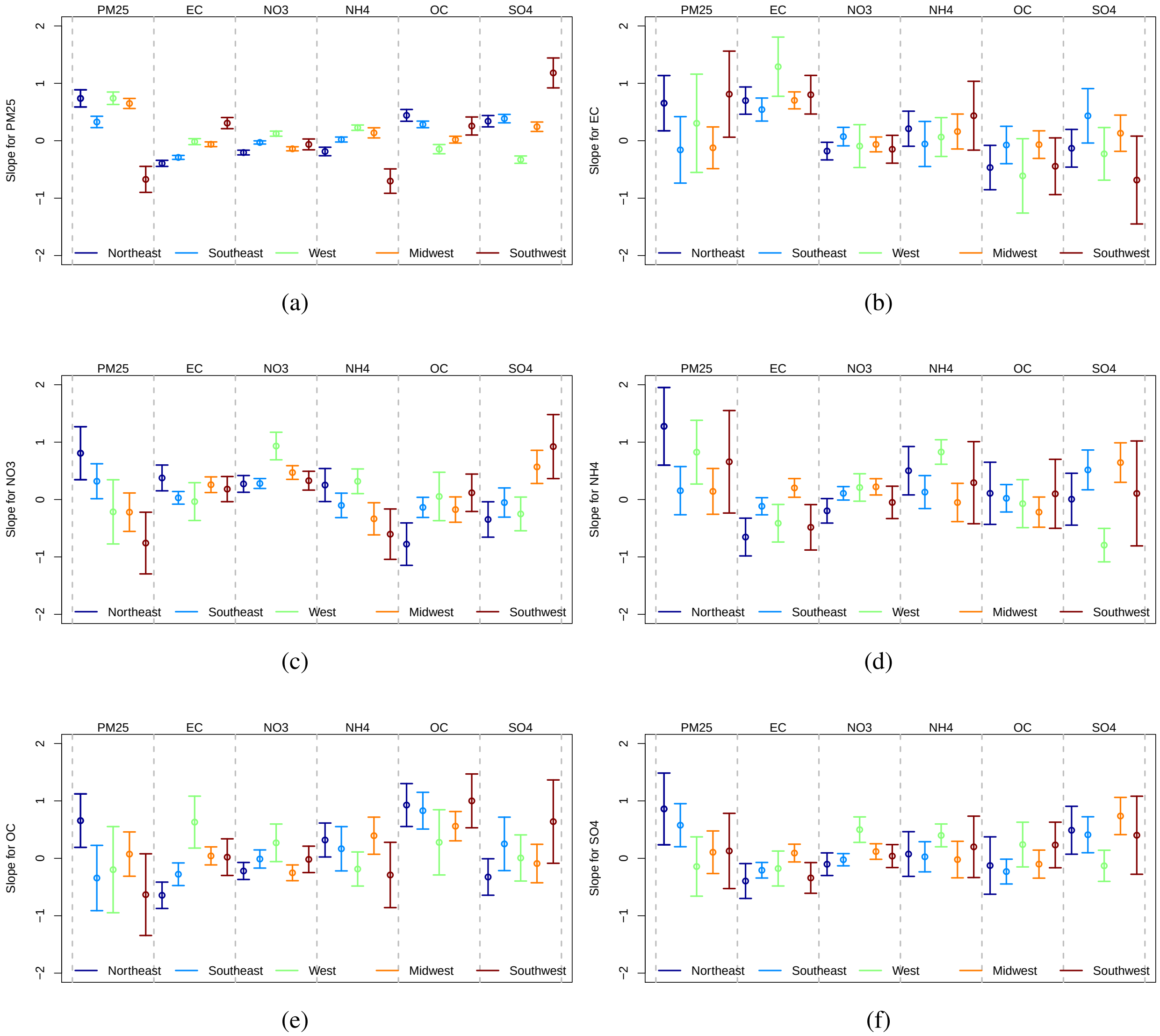


Figure 2: Histograms of station data ($\mu g/m^3$) before and after log transforms.

S.3 Exploratory Analysis

We begin data exploration by regressing each response variable on the spectral covariates separately for 5 US regions using the data from Jul-Sept, 2011. The spectral covariates in our exploratory analysis are computed using 10 equal-width bins of frequency in the interval $[0, 2\pi)$ as described in Section 4.1. The associations at different spatial scales are compared across regions, for example, Figure 3 shows the estimated regression coefficients corresponding to spatial scale $>$120km (or frequency $< \pi/5$). The differences of the associations estimated by region are largely insignificant; this justifies our model assumption of spatially constant regression coefficients. Although in a few cases regional differences are observed, for example, the regression coefficients in the Northeast and Southwest regions are different between station PM$_{2.5}$ and the spectral covariates of $PM_{2.5}$, EC, NH_4, and SO_4 pollutants (Figure 3a).



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3: **Estimated associations at >120km spatial scale by region**. Regression coefficients and their 95% confidence intervals are estimated by regressing each station pollutant on the spectral covariates of each CMAQ pollutant.

We then analyze the residuals from a regression model with all regions combined to investigate the temporal, spatial and cross-species dependence. Table 1 shows the coefficient estimates for the combined regions. For assessing time dependence, we compute the autocorrelation function (ACF) for each pollutant at each site. At lag-1, only 160 stations measuring $PM_{2.5}$ have enough daily observations to estimate the correlation and 150 of them are significant, while at lag-2 a similar number of autocorrelation are computed and only 56% of them are significant; however, for

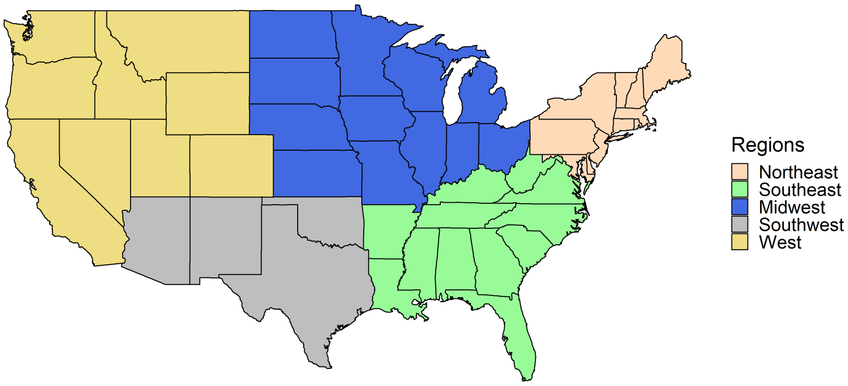


Figure 4: **Map of the regions** used for estimation in Figure 3

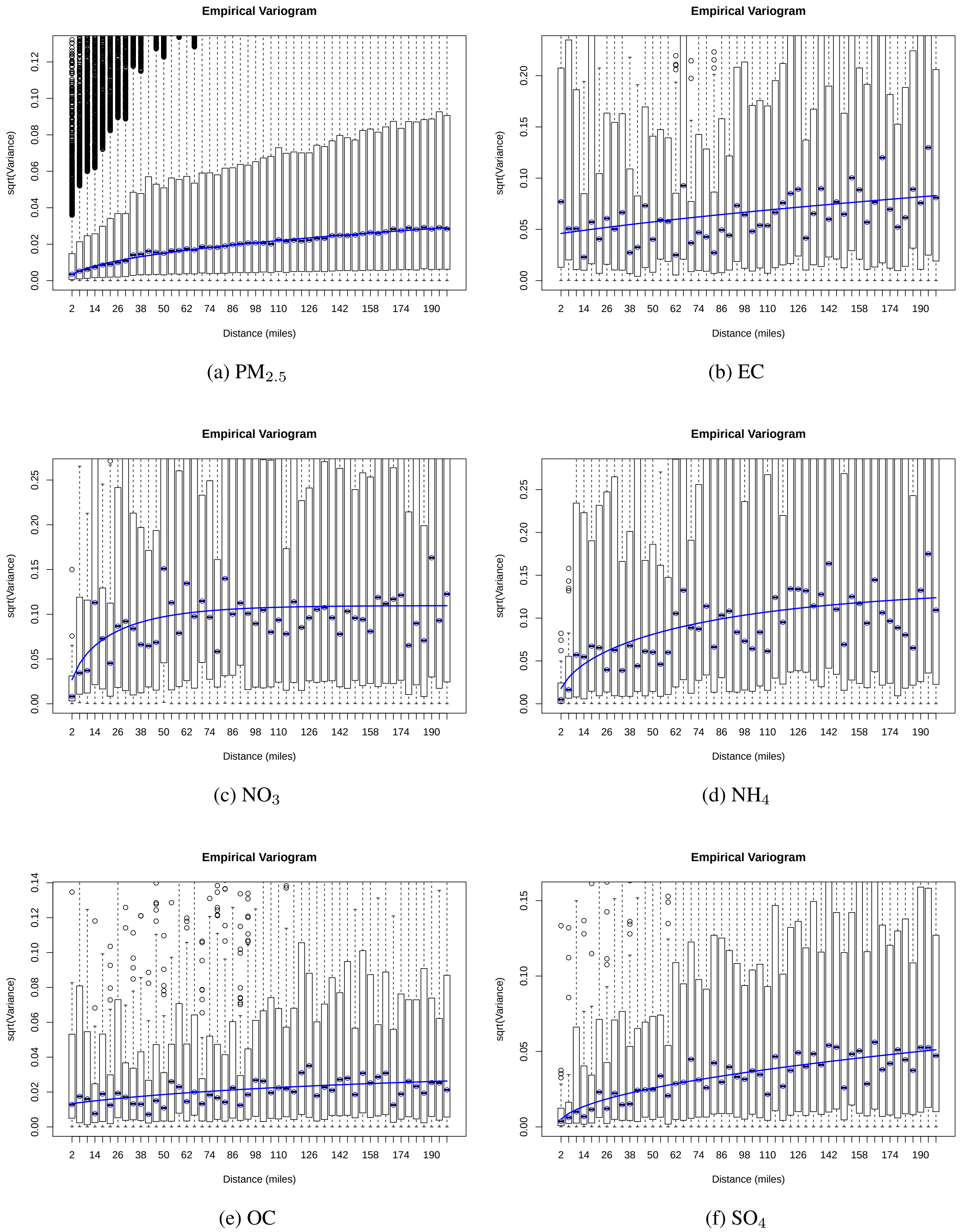
lag-1 and -2 only three stations measuring constituent species provided enough daily observations for estimating the autocorrelations. For lag-3 and beyond, each lag has only a small proportion ($<15\%$) of autocorrelations that are significant. All assessments are based on a significance level of 5%. Although station $\mathrm{PM}_{2.5}$ data indicate non-negligible temporal dependence, however all constituent species lack daily data to estimate such dependence. Therefore, we treat the residuals as independent replicates in time.

The variogram is a common exploratory tool for assessing spatial dependence. It plots the pairwise squared differences as a function of distance. For each pollutant and the days with more than 20 observations, we compute the daily empirical variogram from the residuals and fit an exponential model using weighted least squares to the combined daily empirical variograms. Figure 5 shows the daily empirical variograms combined over Jul-Sept, 2011 and the fitted variogram to the median pairwise squared differences of each distance bin.

Our proposed model employs spatial cross dependence among pollutants to improve spatial predictions. Here we use cross-correlogram to assess such dependence among pollutants. For each pair of pollutants, we compute the empirical cross-correlogram on the residuals for days with more than 50 observations for each pollutant; they are then combined by distance. Figure 6 plots the median correlation of each distance bin against distance. We see existing cross-correlation among pollutants, for example between total PM and the constituent species, that could be exploited for improving predictions.

Table 1: **Coefficient estimates** (multiplied by 100) from regressing station data on spectral covariates with cross-species. Both the station data and spectral covariates are standardized. The coefficients measure the association between station and CMAQ data at multiple spatial scales. The spatial scale (second row) corresponding to the spectral covariate $\tilde{X}_l$ is converted from frequency by using 12km$\times 2\pi/\delta_{l^*}$, where 12km is the CMAQ grid size and δ_{l^*} is the midpoint of the frequency interval $[\delta_l, \delta_{l+1})$. The significant coefficients at a 5% significance level are highlighted in blue.

		$\tilde{X}_1$	$\tilde{X}_2$	$\tilde{X}_3$	$\tilde{X}_4$	$\tilde{X}_5$	$\tilde{X}_6$	$\tilde{X}_7$	$\tilde{X}_8$
Scale(km)		240	80	48	34	27	22	18	16
$PM_{2.5}$	$PM_{2.5}$	56.1	1.2	6.3	-2.7	4.0	1.7	-4.0	2.3
	EC	-14.6	4.5	2.2	4.7	-1.5	2.1	-0.2	-1.9
	NO_3	-8.4	-2.9	-0.1	0.3	-0.8	-1.4	0.6	0.2
	NH_4	10.1	4.9	1.8	-1.5	1.7	-0.3	0.4	-0.4
	OC	12.9	-1.4	-7.6	0.0	0.0	-3.3	2.4	-1.8
	SO_4	16.5	-2.0	-0.5	1.8	-1.5	1.0	1.5	-0.8
EC	$PM_{2.5}$	10.4	-1.5	-4.9	2.7	-1.4	-12.8	-6.0	3.6
	EC	59.0	1.2	3.9	7.8	-3.9	11.9	-1.3	-11.9
	NO_3	-2.0	7.4	1.1	-0.1	0.5	-1.2	0.9	1.0
	NH_4	15.7	-3.9	-1.2	-3.2	5.1	1.5	-0.3	-2.5
	OC	-16.6	6.6	-6.9	-3.9	4.3	1.3	11.2	5.2
	SO_4	7.5	5.1	9.1	1.8	3.3	-2.1	-0.8	-4.5
NO_3	$PM_{2.5}$	44.2	27.5	3.2	16.3	13.6	-1.0	10.4	0.9
	EC	17.8	-9.4	-3.1	11.0	-8.7	0.4	7.3	-0.1
	NO_3	45.9	0.3	-0.5	-1.6	1.0	-1.3	-2.7	0.4
	NH_4	8.5	8.9	5.3	-0.8	3.3	0.5	0.7	-2.9
	OC	-43.3	-0.3	-2.8	-15.2	-1.0	2.1	-7.0	1.4
	SO_4	-22.4	-6.4	-1.2	0.0	-15.8	-5.1	-7.1	-3.0
NH_4	$PM_{2.5}$	84.1	40.1	-4.0	5.6	-1.1	-3.6	13.1	4.7
	EC	-21.0	-1.4	13.3	6.0	0.5	4.2	6.9	0.4
	NO_3	9.6	-3.2	-1.9	-1.2	1.5	-3.6	-1.8	0.3
	NH_4	28.4	5.4	4.6	-3.0	3.7	-3.1	-0.3	-2.5
	OC	-23.0	-18.5	-7.1	-1.6	3.0	2.0	-12.9	-3.9
	SO_4	6.7	-9.3	-0.2	2.0	-11.6	-3.3	-2.9	-4.9
OC	$PM_{2.5}$	-4.5	-3.6	-6.8	-2.8	-15.6	-6.7	3.7	8.0
	EC	-17.8	-9.2	3.4	4.9	0.3	10.8	-8.3	-10.4
	NO_3	-5.1	0.5	2.4	0.6	3.3	-3.3	2.3	-1.1
	NH_4	17.6	2.8	-1.4	-6.5	4.3	2.8	-1.0	2.2
	OC	76.0	15.7	-0.4	0.1	7.7	3.7	4.7	3.6
	SO_4	6.2	2.2	6.7	8.2	2.7	-5.4	1.4	-9.5
SO_4	$PM_{2.5}$	31.5	17.2	1.5	7.2	-3.7	-4.7	10.7	5.0
	EC	-16.0	-2.2	4.8	-1.2	5.7	0.5	4.4	3.4
	NO_3	7.1	-6.1	-1.7	-0.3	1.3	-1.7	-2.5	-0.9
	NH_4	10.2	7.5	2.1	-2.5	-0.1	-0.4	0.2	-2.6
	OC	-5.7	-2.9	-5.1	-0.4	-1.9	6.0	-12.1	-4.5
	SO_4	52.6	-1.9	3.0	1.1	-4.6	3.5	-1.6	-1.3



(a) $PM_{2.5}$

(b) EC

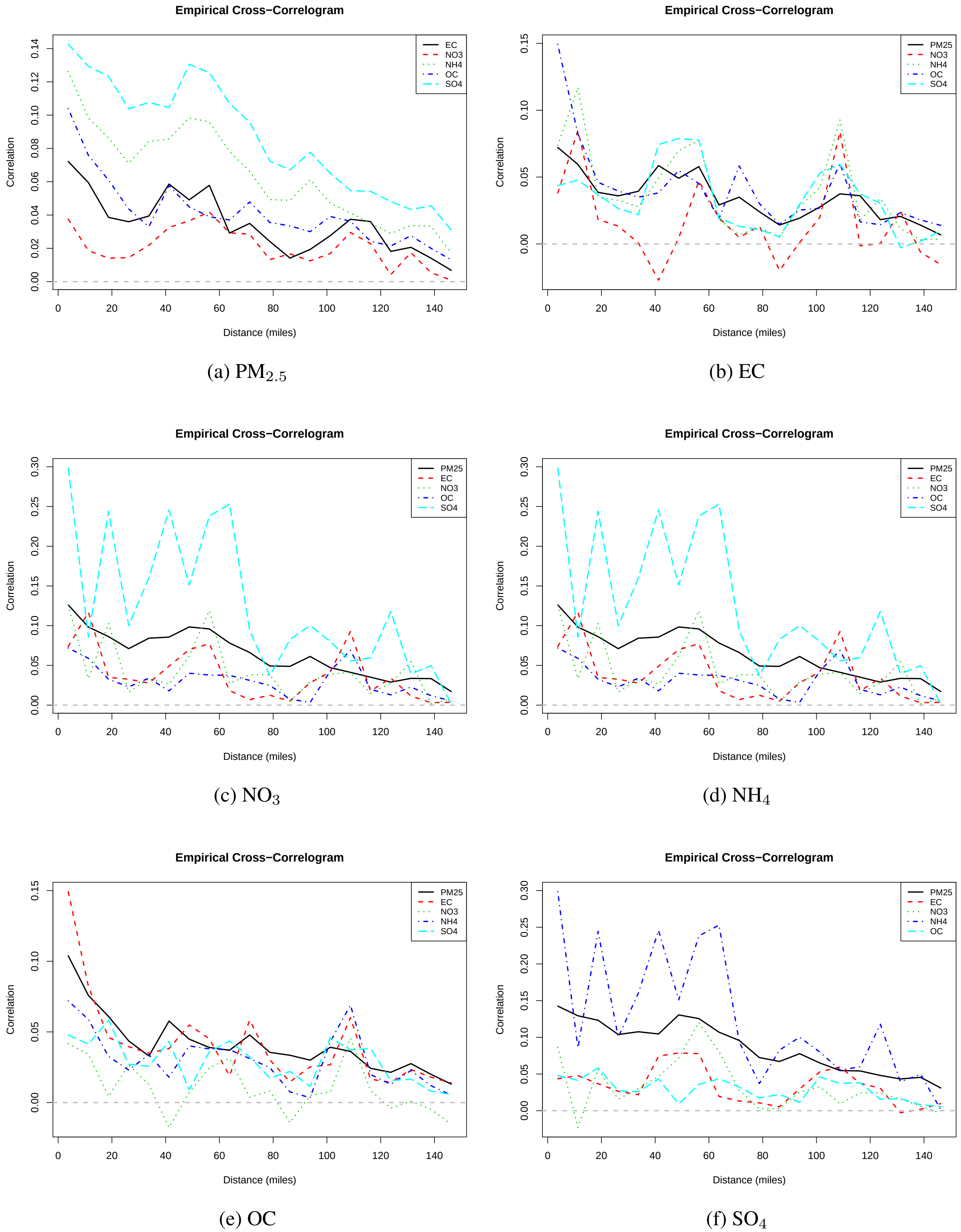
(c) NO_3

(d) NH_4

(e) OC

(f) SO_4

Figure 5: **Empirical variograms** The boxplot shows the daily empirical vaiograms combined by distance, and the blue line shows the fitted vaiogram to the median within each bin assuming an exponential model.



(a) $PM_{2.5}$

(b) EC

(c) NO_3

(d) NH_4

(e) OC

(f) SO_4

Figure 6: **Empirical cross-correlograms.** The median correlation plotted against distance for each pair of pollutants. Zero correlation is plotted in gray for reference.

S.4 Simulation Study

We conduct a simulation study to assess the inference performance of parallel MCMC for multivariate spatial data. We simulate multivariate spatial process with 3 response variables at misaligned locations for 10 time points with time as independent replicates. We fit the model using the MCMC algorithm with priors as specified in Section 3.5 for each time point, then combine the MCMC samples from the subposteriors to make inference for model parameters. The fitted model is then used to make prediction at hold-out locations to access prediction performance.

At each time point, data are drawn from $Y_k(\boldsymbol{s}) = \beta_{k0} + \beta_{k1} X_k(\boldsymbol{s}) + w_(\boldsymbol{s}) + \epsilon_k(\boldsymbol{s})$ for $k = 1, 2, 3$ in the unit square $[0, 1]^2$ spatial domain, where $w(\boldsymbol{s}) = [w_1(\boldsymbol{s}), w_2(\boldsymbol{s}), w_3(\boldsymbol{s})]$ follows the linear model of coregionalization (LMC). The multivariate response is simulated at 100 locations, and we randomly select 50 locations for each variable to create misaligned observations, hence the total number of observations at each time point is $N = 150$. We then repeat the above simulation 10 times to create independent replicates of $\boldsymbol{Y}(\boldsymbol{s}) = [Y_1(\boldsymbol{s}), Y_2(\boldsymbol{s}), Y_3(\boldsymbol{s})]$. The regression coefficient values are $\boldsymbol{\beta_0} = (1, -1, -1)$, $\boldsymbol{\beta_1} = (-1, 1, -1)$, and the variance-covariance matrix C for the cross dependence matrix is a 3×3 matrix with elements $C_{11} = 1$, $C_{12} = 1$, $C_{13} = -1$, $C_{22} = 3$, $C_{23} = -0.5$ and $C_{33} = 1.3125$. We use the exponential covariance function with a common spatial decay parameter $\phi = 6$ to simulate the independent spatial process $v_k(\boldsymbol{s})$. Measurement errors are $\tau_k = 0.1$ for $k = 1, ..., 3$.

Figure 7 shows the distributions of posterior mean from 100 simulated data sets with the coverage based 95% credible interval labeled obove each boxplot. The regression coefficients and the variance-covariance parameter estimates are unbiased with reasonable coverage, while the residual variance and spatial decay estimates have small bias with lower coverage due to non-normality. Most importantly, the coverage of the 95% prediction intervals (not shown in Figure 7) is between 93% and 96%.

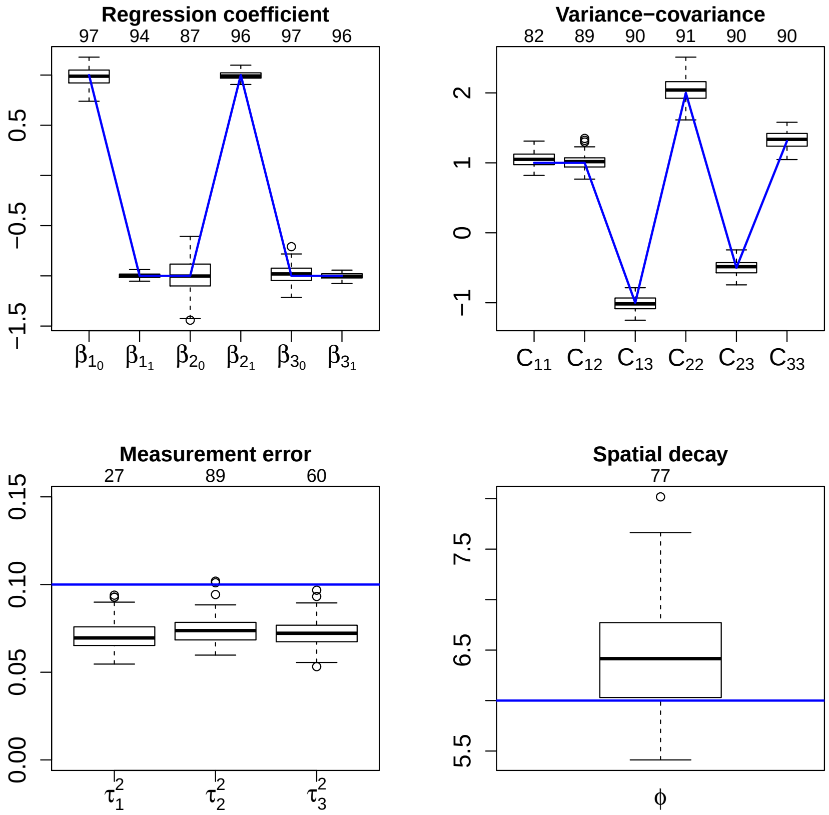


Figure 7: Distributions of posterior means from 100 simulated data sets. The coverage of 95% posterior intervals multiplied 100 is shown above each boxplot.