A data fusion approach for spatial analysis of speciated PM2.5 across time

Alan E. Gelfand

Duke University

(Joint work with Colin Rundel, Erin Schliep, David Holland)

Introduction

- $PM_{2.5}$ exposure is linked to adverse health effects such as lung cancer and cardiovascular disease.
- $PM_{2.5}$ is a complex mixture of different species whose composition varies in space and time.
- EPA, other federal, state, and local organizations altogether monitor, on a national scale, total $PM_{2.5}$ along with its primary species.
- Here, data from three separate monitoring station networks as well as output from a deterministic atmospheric computer model.
- We offer a novel multi-level speciated $PM_{2.5}$ model which fuses these data sources

Introduction cont.

- Epidemiologic and exposure studies on the health effects of total fine particulate matter ($PM_{2.5}$) have led the US EPA to establish a mass-based ambient air quality standard for this contaminant.
- The composition of $\mathrm{PM}_{2.5}$ is a complex mixture of different species and composition varies with season and location
- Useful to understand which of the different species are most strongly connected to various adverse health effects
- So, a need to develop *good* predictions of speciated $\mathrm{PM}_{2.5}$ over space and time.

Our Contribution

Multi-level speciated $PM_{2.5}$ model with following novel features:

- (1) it fuses data from three monitoring station networks: (i) the urban Chemical Speciation Network (CSN), (ii) the large-scale $PM_{2.5}$ Federal Reference Monitoring network (FRM), and (iii) the rural speciated $PM_{2.5}$ Interagency Monitoring of Protected Visual Environments (IMPROVE) network, with gridded 12 km output from the Community Multi-scale Air Quality (CMAQ) numerical atmospheric model
- (2) it models each of the five primary species of $PM_{2.5}$ through fusion of CSN, IMPROVE, and CMAQ and models total $PM_{2.5}$ through FRM, CSN, IMPROVE, and CMAQ
- (3) it introduces species level measurement error models as well as total $PM_{2.5}$ measurement error models, all varying around the respective *true* levels
- (4) the model for the true levels incorporates an 'other' species component to ensure that the true totals are physically consistent

The Data

- Four distinct sources: monitoring data from the CSN, IMPROVE, and FRM monitoring networks and gridded numerical model output from the CMAQ model.
- We model total $PM_{2.5}$ (in $\mu g/m^3$) and the five major $PM_{2.5}$ species: sulfate, nitrate, total carbonaceous matter, ammonium, and fine soil or crustal material
- For the IMPROVE network, ammonium concentration is *inferred* based on nitrate and sulfate concentrations
- EPA established CSN to provide national measurements of speciated $PM_{2.5}$ data at mostly urban locations.
- CSN quantifies daily total $PM_{2.5}$ and most of the species. 221 monitoring locations. Speciated data are collected on a daily, three day, or six day cycle depending on the location.
- The IMPROVE network consists of 172 sites at mostly rural, remote locations. It reports observations of daily total $PM_{2.5}$ and speciated components every 3 days
- CSN has an urban focus, IMPROVE focuses on monitoring remote environments

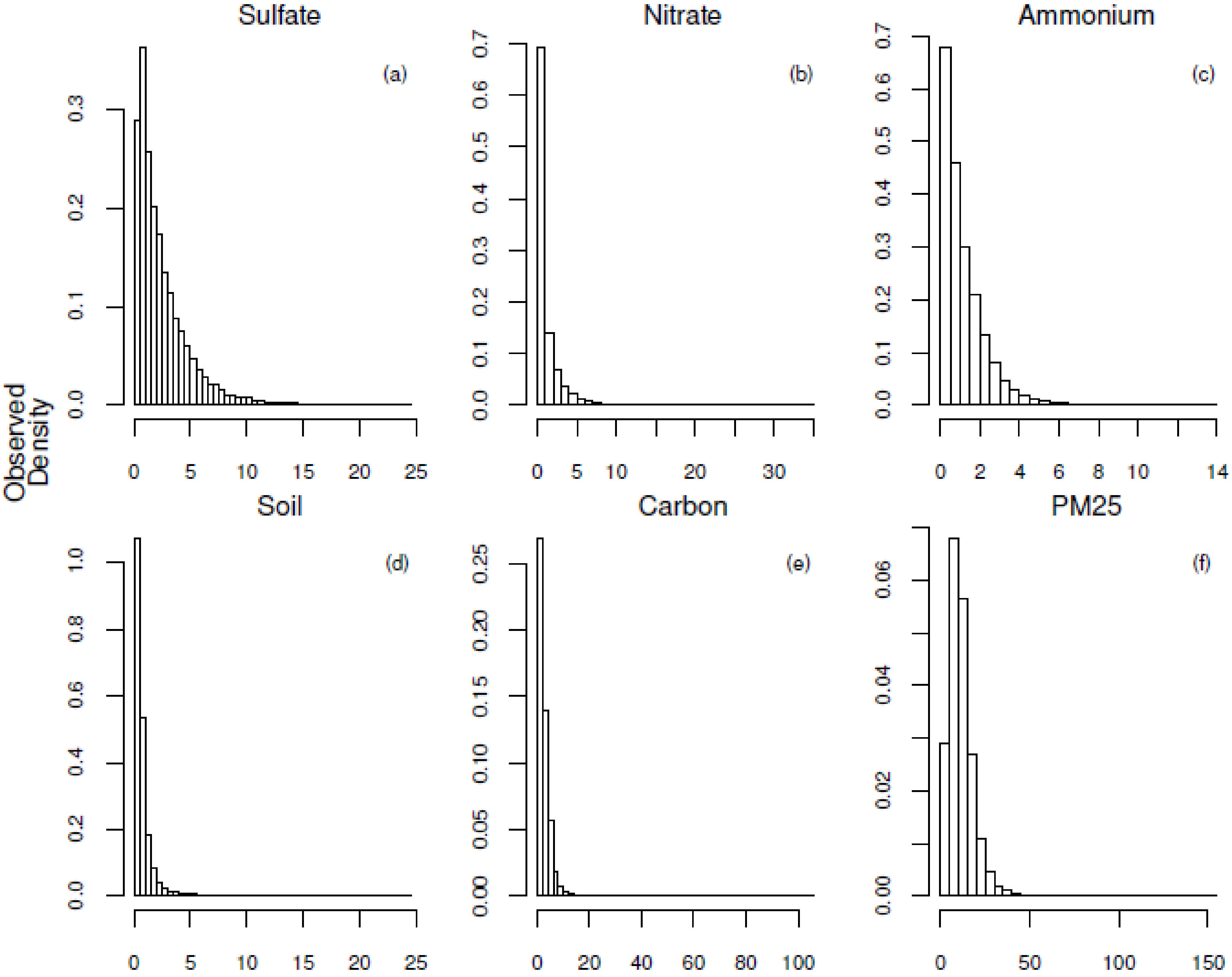

Figure 3. Histograms of the observed species and total $PM_{2.5}$ concentrations across the CSN and IMPROVE networks.

The data cont.

- We supplement the CSN and IMPROVE data with total $PM_{2.5}$ from the large FRM network consisting of 949 sites reporting on a daily, three day, or six day monitoring cycle.
- This network monitors a variety of urban, suburban, and rural environments with the major focus on urban areas.
- FRM does **not** monitor speciated $PM_{2.5}$ but the total particulate mass values help with regard to the constraint that the sum of the true levels for the 5 major species does not exceed the true total mass concentration.

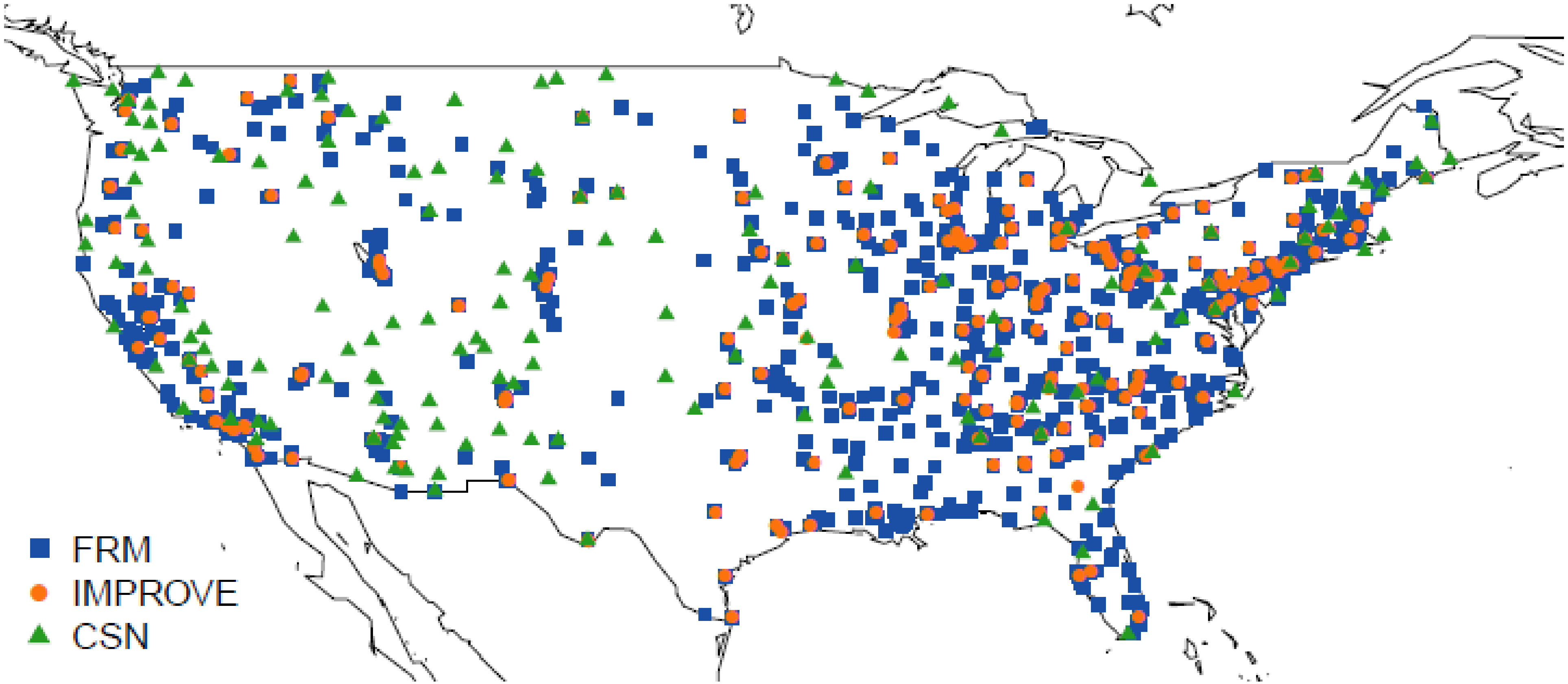

Figure 1: Location of Chemical Speciation Network (CSN), Interagency Monitoring of Protected Visual Environments (IMPROVE), and Federal Reference Method (FRM) monitoring stations in 2007.

The data cont.

- CMAQ model output employs Weather Research and Forecast model meteorology and gridded emissions of primary $PM_{2.5}$ and precursors to secondary $PM_{2.5}$.
- The National Emissions Inventory is the primary source for the emissions data.
- Aerosol transport, atmospheric chemistry, and secondary $PM_{2.5}$ formation are simulated to provide the CMAQ $PM_{2.5}$ species concentrations.
- Illustratively, we work with data from the year 2007 using 12 km grids, covering the conterminous US.
- For each species, can obtain association between CMAQ and monitoring data.
- The strongest pairwise correlations are: ammonium-sulfate 0.857, ammonium-nitrate 0.599, carbon-sulfate 0.469, and ammonium-carbon 0.462.

Some modeling challenges

- Although CSN, IMPROVE, and FRM mostly report on the same sampling schedule there is considerable temporal misalignment among monitoring sites.
- So, we model $PM_{2.5}$ total and species weekly, averaging all CMAQ output and monitoring observations within each week.
- An additional issue: total recorded particulate mass can be less than the sum of the 5 major species defined above.
- For 2007, we show the incidence of this issue over all weeks by each monitoring network.
- However, the amount of excess mass is generally low: 80% of occurrences having less than 1 $\mu g/m^3$ of excess mass, 98% having less than 5 $\mu g/m^3$.

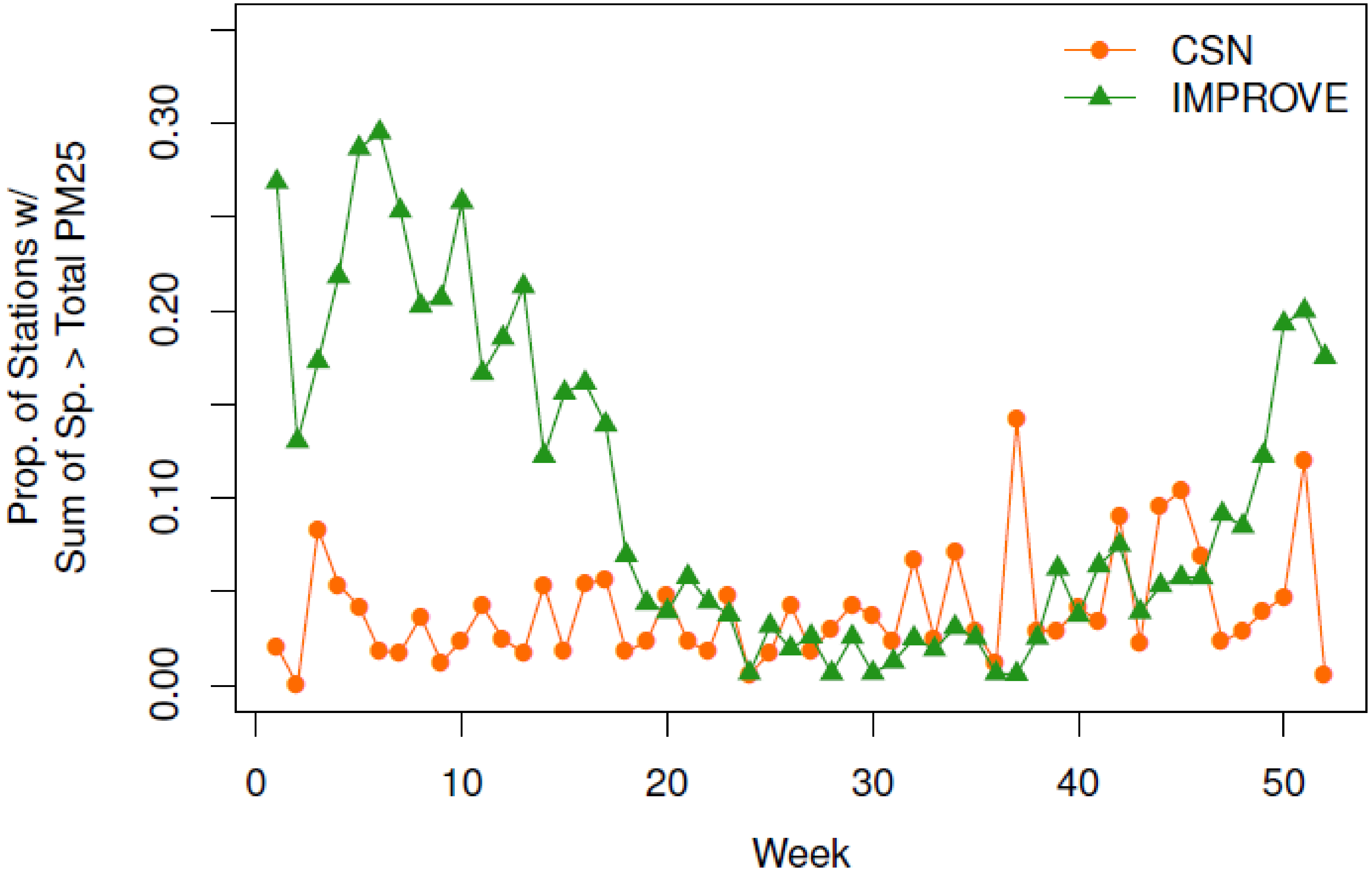

Figure 2. Comparison of the proportion of stations reporting a sum of $PM_{2.5}$ species that exceeds the observed total $PM_{2.5}$ for each week and network in 2007.

Statistical Models

The Data Stage:

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

cont.

- We assume that all observed values are noisy observations of an associated underlying and unobserved true spatial field.
- Let $Z_t^i(\boldsymbol{s})$ be the true level of $PM_{2.5}$ species i at time t for location $\boldsymbol{s}$.
- Additionally, at time t and location $\boldsymbol{s}$, let $Z_t^o(\boldsymbol{s})$ denote the true level of the total of all of the other unmeasured species and let $Z_t^{tot}(\boldsymbol{s})$ denote the true total.
- We have $\sum_{i=1}^{5} Z_t^i(\boldsymbol{s}) + Z_t^o(\boldsymbol{s}) = Z_t^{tot}(\boldsymbol{s})$.
- We model all Z's to be ≥ 0 with probability 1.

cont.

- Then, for $i = 1, 2, ..., 5$,

$$\begin{aligned} C_t^i(\boldsymbol{s}) &= Z_t^i(\boldsymbol{s}) + \epsilon_{C,t}^i(\boldsymbol{s}) \\ I_t^i(\boldsymbol{s}) &= Z_t^i(\boldsymbol{s}) + \epsilon_{I,t}^i(\boldsymbol{s}) \end{aligned}$$

  and

$$\begin{aligned} C_t^{tot}(\boldsymbol{s}) &= Z_t^{tot}(\boldsymbol{s}) + \epsilon_{C,t}^{tot}(\boldsymbol{s}) \\ I_t^{tot}(\boldsymbol{s}) &= Z_t^{tot}(\boldsymbol{s}) + \epsilon_{I,t}^{tot}(\boldsymbol{s}) \\ F_t^{tot}(\boldsymbol{s}) &= Z_t^{tot}(\boldsymbol{s}) + \epsilon_{F,t}^{tot}(\boldsymbol{s}). \end{aligned}$$

- The measurement error formulation allows the flexibility of weekly variance components. That is, all of the ϵ's are independent with distinct variance components.
- Altogether there are 13 variance components; 5×2 for the species plus 3 for the totals for each week.
- Note that there is no observed "o" so no additional measurement error modeling is needed.

cont.

Modeling the 'truth'

- Next, we model the Zs.
- Due to the sum constraint, we model the Z^is and Z^os yielding implicit realizations for the Z^{tot}s.
- We build regression models for the Z^is in the form of downscalers to incorporate CMAQ output in the model.
- So, we take advantage of the fact that there are 97,416 CMAQ grid cells for the continental U.S. while we have only 221, 172, and 949 monitoring sites for CRM, IMPROVE, and FRM, respectively.
- Downscaling is latent (i.e. the Z's are not observed). We are downscaling CMAQ to the *true* species concentrations, not directly to the data.

cont.

- An additional constraint - all the Zs to be nonnegative.
- Two common ways: (i) a Tobit transformation ($U = \mathtt{max}(0, \tilde{U})$) where $\tilde{U}$ is a normally distributed random variable) or (ii) through a log normal.
- Tobit seems more sensible (one-tail distribution); also yields better behaved downscalers than those on the exponential scale (required with the log normal).
- Also, model comparison, holding out stations, leads to preference for the Tobit model.

Tobit details

- The Tobit model is implemented by defining

  $$Z_t^i(s) = \max\left(0, \widetilde{Z}_t^i(s)\right)$$

  where

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

- Here, $\beta_{0,t}^i$ and $\beta_{1,t}^i$ serve as additive and multiplicative bias adjustments to the CMAQ prediction with $\beta_{0,t}^i(\boldsymbol{s})$ providing local spatial adjustment to the intercept.
- For the grid block B_s, containing location s, the weekly average CMAQ output on week t is denoted by $Q_t^i(B_s)$.
- We model $\beta_{0,t}^i(\boldsymbol{s})$ as a Gaussian process with exponential covariance

  $$\beta_{0,t}^i(\boldsymbol{s}) = \sigma_t^i\, w_t^i(\boldsymbol{s})$$

- Spatially varying slopes - $\beta_{1,t}^i(\boldsymbol{s})$ (centered around $\beta_{1,t}^i$) correlated with the $\beta_{0,t}^i(\boldsymbol{s})$?

The 'other' species

- We model the true value of 'other' species, $Z_t^o(s)$, in the same way as the primary species,

$$
\begin{aligned}
\widetilde{Z}_t^o(\mathbf{s}) &= \beta_{0,t}^o + \beta_{0,t}^o(\mathbf{s}) + \beta_{1,t}^o\, Q_t^o(B_{\mathbf{s}}) \\
Z_t^o(s) &= \max\left(0, \widetilde{Z}_t^o(s)\right),
\end{aligned}
$$

  with $\beta_{0,t}^o(\mathbf{s})$ being a zero mean Gaussian process with exponential covariance parameterized by ϕ_t^o and $\sigma_{o,t}$ as the range range and scale parameters respectively.

- We estimate ‘other’ for CMAQ using

$$
Q_t^o(B_{\mathbf{s}}) = Q_t^{tot}(B_{\mathbf{s}}) - \sum_{i=1}^{5} Q_t^i(B_{\mathbf{s}}).
$$

- As with the species models, the CMAQ 'other' data help constrain the true values of 'other' in the model, with the Tobit specification ensuring that $Z_t^o(\mathbf{s})$ is always positive.

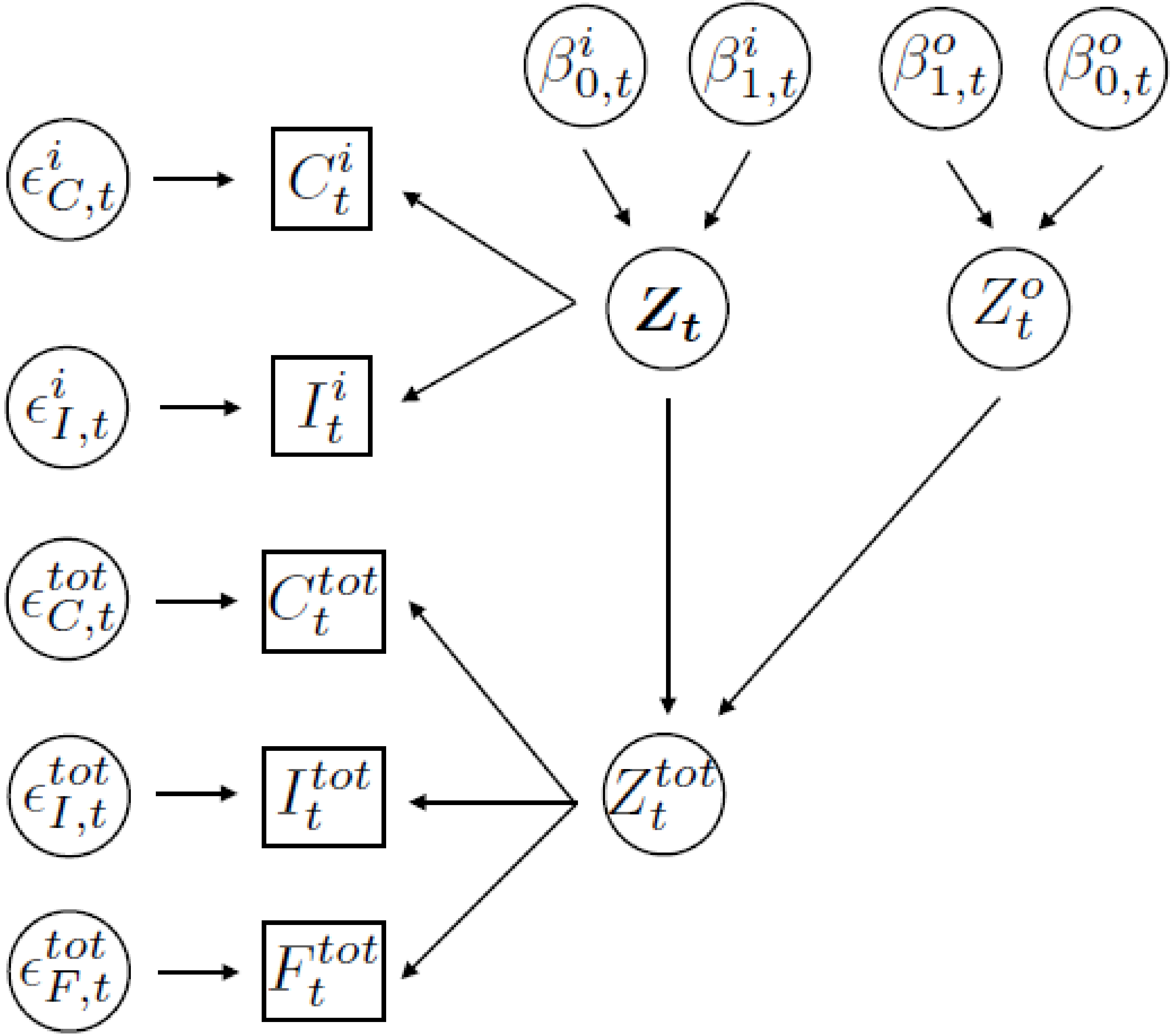

Figure 4. Diagram of model framework for total and speciated $\mathrm{PM}_{2.5}$. Arrows indicate the dependence between model parameters.

Modeling cont.

- For a given t, the joint distribution of all of the variables, ignoring parameters, can be written concisely as the product of all species i (including "other") and locations s.

$$\prod_{s} \left[C_t^{tot}(s), I_t^{tot}(s), F_t^{tot}(s) | Z_t^{tot}(s)\right]$$

$$\prod_{i} \left[C_t^{i}(\boldsymbol{s}), I_t^{i}(\boldsymbol{s}) | Z_t^{i}(\boldsymbol{s})\right] \left[Z_t^{i}(s) \mid \tilde{Z}_t^{i}(s)\right] \left[\tilde{Z}_t^{i}(s) \mid Q_t^{i}(B_s)\right]$$

- Simultaneous univariate downscaler models for each species subject to the summation constraint that their mass plus 'other' must be equal to the true value of total $PM_{2.5}$.
- The summation constraints enable introduction of the FRM data in addition to the CSN and IMPROVE data.

Species dependence

- Dependence between the $\text{PM}_{2.5}$ species using the multivariate downscaling?
- We explored bivariate dependence between sulfate and ammonium, the two species with highest correlation by jointly modeling the local spatial adjustment terms ($\beta_0^1(\boldsymbol{s})$ and $\beta_0^3(\boldsymbol{s})$) for these species using a coregionalization approach,

$$\begin{pmatrix} \beta_0^1(\boldsymbol{s}) \\ \beta_0^3(\boldsymbol{s}) \end{pmatrix} = \boldsymbol{A} \begin{pmatrix} w_0(\boldsymbol{s}) \\ w_1(\boldsymbol{s}) \end{pmatrix}$$

where $w_0(\boldsymbol{s})$ and $w_1(\boldsymbol{s})$ are mean zero Gaussian processes with exponential covariance structure.

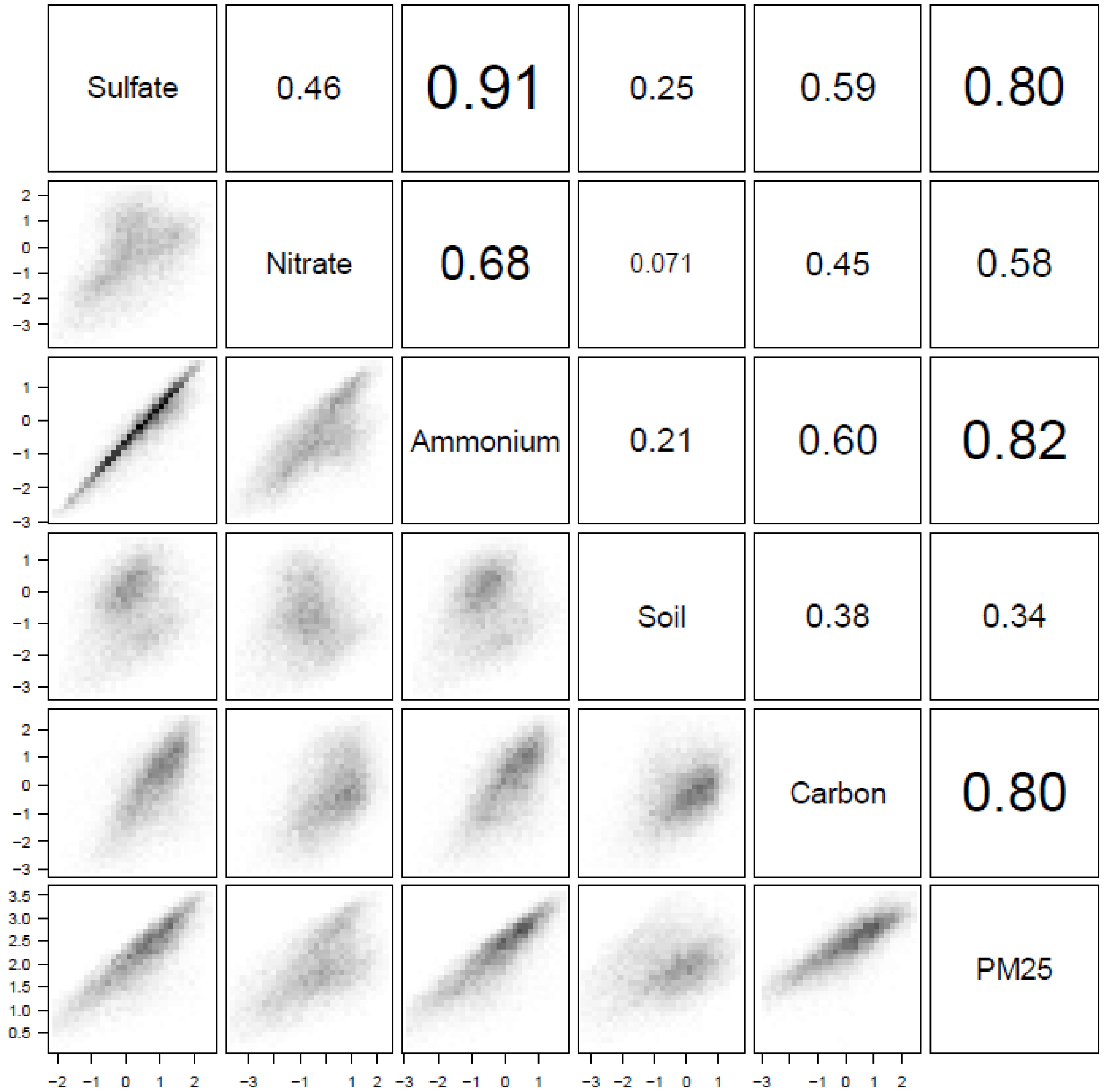

Figure 5: Pairs plot of $PM_{2.5}$ species and total mass. Cells above the diagonal report the Pearson correlation coefficient between each pair of variables, while cells below the diagonal contain bivariate histograms using rectangular binning.

Priors

- Noninformative and weakly informative priors that preserve conjugacy where possible.
- Spatial models are fitted separately for each t and in parallel.
- The measurement error terms, $\epsilon_t^i(\boldsymbol{s})$, for each species and total $\mathrm{PM}_{2.5}$ are assumed to be normal with mean zero and variance hyper-parameter $\tau^2_{\epsilon_t^i}(s)$ which in turn have an inverse gamma(2,2) hyper-prior.
- Similarly, σ_t^i, the variance parameters for the covariance functions for $\beta_0^i(\boldsymbol{s})$, are also given inverse gamma(2,2) prior distributions. Little sensitivity.
- The $\beta_{0,t}^i$ and $\beta_{1,t}^i$ are given normal priors with mean zero and a large variance (500).
- Finally, the spatial range parameters, ϕ_t^i, are given uniform prior distributions with support for the effective range to be between 0 and half the maximum distance between stations.

Fitting

- The model is fit using a hybrid MCMC with a Metropolis-Hastings within Gibbs sampler.
- Parameters $\tau^2_{\epsilon^i_t}$, σ^i_t are updated using Gibbs steps.
- The introduction of the Tobit transformation breaks conjugacy for parameters β^i_0, β^i_1, and $\beta^i_0(\boldsymbol{s})$. Metropolis-Hastings updates employed.
- All ϕ^i_ts require a Metropolis-Hastings step as well.
- For the parameters updated via Metropolis-Hastings we use a random walk proposal distribution with proposal variances tuned to achieve univariate acceptance rates close to 40%.

Prediction

- Prediction is carried out through generation of posterior predictive samples using composition.
- Of interest, for a new location $\boldsymbol{s}'$ are the posterior predictive distributions for the Z's, the C's, the I's, and the F's.
- The Z's are used for kriging, i.e., for interpolation of species levels in a given week at a new location.
- The predictive distributions for the C's and F's can be used for model validation and model choice.
- That is, using hold out data, comparison can be made with the observed values using, e.g., predictive mean square error (PMSE), empirical coverage vs. nominal coverage for predictive intervals, and continuous rank probability scores (CRPS).
- The Z's can not be used for this purpose; they do not carry the uncertainty associated with observed data.
- Prediction for a fine grid of locations across North America as well as aggregation of these predictions seasonally and over the entire year.

Model comparison

- We randomly select of 10% of CSN and IMPROVE stations as a validation set to assess the out-of-sample predictive performance of the models.
- Also include all locations which do not contain a complete set of observations.
- For each validation location s and week t we sample the posterior predictive distributions of $C_t^i(\boldsymbol{s})$ or $I_t^i(\boldsymbol{s})$, depending the location's network membership, for all five species, other, and total PM$_{2.5}$ ($i \in \{1, \ldots, 5, o, tot\}$).
- Comparison using root mean square error, continuous ranked probability score, and the empirical coverage of the nominal 90% predictive interval.

The criteria

- $$\text{RMSE} = \sqrt{\frac{1}{\sum_t n_t} \sum_{t=1}^{52} \sum_{r=1}^{n_t} \left(\widehat{Y_t}(\boldsymbol{s}_r) - Y_t(\boldsymbol{s}_r) \right)^2 \cdot \mathrm{I}\left(Y_t(s_r) \right)},$$

  where n_t is the number of validation locations at time t and $\boldsymbol{s}_r$ is the rth validation location. For t and $\boldsymbol{s}_r$, $\widehat{Y_t}(\boldsymbol{s}_r)$ is the posterior predictive mean and $Y_t(\boldsymbol{s}_r)$ is the observed value.

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

  where F is the empirical cumulative predictive distribution function and y is the observed value. The reported CRPS values represent the average CRPS values over all validation locations for each time point.

- Three tobit models: (i) the species are downscaled independently, (ii) species dependence between sulfate and ammonium, and (iii) model without CMAQ

TABLES

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

Table 1. Model validation results of posterior predictive means from joint downscaler model variants for randomly selected hold out locations. These include root mean square predictive error (RMSPE), average continuous rank probability score (CRPS), empirical 90% coverages (EmpCov), and out of sample R^2 (OoS R^2).

Results of the Tobit analysis

- Clear seasonal spatial patterns of speciated and total $PM_{2.5}$.
- Sulfate levels higher in the warm summer months and account for the largest fraction of total $PM_{2.5}$ in the eastern US.
- Nitrate concentrations higher in the winter, particularly in the upper mid-west due to increased ammonia availability in that region.
- Ammonium levels exhibit strong seasonal variation; highest concentrations occur during the summer in the central eastern US.
- High predicted soil levels during the spring and summer seasons in the southwest US due to low soil moisture and high wind speeds in this region.
- Highest levels of total carbon in the upper northwest during the summer months.

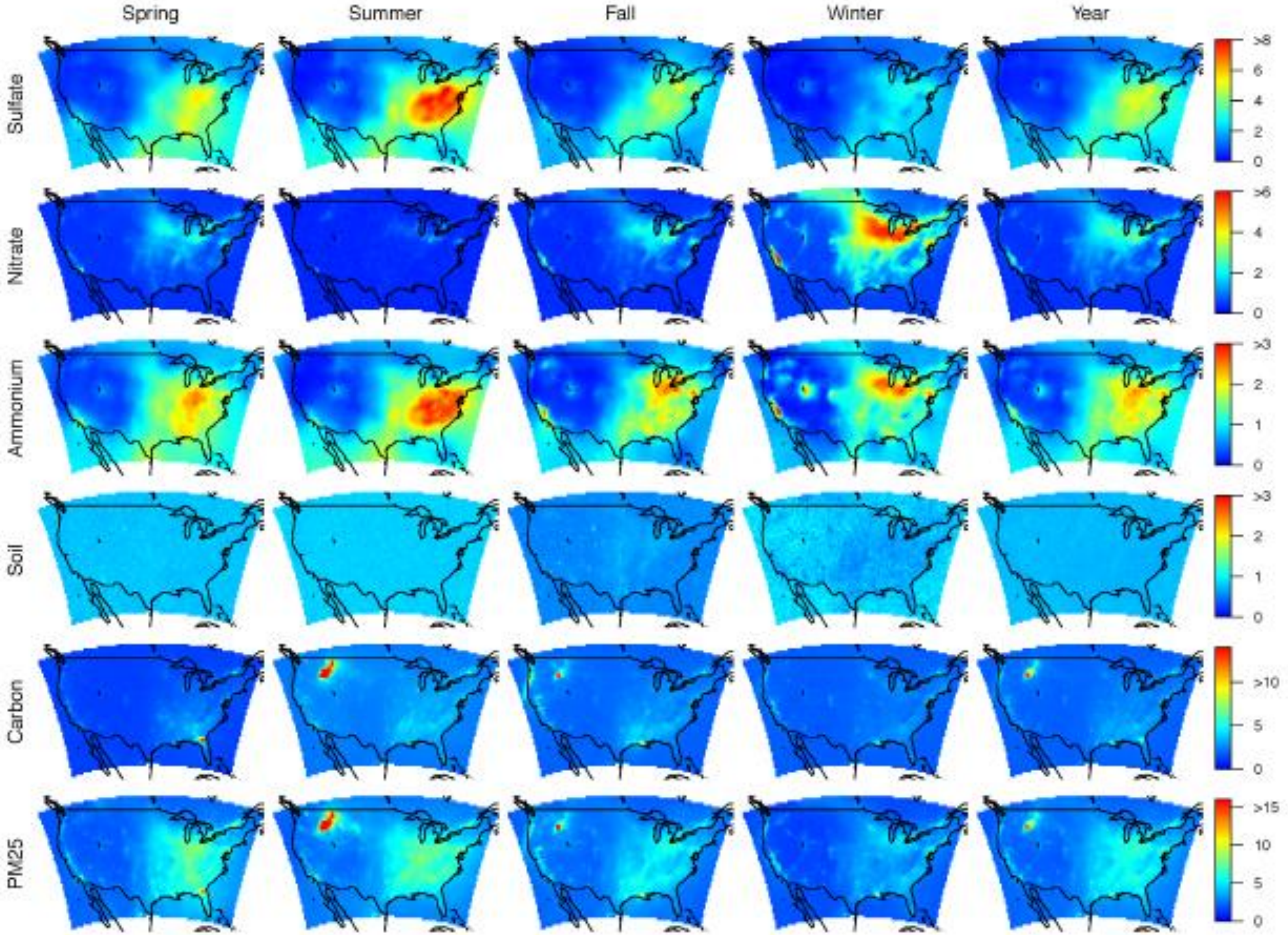

Figure 5. Maps of the seasonal and yearly average of posterior mean predicted $PM_{2.5}$ species and total $PM_{2.5}$ from the tobit with species dependence model.

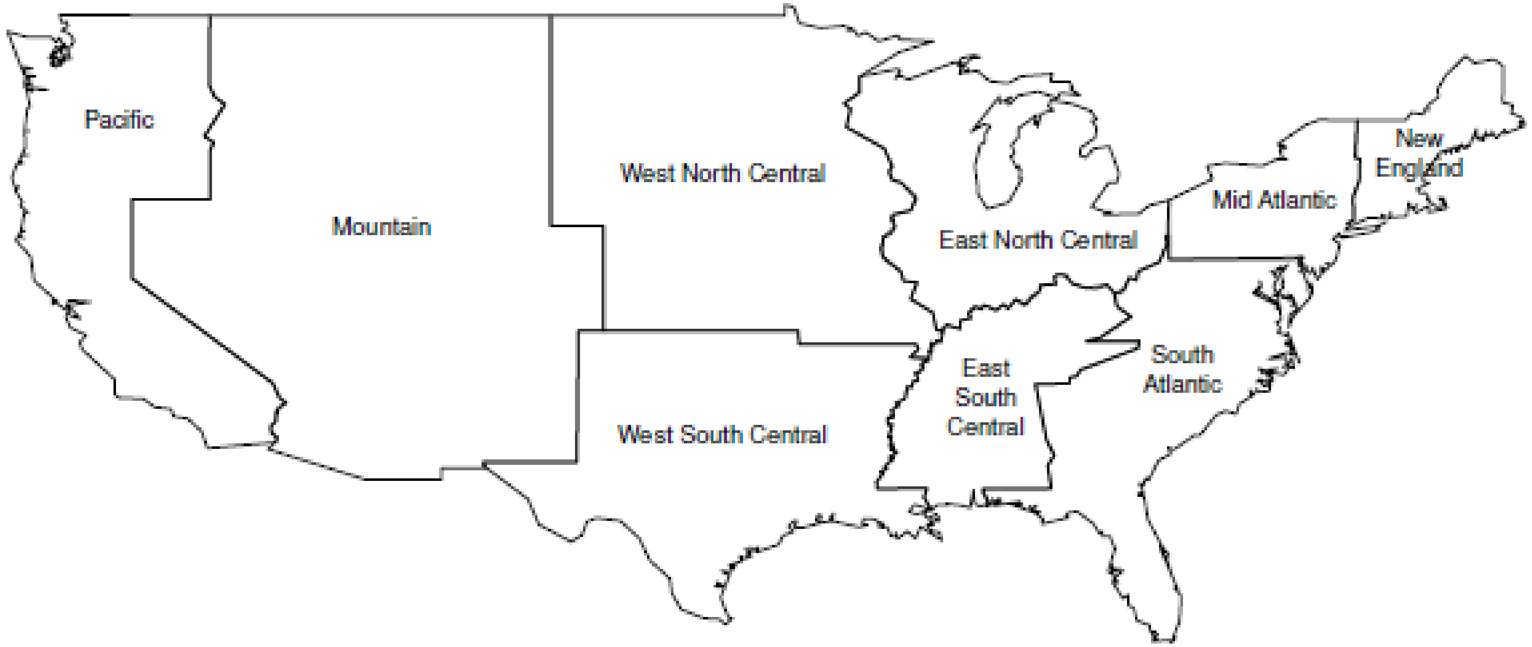

Figure 7. Map of the regions used to aggregate predictions based on regions used in Choi *et al.* (2009).

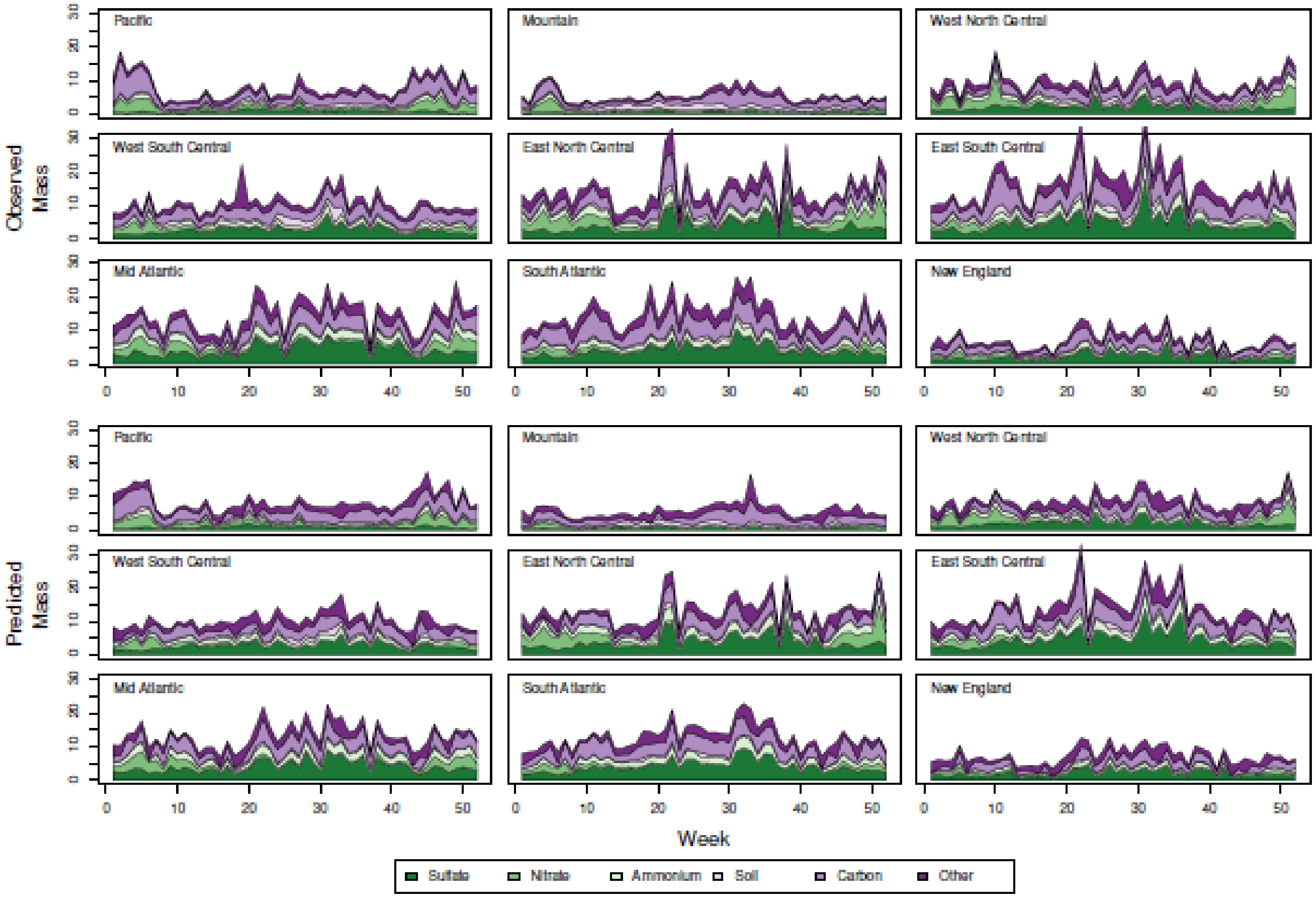

Figure 6. Stacked timeseries plots of the observed mass at CSN and IMPROVE monitoring locations aggregated to multistate regions defined in Figure 7 and the posterior mean predicted $PM_{2.5}$ species from the tobit with species dependence model aggregated to the same regions.

Summary

- We have presented a complex downscaling model for for speciated $PM_{2.5}$ which accomplishes a fusion of four data sources.
- The model specifies a latent ‘truth’ where the downscaling is done and then introduces measurement error to accommodate challenges with the monitoring data.
- We have modeled each species at its own magnitude and incorporated an 'other' component in order to provide coherency between species and total.
- We fit independent weekly models and provide summaries for the entire continental U.S.
- We show that downscaling substantially improves upon what is essentially ordinary kriging at the species level.

Future work

- More and more speciated data is being collected.
- Can enable assessment of change in speciation across years as well as across space. Are exposures to some species increasing while others are perhaps stable or decreasing?
- There are some cities which collect their own $PM_{2.5}$ component data. Hence, further validation of our model as well as potential refinement.
- Lastly, need to link exposure to adverse health outcomes.
- With a better understanding of the nature of species-level exposure, we can consider different health outcomes; which species are more influential with regard to the differing outcomes?