A Data Fusion Approach for Space-Time Analysis of Speciated $PM_{2.5}$

Colin Rundel

Duke University

August 3, 2014

Background

Fine particulate matter ($PM_{2.5}$) is an EPA regulated air pollutant linked to a variety of adverse health effects

- Classified based on particle size (< 2.5 μm diameter)
- Major species: Sulfate, Nitrate, Ammonium, Soil, Carbon.
- Minor species: trace elements (K, Mg, Ca), heavy metals (Cu, Fe), etc.
- Complex spatio-temporal dependence between species

Data (2007)

Speciated $PM_{2.5}$ Sources

- Chemical Speciation Network (CSN) - 221 stations
- Interagency Monitoring of Protected Visual Environments (IMPROVE) - 172 stations

Total $PM_{2.5}$ Sources

- Federal Reference Method (FRM) - 949 stations

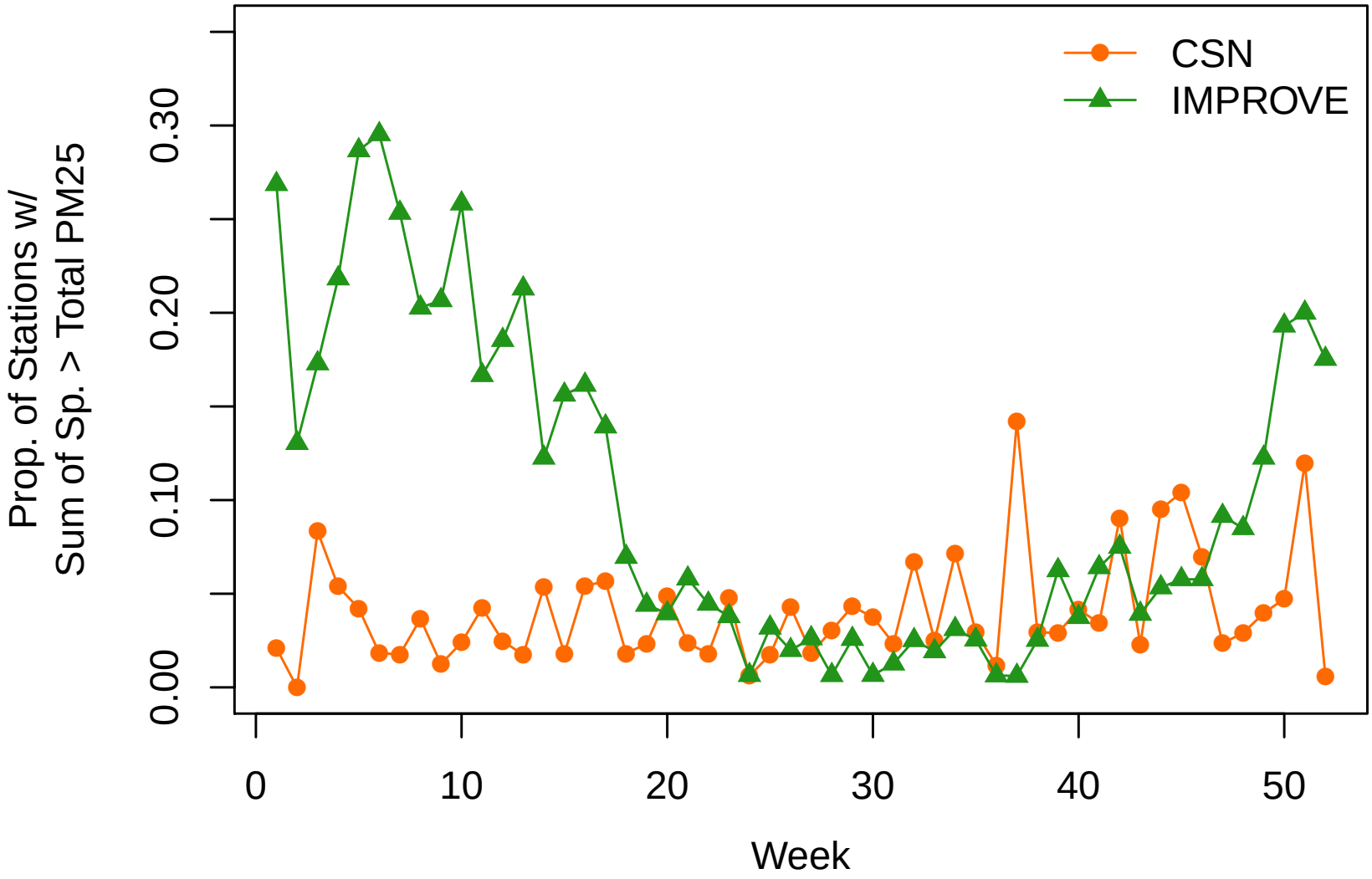
Model Output

- Community Multi-scale Air Quality (CMAQ) - 12 km grid

Data Issues

- Monitoring frequency
- Total vs Sum of Species

Total $PM_{2.5}$ vs Sum of Species



Species Model Details

For the 5 major species (Sulfate, Nitrate, Ammonium, Soil, Carbon) and the two networks (CSN, IMPROVE):

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

where $Z_t^i(\boldsymbol{s})$ are the latent "true" concentrations of species i at time t and locations $\boldsymbol{s}$,

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

Total $PM_{2.5}$ Model Details

For total $PM_{2.5}$ from the three networks (CSN, IMPROVE, FRM):

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

where $Z_t^{tot}(\boldsymbol{s})$ are the latent "true" concentration of total $PM_{2.5}$ at time t and locations $\boldsymbol{s}$, which is given by the sum of the major species and the "other" species concentrations.

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

$$Z_t^o(s) = \max\left(0,\ \widetilde{Z}_t^o(s)\right) \qquad \widetilde{Z}_t^o(\boldsymbol{s}) = \beta_{0,t}^o + \beta_{0,t}^o(\boldsymbol{s}) + \beta_{1,t}^o\, Q_t^o(B_{\boldsymbol{s}})$$

Spatial Dependence

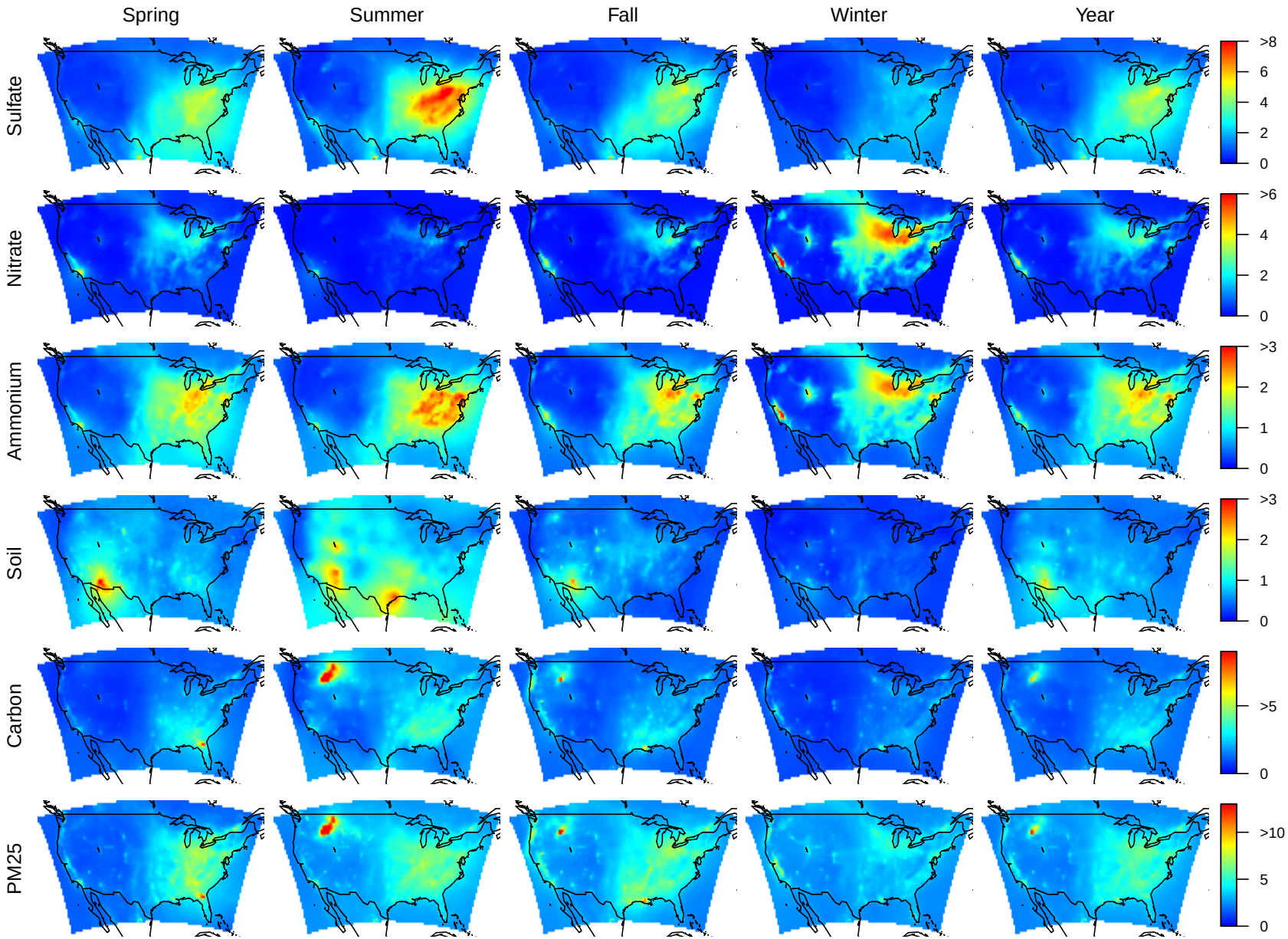
Spatial dependence enters the model through the $\beta^i_{0,t}(s)$ parameters for $i \in \{o, 1, 2, 3, 4, 5\}$.

$$\beta^i_{0,t}(\boldsymbol{s}) = \sigma^i_t \; w^i_t(\boldsymbol{s})$$

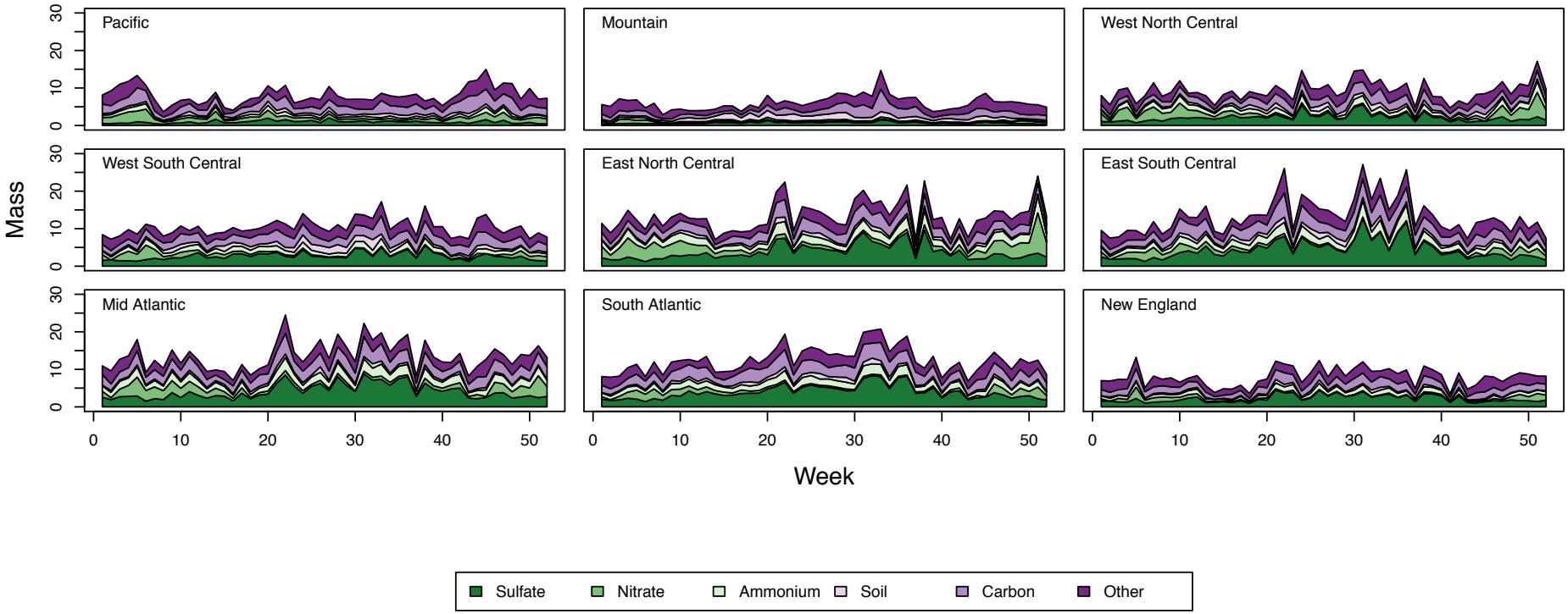
where $w^i_t(\boldsymbol{s})$ are zero mean, variance 1, Gaussian processes with exponential correlation given by

$$\text{corr}(w^i_t(\boldsymbol{s}), w^i_t(\boldsymbol{s}')) = \exp(-\phi^i_t|\boldsymbol{s} - \boldsymbol{s}'|)$$

Model results



Model results



Model Validation

		Sulfate	Nitrate	Ammonium	Soil	Carbon	PM25
RMSE	Tobit w/o CMAQ	1.347	2.257	0.858	1.363	3.073	6.298
	Tobit	1.151	1.641	0.724	1.307	2.851	5.393
CRPS	Tobit w/o CMAQ	0.639	0.758	0.374	0.468	1.064	3.023
	Tobit	0.554	0.558	0.329	0.438	0.885	2.452
EmpCov	Tobit w/o CMAQ	0.935	0.931	0.933	0.907	0.923	0.924
	Tobit	0.920	0.930	0.924	0.915	0.921	0.906

Validation based on randomly selecting 10% of stations as hold outs.

Run times

Total run time for model fitting (50,000 iterations):

- CPU - 7.7 hours
- CPU+GPU - 4.8 hours

$\times$ 52 weeks

Total run time for model prediction at 5950 locations (1,000 iterations):

- CPU - 7.2 hours
- CPU+GPU - 4.3 hours

$\times$ 52 weeks

One run takes about 775 hours total on CPU alone, 473 on CPU and GPU.

Model fitting performance

Parameter	CPU (secs)	CPU+GPU (secs)	Rel. Performance
$\beta_0,\ \beta_1$	0.00029	0.00030	0.97
$\beta_0(s)$	0.09205	0.09132	1.00
σ	0.00383	0.00385	0.99
ϕ	0.46084	0.25174	1.83
ϵ	0.00003	0.00003	1.00
Total	0.55708	0.34729	1.60

Acknowledgments

- Alan Gelfand - Duke
- Dave Holland - EPA
- Erin Schliep - Duke
- Wyat Apel - NERL

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

Information & Questions

Email : rundel@gmail.com

Presentation : `http://github.com/rundel/Presentations/`

Paper : Rundel C., Schliep E., Holland D., Gelfand A. (2014) A data fusion approach for space-time analysis of speciated PM2.5. *The Annals of Applied Statistics*. *In submission*

Information & Questions

Email : rundel@gmail.com

Presentation : `http://github.com/rundel/Presentations/`

Paper : Rundel C., Schliep E., Holland D., Gelfand A. (2014) A data fusion approach for space-time analysis of speciated PM2.5. *The Annals of Applied Statistics*. *In submission*

Questions?

Performance

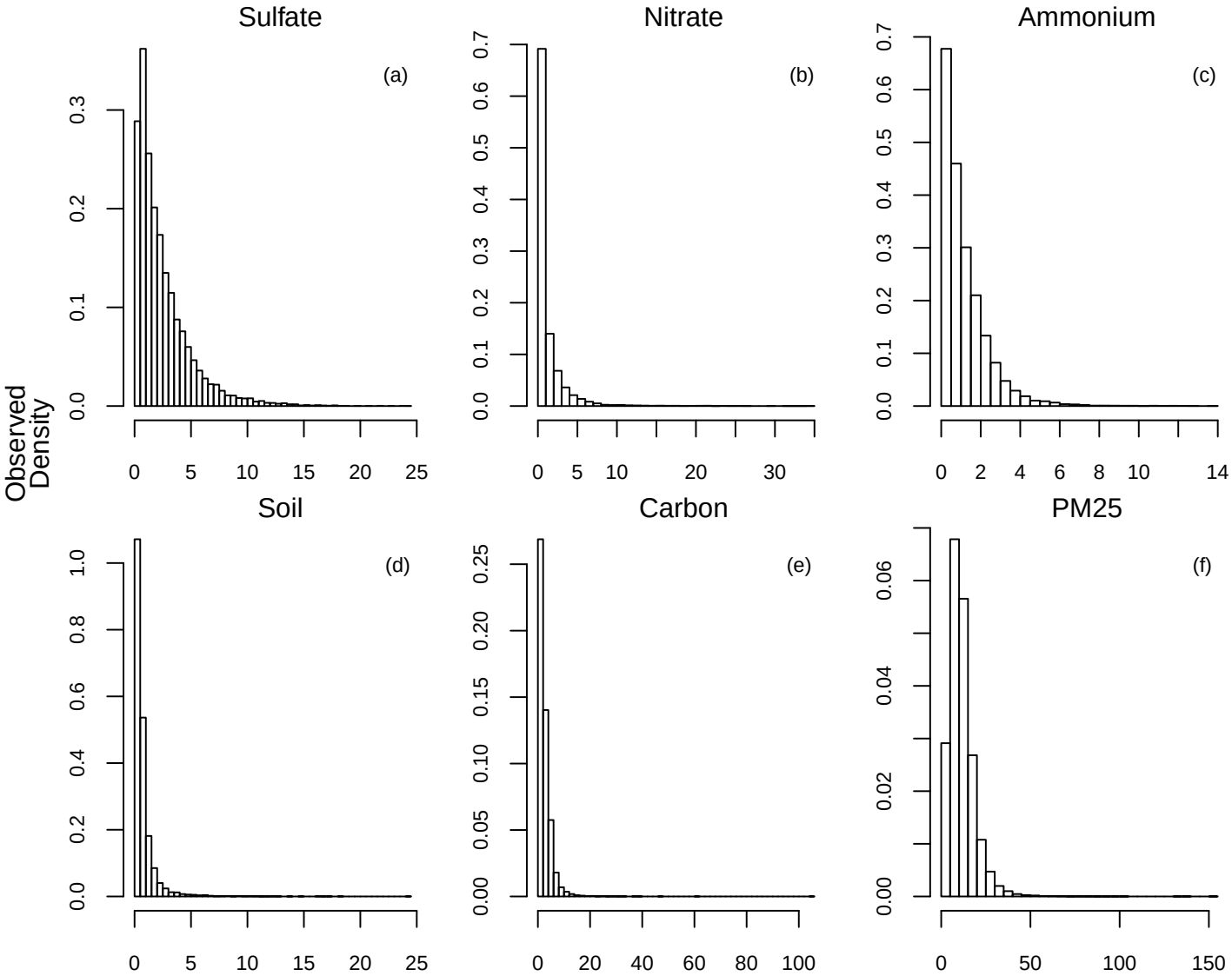
System Specs:

- 4 core Intel i5-2500K @ 3.30 GHz
- 16 GB DDR3 @ 1333 MHz
- GeForce GTX 460

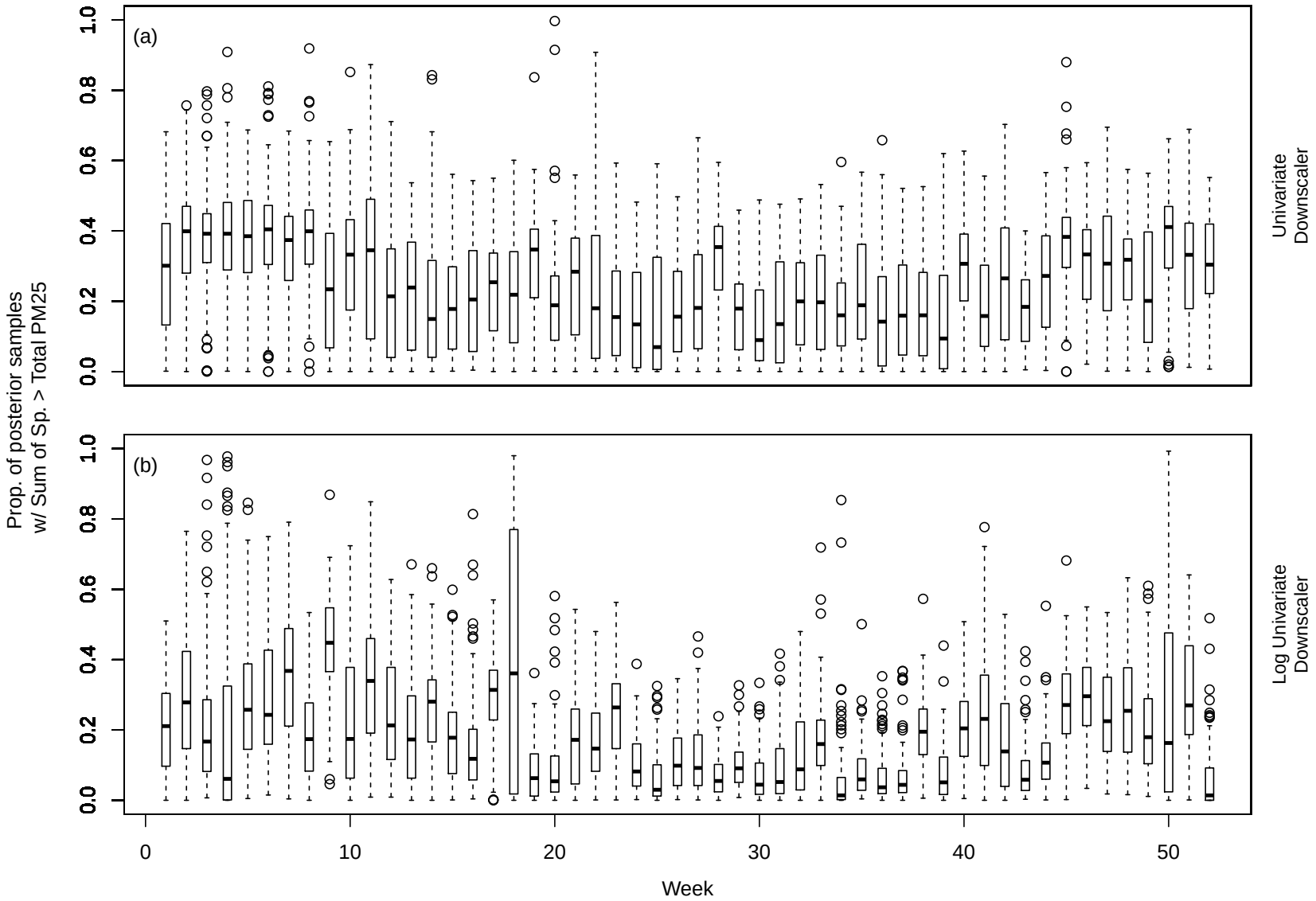
Software Specs:

- Ubuntu 13.10
- OpenBlas 0.2.8
- CUDA 6.0RC
- Magma 1.4.1

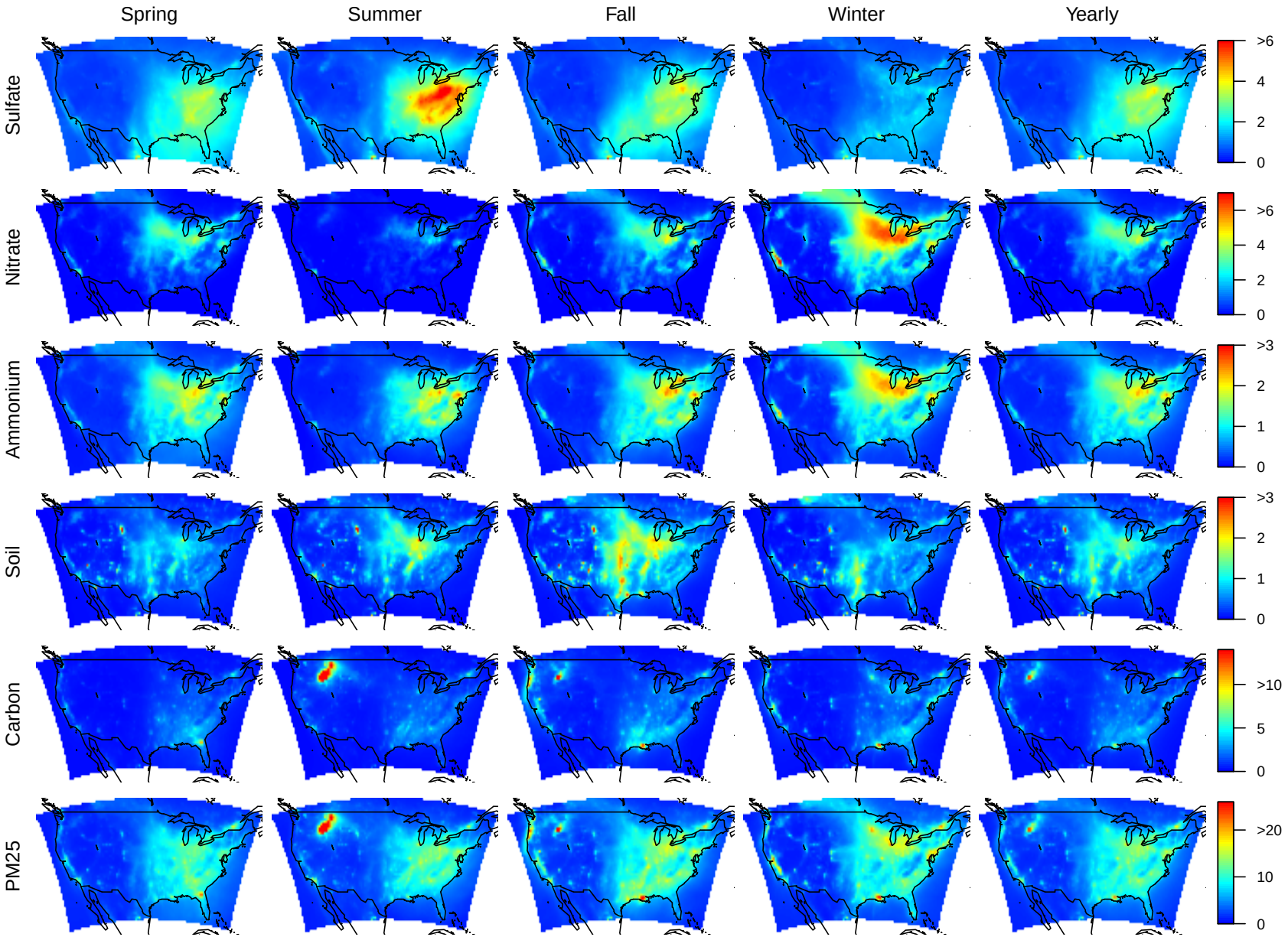
Species Distributions



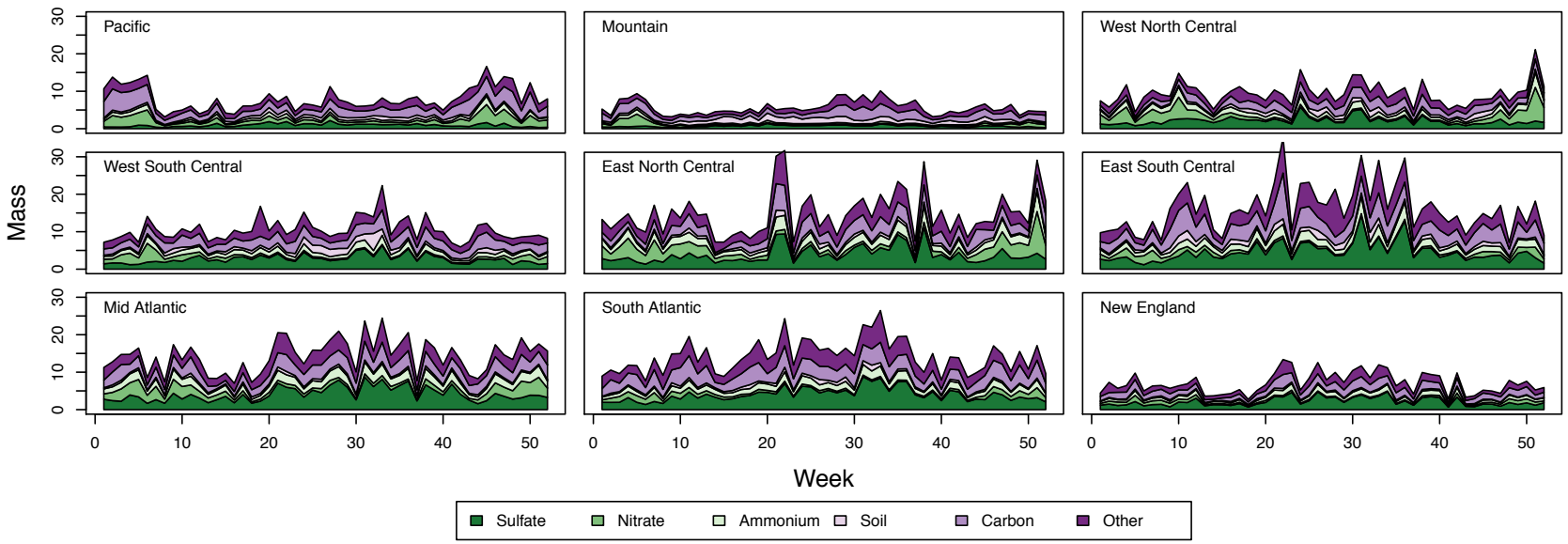
Univariate Model Exceedance



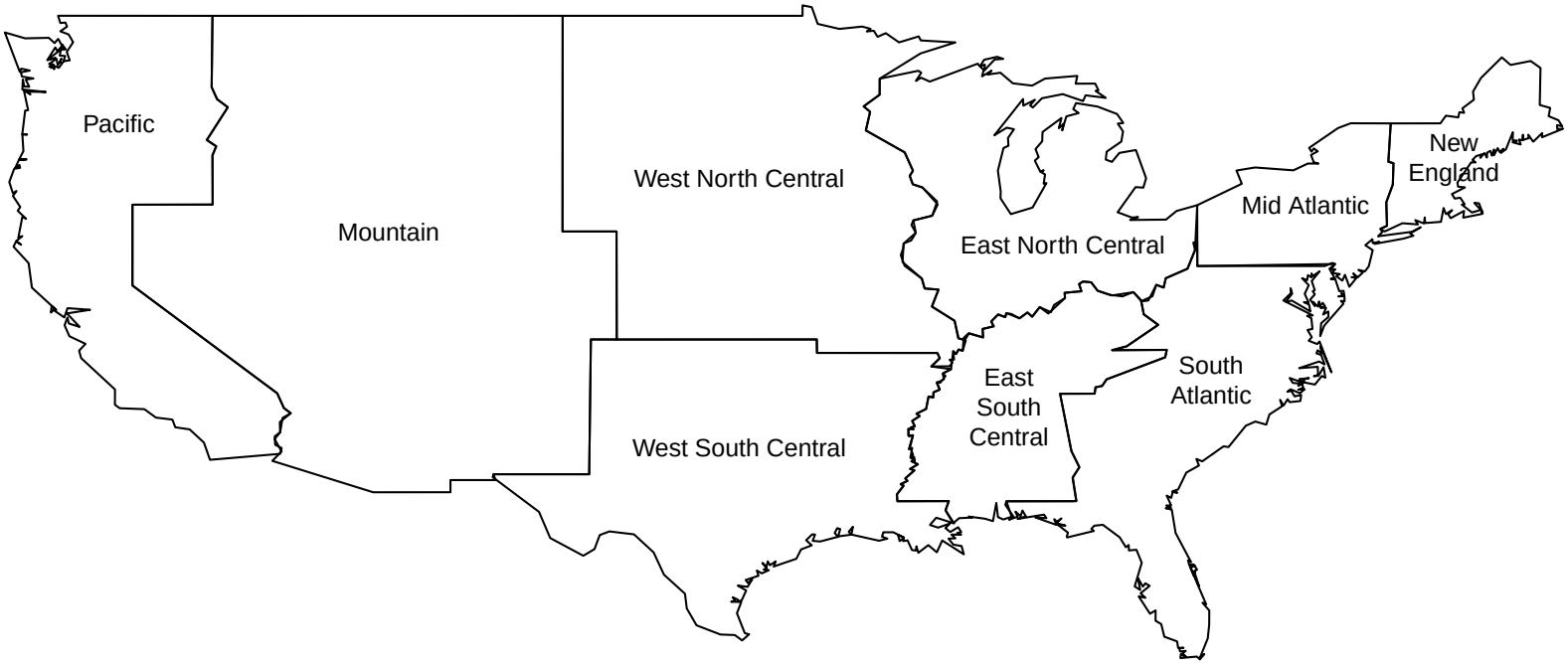
CMAQ Maps


Spring
Summer
Fall
Winter
Yearly
Sulfate
Nitrate
Ammonium
Soil
Carbon
PM25
>6
4
2
0
>6
4
2
0
>3
2
1
0
>3
2
1
0
>10
5
0
>20
10
0

Data TS


Pacific
Mountain
West North Central
West South Central
East North Central
East South Central
Mid Atlantic
South Atlantic
New England
Mass
Week
0
10
20
30
40
50
Sulfate
Nitrate
Ammonium
Soil
Carbon
Other

Regions


Pacific
Mountain
West North Central
East North Central
Mid Atlantic
New England
West South Central
East South Central
South Atlantic