Fusing point and areal level space-time data with application to wet deposition

Sujit Sahu
http://www.soton.ac.uk/∼sks
Southampton Statistical Sciences Research Institute,
Joint work with Alan Gelfand and David Holland (USEPA)

Warwick: March 2009

Modelling Ingredients and Innovations

- Combine observed data with output from computer simulation model. Data Assimilation
- Modelling large space-time data, about 1.8 million.
- Change of support problem: reconcile the differences between point and areal data using multi-scale random effects.
- Missing data, covariate effects, point mass at zero.
- Auto-regressive (AR) process in time and
- Conditional Auto-Regressive (CAR) in space.
- Aggregation in time and space; spatial interpolation.

All done using Bayesian Hierarchical Modelling and MCMC

Modelling Ingredients and Innovations

- Combine observed data with output from computer simulation model. Data Assimilation
- Modelling large space-time data, about 1.8 million.
- Change of support problem: reconcile the differences between point and areal data using multi-scale random effects.
- Missing data, covariate effects, point mass at zero.
- Auto-regressive (AR) process in time and
- Conditional Auto-Regressive (CAR) in space.
- Aggregation in time and space; spatial interpolation.

All done using Bayesian Hierarchical Modelling and MCMC

Modelling Ingredients and Innovations

- Combine observed data with output from computer simulation model. Data Assimilation
- Modelling large space-time data, about 1.8 million.
- Change of support problem: reconcile the differences between point and areal data using multi-scale random effects.
- Missing data, covariate effects, point mass at zero.
- Auto-regressive (AR) process in time *and*
- Conditional Auto-Regressive (CAR) in space.
- Aggregation in time and space; spatial interpolation.

All done using Bayesian Hierarchical Modelling and MCMC

Modelling Ingredients and Innovations

- Combine observed data with output from computer simulation model. Data Assimilation
- Modelling large space-time data, about 1.8 million.
- Change of support problem: reconcile the differences between point and areal data using multi-scale random effects.
- Missing data, covariate effects, point mass at zero.
- Auto-regressive (AR) process in time and
- Conditional Auto-Regressive (CAR) in space.
- Aggregation in time and space; spatial interpolation.

All done using Bayesian Hierarchical Modelling and MCMC

Modelling Ingredients and Innovations

- Combine observed data with output from computer simulation model. Data Assimilation
- Modelling large space-time data, about 1.8 million.
- Change of support problem: reconcile the differences between point and areal data using multi-scale random effects.
- Missing data, covariate effects, point mass at zero.
- Auto-regressive (AR) process in time *and*
- Conditional Auto-Regressive (CAR) in space.
- Aggregation in time and space; spatial interpolation.

All done using Bayesian Hierarchical Modelling and MCMC

Modelling Ingredients and Innovations

- Combine observed data with output from computer simulation model. Data Assimilation
- Modelling large space-time data, about 1.8 million.
- Change of support problem: reconcile the differences between point and areal data using multi-scale random effects.
- Missing data, covariate effects, point mass at zero.
- Auto-regressive (AR) process in time *and*
- Conditional Auto-Regressive (CAR) in space.
- Aggregation in time and space; spatial interpolation.

All done using Bayesian Hierarchical Modelling and MCMC

Modelling Ingredients and Innovations

- Combine observed data with output from computer simulation model. Data Assimilation
- Modelling large space-time data, about 1.8 million.
- Change of support problem: reconcile the differences between point and areal data using multi-scale random effects.
- Missing data, covariate effects, point mass at zero.
- Auto-regressive (AR) process in time *and*
- Conditional Auto-Regressive (CAR) in space.
- Aggregation in time and space; spatial interpolation.

All done using Bayesian Hierarchical Modelling and MCMC

Challenges in Implementation (MCMC)

- High dimensional ($\approx$34K) latent spatial processes.
- Spatially varying slope and intercepts.
- Parameter identifiability in hierarchical models.
- Constrained parameter spaces.
- Selection of model tuning parameters.
- Complexity in coding, monitoring and diagnosing MCMC convergence.
- Spatial prediction on a large grid (Eastern US).

Friendly Agents

- Conjugate sampling.
- Some proper prior processes.
- Cheap local CAR updates (Gibbs Sampling).

Challenges in Implementation (MCMC)

- High dimensional ($\approx$34K) latent spatial processes.
- Spatially varying slope and intercepts.
- Parameter identifiability in hierarchical models.
- Constrained parameter spaces.
- Selection of model tuning parameters.
- Complexity in coding, monitoring and diagnosing MCMC convergence.
- Spatial prediction on a large grid (Eastern US).

Friendly Agents

- Conjugate sampling.
- Some proper prior processes.
- Cheap local CAR updates (Gibbs Sampling).

Challenges in Implementation (MCMC)

- High dimensional ($\approx$34K) latent spatial processes.
- Spatially varying slope and intercepts.
- Parameter identifiability in hierarchical models.
- Constrained parameter spaces.
- Selection of model tuning parameters.
- Complexity in coding, monitoring and diagnosing MCMC convergence.
- Spatial prediction on a large grid (Eastern US).

Friendly Agents

- Conjugate sampling.
- Some proper prior processes.
- Cheap local CAR updates (Gibbs Sampling).

Challenges in Implementation (MCMC)

- High dimensional ($\approx$34K) latent spatial processes.
- Spatially varying slope and intercepts.
- Parameter identifiability in hierarchical models.
- Constrained parameter spaces.
- Selection of model tuning parameters.
- Complexity in coding, monitoring and diagnosing MCMC convergence.
- Spatial prediction on a large grid (Eastern US).

Friendly Agents

- Conjugate sampling.
- Some proper prior processes.
- Cheap local CAR updates (Gibbs Sampling).

Challenges in Implementation (MCMC)

- High dimensional ($\approx$34K) latent spatial processes.
- Spatially varying slope and intercepts.
- Parameter identifiability in hierarchical models.
- Constrained parameter spaces.
- Selection of model tuning parameters.
- Complexity in coding, monitoring and diagnosing MCMC convergence.
- Spatial prediction on a large grid (Eastern US).

Friendly Agents

- Conjugate sampling.
- Some proper prior processes.
- Cheap local CAR updates (Gibbs Sampling).

Challenges in Implementation (MCMC)

- High dimensional ($\approx$34K) latent spatial processes.
- Spatially varying slope and intercepts.
- Parameter identifiability in hierarchical models.
- Constrained parameter spaces.
- Selection of model tuning parameters.
- Complexity in coding, monitoring and diagnosing MCMC convergence.
- Spatial prediction on a large grid (Eastern US).

Friendly Agents

- Conjugate sampling.
- Some proper prior processes.
- Cheap local CAR updates (Gibbs Sampling).

Challenges in Implementation (MCMC)

- High dimensional ($\approx$34K) latent spatial processes.
- Spatially varying slope and intercepts.
- Parameter identifiability in hierarchical models.
- Constrained parameter spaces.
- Selection of model tuning parameters.
- Complexity in coding, monitoring and diagnosing MCMC convergence.
- Spatial prediction on a large grid (Eastern US).

Friendly Agents

- Conjugate sampling.
- Some proper prior processes.
- Cheap local CAR updates (Gibbs Sampling).

Challenges in Implementation (MCMC)

- High dimensional ($\approx$34K) latent spatial processes.
- Spatially varying slope and intercepts.
- Parameter identifiability in hierarchical models.
- Constrained parameter spaces.
- Selection of model tuning parameters.
- Complexity in coding, monitoring and diagnosing MCMC convergence.
- Spatial prediction on a large grid (Eastern US).

Friendly Agents

- Conjugate sampling.
- Some proper prior processes.
- Cheap local CAR updates (Gibbs Sampling).

Challenges in Implementation (MCMC)

- High dimensional ($\approx$34K) latent spatial processes.
- Spatially varying slope and intercepts.
- Parameter identifiability in hierarchical models.
- Constrained parameter spaces.
- Selection of model tuning parameters.
- Complexity in coding, monitoring and diagnosing MCMC convergence.
- Spatial prediction on a large grid (Eastern US).

Friendly Agents

- Conjugate sampling.
- Some proper prior processes.
- Cheap local CAR updates (Gibbs Sampling).

Chemical Deposition

- Combustion of fossil fuel produces various chemicals including sulfate and nitrate gases.
- In the eastern U.S., most SO_2, and NO_x release attributed to power plants.
- Emitted to the air; wet deposition and dry deposition; interest in total deposition.
- Deposition means return to the earth's surface by means of precipitation (rain or snow) for example.
- Wet Deposition = Precipitation $\times$ Concentration.
- Wet deposition is responsible for damage to lakes, forests, and streams.

Chemical Deposition

- Combustion of fossil fuel produces various chemicals including sulfate and nitrate gases.
- In the eastern U.S., most SO_2, and NO_x release attributed to power plants.
- Emitted to the air; wet deposition and dry deposition; interest in total deposition.
- Deposition means return to the earth's surface by means of precipitation (rain or snow) for example.
- Wet Deposition = Precipitation $\times$ Concentration.
- Wet deposition is responsible for damage to lakes, forests, and streams.

Chemical Deposition

- Combustion of fossil fuel produces various chemicals including sulfate and nitrate gases.
- In the eastern U.S., most SO_2, and NO_x release attributed to power plants.
- Emitted to the air; wet deposition and dry deposition; interest in total deposition.
- Deposition means return to the earth's surface by means of precipitation (rain or snow) for example.
- Wet Deposition = Precipitation $\times$ Concentration.
- Wet deposition is responsible for damage to lakes, forests, and streams.

Chemical Deposition

- Combustion of fossil fuel produces various chemicals including sulfate and nitrate gases.
- In the eastern U.S., most SO_2, and NO_x release attributed to power plants.
- Emitted to the air; wet deposition and dry deposition; interest in total deposition.
- Deposition means return to the earth's surface by means of precipitation (rain or snow) for example.
- Wet Deposition = Precipitation $\times$ Concentration.
- Wet deposition is responsible for damage to lakes, forests, and streams.

Chemical Deposition

- Combustion of fossil fuel produces various chemicals including sulfate and nitrate gases.
- In the eastern U.S., most SO_2, and NO_x release attributed to power plants.
- Emitted to the air; wet deposition and dry deposition; interest in total deposition.
- Deposition means return to the earth's surface by means of precipitation (rain or snow) for example.
- Wet Deposition = Precipitation $\times$ Concentration.
- Wet deposition is responsible for damage to lakes, forests, and streams.

Chemical Deposition

- Combustion of fossil fuel produces various chemicals including sulfate and nitrate gases.
- In the eastern U.S., most SO_2, and NO_x release attributed to power plants.
- Emitted to the air; wet deposition and dry deposition; interest in total deposition.
- Deposition means return to the earth's surface by means of precipitation (rain or snow) for example.
- Wet Deposition = Precipitation $\times$ Concentration.
- Wet deposition is responsible for damage to lakes, forests, and streams.

The CMAQ model

- CMAQ = Community Multi-scale Air Quality Model.
- A computer simulation model which produces "averaged" output on 36km, 12km (used here), and now 4km grid cells.
- Uses variables such as power station emission volumes, meteorological data, land-use, etc. with atmospheric science (appropriate differential equations) to predict deposition levels. Not driven by monitoring station data.
- Predictions are biased but no missing data; monitoring data provide more accurate deposition but "missingness".

The CMAQ model

- CMAQ = Community Multi-scale Air Quality Model.
- A computer simulation model which produces "averaged" output on 36km, 12km (used here), and now 4km grid cells.
- Uses variables such as power station emission volumes, meteorological data, land-use, etc. with atmospheric science (appropriate differential equations) to predict deposition levels. Not driven by monitoring station data.
- Predictions are biased but no missing data; monitoring data provide more accurate deposition but "missingness".

The CMAQ model

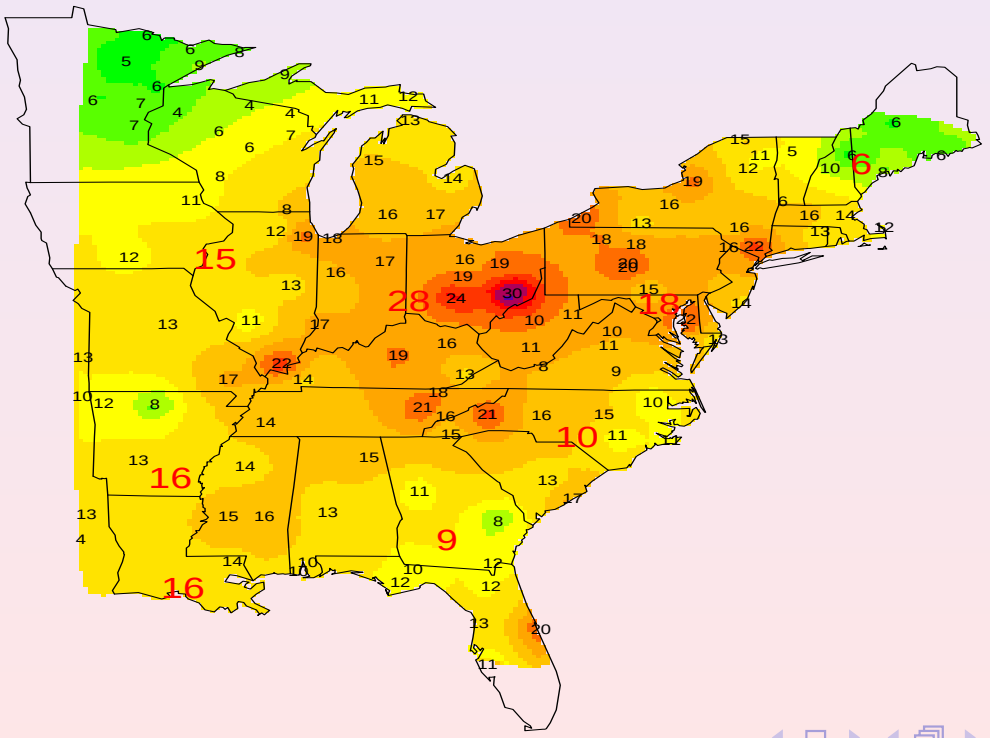
- CMAQ = Community Multi-scale Air Quality Model.
- A computer simulation model which produces "averaged" output on 36km, 12km (used here), and now 4km grid cells.
- Uses variables such as power station emission volumes, meteorological data, land-use, etc. with atmospheric science (appropriate differential equations) to predict deposition levels. Not driven by monitoring station data.
- Predictions are biased but no missing data; monitoring data provide more accurate deposition but "missingness".

The CMAQ model

- CMAQ = Community Multi-scale Air Quality Model.
- A computer simulation model which produces "averaged" output on 36km, 12km (used here), and now 4km grid cells.
- Uses variables such as power station emission volumes, meteorological data, land-use, etc. with atmospheric science (appropriate differential equations) to predict deposition levels. Not driven by monitoring station data.
- Predictions are biased but no missing data; monitoring data provide more accurate deposition but "missingness".

National Atmospheric Deposition Program (NADP)

- NADP collects point-referenced data at several sites.
- They then use simple interpolation to produce maps.



Inverse Distance weighting (IDW)

- "Poor person's" methodology:
- Value at a new site = weighted mean of observations,
- Weights inversely proportional distance^2.

Problems

- Not model based!
- Unable to accommodate known covariate - precipitation!
- Unable to handle 0's and missing data.
- Can't fuse with model output data.
- With dynamic data, can only do independent weekly or aggregated annually.

No associated uncertainty maps!

Inverse Distance weighting (IDW)

- "Poor person's" methodology:
- Value at a new site = weighted mean of observations,
- Weights inversely proportional distance^2.

Problems

- Not model based!
- Unable to accommodate known covariate - precipitation!
- Unable to handle 0's and missing data.
- Can't fuse with model output data.
- With dynamic data, can only do independent weekly or aggregated annually.

No associated uncertainty maps!

Inverse Distance weighting (IDW)

- "Poor person's" methodology:
- Value at a new site = weighted mean of observations,
- Weights inversely proportional distance^2.

Problems

- Not model based!
- Unable to accommodate known covariate - precipitation!
- Unable to handle 0's and missing data.
- Can't fuse with model output data.
- With dynamic data, can only do independent weekly or aggregated annually.

No associated uncertainty maps!

Inverse Distance weighting (IDW)

- "Poor person's" methodology:
- Value at a new site = weighted mean of observations,
- Weights inversely proportional distance^2.

Problems

- Not model based!
- Unable to accommodate known covariate - precipitation!
- Unable to handle 0's and missing data.
- Can't fuse with model output data.
- With dynamic data, can only do independent weekly or aggregated annually.

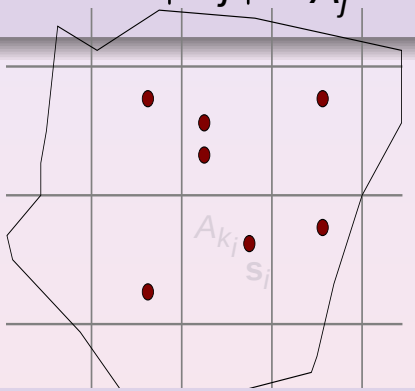
No associated uncertainty maps!

Change of support problem

Fuentes and Raftery, 2005

- Need to upscale (block-average) point level $Z(\mathbf{s}, t)$ to obtain grid level $Z(A_j, t)$.

$$Z(A_j, t) = \frac{1}{|A_j|} \int_{A_j} Z(\mathbf{s}, t)\, d\mathbf{s}, \tag{1}$$



Many more A's than s's

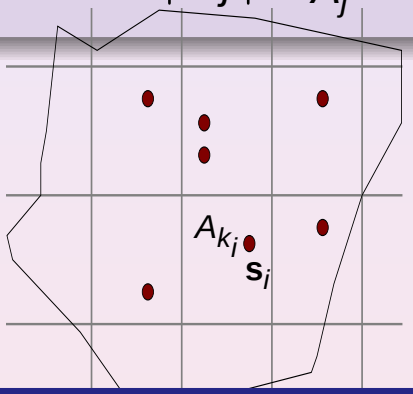
- We use Measurement Error model (MEM) at point level centred around grid level values.
- Make inference at the point level by downscaling.
- Huge computational advantages, **NO** integration (1).

Change of support problem

Fuentes and Raftery, 2005

- Need to upscale (block-average) point level $Z(\mathbf{s}, t)$ to obtain grid level $Z(A_j, t)$.

$$Z(A_j, t) = \frac{1}{|A_j|} \int_{A_j} Z(\mathbf{s}, t)\, d\mathbf{s}, \tag{1}$$



Many more A's than s's

- We use Measurement Error model (MEM) at point level centred around grid level values.
- Make inference at the point level by downscaling.
- Huge computational advantages, **NO** integration (1).

Our Contribution

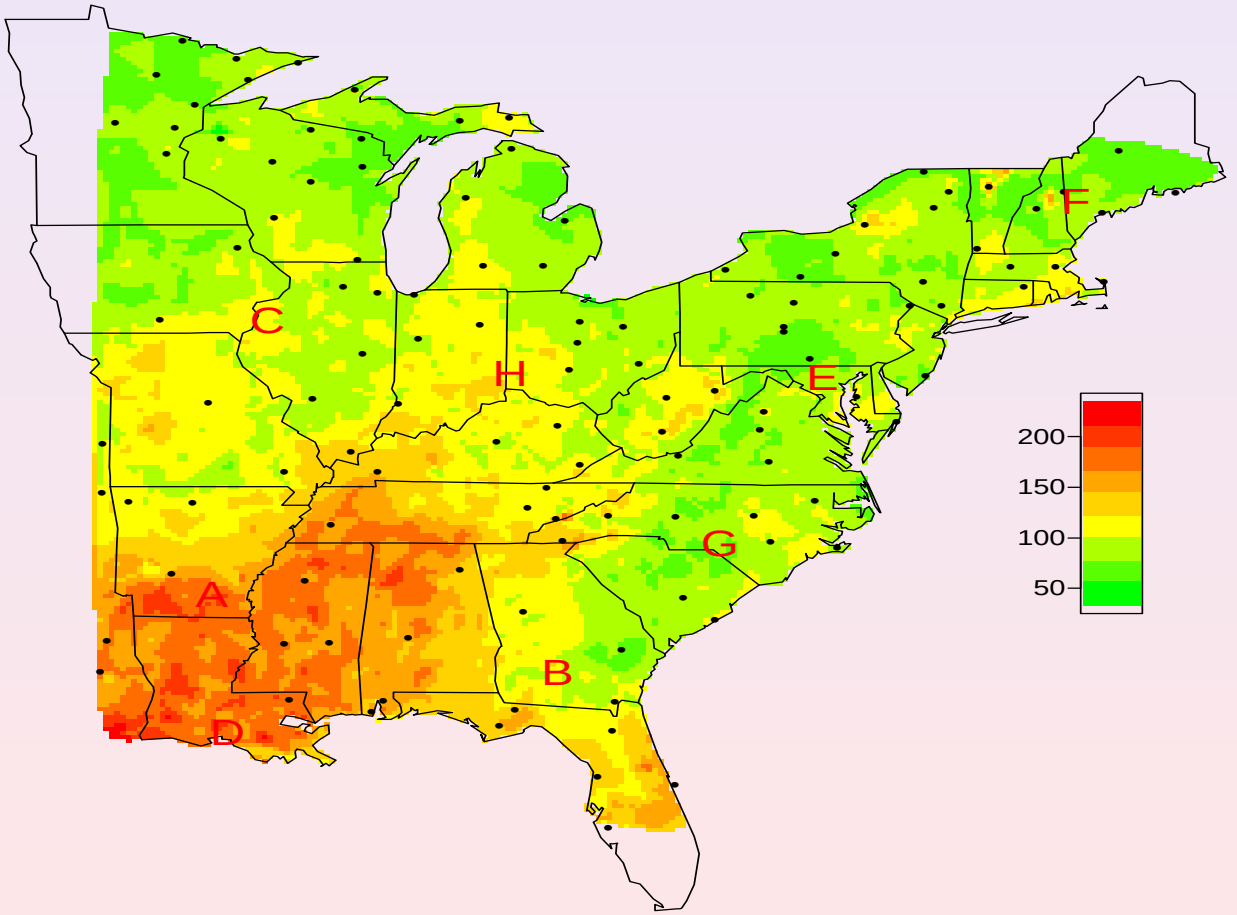
- Model data on weekly time intervals.
- Fuse with gridded weekly CMAQ output.
- Use weekly precipitation information, available from other monitoring networks.
- Interpolate in space, predict in time
- Obtain quarterly and annual maps.
- Reveal spatial pattern in deposition.

Our data set

Data

- Weekly deposition data from 128 sites in the eastern U.S. for the year 2001.
- Use 120 sites to estimate, remaining 8 to validate.
- Weekly CMAQ output from $J = 33,390$ grid cells (about 1.8 million values!)
- Weekly precipitation data from 2827 predictive sites.

Annual Precipitation



Exploratory Analyses

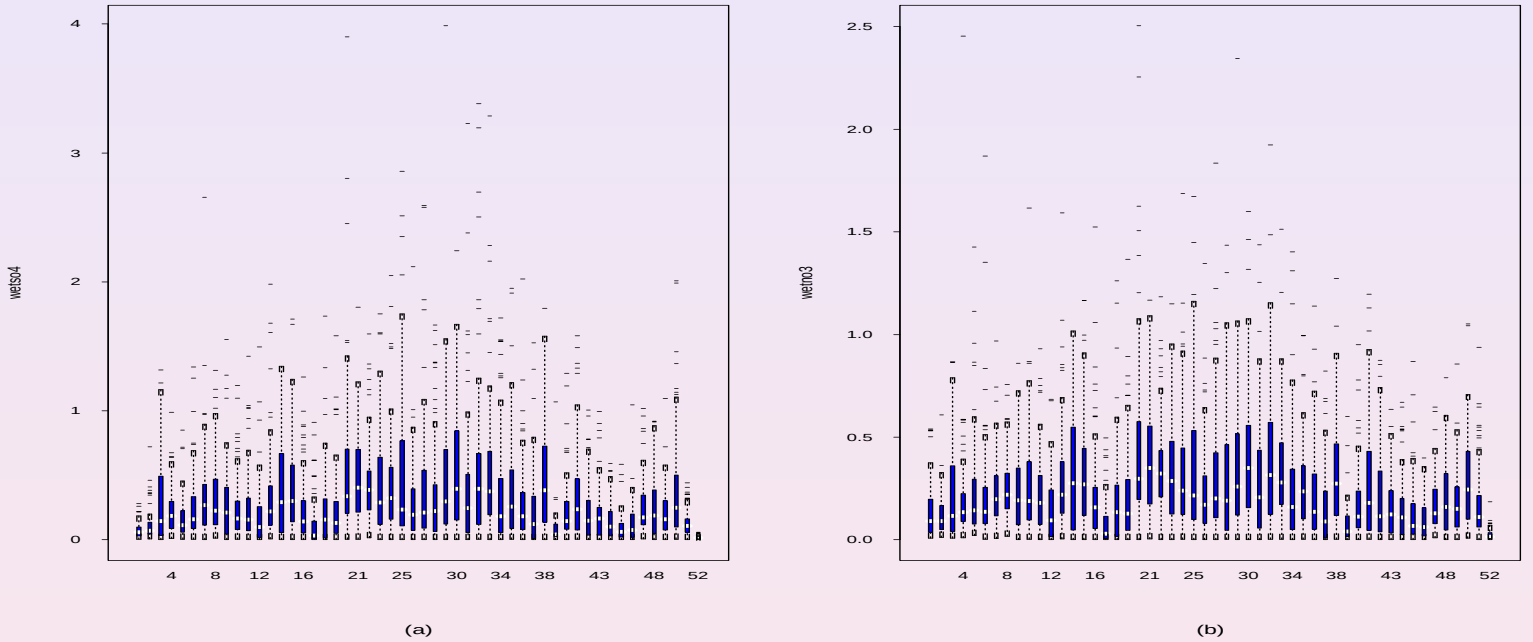


Figure: Boxplot of weekly depositions: (a) sulfate and (b) nitrate.

Exploratory Analyses ...

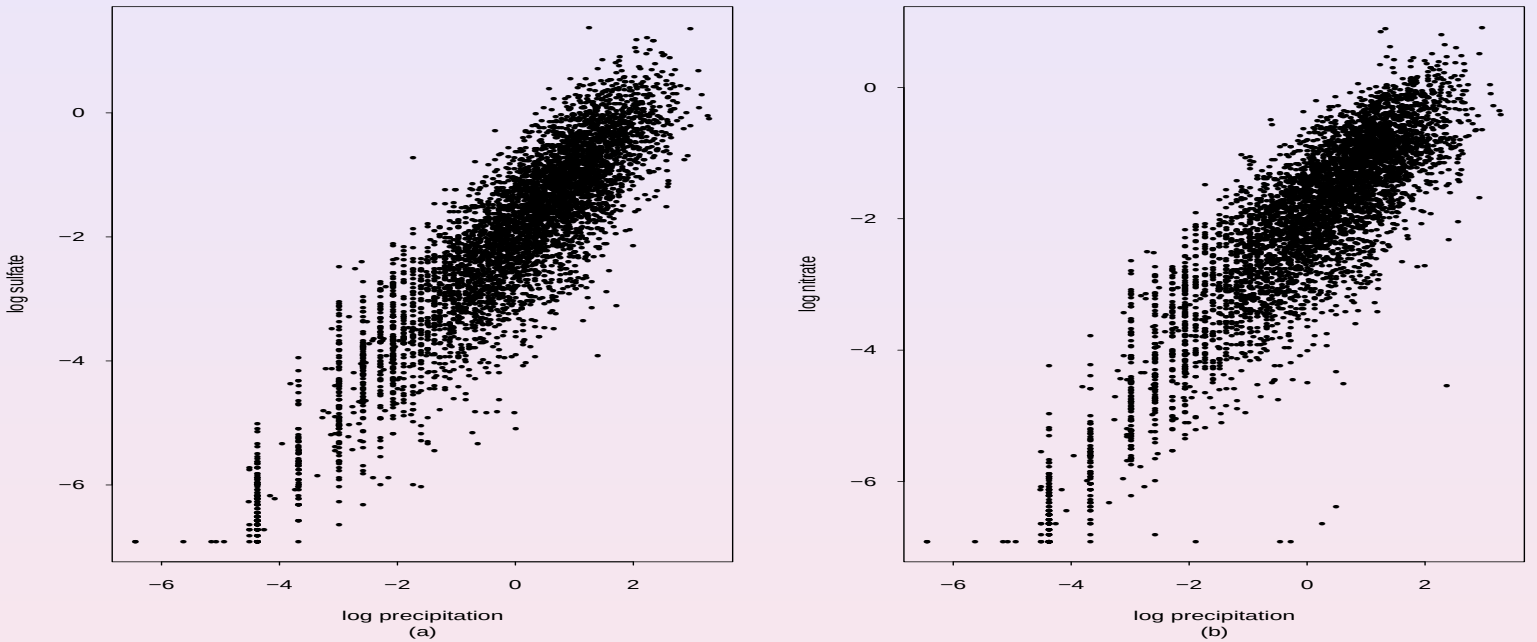


Figure: Deposition against precipitation (both on the log scale): (a) sulfate and (b) nitrate.

Exploratory Analyses ...

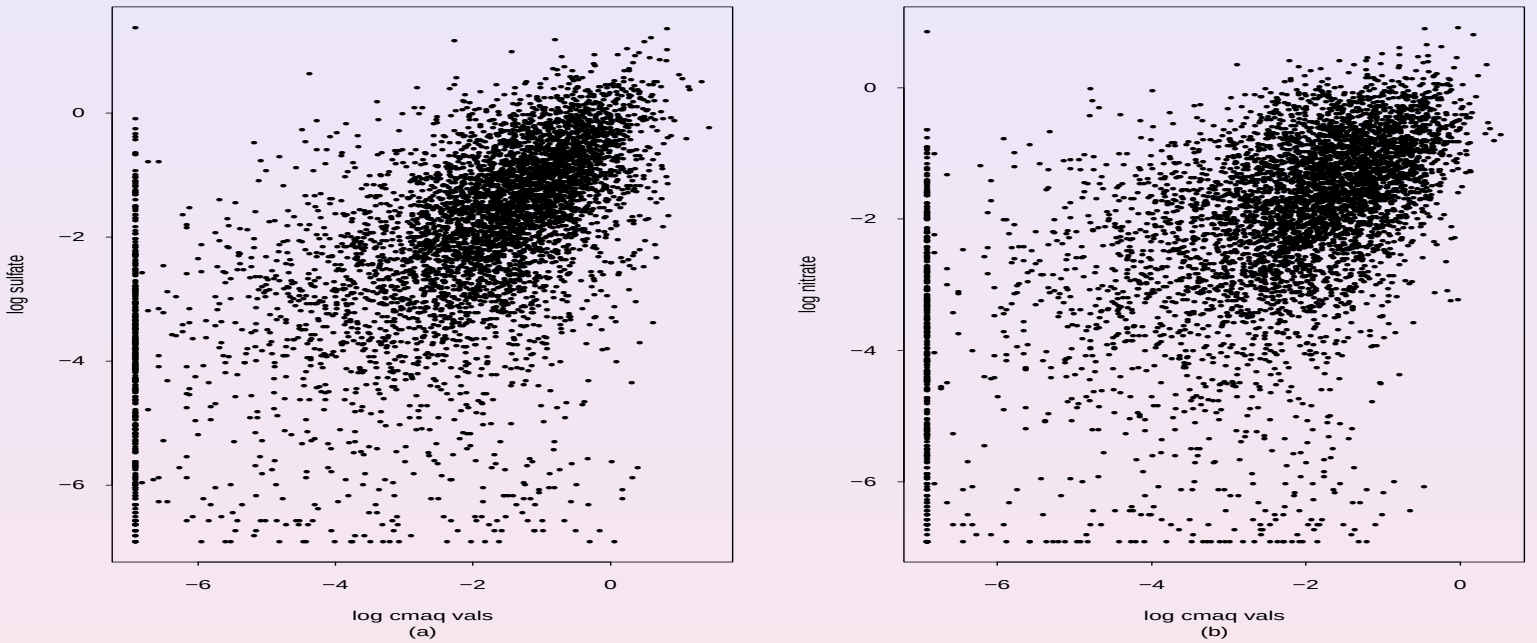


Figure: Deposition at the NADP sites against the CMAQ values in the grid cell covering the corresponding NADP site on the log scale: (a) sulfate and (b) nitrate.

Modelling Requirements

- No deposition, $Z(\mathbf{s}_i, t)$, without precipitation, $P(\mathbf{s}_i, t)$;
- enforced by a latent atmospheric space-time process $V(\mathbf{s}_i, t)$ below, $i = 1, \ldots, n = 120$, and for each week t, $t = 1, \ldots, 52$.
- Similarly, model CMAQ output, $Q(A_j, t)$ for each grid cell A_j, $j = 1, \ldots, J = 33,390$ and for each week t,
- enforced using a latent atmospheric areal process $\tilde{V}(A_j, t)$.
- Model everything on the log-scale;
- latent processes take care of point masses at zero. Avoid **log(**0**)** problems.

First stage models

Precipitation model

$$P(\mathbf{s}_i, t) = \begin{cases} \exp\left(U(\mathbf{s}_i, t)\right) & \text{if } V(\mathbf{s}_i, t) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

Deposition model

$$Z(\mathbf{s}_i, t) = \begin{cases} \exp\left(Y(\mathbf{s}_i, t)\right) & \text{if } V(\mathbf{s}_i, t) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Model for CMAQ output

$$Q(A_j, t) = \begin{cases} \exp\left(X(A_j, t)\right) & \text{if } \tilde{V}(A_j, t) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Clarification

- P's, Z's, and Q's are the observed precipitation, NADP deposition, and CMAQ deposition, respectively.
- $V(\mathbf{s}, t)$ is a conceptual point level latent atmospheric process which drives $P(\mathbf{s}, t)$ and $Z(\mathbf{s}, t)$.
- $P(\mathbf{s}, t)$ and $Z(\mathbf{s}, t) = 0$ if $V(\mathbf{s}, t) \leq 0$.
- $U(\mathbf{s}, t)$ and $Y(\mathbf{s}, t)$ are log precipitation and deposition, respectively.
- Models below will specify their values when $V(\mathbf{s}, t) \leq 0$ or if $P(\mathbf{s}, t)$ or $Z(\mathbf{s}, t)$ are missing.
- $\tilde{V}(A_j, t)$ is a conceptual areal level latent atmospheric process which drives $Q(A_j, t)$.
- $X(A_j, t)$ is log CMAQ output where modelling below will specify its values when $\tilde{V}(A_j, t) \leq 0$.

The first stage likelihood

$$f(\mathbf{P}, \mathbf{Z}, \mathbf{Q}|\mathbf{U}, \mathbf{Y}, \mathbf{X}, \mathbf{V}, \tilde{\mathbf{V}}) = f(\mathbf{P}|\mathbf{U}, \mathbf{V}) \times f(\mathbf{Z}|\mathbf{Y}, \mathbf{V}) \times f(\mathbf{Q}|\mathbf{X}, \tilde{\mathbf{V}})$$

which takes the form

$$\prod_{t=1}^{T} \left[\prod_{i=1}^{n} \left\{ 1_{\exp\left(u(\mathbf{s}_i, t)\right)}^{1_{\exp\left(y(\mathbf{s}_i, t)\right)}^{I\left(v(\mathbf{s}_i, t) > 0\right)}} \right\} \prod_{j=1}^{J} \left\{ 1_{\exp\left(x(A_j, t)\right)}^{I\left(\tilde{v}(A_j, t) > 0\right)} \right\} \right]$$

where 1_x denotes a degenerate distribution with point mass at x and $I(\cdot)$ is the indicator function.

Deposition Model

$$\begin{aligned} Y(\mathbf{s}_i, t) &= \beta_0 + \beta_1 U(\mathbf{s}_i, t) + \beta_2 V(\mathbf{s}_i, t) \\ &\quad + (b_0 + b(\mathbf{s}_i)) X(A_{k_i}, t) \\ &\quad + \eta(\mathbf{s}_i, t) + \epsilon(\mathbf{s}_i, t). \end{aligned}$$

- Spatially varying coefficients, $\mathbf{b} = (b(\mathbf{s}_1), \ldots, b(\mathbf{s}_n))'$ is a Gaussian process (GP).
- Spatio-temporal intercept $\boldsymbol{\eta}_t = (\eta(\mathbf{s}_1, t), \ldots, \eta(\mathbf{s}_n, t))'$ is a GP independent in time.
- Allow for spatially varying calibration of CMAQ. Could imagine common $\eta(\mathbf{s}_i)$.
- $\epsilon(\mathbf{s}_i, t) \sim N(0, \sigma^2_\epsilon)$, providing the nugget effect.

The second stage models ...

Precipitation

$$U(\mathbf{s}_i, t) = \alpha_0 + \alpha_1 V(\mathbf{s}_i, t) + \delta(\mathbf{s}_i, t),$$

where $\boldsymbol{\delta}_t = (\delta(\mathbf{s}_1, t), \ldots, \delta(\mathbf{s}_n, t))'$ is a GP independent in time.

CMAQ output

$$X(A_j, t) = \gamma_0 + \gamma_1 \tilde{V}(A_j, t) + \psi(A_j, t), \quad j = 1, \ldots, J.$$

where $\psi(A_j, t) \sim N(0, \sigma^2_\psi)$, independently.

Specification of latent processes

Measurement Error Model (MEM)

$$V(\mathbf{s}_i, t) \sim N(\tilde{V}(A_{k_i}, t), \sigma_v^2),\ i = 1, \ldots, n, t = 1, \ldots, T.$$

The process $\tilde{V}(A_j, t)$ is AR in time and CAR in space

$$\tilde{V}(A_j, t) = \rho \tilde{V}(A_j, t-1) + \zeta(A_j, t),$$

$$\zeta(A_j, t) \sim N\left(\sum_{i=1}^{J} h_{ji}\zeta(A_i, t), \frac{\sigma_\zeta^2}{m_j}\right),$$

Let ∂_j define the m_j neighboring grid cells of the cell A_j.

$$h_{ji} = \begin{cases} \frac{1}{m_j} & \text{if } i \in \partial_j \\ 0 & \text{otherwise.} \end{cases}$$

Clarification

- Note that we can have $Z > 0$, $Q = 0$ and vice versa. Therefore V and $\tilde{V}$ can have opposite signs.
- This arises because we are modelling at two different scales - need processes at two different scales.
- We can view $V(\mathbf{s}, t) - \tilde{V}(A, t)$ as a deviation from the areal average. We assume these realised deviations are independent across space and time.
- We have a conditional model for V and X given $\tilde{V}$. The resulting marginal model for U and Y given $\tilde{V}$ is *multi-scale* - additive random effects at two scales.

Completing the second stage specification

Initial condition: $\tilde{V}(A_j, 0) = \frac{1}{T}\sum_{t=1}^{T} X(A_j, t)$.
Let D be diagonal with entries σ_ζ^2/m_j. $f(\tilde{\mathbf{V}}_t|\tilde{\mathbf{V}}_{t-1}, \rho, \sigma_\zeta^2) \propto$

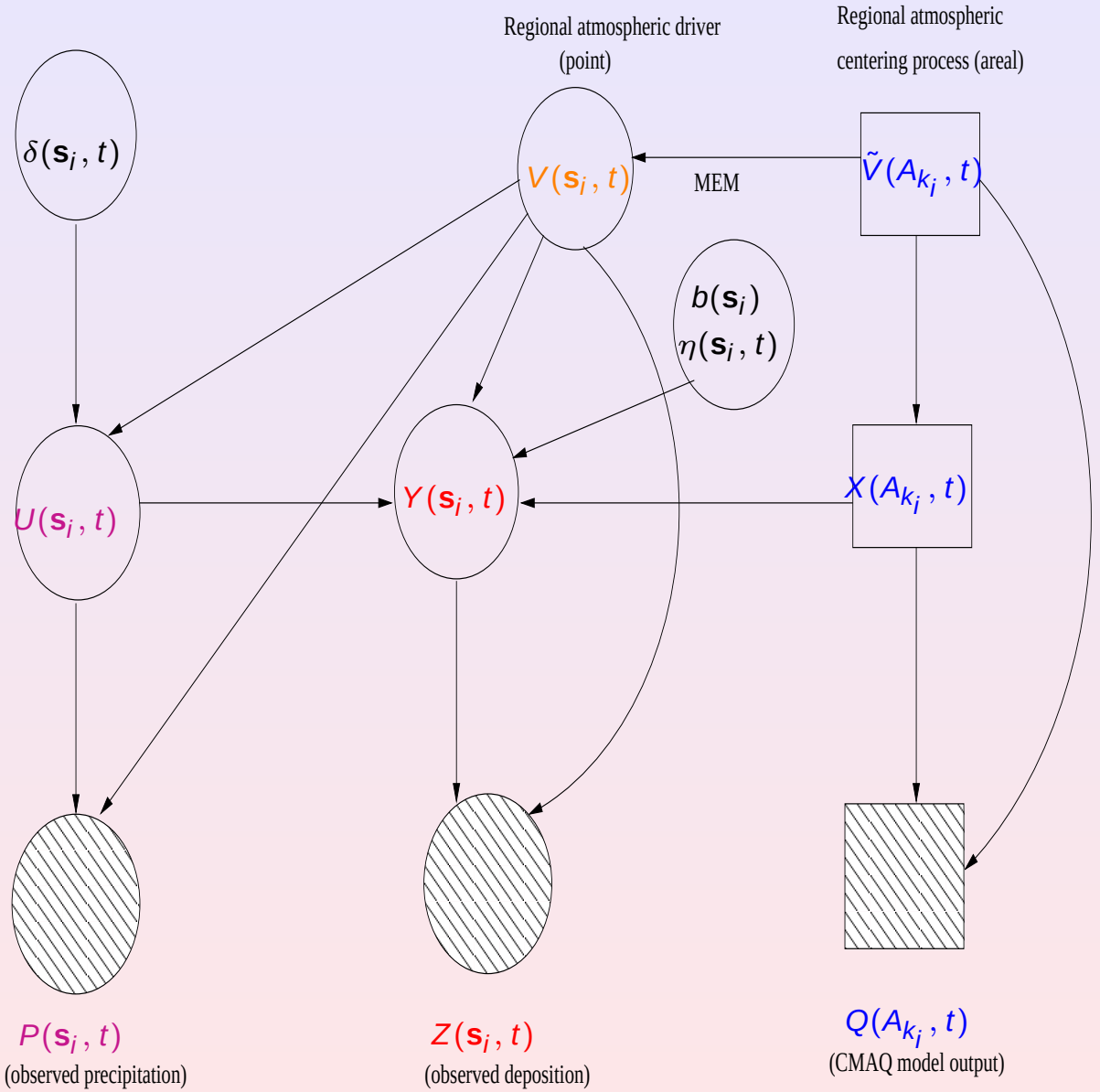
$$\exp\left\{-\frac{1}{2}\left(\tilde{\mathbf{V}}_t - \rho\tilde{\mathbf{V}}_{t-1}\right)' D^{-1}(I-H)\left(\tilde{\mathbf{V}}_t - \rho\tilde{\mathbf{V}}_{t-1}\right)\right\}.$$

The second stage specification is given by:

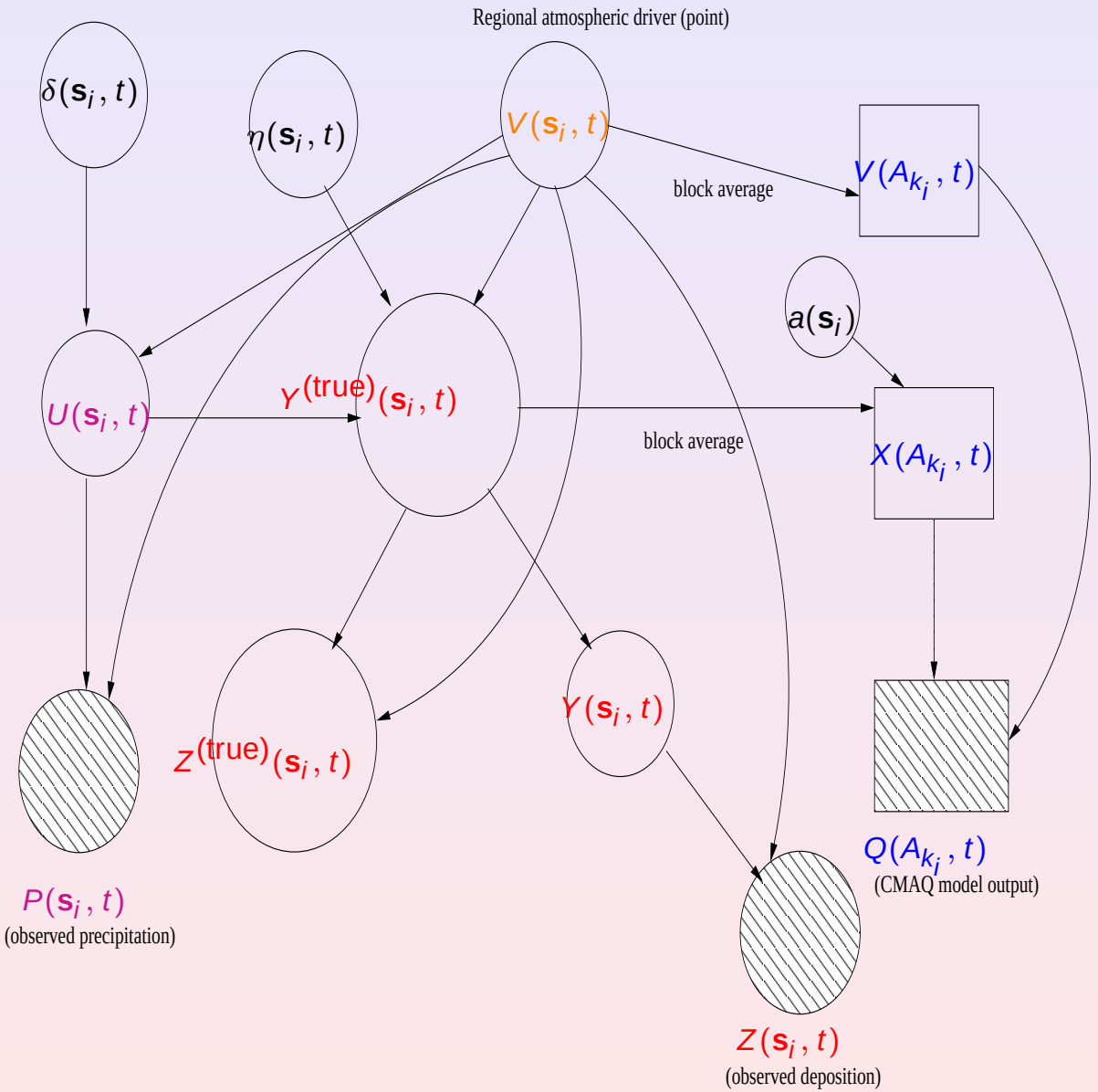
$$\prod_{t=1}^{T}\Big[f(\mathbf{Y}_t|\mathbf{U}_t, \mathbf{V}_t, \mathbf{X}_t, \eta_t, \mathbf{b}, \theta) \times f(\eta_t|\theta)\, f(\mathbf{U}_t|\mathbf{V}_t, \theta) \times f(\mathbf{V}_t|\tilde{\mathbf{V}}_t^{(1)}, \theta) \times f(\mathbf{X}_t|\tilde{\mathbf{V}}_t, \theta) f(\tilde{\mathbf{V}}_t|\tilde{\mathbf{V}}_{t-1}, \theta)\Big]\, f(\mathbf{b}|\theta).$$

$\boldsymbol{\theta} =$
$(\alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2, b_0, \gamma_0, \gamma_1, \rho, \sigma_\delta^2, \sigma_b^2, \sigma_\eta^2, \sigma_\epsilon^2, \sigma_\psi^2, \sigma_v^2, \sigma_\zeta^2)$.

Graphical representation of the model.



Graphical representation of a fusion model.



Predictions

At a new site $\mathbf{s}'$ and time t' we need $Z(\mathbf{s}', t')$ which depends on $Y(\mathbf{s}', t')$. If $P(\mathbf{s}', t') = 0$ then $Z(\mathbf{s}', t') = 0$.

Suppose otherwise.

- Bayesian predictive distributions:

$$\pi(z_{\text{pred}}|z_{\text{obs}}) = \int \pi(z_{\text{pred}}|\text{par})\,\pi(\text{par}|z_{\text{obs}})d\,\text{par}.$$

- Need to simulate $Y(\mathbf{s}', t')$.
- $V(\mathbf{s}', t') \sim N(\tilde{V}(A', t'), \sigma_v^2)$.
- $U(\mathbf{s}', t'), \eta(\mathbf{s}', t')$ and $b(\mathbf{s}')$ are simulated from the conditional distributions at $\mathbf{s}'$ given $\mathbf{s}_1, \ldots, \mathbf{s}_n$
- $X(A', t') = \log Q(A'', t')$ if $Q(A', t') > 0$, otherwise updated in the MCMC.
- More details in the paper.

Predictions

At a new site $\mathbf{s}'$ and time t' we need $Z(\mathbf{s}', t')$ which depends on $Y(\mathbf{s}', t')$. If $P(\mathbf{s}', t') = 0$ then $Z(\mathbf{s}', t') = 0$.

Suppose otherwise.

- Bayesian predictive distributions:

$$\pi(z_{\text{pred}}|z_{\text{obs}}) = \int \pi(z_{\text{pred}}|\text{par})\,\pi(\text{par}|z_{\text{obs}})d\text{par}.$$

- Need to simulate $Y(\mathbf{s}', t')$.
- $V(\mathbf{s}', t') \sim N(\tilde{V}(A', t'), \sigma_v^2)$.
- $U(\mathbf{s}', t'), \eta(\mathbf{s}', t')$ and $b(\mathbf{s}')$ are simulated from the conditional distributions at $\mathbf{s}'$ given $\mathbf{s}_1, \ldots, \mathbf{s}_n$
- $X(A', t') = \log Q(A'', t')$ if $Q(A', t') > 0$, otherwise updated in the MCMC.
- More details in the paper.

Predictions

At a new site $\mathbf{s}'$ and time t' we need $Z(\mathbf{s}', t')$ which depends on $Y(\mathbf{s}', t')$. If $P(\mathbf{s}', t') = 0$ then $Z(\mathbf{s}', t') = 0$.

Suppose otherwise.

- Bayesian predictive distributions:

$$\pi(z_{\text{pred}}|z_{\text{obs}}) = \int \pi(z_{\text{pred}}|\text{par})\,\pi(\text{par}|z_{\text{obs}})d\text{par}.$$

- Need to simulate $Y(\mathbf{s}', t')$.
- $V(\mathbf{s}', t') \sim N(\tilde{V}(A', t'), \sigma_v^2)$.
- $U(\mathbf{s}', t'), \eta(\mathbf{s}', t')$ and $b(\mathbf{s}')$ are simulated from the conditional distributions at $\mathbf{s}'$ given $\mathbf{s}_1, \ldots, \mathbf{s}_n$
- $X(A', t') = \log Q(A'', t')$ if $Q(A', t') > 0$, otherwise updated in the MCMC.
- More details in the paper.

Choosing ϕ_η, ϕ_δ and ϕ_b

Estimation is challenging due to weak identifiability of variances and ranges. So, we fix the ϕ's (spatial decay parameters).

Use validation mean-square error

$$\text{VMSE} = \frac{1}{n_v} \sum_{i=1}^{8} \sum_{t=1}^{52} \left(Z(\mathbf{s}_i^*, t) - \hat{Z}(\mathbf{s}_i^*, t) \right)^2 I(\text{obs})$$

where $I(\text{obs}) = 1$ if $Z(\mathbf{s}_i^*, t)$ has been observed and 0 otherwise, and $n_v = \sum_{i=1}^{8} \sum_{t=1}^{52} I(\text{obs})$. Here $n_v = 407$.

Optimal ranges were 500, 1000 and 500 kilometers, respectively for the exponential covariance function. VMSE is not sensitive near these values. May be we can use importance sampling to compute VMSE for different ϕ's.

Choosing ϕ_η, ϕ_δ and ϕ_b

Estimation is challenging due to weak identifiability of variances and ranges. So, we fix the ϕ's (spatial decay parameters).

Use validation mean-square error

$$\text{VMSE} = \frac{1}{n_v} \sum_{i=1}^{8} \sum_{t=1}^{52} \left(Z(\mathbf{s}_i^*, t) - \hat{Z}(\mathbf{s}_i^*, t) \right)^2 I(\text{obs})$$

where $I(\text{obs}) = 1$ if $Z(\mathbf{s}_i^*, t)$ has been observed and 0 otherwise, and $n_v = \sum_{i=1}^{8} \sum_{t=1}^{52} I(\text{obs})$. Here $n_v = 407$.

Optimal ranges were 500, 1000 and 500 kilometers, respectively for the exponential covariance function. VMSE is not sensitive near these values. May be we can use importance sampling to compute VMSE for different ϕ's.

Validation

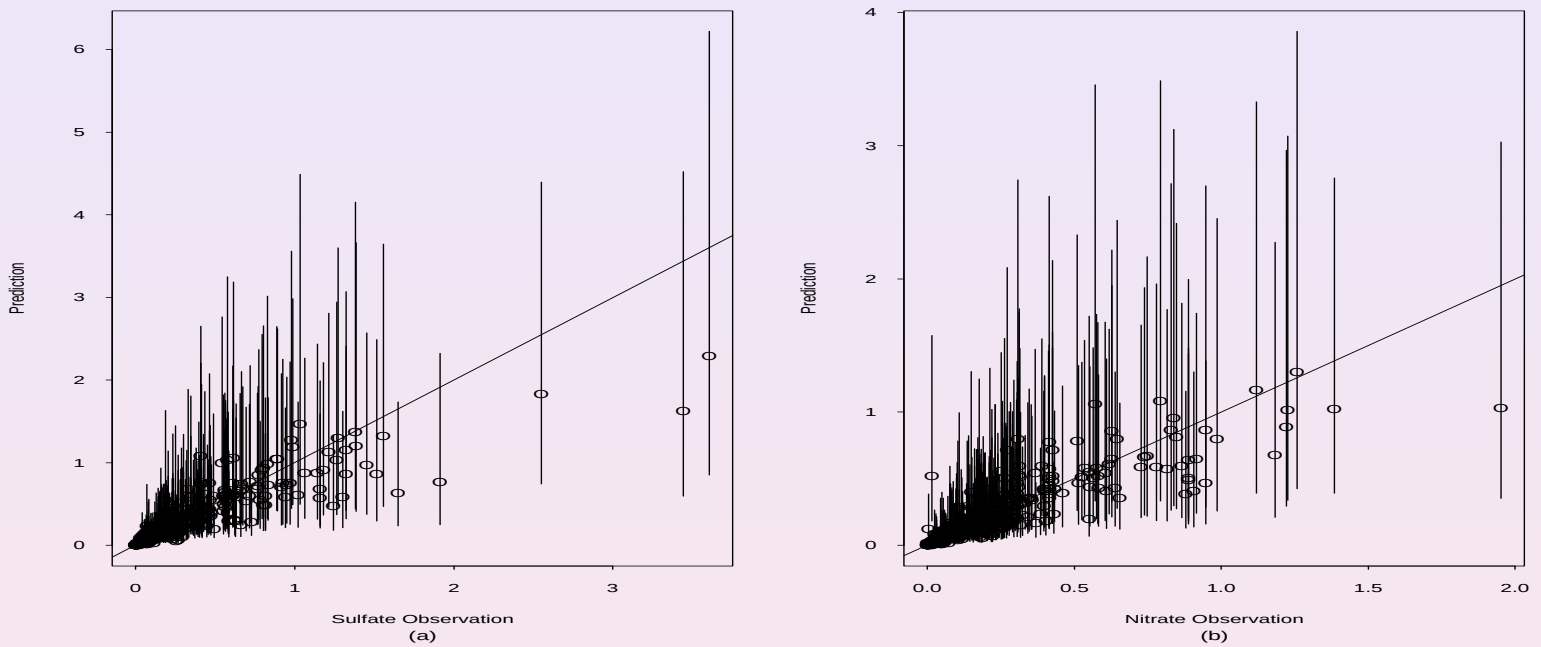
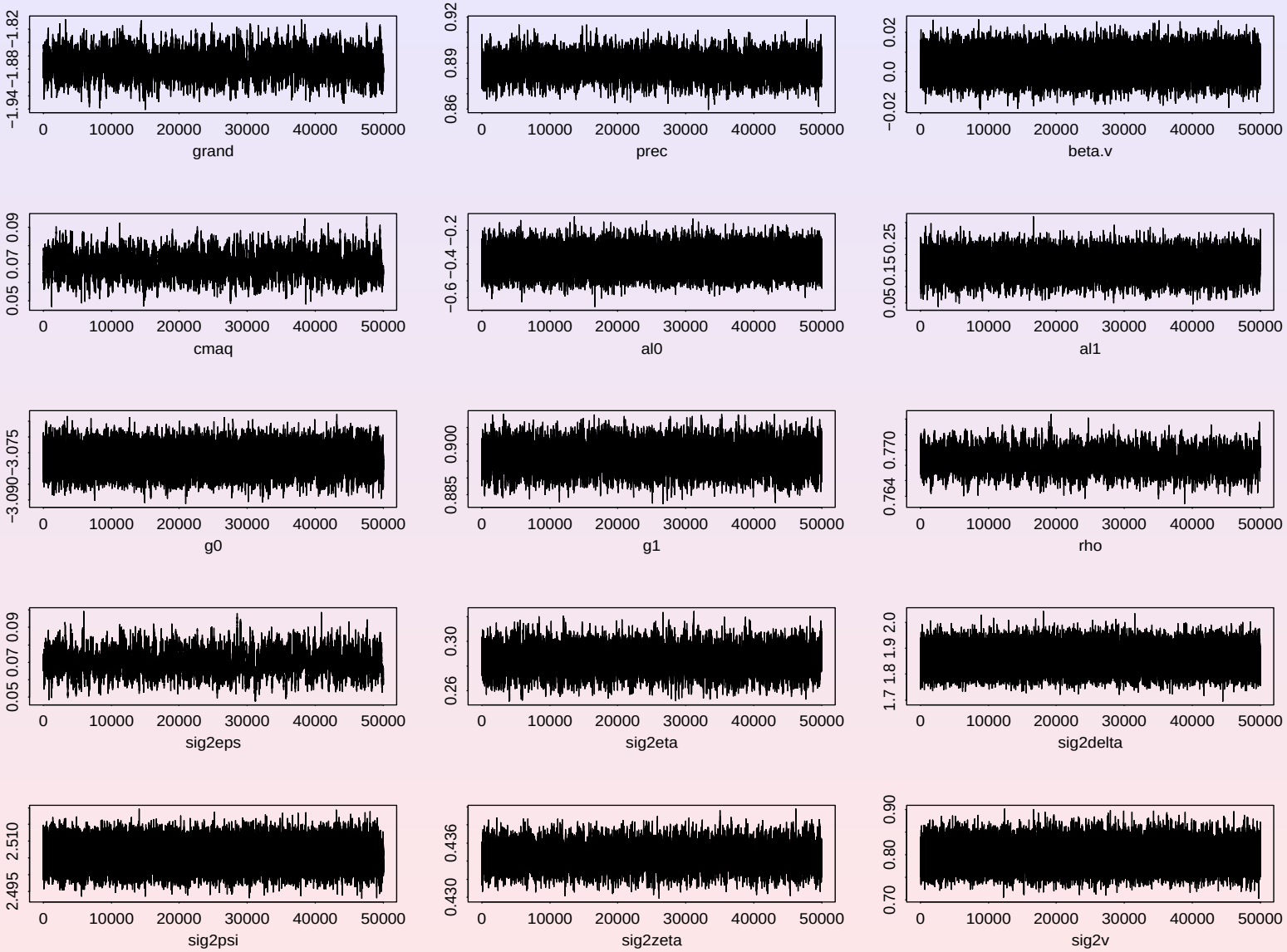
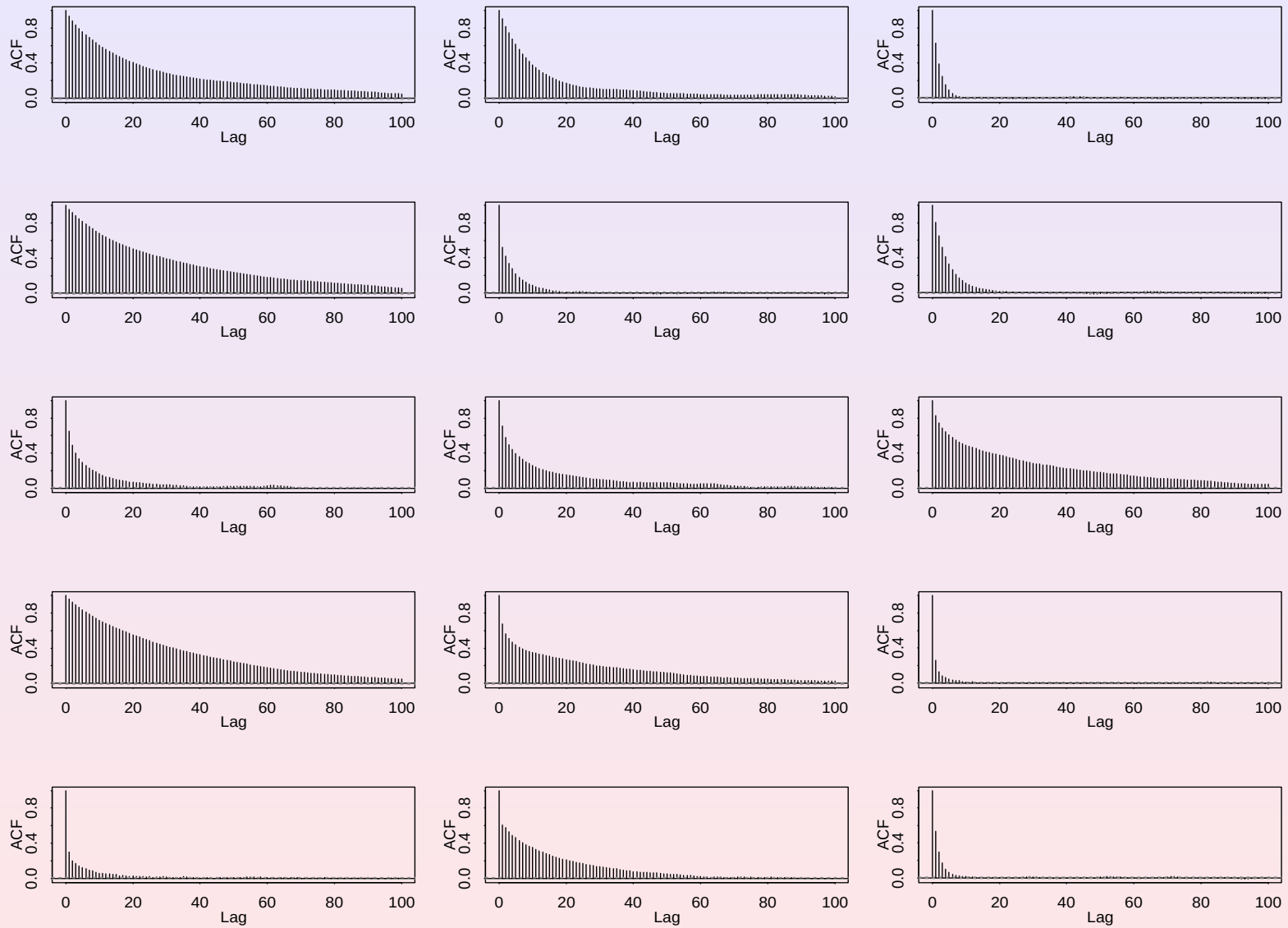


Figure: Validation versus the observed values at the 8 reserved sites. Validation prediction intervals are plotted as vertical lines. (a) sulfate and (b) nitrate.

MCMC Diagnostics



MCMC Diagnostics ...



Spatially varying slopes?

- Do we need the spatially varying $b(\mathbf{s})$'s?
- Only a few of the $b(\mathbf{s}_i)$ are significant; they are small relative to their standard errors.
- Importance of precipitation and the spatially varying intercept makes it difficult to find spatially varying contribution of CMAQ.
- Fusion approaches also have not found spatially varying intercepts.
- Still can see space-time bias in CMAQ by comparing model predictions with CMAQ output.

Parameter Estimates

	Sulfate			Nitrate		
	mean	sd	95%CI	mean	sd	95%CI
α_0	–0.4497	0.0871	(–0.6189, –0.2733)	–0.3548	0.0596	(–0.4695, -0.2369)
α_1	0.1787	0.0379	(0.1017, 0.2499)	0.1522	0.0336	(0.0843, 0.2161)
β_0	–1.9414	0.0196	(–1.9784, –1.9012)	–1.9976	0.0192	(–2.0344, –1.9605)
β_1	0.9103	0.0067	(0.8972, 0.9240)	0.8412	0.0070	(0.8274, 0.8553)
β_2	0.0029	0.0062	(–0.0091, 0.0151)	0.0040	0.0060	(–0.0078, 0.0159)
b_0	0.0490	0.0053	(0.0386, 0.0599)	0.0535	0.0062	(0.0409, 0.0652)
γ_0	–3.0768	0.0035	(–3.0836, -3.0700)	–3.2177	0.0033	(–3.2242, –3.2112)
γ_1	0.8957	0.0034	(0.8891, 0.9025)	0.7368	0.0033	(0.7303, 0.7433)
ρ	0.7688	0.0012	(0.7664, 0.7712)	0.7492	0.0013	(0.7468, 0.7517)
σ^2_δ	2.6438	0.0602	(2.5254, 2.7631)	1.8694	0.0387	(1.7942, 1.9476)
σ^2_η	0.2812	0.0101	(0.2616, 0.3010)	0.3354	0.0105	(0.3149, 0.3564)
σ^2_ϵ	0.0718	0.0057	(0.0607, 0.0832)	0.0727	0.0074	(0.0588, 0.0878)
σ^2_ψ	2.5062	0.0033	(2.4997, 2.5127)	2.2148	0.0028	(2.2092, 2.2203)
σ^2_ν	0.8087	0.0259	(0.7601, 0.8620)	0.7821	0.0237	(0.7366, 0.8290)
σ^2_ζ	0.4345	0.0011	(0.4322, 0.4367)	0.4340	0.0012	(0.4316, 0.4363)

Annual Sulfate Deposition

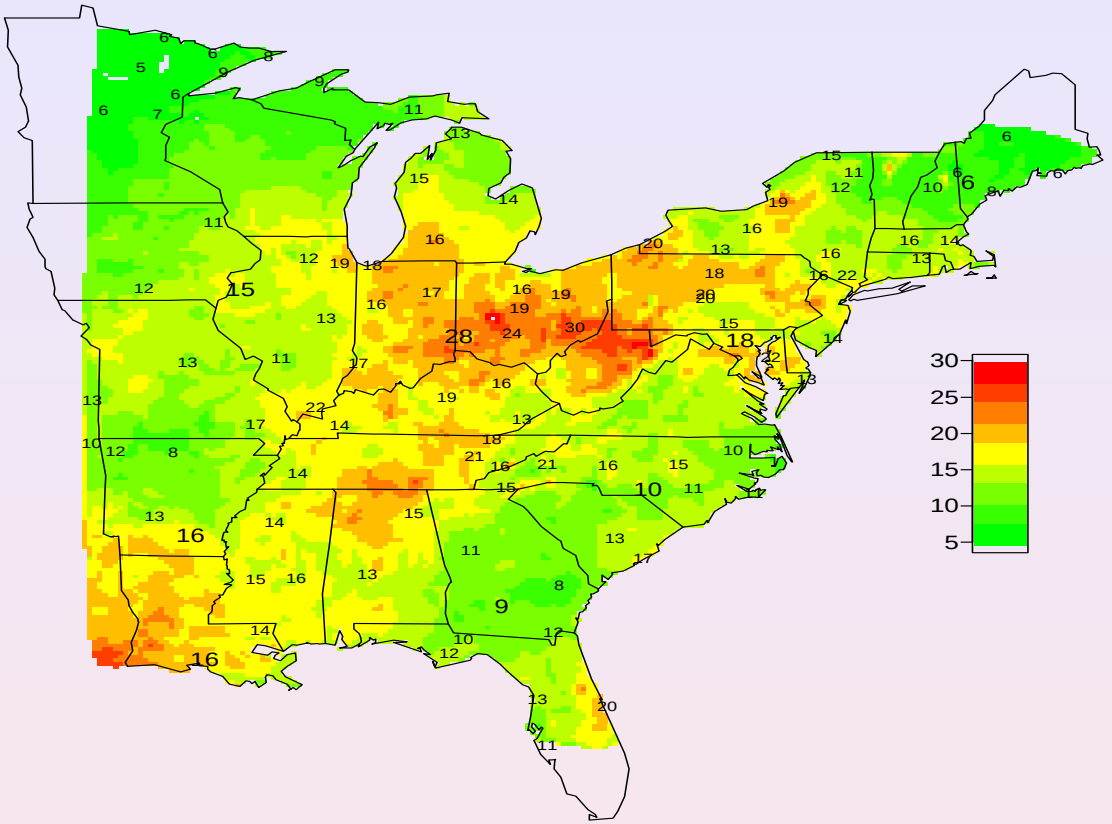
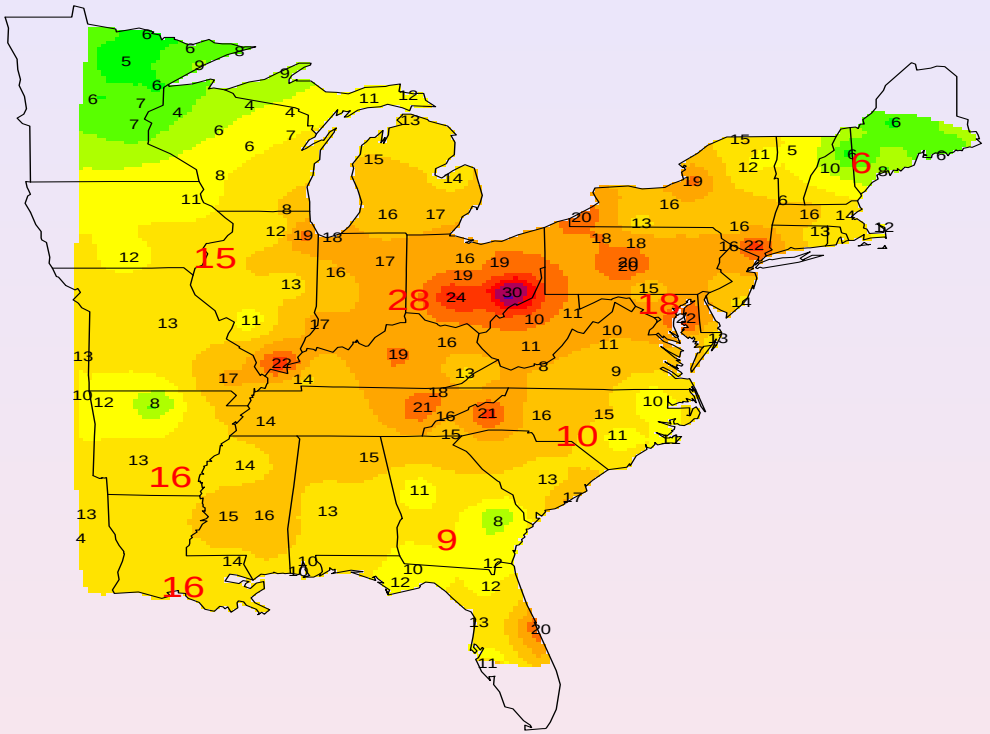


Figure: Model predicted map of annual sulfate deposition in 2001. The observed annual totals are labeled; a larger font size is used for the validation sites.

Compared with IDW map



Annual Nitrate Deposition

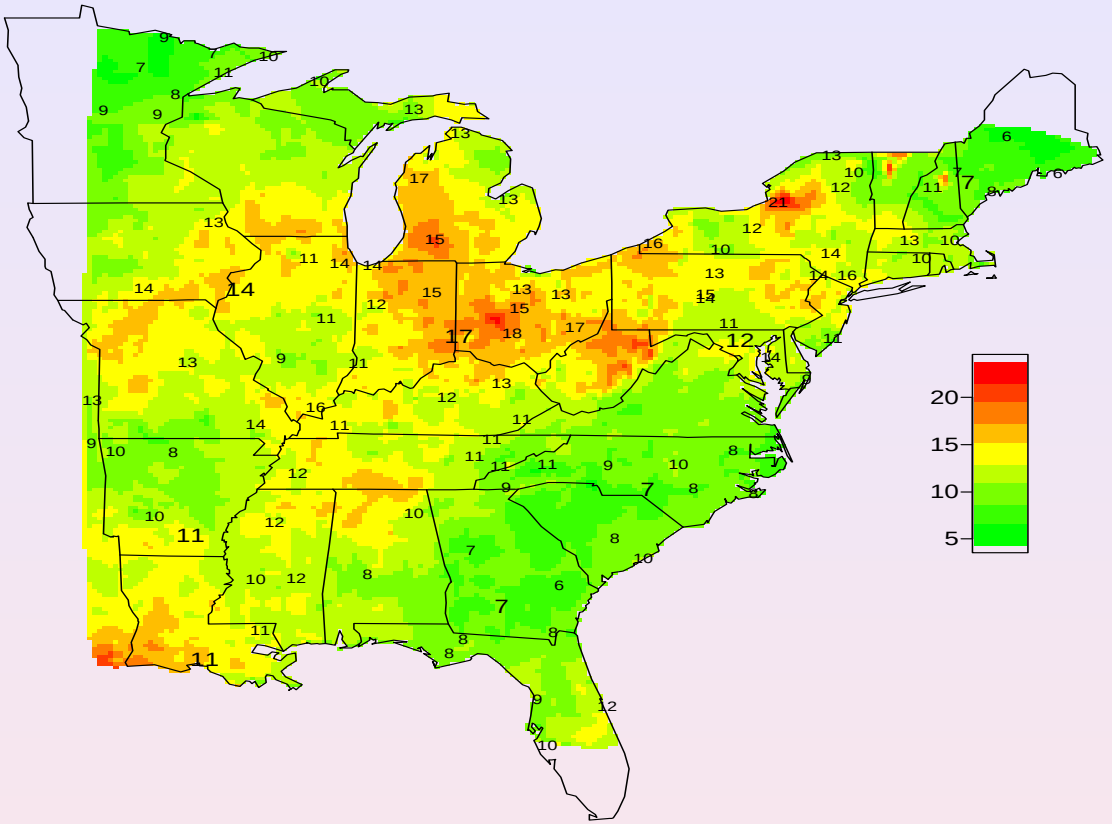


Figure: Model predicted map of annual nitrate deposition in 2001. The observed annual totals are labeled; a larger font size is used for the validation sites.

Map of the length of 95% prediction intervals

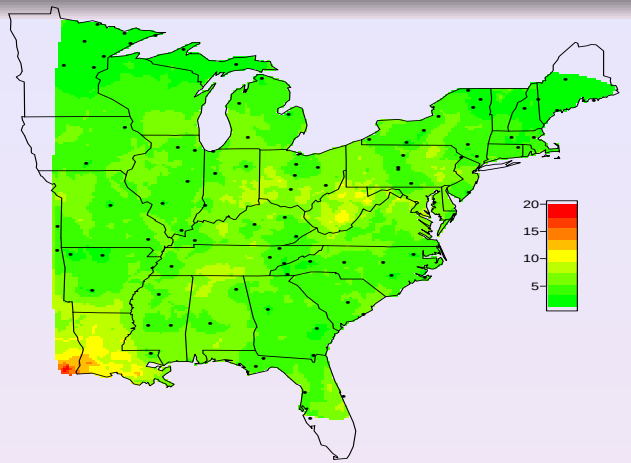


Figure: Uncertainty map of annual sulfate deposition.

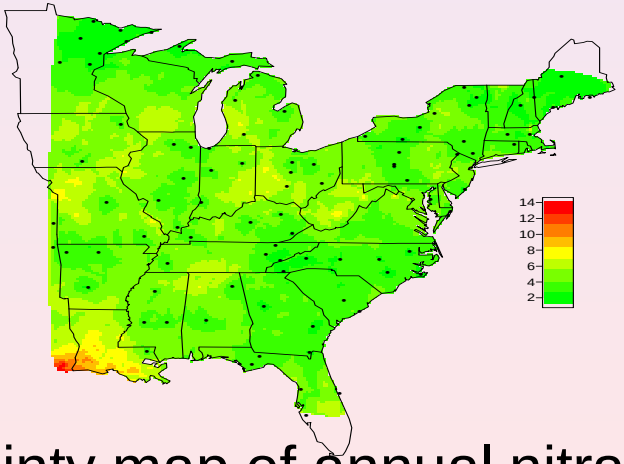


Figure: Uncertainty map of annual nitrate deposition.

Quarterly Sulfate Deposition

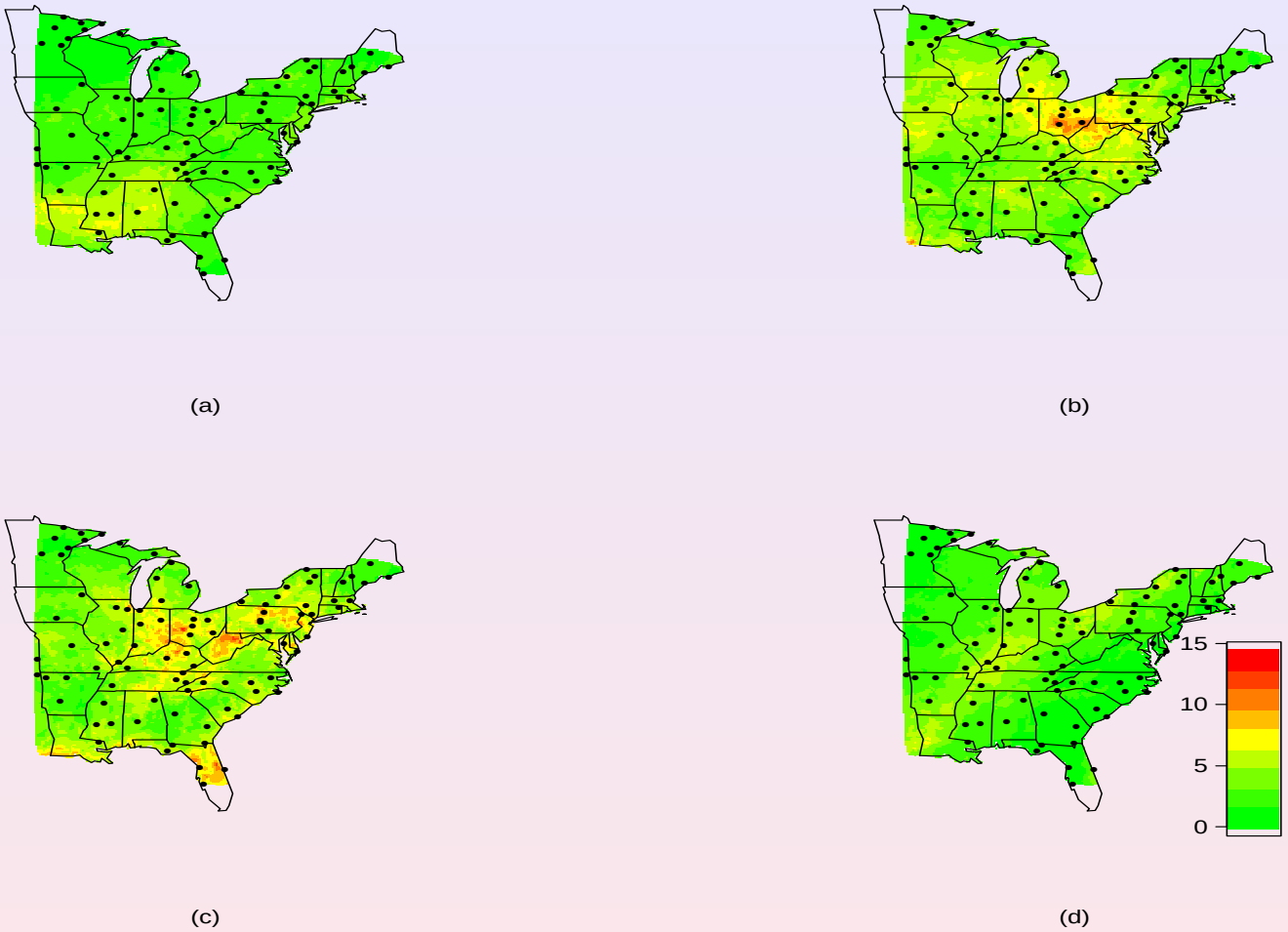


Figure: (a) Jan–Mar, (b) Apr-Jun, (c) Jul–Sep, (d) Oct-Dec.

Discussion

- Novel spatio-temporal model for fusing point and areal data which validates well.
- Inference can be provided for any spatial or temporal aggregation.
- There are models in between IDW and ours but may sacrifice the features we accommodate.
- Preferable to fusion using block averaging since number of modelled grid cells is much greater than number of monitoring sites, even worse if across time.
- Future: With data from more than one year we can study trends in deposition with regard to regulatory assessment.
- Future: Develop model for dry deposition, hence total deposition.

Paper download...

- Google Search Sujit Sahu to download the paper.
- One-day meeting on Environmental and Spatial Statistics sponsored by the Royal Statistical Society on June 19, 2009 in Southampton. http://www.s3ri.soton.ac.uk/courses/environmental/.
- A 3-day short course before that...