TIES conference. June 18, 2002.

Statistical Assessment of Air Quality Numerical Models.

Montserrat Fuentes

Statistics Department NCSU

and

US Environmental Protection Agency

fuentes@stat.ncsu.edu

http://www.stat.ncsu.edu/∼fuentes

In collaboration with:

Adrian Raftery (University of Washington)

Slide 1

Motivation

OUR GOAL:

Evaluation of physically based computer models for air quality applications is crucial to assist in control strategy selection. The high risk of getting the wrong control strategy has costly economic and social consequences.

The main objective of our work is to statistically assess the performance of air quality models.

The objective comparison of modeled concentrations with observed field data provides a means for assessing model performance.

Slide 2

THE PROBLEM:

To statistically assess the performance of air quality models we need measures of how well the model output and real data agree.

- An approach is to use spatio-temporal models for monitoring data to provide estimates of average concentrations over grid cells corresponding to model prediction (Dennis et al. (1990), Sampson and Guttorp (1998)). This approach is reasonable when the monitoring data are dense enough that we can fit an appropriate spatio-temporal model to the data. In situations with few and **sparse data points** that show a **lack of stationary** (eg. CASTNet), the interpolated grid square averages would be poor because of the sparseness of the network, so treating them as ground truth for model evaluation would be questionable.

Slide 3

A related problem is that the comparison does not take into account the **uncertainty** in the interpolated values.

Slide 4

THE SOLUTION:

We develop a new approach to the model evaluation problem, and show how it can also be used to remove the bias in model output.

We specify a simple model for both numerical models predictions and field data in terms of the unobserved ground truth, and estimate it in a Bayesian way.

Solutions to all the problems considered here follow directly. Model evaluation then consists of comparing the field observations with their predictive distributions given the output of numerical models. Bias removal follows from estimation of the bias parameters in the air quality model.

The resulting approach takes account of and estimates the bias in the atmospheric models, the lack of stationarity in the data, the ways in which spatial structure and dependence change with locations, the change of support problem, and the uncertainty about these factors.

Slide 5

OUTLINE

1. Background on Spatial processes.
2. New model for nonstationary environmental processes.
3. Statistical assessment of model performance.
4. Application.

Slide 6

1. Background on Spatial Statistics

Consider a stochastic process $\{Z(\mathbf{s}),\ \mathbf{s} \in D\}$ where D is a subset of $\mathcal{R}^d$ (d-dimensional Euclidean space). For example, $Z(\mathbf{s})$ may represent the concentration of SO_2 at a specific location $\mathbf{s}$. Let

$$\mu(\mathbf{s}) = \mathrm{E}\{Z(\mathbf{s})\}, \quad \mathbf{s} \in D,$$

denote the mean value at location $\mathbf{s}$. We also assume that the variance of $Z(\mathbf{s})$ exists for all $\mathbf{s} \in D$.

- Z is *second-order stationary* if $\mu(s) \equiv \mu$ and

$$\mathrm{cov}\{Z(\mathbf{s}_1), Z(\mathbf{s}_2)\} = C(\mathbf{s}_1 - \mathbf{s}_2)$$

where $C(\mathbf{s})$ is the covariance function.

Slide 7

The **covariance** provides a measure of spatial correlation by describing how sample data are related with distance and direction.

A *Gaussian* process which is second-order stationary is also strictly stationary.

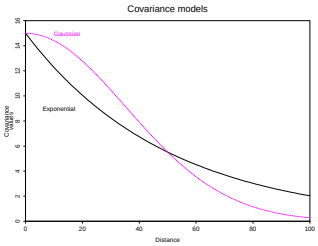


Figure 1: Covariance models: Exponential and Gaussian.

Slide 8

Variogram

- Geostatistical data typically exhibit **small-scale variation** that may be modeled as spatial autocorrelation and incorporated into estimation procedures. The **variogram** provides a measure of spatial correlation by describing how sample data are related with distance and direction.

$$Variogram(\mathbf{x}-\mathbf{y}) = \frac{1}{2}\text{Var}\left\{Z(\mathbf{x}) - Z(\mathbf{y})\right\}$$

$$= \text{cov}(\mathbf{0}) - \text{cov}(\mathbf{x}-\mathbf{y})$$

where Z is a second-order stationary process.

Slide 9

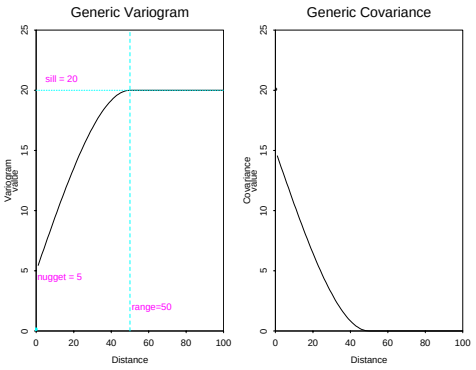


Figure 2: Variogram and Covariance.

Slide 10

2. New Model for nonstationarity.

Fuentes (2001, Environmetrics), Fuentes, (2002, Biometrika)

We represent a nonstationary process Z observed on a region D as a MIXTURE of orthogonal local stationary processes.

$$Z(\mathbf{x}) = \sum_{i=1}^{k} Z_i(\mathbf{x}) w_i(\mathbf{x})$$

where $S_1, \ldots, S_k$ are well-defined subregions that cover D, and Z_i is a local stationary process in the subregion S_i, $w_i(\mathbf{x})$ is a positive kernel function centered at the centroid of S_i.

The nonstationary covariance of Z is defined in terms of the local stationary covariances of the processes Z_i for $i = 1, \ldots, k$,

$$\text{cov}(Z(\mathbf{x}), Z(\mathbf{y})) = \sum_{i=1}^{k} w_i(\mathbf{x}) w_i(\mathbf{y}) \text{cov}(Z_i(\mathbf{x}), Z_i(\mathbf{y})).$$

Slide 11

Assuming a stationary covariance structure $C_{\boldsymbol{\theta}_i}(\mathbf{x}-\mathbf{y})$ with parameter $\boldsymbol{\theta}_i$ for each Z_i we obtain,

$$\text{cov}(Z(\mathbf{x}), Z(\mathbf{y})) = \sum_{i=1}^{k} w_i(\mathbf{x}) w_i(\mathbf{y}) C_{\boldsymbol{\theta}_i}(\mathbf{x}-\mathbf{y}).$$

Generally $\boldsymbol{\theta}_i$ is a vector with the local values of the covariance parameters of Z in region S_i. The covariance parameters could also vary continuously over the domain of interest.

Next, we introduce a generalization of this nonstationary model using an integral representation of the process (instead of a sum), which allows the covariance parameters to vary continuously on the domain.

Slide 12

Generalization of the nonstationary model

We represent a nonstationary process Z observed on a region D as a convolution of local stationary processes:

$$Z(\mathbf{s}) = \int_D K(\mathbf{s}-\mathbf{u}) Z_{\boldsymbol{\theta}(\mathbf{u})}(\mathbf{s}) d\mathbf{u}.$$

where K is a kernel function and $Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s}),\ \mathbf{s} \in D$ is a family of (independent) stationary processes indexed by $\boldsymbol{\theta}$.

The covariance $C(\mathbf{s_1}, \mathbf{s_2}; \boldsymbol{\theta})$ of Z is a convolution of the local covariances $C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s_1} - \mathbf{s_2})$,

$$C(\mathbf{s_1}, \mathbf{s_2}; \boldsymbol{\theta}) =$$

$$\int_D K(\mathbf{s_1} - \mathbf{s}) K(\mathbf{s_2} - \mathbf{s}) C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s_1} - \mathbf{s_2}) d\mathbf{s}.$$

Slide 13

Two approaches for spatial interpolation

- A geostatistical approach for interpolation. The predicted value of Z at location $\mathbf{s}_0$ is obtained using the traditional kriging equations with the estimated covariance $\hat{C}$.
- A Bayesian approach for spatial interpolation is recommended, to take into account the uncertainty about the covariance parameters. The quantity of interest is the predictive posterior distribution (ppd) for $Z(\mathbf{x}_0)$ given $\mathbf{Z}$, the observed values of the process Z. The ppd is obtained by integrating out the covariance parameters,

$$p(Z(\mathbf{x}_0)|\mathbf{Z}) \propto \int p(Z(\mathbf{x}_0)|\boldsymbol{\theta}\, \mathbf{Z}) p(\boldsymbol{\theta}|\mathbf{Z}) d\boldsymbol{\theta}$$

Slide 14

Hierarchical Bayesian approach

The parameter function $\boldsymbol{\theta}(\mathbf{u})$ for $\mathbf{u} \in \mathbf{D}$, measures the lack of stationarity of the process Z. If would be natural to treat $\boldsymbol{\theta}(\mathbf{u})$ as a stochastic process, with correlated errors.

We consider a hierarchical Bayesian approach to model and take into the account the spatial structure of the parameter $\boldsymbol{\theta}(\mathbf{u})$ in the prediction of Z.

Stage 1:

The process Z is as a convolution of local stationary processes:

$$Z(\mathbf{s}) = \int_D K(\mathbf{s}-\mathbf{u}) Z_{\boldsymbol{\theta}(\mathbf{u})}(\mathbf{s}) d\mathbf{u}.$$

where K is a kernel function and $Z_{\boldsymbol{\theta}}(\mathbf{s}),\ \mathbf{s} \in D$ is a family of (independent) stationary Gaussian processes indexed by $\boldsymbol{\theta}$. The kernel K has a bandwidth h.

Slide 15

Thus, the distribution of Z given $\boldsymbol{\theta}$ and h is Gaussian:

$$[Z|\boldsymbol{\theta},\ h] \text{ is Gaussian}$$

Stage 2:

We propose the following model for the parameter function $\boldsymbol{\theta}$

$$\boldsymbol{\theta}(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon_{\boldsymbol{\theta}}(\mathbf{s}) \qquad (1)$$

The process $\epsilon_{\boldsymbol{\theta}}(\mathbf{u})$ represents some spatially correlated zero-mean noise, it has zero-mean and a stationary covariance with parameters $_0$. The function μ represents the large scale structure, is a polynomial with coefficients $\boldsymbol{\beta}_0$. The vector parameter $\boldsymbol{\beta}_0$ is unknown.

Thus, we have in stage 2:

$$[\boldsymbol{\theta}|\boldsymbol{\beta}_0,\ _0] \text{ is Gaussian}$$

Slide 16

If the goal is to predict Z at a location $\mathbf{x}_0$, the Bayesian solution is the predictive distribution of $Z(\mathbf{x}_0)$ given the observations $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_N))$,

$$p(Z(\mathbf{x}_0)|\mathbf{Z}) \propto$$

$$\int p(Z(\mathbf{x}_0)|\mathbf{Z}, \boldsymbol{\theta}(\mathbf{u}), h)\, p(\boldsymbol{\theta}(\mathbf{u}), h|\mathbf{Z})\, dh\, d\boldsymbol{\theta}.$$

an *Gibbs sampling* approach is used to simulate m values from the posterior of the parameters $\boldsymbol{\theta}$ and h (bandwidth of kernel). Thus, the predictive distribution is approximated by the *Rao-Blackwellized estimator*:

$$p(Z(\mathbf{x}_0)|\mathbf{Z}) =$$

$$\frac{1}{m}\sum_{i=1}^{m} p(Z(\mathbf{x}_0)|\mathbf{Z}, \boldsymbol{\theta}(\mathbf{u})^{(i)} \text{ for } \mathbf{u} \in D, h^{(i)})$$

where $\boldsymbol{\theta}(\mathbf{u})^{(i)}$ for $\mathbf{u} \in D$, and $h^{(i)}$, constitute the i-th draw from the posterior distribution.

Slide 17

3. Evaluation of numerical models

(Fuentes and Raftery, 2001)

Two sources of information for air fluxes:

I. Point Measures of Pollutant Concentrations

Atmospheric deposition takes place via two pathways: wet deposition and dry deposition. Wet deposition rates of acidic species across the United States have been well documented over the last 10 to 15 years; however, comparable information is *unavailable* for dry deposition rates. Since 1990 EPA operates approximately 50 sites through US to establish spatial patterns of deposition and concentration.

Slide 18

II. Regional Models (Models-3) Estimated Concentrations

The present generation of regional scale air quality models can consider land cover, plant growth rate, topography, and other factors in estimating pollutant concentrations and fluxes in a grid.

Slide 19

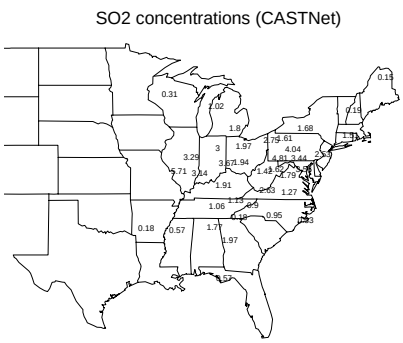


Figure 3: CASTNet weekly concentrations of SO_2), for the week of July 11, 1995.

Slide 20

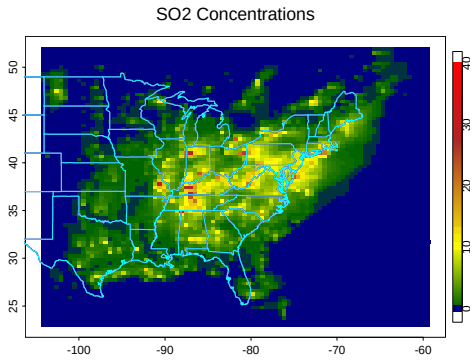
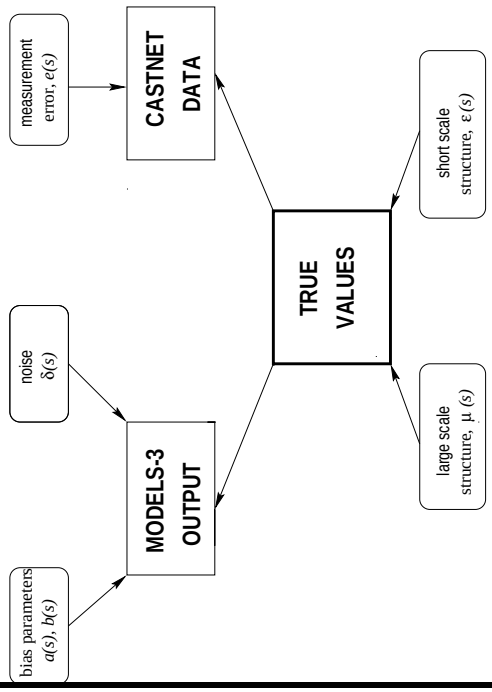


Figure 4: Output of Models-3, weekly average of SO_2 concentrations (ppb), for the week of July 11, 1995. The resolution is 36km $\times$ 36km.

Slide 21



Slide 22

We should not treat CASTNet ($\hat{Z}$) measurements as the "ground truth". We assume there is some smooth underlying (but unobserved) field $Z(\mathbf{s})$, where $Z(\mathbf{s})$ measures the "true" concentration of the pollutant at location $\mathbf{s}$. We write

$$\hat{Z}(\mathbf{s}) = Z(\mathbf{s}) + e(\mathbf{s})$$

where $e(\mathbf{s}) \sim N(0, \sigma_e^2)$ represents the measurement error (nugget) at location $\mathbf{s}$.

Since the output of Models-3 ($\tilde{Z}$) are not point measurements but areal estimations in subregions $B_1, \ldots, B_N$ that cover the domain, D, we have:

$$\tilde{Z}(B_1) = a(B_1) + b \int_{B_1} Z(\mathbf{s}) d\mathbf{s} + \delta(B_1)$$

Slide 23

The true underlying process Z is a spatial process with a nonstationary covariance,

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s})$$

where $Z(\mathbf{s})$ has a spatial trend, $\mu(\mathbf{s})$, that is a function of some metereological and geographic covariates $f_1, \ldots, f_p$ that are know functions at some locations $\mathbf{s}$, with unknown coefficients $\boldsymbol{\beta}$:

$$\mu(\mathbf{s}) = \sum \beta_i f_i(\mathbf{s})$$

We assume $Z(\mathbf{s})$ has zero-mean correlated errors $\epsilon(\mathbf{s})$. The process $\epsilon(\mathbf{s})$ has a nonstationary covariance with parameter vector $\boldsymbol{\theta}$ that might change with location.

Slide 24

- Evaluation of MODELS-3:

  For model evaluation we simulate values of CASTNet given models-3, from the following posterior predictive distribution:

$$P(\hat{Z}|\tilde{Z}, a = 0, b = 1).$$

Slide 25

- Estimating Bias of Models-3:

  For bias removal we simulate values of the parameters a and b from the posterior distribution:

$$P(a, b|\hat{Z}, \tilde{Z}).$$

Slide 26

- Air Quality mapping by combining CASTNet and MODELS-3

  If the goal is to get more reliable maps of air pollution, we could predict the value of Z (the truth) at location $\mathbf{x}_0$ given ALL the data (CASTNet and Models-3), thus we need the predictive distribution of $Z(\mathbf{x}_0)$ given the observations ($\hat{Z}$ and $\tilde{Z}$)

  For spatial prediction we simulate values of Z from the posterior predictive distribution:

$$P(Z|\hat{Z}, \tilde{Z}).$$

Slide 27

4. Application

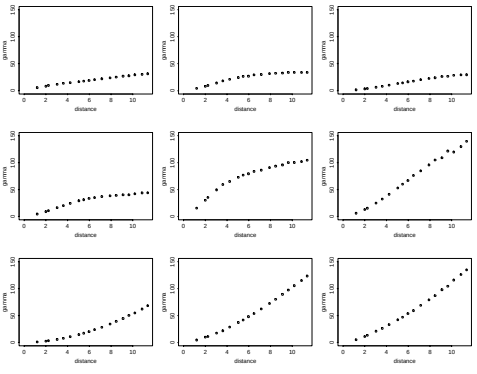


Figure 5: We divided the domain shown in the previous figure in 9 subregions and we calculated the empirical semivariograms. There is clear evidence of lack of stationarity.

Slide 28

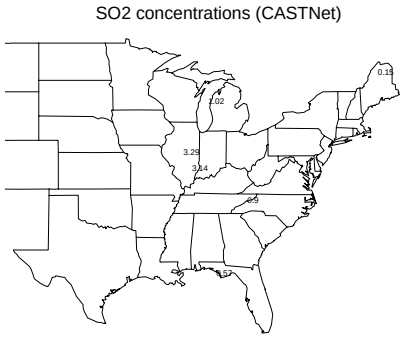


Figure 6: Weekly average of SO_2 concentrations (ppb) at 6 selected CASTNet sites, for the week of July 11, 1995.

Slide 29

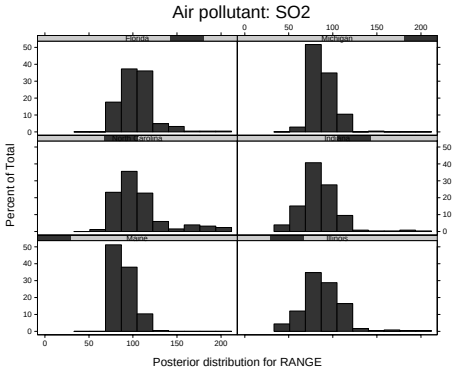


Figure 7: Posterior distributions for the range parameter (km) of the covariance for Models-3 SO_2 concentrations, for the week starting July 11, 1995. At 6 selected locations.

Slide 30

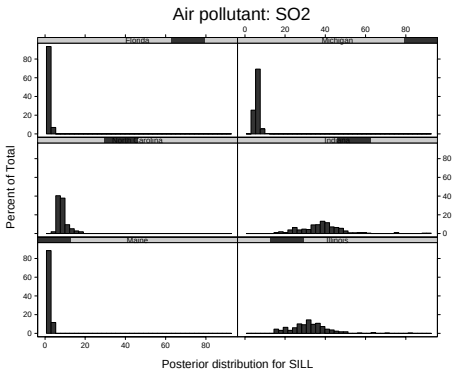


Figure 8: Posterior distributions for the sill parameter of the covariance for Models-3 SO_2 concentrations, for the week starting July 11, 1995. At the 6 selected locations.

Slide 31

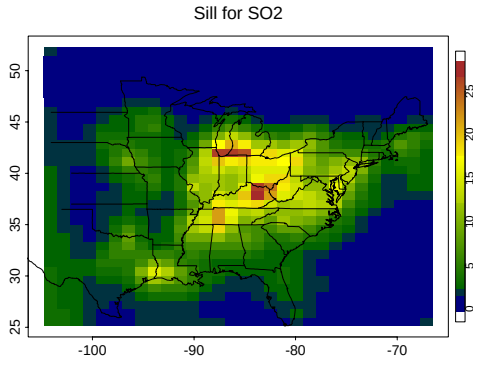


Figure 9: Map of the modes of the posterior distributions for the sill parameter of the covariance for Models-3 SO_2 concentrations, for the week starting July 11, 1995.

Slide 32

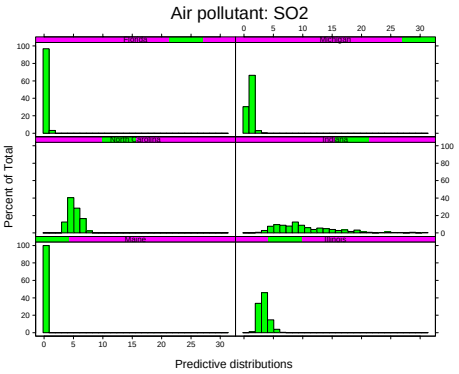


Figure 10: Predictive distributions for the Models-3 SO_2 concentrations, at 6 selected locations where we have CASTNet measurements, for the week starting July 11, 1995.

Slide 33

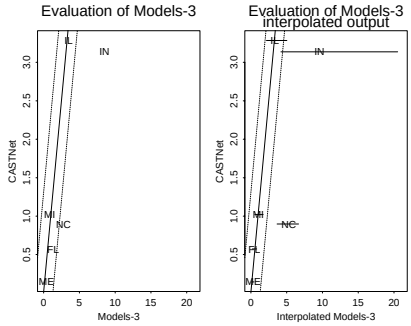


Figure 11: The graph on the left shows CASTNet measurements for the week starting July 11, 1995, versus the values of Models-3 for the pixels that are the closest to each CASTNet site, without considering the change of support. The graph on the right shows the CASTNet measurements versus the modes and 90% credible intervals of the predictive Bayesian distributions derived from Models-3 at the CASTNet locations.The dotted lines indicate a 90% confidence region for the CASTNet values.

Slide 34

Site	CASTNet	Models-3	90%	C. I.
ME	0.15	0.33	0.10	0.43
IL	3.29	3.33	2.17	5.03
NC	0.90	5.32	3.67	6.67
IN	3.14	9.59	4.20	20.50
FL	0.57	0.52	0.20	0.80
MI	1.02	1.04	0.53	1.70

TABLE 1. Column 2 are the CASTNet values ($\hat{Z}$). Columns 3-5 show the modes and the corresponding 90% credible intervals of the posterior predictive distribution $P(\hat{Z}|\tilde{Z}, a = 0, b = 1)$ for model validation.

Slide 35

We could remove the bias in the interpolated Models-3 values by taking into account the additive bias measured by $a(\mathbf{x})$ (a polynomial of degree 4 with coefficients a_0) and the multiplicative bias b.

We simulate values of a_0 and b from the posterior distribution $P(a, b|\hat{Z}, \tilde{Z})$, at each site, and we obtained the following adjusted Models-3 values (adjusted value $= ((\text{Models } 3) - a)/b$) at the 6 selected sites: 0.22, 2.90, 1.67, 2.36, 0.96, and 1.00.

These values are again similar to CASTNet, especially considering that the uncertainty about CASTNet is approximately 0.8ppb.

Slide 36

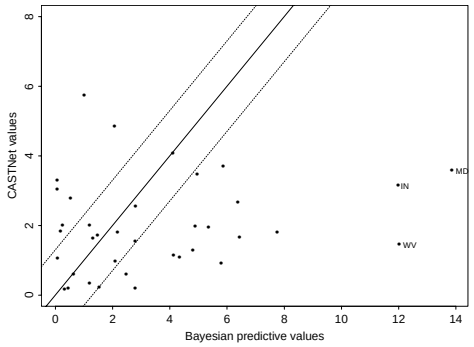


Figure 12: CASTNet values of SO_2 versus the mean of the predictive posterior distribution at each site.

Slide 37

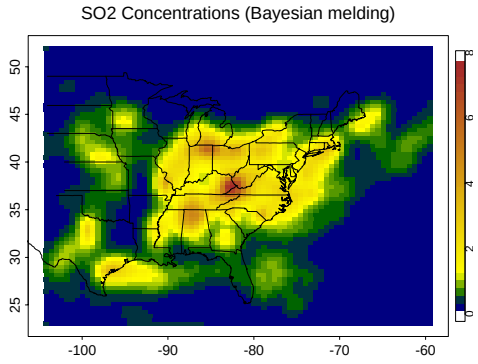


Figure 13: Predicted SO_2 concentrations via a Bayesian melding approach to combine CASTNet and Models-3 data. This graph shows the mean of the posterior predictive distribution for the SO_2 concentrations.

Slide 38

Conclusions

- Nonstationary processes can be modeled as a convolution of local stationary processes. The nonstationary covariance is a convolution of the local stationary covariances.
- Uncertainty in the specification of the non-stationarity is incorporated by obtaining the predictive distribution through a Hierarchical Bayesian model.
- We evaluate air quality models by obtaining the posterior predictive distribution of the measurements at the monitoring sites given the numerical models output.
- We remove the bias the air quality models by obtaining the posterior distribution of the bias parameters given the measurements at the monitoring sites and the numerical models output.

Slide 39

- We combine data using conditionally specified spatial models through a Bayesian melding approach.
- The approach presented in this paper gave us a good understanding of the spatial structure of the "true" concentrations of SO_2. This information can be very useful for designing future data collection. Part of our future work is to use the findings presented here for monitoring network design.

Slide 40