Spatial bias modeling with application to assessing remotely-sensed aerosol as a proxy for particulate matter

Chris Paciorek
Department of Biostatistics
Harvard School of Public Health
www.biostat.harvard.edu/~paciorek

with Yang Liu, HSPH Environmental Health

Research supported by HEI 4746-RFA05-2/06-7

Setting

- Proxy information is increasingly common in environmental science and other applications
  - Deterministic model output
    - Climate models
    - Atmospheric chemistry models
    - Meteorological models
  - Remote sensing information
    - Pollutant concentrations
    - Meteorological variables
    - Land use
  - Proxy data such as biomarkers
- Understanding the discrepancies (biases) between the proxy and the process of interest is critical, but not adequately explored scientifically or statistically.

Satellite AOD as a proxy for $PM_{2.5}$

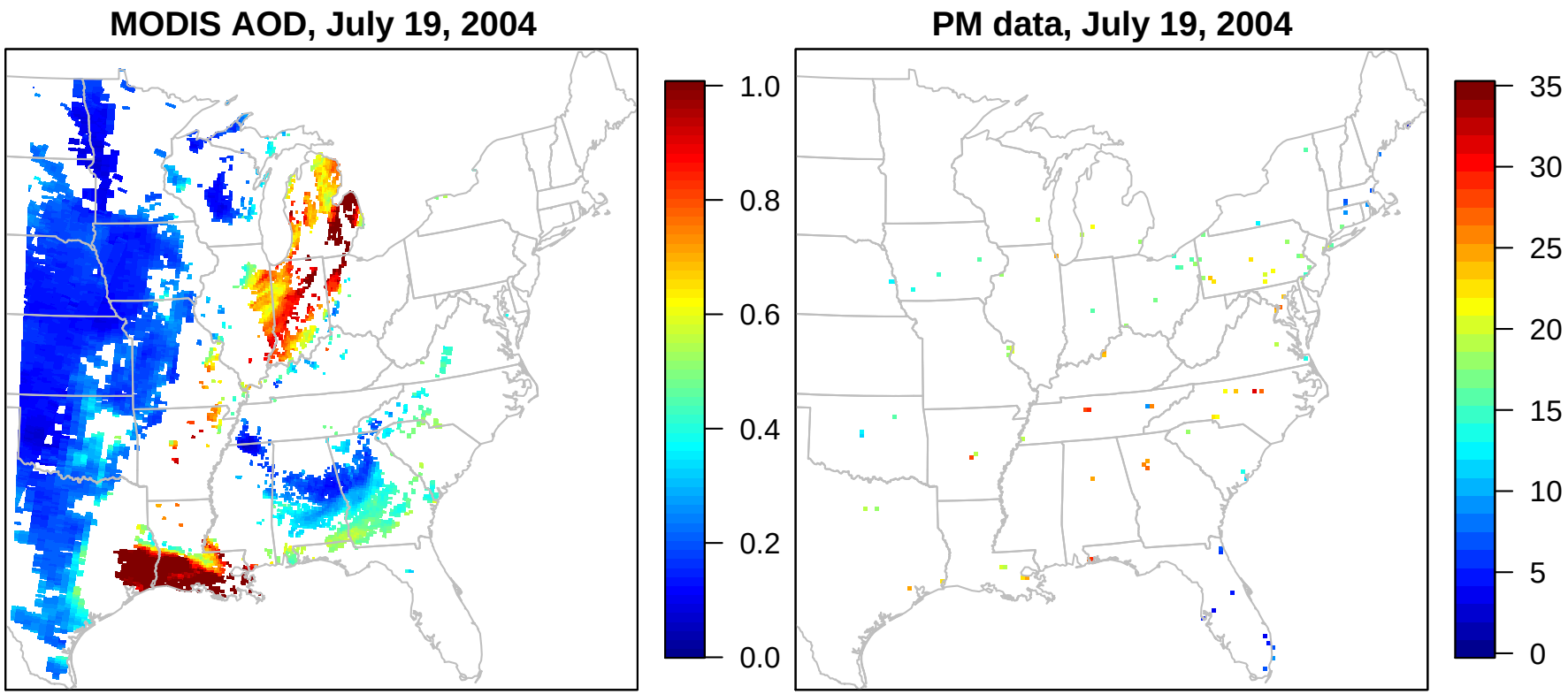
Aerosol optical depth:

- Integrated vertical column measurement based on light reflecting off the earth surface
- Algorithms for separating surface reflectance from aerosol effect
- Pixels provide nominal resolution of 4 (GOES), 10 (MODIS), 18 (MISR) km
- Clouds and orbit patterns lead to unavailable retrievals

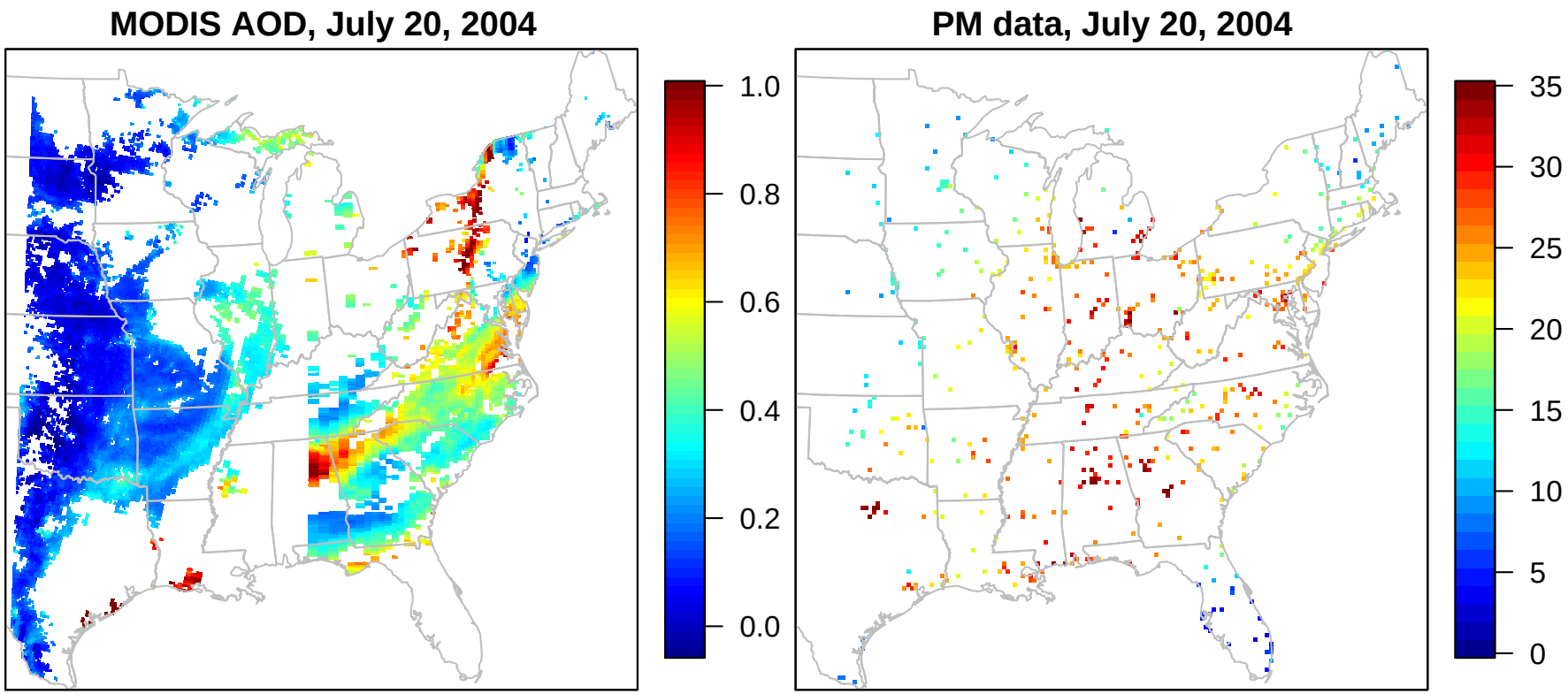
Correlations of matched daily AOD and $PM_{2.5}$ (24-hour average), eastern U.S., 2004

- MODIS (10:30 am snapshot): 0.60
- MISR (10:30 am snapshot): 0.50
- GOES (average of half-hourly retrievals): 0.38

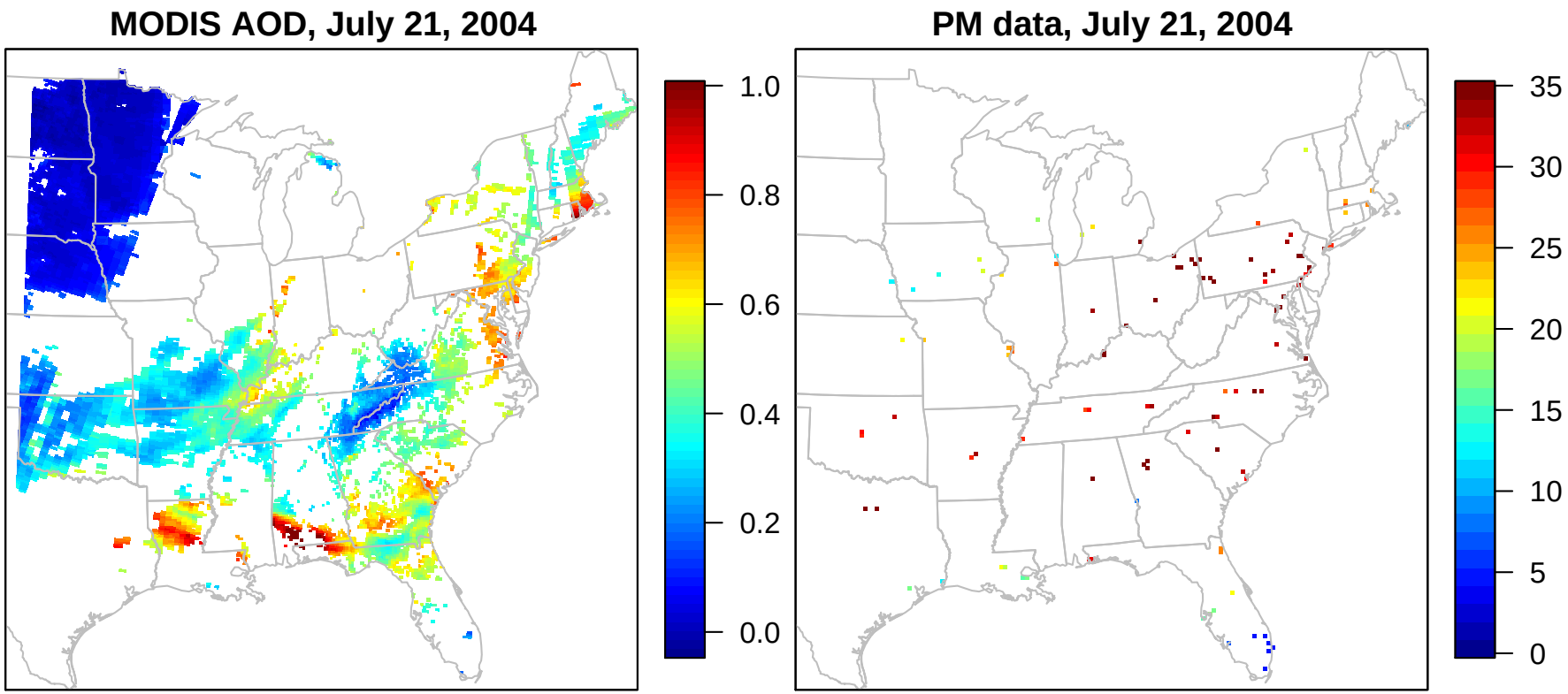
Daily Comparison (1)



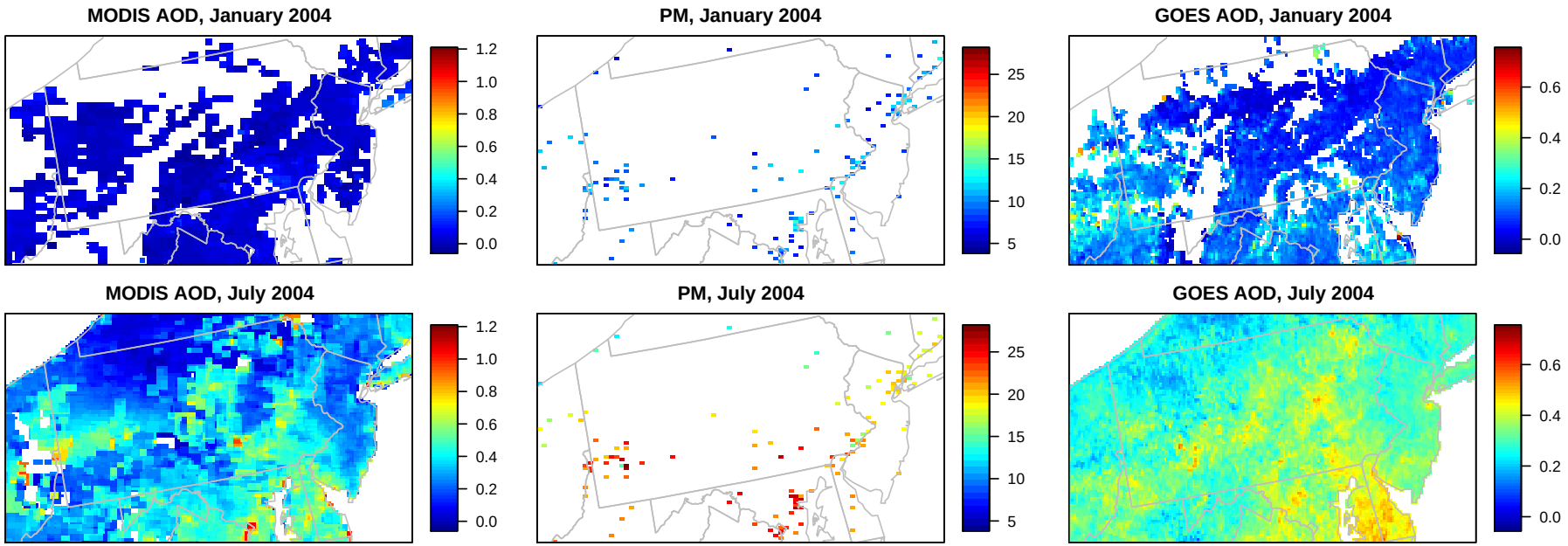
Daily Comparison (2)



Daily Comparison (3)



Monthly Comparison



Key Questions

- Scientific:
  - Can AOD help predict spatial patterns in PM?
    - Are there specific spatial scales for which AOD is helpful (or not helpful)?
  - Does including AOD in the model improve predictions of PM concentrations conditional on other information?
    - GIS-based covariates
    - Spatial smoothing
    - Meteorological covariates
- Statistical
  - How can we use statistical modeling to better understand relationships between proxy information and processes of interest?

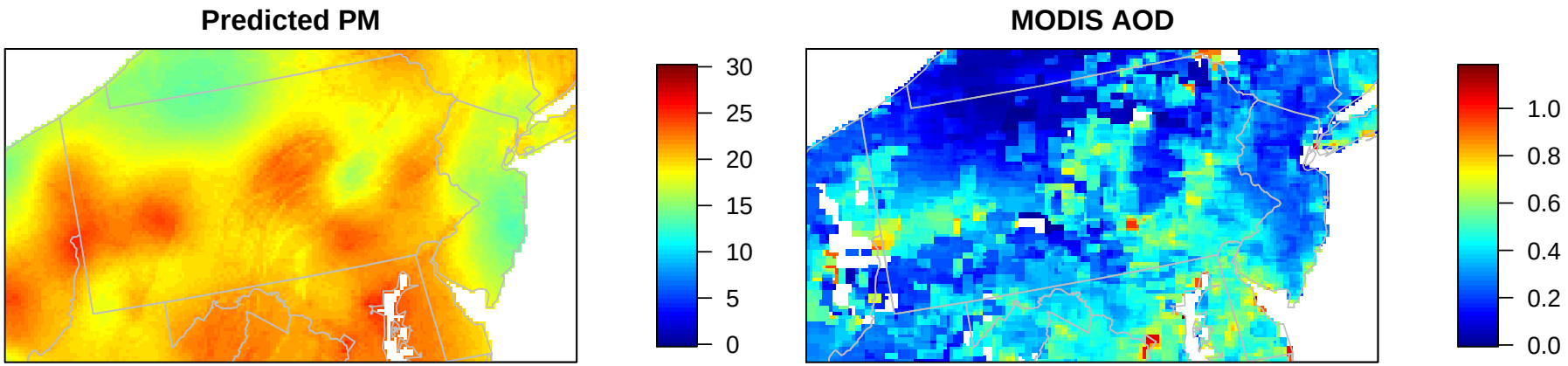
A Basic Data Fusion Model

- Fuentes and Raftery (2005, Biometrics) proposed treating the proxy as a second data source.
- A basic model:

$$\begin{aligned} Y_i &\sim \mathcal{N}(P(s_i), \sigma_y^2) \\ A_m &\sim \mathcal{N}(\beta_0 + \beta_1 P(s_m), \sigma_a^2) \\ P(\cdot) &\sim \mathcal{GP}(\mu(\cdot), C(\cdot,\cdot)) \end{aligned}$$

  where Y is the gold-standard data, A is the proxy information source, and $P(\cdot)$ is the latent process of interest.
- This model treats the proxy as reflecting the latent process with additive bias, β_0, and multiplicative bias, β_1, plus white noise error.

Implications of Simple Bias Structures



Key question: Are our predictions of the process of interest distorted by unrelated patterns in the proxy.

Flexible Spatial Bias Modeling

- Consider additive bias as a spatial process, $\phi(\cdot)$:

$$\begin{aligned} Y(s_i) &\sim \mathcal{N}(P(s_i), \sigma_y^2) \\ A(s_i) &\sim \mathcal{N}(\phi(s_i) + \beta_1 P(s_i), \sigma_a^2) \\ P(\cdot) &\sim \mathcal{GP}(\mu_P(\cdot), C_P(\cdot,\cdot)) \\ \phi(\cdot) &\sim \mathcal{GP}(\mu_\phi(\cdot), C_\phi(\cdot,\cdot)) \end{aligned}$$

- We can explore the relationship of the proxy and gold standard through analysis of the spatial scales of $\phi(\cdot)$.
- Depending on the scale of $\phi(\cdot)$, we might call it either 'bias' or 'systematic error' (residual spatial correlation).
- One can include covariates in the various mean terms, μ_P and μ_ϕ.

Additional Comments on the Flexible Model

- One can view the model in the form of a coregionalization model

$$\begin{pmatrix} Y \\ A \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \beta_1 & 1 \end{pmatrix} \begin{pmatrix} P \\ \phi \end{pmatrix} + \begin{pmatrix} \epsilon_y \\ \epsilon_a \end{pmatrix}$$

- Or as a factor analysis, with spatial factors, P and ϕ, and constrained loadings.
- Treating the multiplicative bias, β_1, as a spatial process causes identifiability issues, but subject matter knowledge might help in setting up models for $\phi(\cdot)$ and $\beta_1(\cdot)$.

Bias Scenarios

- $\phi(\cdot)$ very smooth (large-scale variation only):
  - Proxy and gold standard show similar patterns at small and moderate scales, but there is a large-scale bias that causes an offset between proxy and gold standard.
  - $\phi(\cdot)$ is a large-scale bias correction term that should be estimable with a moderate amount of gold standard data.
- $\phi(\cdot)$ wiggly but with little large-scale variation (small-scale variation only):
  - Proxy and gold standard show similar large-scale patterns but small-scale variation in proxy unrelated to gold standard.
  - $\phi(\cdot)$ is small-scale bias, or equivalently, spatially-correlated error in the proxy.
  - Without dense data, bias cannot be corrected for; model treats this as error that is uninformative about the process of interest.
- $\phi(\cdot)$ with both large- and small-scale variation, $\beta_1 \approx 0$:
  - Little correspondence between proxy and process of interest at any scale.
  - Proxy best described by a separate latent process.

Bias Diagnostics

- Jun and Stein (2004; Atmos. Env.) consider scales of model error $(Y - A)$ relative to observations (Y) and model output (A):

$$R(d) = \frac{\text{Variog}(Y - A)}{\text{Variog}(Y) + \text{Variog}(A)}$$

  where $R(d) = 1$ if the model output captures none of the variability in the observations at scale d.

- We propose a similar diagnostic in the model-based framework as

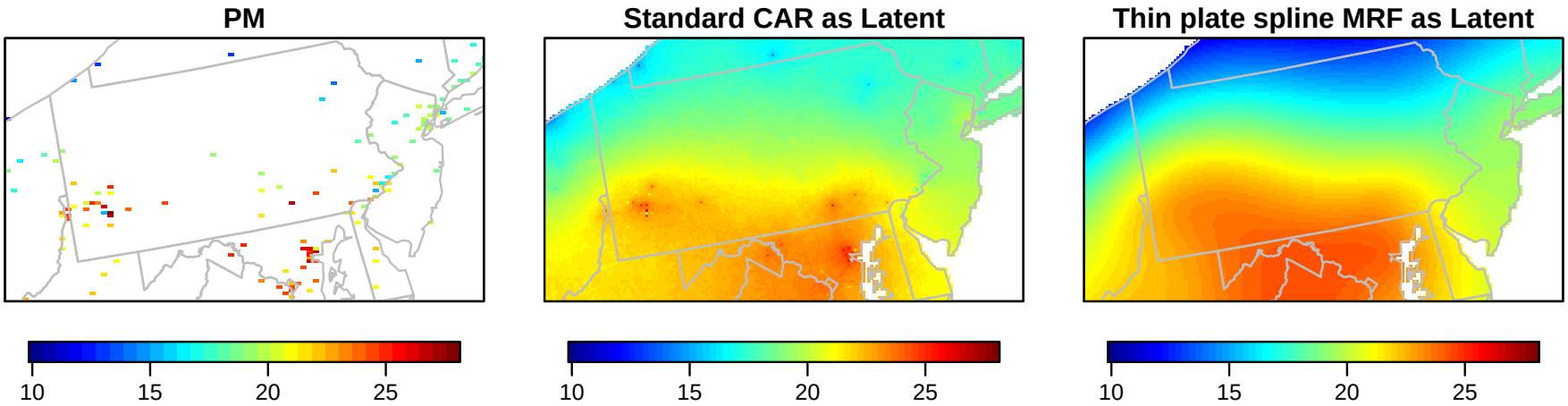
$$R(d) = \frac{\text{Variog}(\phi)}{\text{Variog}(\beta_1 P) + \text{Variog}(\phi + \beta_1 P)}$$

  $R(d)$ is the spatial bias variability as a proportion of the explained variation in the proxy, at scale d.

Computational Issues

- Proxy information is often massive in size, causing difficulty in computing the likelihood for the proxy.
- Some potential approaches include:
  - If ϕ or P is smooth, use reduced rank approaches such as penalized thin plate splines (Ruppert et al. 2003 book; Wood 2006 book)
  - If ϕ or P is wiggly and can be represented on a regular grid, use Markov random field approximations to a thin plate spline (Rue and Held 2005 book, Yue and Speckman, in submission)
  - Other techniques for large spatial datasets may also be useful: covariance tapering (Furrer et al. 2006, JCGS; Kaufman et al. in press, JASA), approximate likelihoods (Stein et al. 2004, JRSSB; Fuentes 2007, JASA)
- One concern in MCMC sampling is that P and ϕ can trade off, so mixing may be troublesome.

Sidenote: MRF Approximation to the Thin Plate Spline



Comparison of Weights for the MRF Models

Standard CAR

$$\begin{array}{ccc} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{array}$$

Thin plate spline MRF approximation

$$\begin{array}{ccccc} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{array}$$

Precision matrix elements for one row, oriented spatially (with respect to the focal grid cell of the row) to indicate neighborhood structure.

Key feature of TPS MRF: Precision matrices are sparse but realizations can be very smooth or very wiggly.

Model Structure

Likelihood for monthly average data:

$$\begin{aligned} \text{PM}_i = y_i &\sim \mathcal{N}(P(s(i)) + \sum_k f_k(z_{k,i}), \sigma^2_{y,i}) \\ \text{AOD}_m = a_m &\sim \mathcal{N}(\phi(s_m) + \beta_1 P(s_m), \sigma^2_{a,m}) \end{aligned}$$

- $f_k(\cdot)$, $k = 1, \ldots, K_f$ are penalized spline functions of within-grid cell covariates:
  - Distance to A1 and A2 roads; distance to local emission sources, weighted by emissions strength

Latent $\text{PM}_{2.5}$ process, $P(s)$, on 4 km grid:

$$P(s_m) = \sum_k h_k(w_k(s_m)) + g(s_m)$$

- $h_k(\cdot)$, $k = 1, \ldots, K_h$ are penalized spline functions of grid cell-scale covariates:
  - Land use, population density, road density, elevation, local emissions

Smooth term structure

Thin plate spline-based smooth terms, evaluated on the grid:

$$\begin{aligned} g &= Zb_g \\ \phi &= Zb_\phi \\ b_g &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_g^2) \\ b_\phi &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\phi^2) \end{aligned}$$

- Z is a reduced rank thin plate spline basis matrix, following Ruppert, Wand, and Carroll (2003), *Semiparametric Regression*.
- $b_{(\cdot)}$ are basis coefficients for the given smooth term.
- Variance components, $\sigma^2_{(\cdot)}$, penalize complexity.
- Regression smooths, $f_k(\cdot)$ and $h_k(\cdot)$, are represented in a similar fashion.

Spatial Misalignment

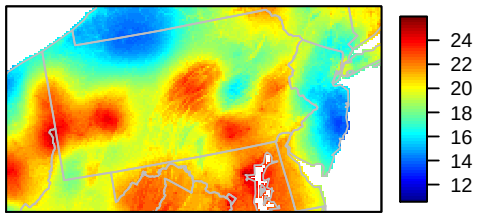
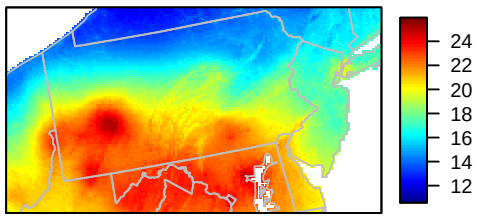
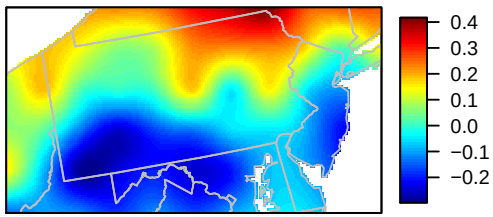
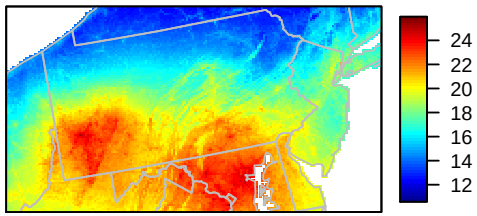
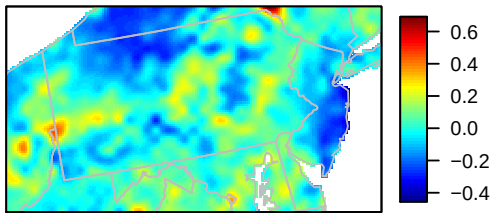
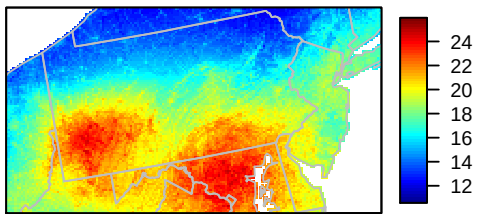
AOD is relatively smooth from pixel to pixel so we choose to handle misalignment in ad hoc way.

- MODIS: pixel locations change from orbit to orbit:
  - We associate each grid cell on each day with pixel centroids and average values for each cell to the month.
- GOES: fixed pixels
  - We average to the month and then compute a weighted average to realign to the 4 km grid.

MCMC Implementation

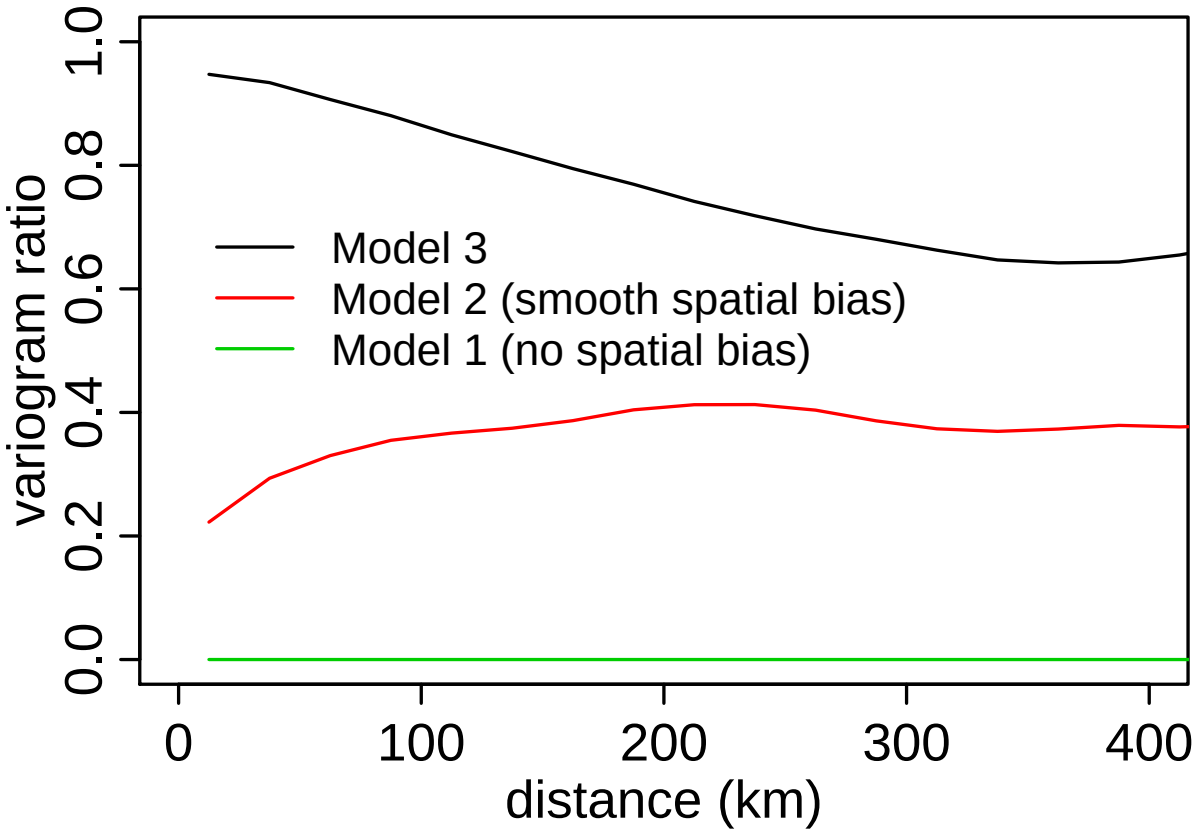
- Because of conditional conjugacy, $\psi = \{b_g, b_\phi, b_f, b_h, \beta_0, \mu\}$ can be sampled from its exact conditional.
- Importance: g, ϕ, f_k, and h_k are all competing to explain the spatial patterns in the data; joint sampling accounts for this dependence.
- Also, there is high dependence between the spline coefficients and their associated variance component (e.g., between b_g and σ_g^2).
  - Therefore, jointly sample: $\{\sigma_g^2, \psi\}$, $\{\sigma_\phi^2, \psi\}$, $\{\{\sigma_{f_k}^2\}_{k=1,\ldots,K_f}, \psi\}$, $\{\{\sigma_{h_k}^2\}_{k=1,\ldots,K_h}, \psi\}$.
  - Joint sampling is done with a Metropolis proposal for the variance component and then sampling ψ from its conditional normal, with a single acceptance decision and a Hastings adjustment needed because we are not sampling from the joint conditional.

Effects of Bias Structure: July 2004, MODIS

Model Structure	Predicted PM	Estimated Spatial Bias
Model 1: No spatial bias term; AOD as a simple proxy with scalar additive and multiplicative bias.		Model 1 assumes no spatial bias
Model 2: Spatial bias constrained to be a somewhat smooth process with a maximum of 55 degrees of freedom (a penalized spline with 55 knots)		
Model 3: Spatial bias relatively unconstrained with a maximum of 755 degrees of freedom.		
Model 4: AOD not used.		AOD not used in Model 4

Spatial Scales of Variation

The proposed variogram ratio is: $R(d) = \frac{\text{Variog}(\phi)}{\text{Variog}(\beta_1 P)+\text{Variog}(\phi+\beta_1 P)}$



Implications for Using AOD to Predict $PM_{2.5}$

- The results suggest there is little common spatial pattern to PM and AOD observations.
- Systematic error is considerable and critical to include, and predictions are very sensitive to assumptions about this error.
- The bias term, $\phi(\cdot)$, varies at both small- and large-scales, with little apparent relationship with the gold standard (β_1 is estimated to be near zero and $R(d)$ is near 1).
- Results for the other 11 months and using GOES AOD or meteorology-adjusted AOD give similar conclusions.
- Despite raw daily correlations between AOD and $PM_{2.5}$, spatial patterns in AOD provide little useful information for predicting spatial patterns in PM.

Further Assessment: Correlations of AOD and PM

	Raw AOD			Calibrated AOD		
	MODIS	MISR	GOES	MODIS	MISR	GOES
	Daily values, eastern U.S.					
Temporal plus spatial variation: Overall correlation of daily values across all sites and days.	0.60	0.50	0.38	0.64	0.57	0.40
Spatial variation only: Average of daily spatial correlations.	0.35	0.30	0.23	0.45	0.32	0.29
	Yearly averages, mid-Atlantic focal region					
Spatial variation only: Correlation of yearly averages.	0.09	0.25	-0.07	0.49	0.22	0.53

Correlations of raw and calibrated daily AOD with matched 24-h PM in 2004 for the eastern U.S. (top portion) and correlations of raw and calibrated yearly-average AOD with yearly-average PM (sites with at least 100 daily PM observations, matched in space to AOD) for our mid-Atlantic focal region in 2004 (bottom portion). Yearly averages reflect all available AOD retrievals and all available 24-h average PM concentrations. Calibrated AOD has been adjusted to account for the effects of planetary boundary layer (PBL) height, relative humidity (RH), season, and regional variation in modifying the relationship between daily AOD and PM. Yearly results exclude one outlying site.

Further Assessment: AOD as a covariate for PM

Likelihood for monthly average $PM_{2.5}$:

$$PM_{it} = y_{it} \sim \mathcal{N}(P_t(s_i) + \sum_k f_k(z_{k,i}), \sigma^2_{it})$$

Latent $PM_{2.5}$ process, $P_t(s)$, on 4 km grid:

$$P_t(s_m) = \mu_t + \beta_{1,t} A_t(s_m) + \sum_k h_k(w_k(s_m, t)) + g_t(s)$$

- Advantages of the covariate approach:
  - Direct estimation of the regression coefficient for AOD
  - Ease of interpretation
- Disadvantages:
  - Doesn't easily handle missing AOD
  - Doesn't address scale issues directly

AOD as a covariate: Cross-validation R^2(MSPE)

Model	Yearly averages		Monthly averages	
	All monitors (n=151)	Pop'n exposure monitors (n=130)	All monitors (n=1793)	Pop'n exposure monitors (n=1542)
	Models including land use, emissions, and meteorological predictors			
No AOD	0.580 (1.04)	0.570 (0.93)	0.827 (2.71)	0.839 (2.48)
With calibrated MODIS AOD	0.573 (1.06)	0.564 (0.94)	0.825 (2.73)	0.839 (2.50)
With calibrated GOES AOD	0.572 (1.06)	0.563 (0.95)	0.825 (2.73)	0.838 (2.50)
	Models without land use, emissions, and meteorological predictors			
No AOD	0.463 (1.33)	0.456 (1.38)	0.794 (3.22)	0.810 (2.94)
With calibrated MODIS AOD	0.467 (1.32)	0.459 (1.17)	0.794 (3.22)	0.810 (2.94)
With calibrated GOES AOD	0.467 (1.33)	0.458 (1.17)	0.794 (3.22)	0.810 (2.94)

Results exclude one outlying site.

Reasons for the Mismatch between AOD and $PM_{2.5}$

- Are the results consistent with the empirical correlations seen in previous studies? Explanations:
  - Differences between temporal and spatial association
  - Matched vs unmatched comparisons
- Some potential causes of spatial variability that interfere with spatial association of AOD and PM:
  - Spatial variability in surface reflectivity
  - Spatial variability in aerosol chemical composition and size distributions.
  - Spatially-coherent missingness due to daily cloud cover, with aggregate effects for longer-term averages
  - Spatial structure in pollution aloft in the atmosphere
  - Spatial structure in pollution at times not captured by the satellite (night-time and hours with no satellite coverage)

General Conclusions

- We need to be more explicit about our assumptions about proxy information and potential bias/spatially-correlated error.
- White noise error, while convenient, may not be appropriate.
  - We know there is error in proxy data sources and in many situations this is likely to be spatially-correlated, and at fine scales.
  - When using results for epidemiology, bias in the health analysis is a key concern.
- Modeling as bias/spatially-correlated error provides for careful assessment of the variation in the proxy, considering scales of concordance and discordance.
  - This approach can enhance simple deterministic model validation, which is often done via scatterplots and R^2 calculations.