# Flexible Spatial Latent Variable Modeling for Proxy Data with Systematic Discrepancy

Chris Paciorek
Department of Statistics; University of California, Berkeley
and
Department of Biostatistics; Harvard School of Public Health

www.biostat.harvard.edu/~paciorek

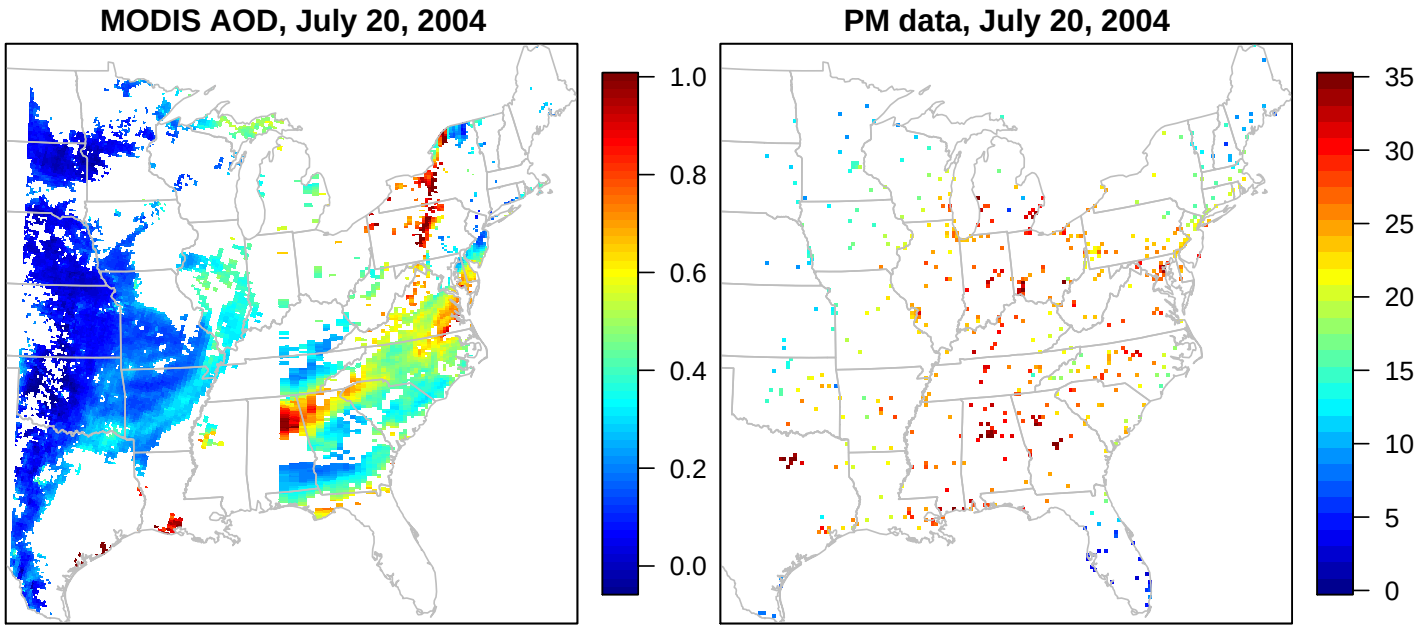
Research supported by HEI 4746-RFA05-2/06-7

June 2010

# Proxy Information in Environmental Applications

- 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, land change

# Combining Information



# Challenges of Proxy Information

- Systematic spatial (and temporal) discrepancy between proxy and truth
  - White noise error structure often implausible
  - This impacts predictions, prediction uncertainty, and assessment of proxy usefulness
  - Ignoring the discrepancy leads to overinterpreting patterns in the proxy
  - Proxy may not directly quantify the process of interest, hence 'discrepancy' rather than 'error' or 'bias'
- Spatial misalignment of gridded proxy information and point-level observations
  - Temporal misalignment can also be an issue
- Proxy datasets are usually very large
  - Standard GP modeling is infeasible

# Prediction of Fine Particulate Matter (PM)

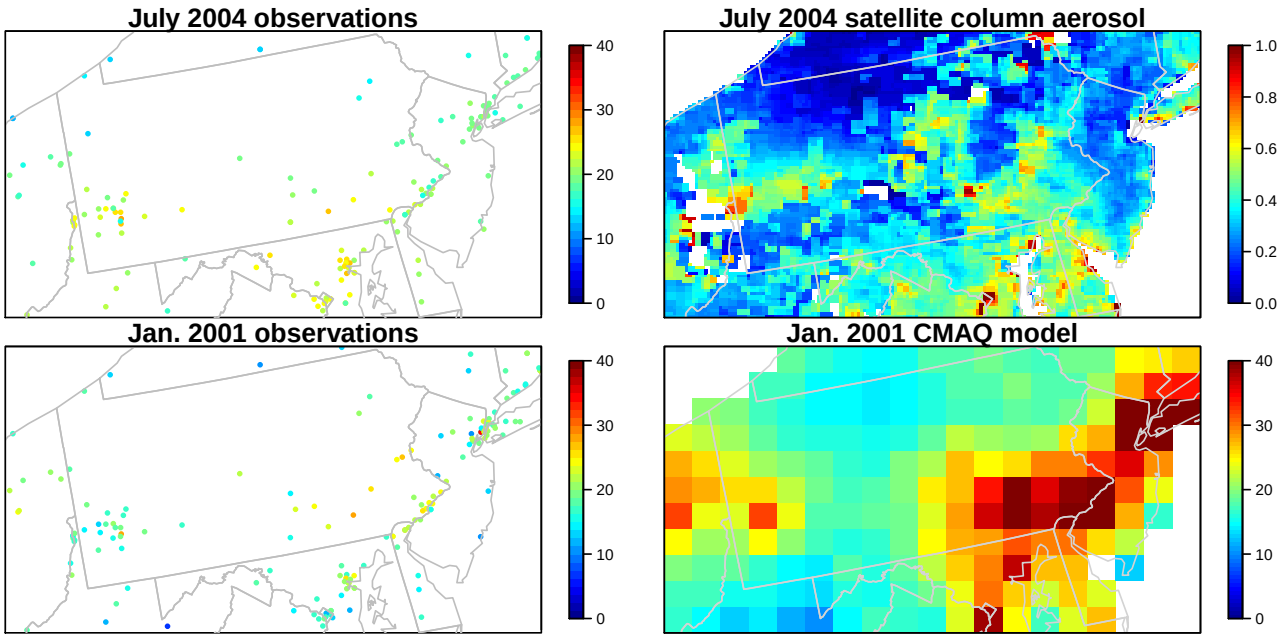
Proxy sources:

- Satellite-derived Aerosol Optical Depth (AOD)
  - Integrated vertical column measurement based on light reflecting off the earth surface
  - Gridded
  - Lots of missing data
- Atmospheric chemistry model output (CMAQ)
  - Gridded, no missing data

Gold standard:

- Ground monitoring network
  - Point-level observations
  - Influenced by local heterogeneity in PM

# PM Information



# A Basic Data Fusion Model

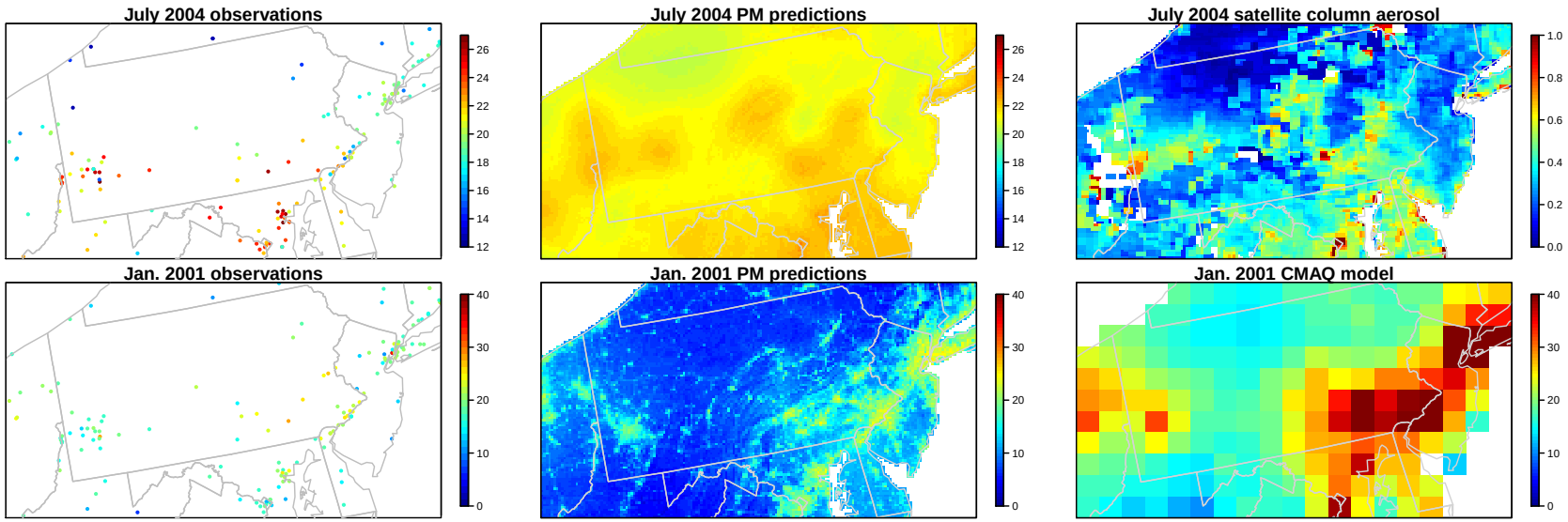
- Fuentes and Raftery (2005, Biometrics) proposed treating the proxy as a second data source.
- A basic model:

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

  where $Y$ is the gold-standard data, $A$ is the proxy information source, and $L(\cdot)$ is the latent process of interest.
- This model treats the proxy as reflecting the latent process with additive bias, $\beta_0(s)$, and multiplicative bias, $\beta_1$, plus white noise error.
  - The additive bias, $\beta_0(s)$, in Fuentes and Raftery (2005) was polynomial in $s$.

# Implications of Simple Bias Structures

## Predictions Based on Non-spatial Bias




Predictions of the process of interest appear to be distorted by unrelated patterns in the proxy.

# Flexible Spatial Discrepancy Modeling

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

$$\begin{aligned} Y &\sim \mathcal{N}(\mu_y(x) + K_y L, \sigma_y^2) \\ A &\sim \mathcal{N}(K_A D + \beta_1 K_A L, \sigma_a^2) \\ L &\sim \text{MRF}(\mu_L(x), Q_L) \\ D &\sim \text{MRF}(\mu_D(x), Q_D) \end{aligned}$$

- Latent processes, $L(\cdot)$ and $D(\cdot)$, are represented on a fine grid.
- We can explore the relationship of the proxy and gold standard through analysis of the spatial scales of $D(\cdot)$.
- $\mu_y(x)$ involves the effect of covariates that explain sub-grid scale variation in the point measurements, while $\mu_L(x)$ and $\mu_D(x)$ are covariate effects on the grid-scale process and the discrepancy term, respectively.

# Bias Scenarios

- $D(\cdot)$ very smooth (large-scale variation only):
  - Proxy and gold standard show similar patterns at small and moderate scales, but there is a large-scale discrepancy that causes an offset between proxy and gold standard.
  - $D(\cdot)$ is a large-scale bias correction term that should be estimable with a moderate amount of gold standard data.
- $D(\cdot)$ 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.
  - $D(\cdot)$ is small-scale discrepancy, or equivalently, spatially-correlated error in the proxy.
  - Without dense data, discrepancy cannot be corrected for; model treats it as error that is uninformative about the latent process.
- $D(\cdot)$ 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.

# A Markov Random Field Model

- Rue and Held (2005) describe a MRF that approximates a thin plate spline (TPS).

**Standard CAR**

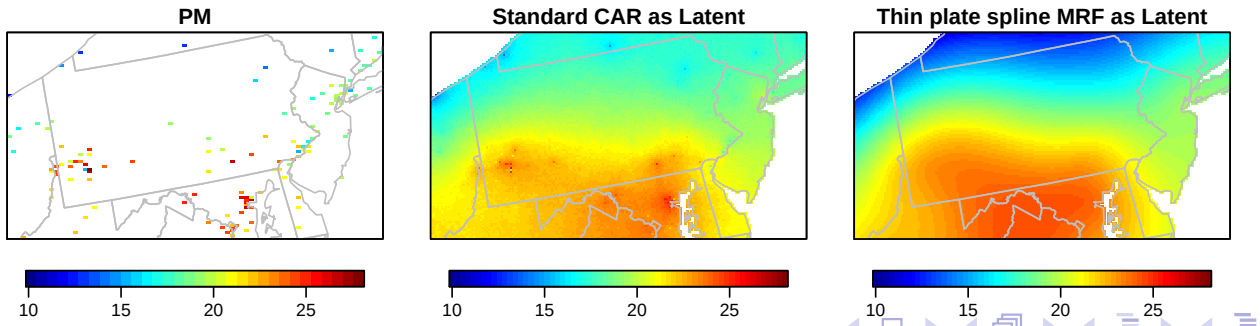
$$\begin{array}{ccc} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{array}$$

**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 of $Q$, oriented spatially (with respect to that row's focal grid cell) to indicate neighborhood structure.

- TPS MRF precision matrices are sparse but realizations can be either globally smooth or just locally smooth.



# Benefits of the MRF Approach

- This TPS approximation can capture smoothly-varying large-scale variation as well as fine-scale spatial patterns.
- Misalignment handled through:
  - Weighted averages of grid cells as approximation to integral
  - Assignment of grid cell value to points, with offset regression terms
- Sparse prior precision matrix provides computational efficiency.

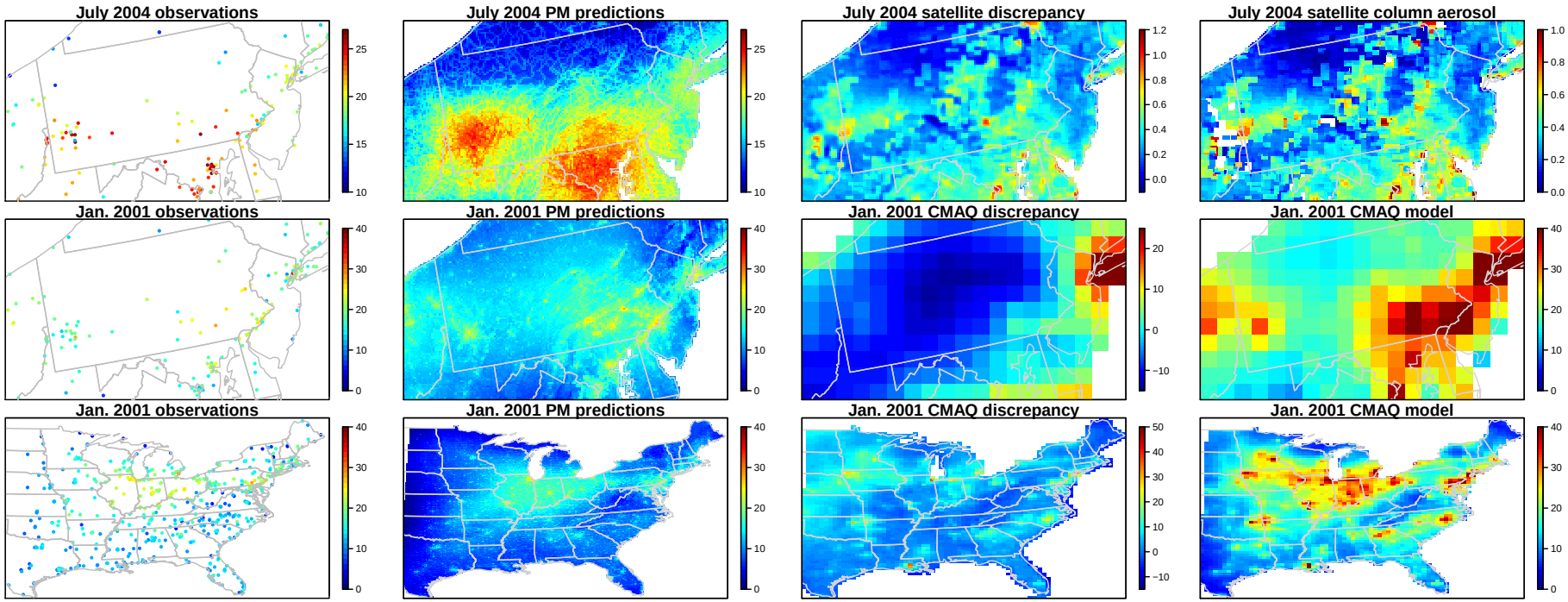
# Computational Strategy

1. Integrate first over $\{D, L\}$, then over $\{\mu_y, \mu_L, \mu_D\}$ so that resulting marginal posterior still involves sparse matrices.
   - Exploit matrix identities to get matrix representations that retain sparsity and avoid $Q^{-1}$.
2. In marginal posterior computations, exploit the sparse structure appropriately.

# Results: Using Proxies to Predict PM

- Satellite AOD:
  - The model fitting suggests there is little common spatial pattern to PM and AOD observations.
    - The discrepancy term, $D(\cdot)$, varies at both small and large scales.
  - As a result the model discounts AOD in predicting PM.
- Atmospheric Chemistry Model (CMAQ):
  - More apparent relationship between CMAQ output and latent PM.
    - The discrepancy term also varies at small and large scales, but more of the variation in the proxy appears to be signal than for AOD.
  - Model still heavily discounts the proxy.

# Predicted PM



# Conclusions (1)

- We need to be more explicit about our assumptions about the error structure of proxies.
  - White noise error, while convenient, is generally not appropriate.
  - Modeling the discrepancy can help to enhance simple deterministic model validation.
    - Standard validation relies on scatterplots and $R^2$ calculations.
    - Modeling the discrepancy allows us to consider scales of concordance and discordance.

# Conclusions(2)

- Distinguishing spatio-temporal signal from spatio-temporal noise is difficult and likely sensitive to modeling assumptions.
  - Additivity assumptions, error structures, spatial field representations.
  - Is there useful information in the proxies that the current model structure is not exploiting?
- Here we had relatively abundant gold standard data, but often this won't be the case and prior assumptions about the correlation structure of the error will be critical.
  - One prominent application is in climate model uncertainty quantification.
  - What can be said about uncertainty in regional climate projections?