Statistics Useful for Deterministic Models: Evaluation, Calibration, Extension, Integration, and Uncertainty Characterization

Chris Paciorek
Department of Biostatistics
Harvard School of Public Health

August 1, 2006

www.biostat.harvard.edu/~paciorek

Uses of Data and Statistics with Deterministic Models

- Preprocessing: data used in various ways to create and parameterize models
- (Joint processing) Data assimilation
- Postprocessing of model output
  - Model evaluation/assessment
  - Model calibration and model averaging
  - Downscaling (extension)
  - Combining model output and data (integration)

Statistical Themes

- Latent processes and variables representing unknown true state of world
- Methods for combining information based on relative uncertainties in information sources
  - Data and model
  - Multiple models
  - Multiple data sources
- Scales of variability (time and space)
- Characterization of uncertainty and accounting for uncertainty in both models and observations
- Upscaling (easy) vs. downscaling (hard)

Outline

- Model evaluation/assessment
- Calibrating parameters in models and averaging models based on data
  - degree of belief in model: relatively high
- Statistical downscaling
- Combining models and data via statistical representations
  - degree of belief in model: relatively low
  - techniques also useful for low resolution reanalysis data or remote sensing data

Sources of uncertainty

- Model output decomposition
  - $O_t = X_t + M_t + P_t + I_t + S_t + T_t + N_t$
  - X_t true state of nature (a spatial field); M_t model error; P_t parameter error; I_t input/starting value error; S_t smoothing error (from gridding); T_t time averaging error; N_t numerical or approximation error
- Observation decomposition
  - $D_t = X_t + T_t + E_t$
  - X_t true state of nature (a spatial field); T_t time averaging error; E_t measurement error

Exploratory empirical evaluation of model output

- model : data, model : model, low resolution data : data
- individual level:
  - time series plots and maps of observations and of model output
  - scatterplots/regression of observations on model output at same time/location
  - plot deviations in space and time to detect spatio-temporal patterns
  - regress deviations (model-observation) on factors that may explain differences
- summary level:
  - calculation of correlations: aggregate over space or time
  - regress correlations on factors that may explain differences
  - plot correlations in space and time to detect spatio-temporal patterns
- may want to consider observation error in your evaluation (e.g., error bars around observations in plots; analyses with observations weighted by their uncertainty)

Space-time mismatch?

- observations are often point locations whereas model output is areal averages

- observations may be time averages (e.g., EPA daily PM) whereas model output might be shorter time aggregations

- Possible solutions:
  - for spatially smooth quantities, ignore spatial mismatch
  - upscaling
    - average the higher resolution data to the lower resolution, potentially accounting for uneven time and spatial spacing
    - statistically smooth high resolution spatial data, then average smoothed surface over model grid box (Meiring et al. 1998)
    - for latter two approaches, estimate uncertainty level in the manipulated data

Comparison of GOES satellite data with EPA PM observations

- half-hourly GOES aerosol (AOD) observations (with many missing) at 4km resolution

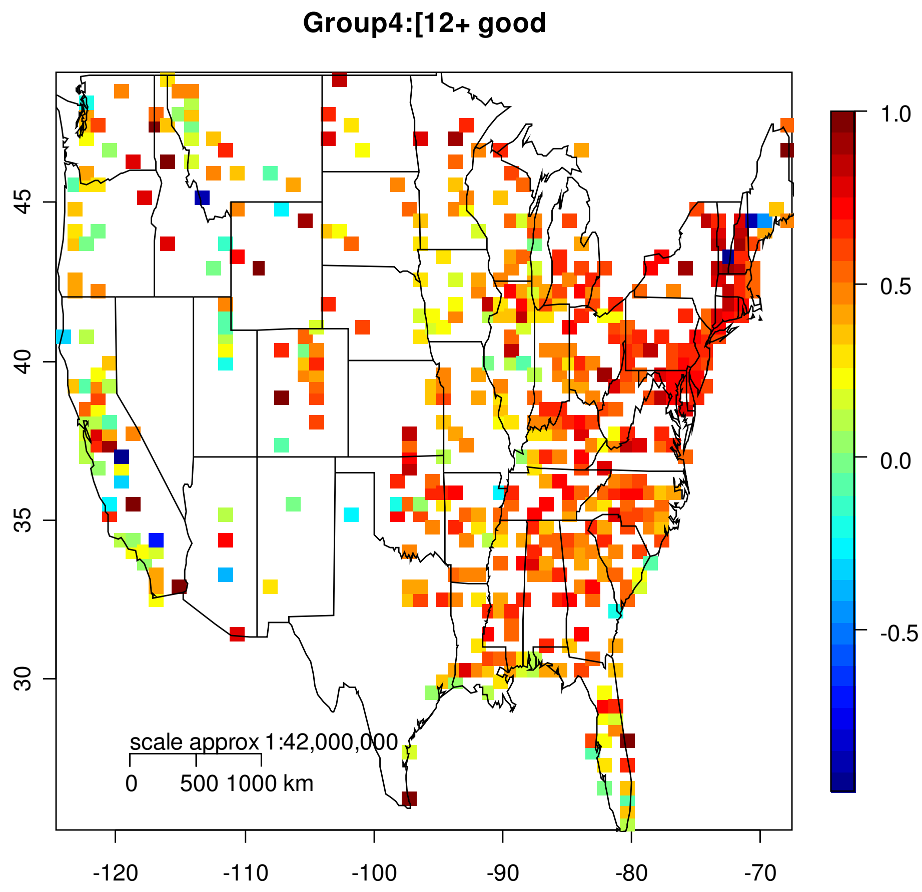
- daily PM observations at point locations

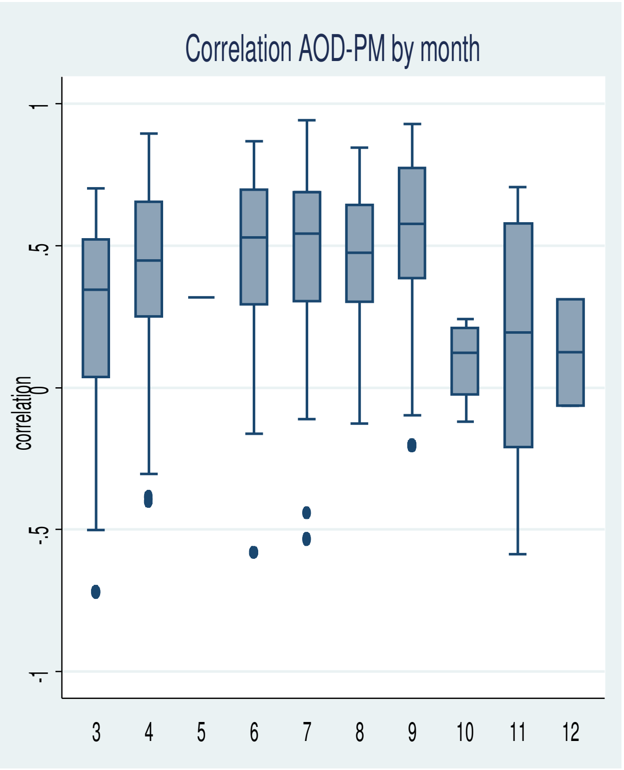
- how strong is the relationship and does the strength of the relationship differ by time and location?

- spatial mismatch: ignored

- temporal mismatch: use time series model to estimate daily AOD accounting for pattern of missing data: $\hat{\mu}_t \neq \bar{D}_t$ but rather a weighted average that upweights observations far from other observations in time (upscaling)

Graphical spatio-temporal comparison





Other approaches to model evaluation

- evaluate space-time correlation structures of data and model output (Jun and Stein 2004)
- build a statistical model that relates model output and data (Fuentes and Raftery 2005)
  - estimate spatio-temporal pattern in bias of model output within statistical model
  - statistical model accounts for data uncertainty and internally calibrates model uncertainty
  - statistical model can build in necessary aggregation to put model and data on same temporal and spatial scale and account for the uncertainty in the aggregated quantities
- more details on building such a model later

Outline

- Model evaluation/assessment
- Calibrating parameters in models and averaging models based on data
  - degree of belief in model: relatively high
- Statistical downscaling
- Combining models and data via statistical representations
  - degree of belief in model: relatively low
  - techniques also useful for low resolution reanalysis data or remote sensing data

Using Data to Improve Models and Model Output

Some degree of trust in the model(s)

- Parameter calibration (Kennedy and O'Hagan 2001)
  - vary parameters and compare fit of model output to data
  - create a posterior distribution over parameter values that reflects uncertainty about parameters based on data
  - $\pi(\theta|D) \propto L(D|M(\theta))\pi(\theta)$
  - average model output over different parameter settings weighted by posterior distribution of parameters
  - a statistical model in which the calibration is done can also provide statistical estimates of remaining model uncertainty

Using Data to Improve Models and Model Output

- Model averaging (Raftery et al. 2005)
  - compare fits of multiple models to data
  - create a posterior distribution over models reflecting model uncertainty based on data
  - $\pi(M_i|D) \propto L(D|M_i)\pi(M_i) = \int L(D|M_i(\theta)\pi(\theta|M_i)\pi(M_i)d\theta$
  - average output from models weighted by posterior probabilities of models
  - $E(f(s,t)) = \sum_i f(s,t|M_i)\pi(M_i|D)$
  - statistical model can account for bias in each model and remaining uncertainty in model average output

Statistical Downscaling

Prediction of fine-resolution features based on coarse-scale information and a statistical model for local effects

- Temporal prediction/extrapolation (probabilistic prediction) for fixed sites at new times
  - regression on model output statistics (MOS) (e.g., Vislocky & Fritsch 1995)
  - weather typing approaches (Bellone et al. 2000, Vrac et al. 2006)
  - stochastic weather generators
- Temporal interpolation for missing time points (e.g., polar-orbiting satellites) (Wikle et al. 2001 - Bayesian model combining reanalysis and finer-resolution satellite data)
- Spatial interpolation at finer scale than observations (e.g., fine-scale PM exposure for epidemiology) (Paciorek, Yanosky, and Suh, in prep.)

Downscaling for temporal prediction/extrapolation

- fixed sites provide data that allow us to related large-scale information to site-specific effects
- e.g., downscaling GCM or reanalysis output to individual sites
- regression on MOS:
  - regression or related techniques (GAM) to relate GCM output variables directly to site specific variables of interest (e.g., precipitation) for training period
  - $Y_{it} = f_i(X_t)$
  - prediction of variables of interest at sites using GCM output variables at new times
  - $Y^*_{it} = f_i(X_{t^*})$

Downscaling for temporal prediction/extrapolation

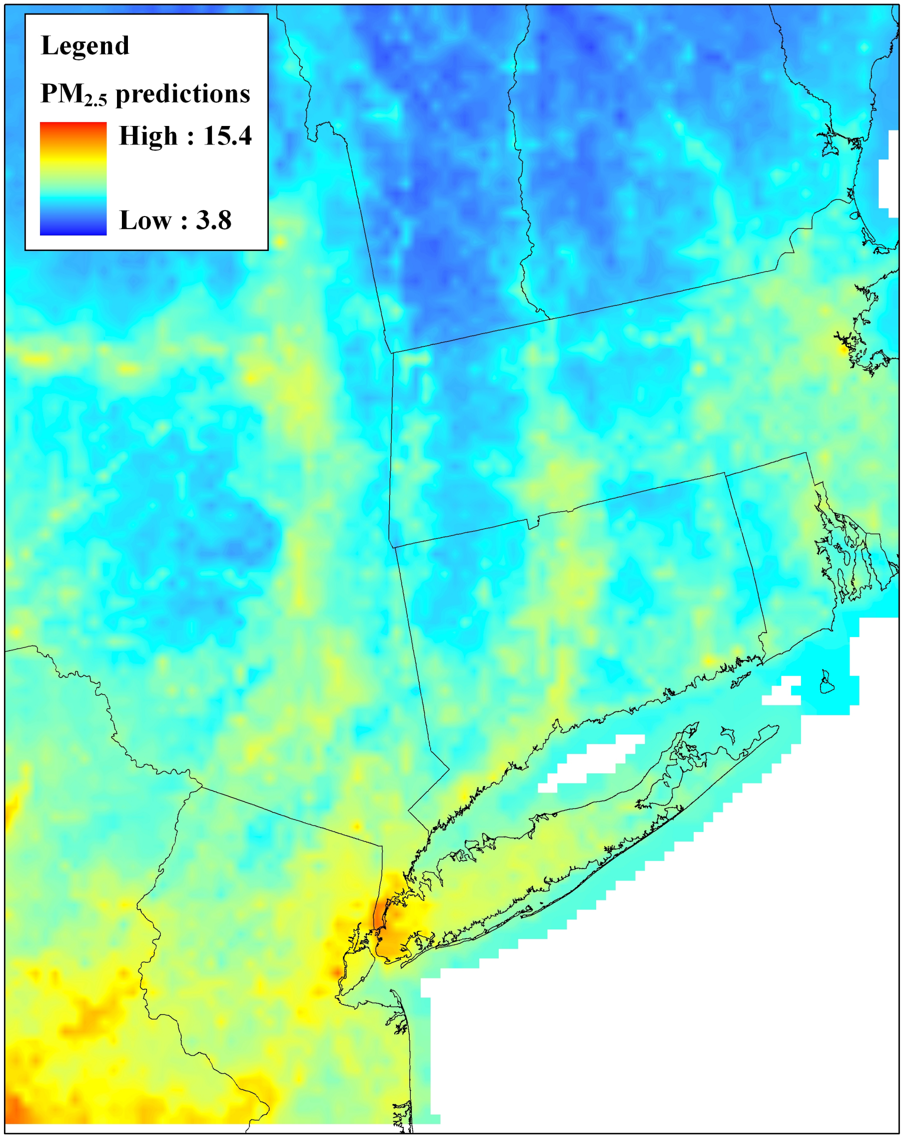
- weather typing (Bellone et al. 2000; Vrac et al. 2006)
  - instead of a giant regression on GCM variables, try to relate GCM variables to a small number of local 'weather' states
  - states defined based on patterns of local variable (e.g., a state of uniform rain; a state with rain in north of region)
  - model weather state transitions as a Markov model influenced by baseline transition probabilities and GCM variables
  - model variable of interest at each site as a regression function of weather state and possibly GCM variables also
  - stage 1: $S_t = f(S_{t-1}, X_t)$ stage 2: $Y_{it} = f_i(S_t)$

- extension of Hughes et al. 1999 approach may allow for spatial interpolation away from fixed sites

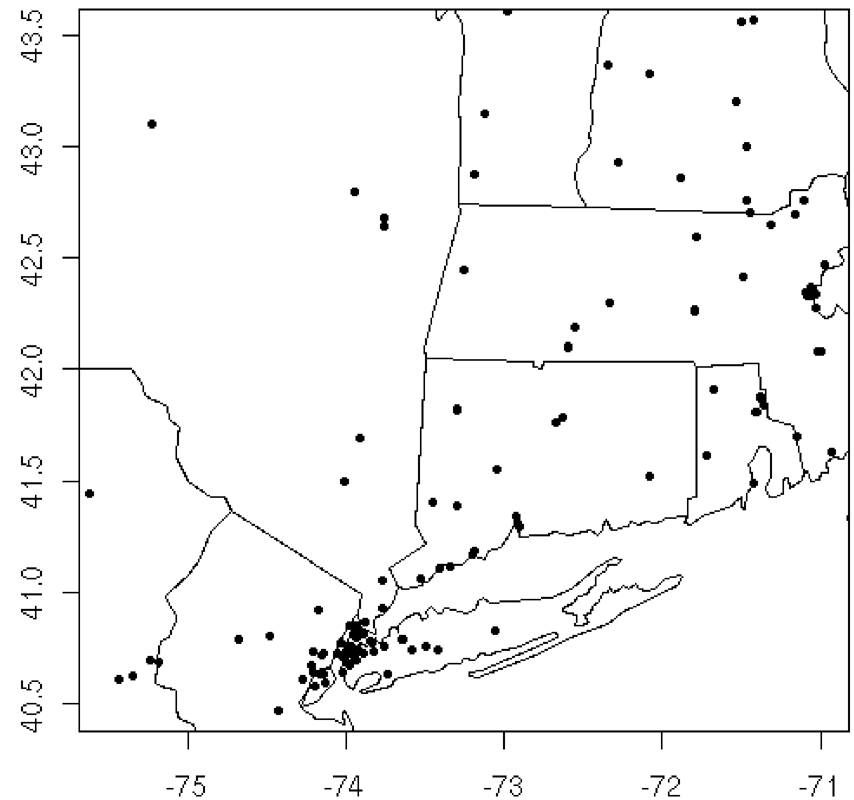
- extrapolation in time relies on assumption that relationships stay constant over time and any changes are caused by changes in the inputs (e.g., GCM variables)

Downscaling for spatial interpolation

- Goal is to predict PM at fine scales for use as exposure in epidemiological models
- Data are EPA PM monitors but pure spatial smoothing is too coarse
- Regression of EPA PM monitoring data on site characteristics and a smooth spatial structure via a generalized additive model: $y_{st} = f_t(s) + \sum_k X_{kt}(s)\beta_k + \epsilon_{st}$
- $g_t(s)$ is spatial smoother that accounts for large scale spatial patterns at time t
- $\sum_k X_{kt}(s)\beta_k$ accounts for local offset based on local characteristics whose effect is assumed to stay constant over time
- possible use of this approach to spatially downscale CMAQ and satellite output for PM prediction



Estimated PM for one month



Monitor locations

Outline

- Model evaluation/assessment
- Calibrating parameters in models and averaging models based on data
  - degree of belief in model: relatively high
- Statistical downscaling
- Combining models and data via statistical representations
  - degree of belief in model: relatively low
  - techniques also useful for low resolution reanalysis data or remote sensing data

Statistical integration/fusion of model output and data

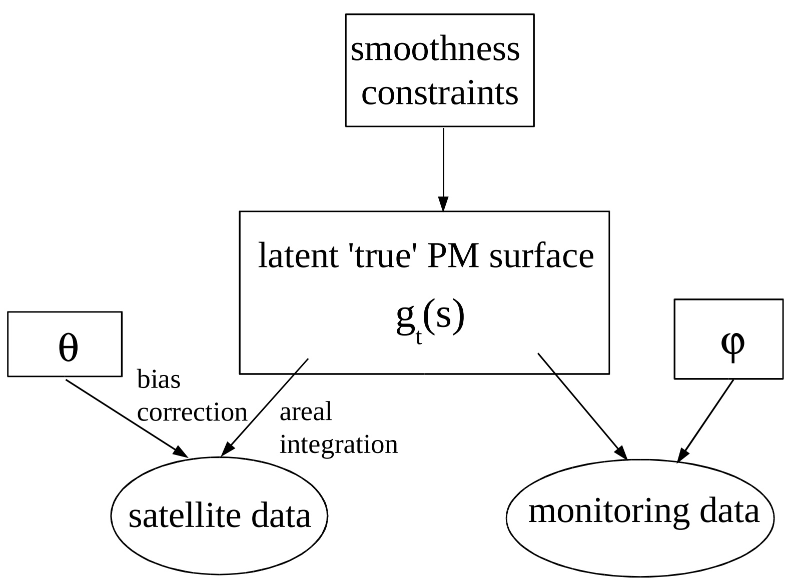
- of greatest potential when trust in model is limited?
- strengths of statistical models that integrate model output and data:
  - best prediction based on all information
  - inherent model evaluation and estimation of model bias
  - account for both model and data uncertainty
  - inherent calibration of uncertainty and uncertainty estimates
  - aggregation consistency can be built into the model
  - model output can be treated as a black box

Possible statistical formulations

- Bayesian statistical model with physical model as prior for latent space-time process
  - $y_{st} = f(s,t) + e_{st} \qquad f(s,t) = a(s,t) + b(s,t)M(s,t)$
- Statistical model for error structure
  - create a spatio-temporal model for $O_{st} - y_{st}$
  - e.g., $O_{st} - y_{st} = f(s,t) + e_{st}$
  - add modelled error back to physical model output to correct the physical model
  - spatio-temporal structure of errors may be simpler than of nature

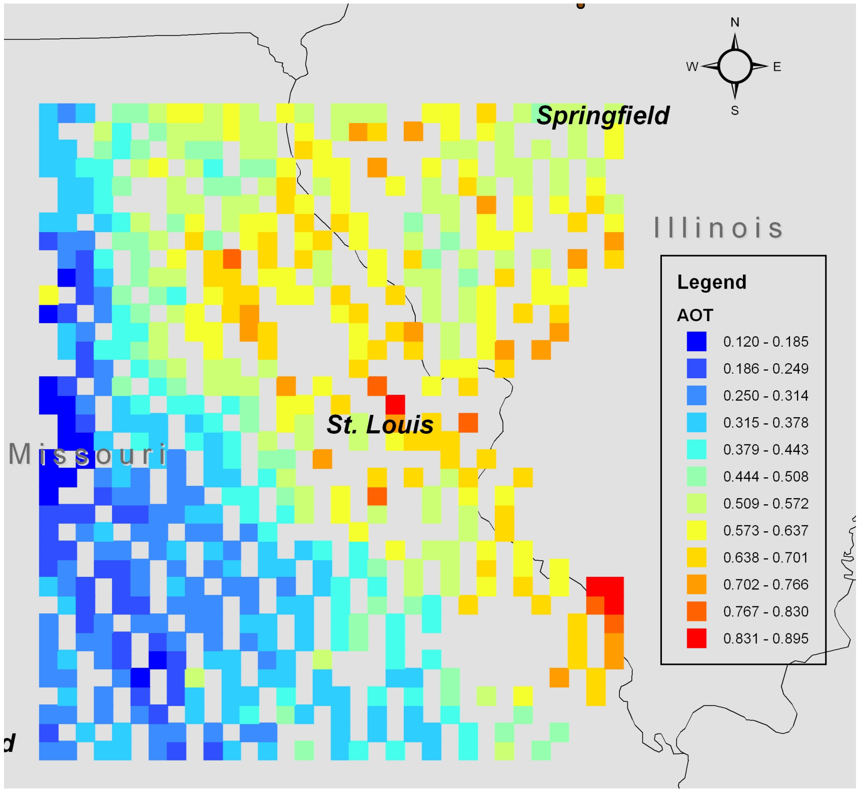
Possible statistical formulations

- Bayesian melding: Bayesian statistical model with observations and physical model treated as 'data' (Fuentes and Raftery 2005)
  - $y_{st} = f(s,t) + e_{st}$
  - $O(area)_t = \int (a(s,t) + b(s,t)f(s,t) + \epsilon(s,t))\, ds$
  - prior distribution for $f(s,t)$, unknown latent process ('true' state of nature)
  - integration accounts for areal aggregation

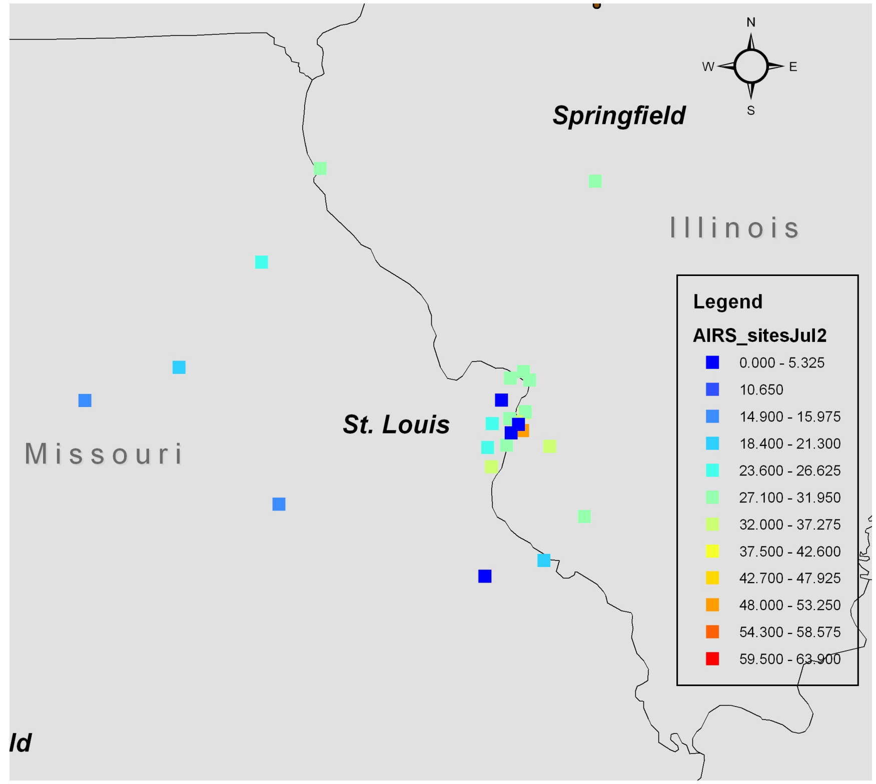


Bayesian melding: Bells and whistles

- statistical technique for combining information sources
- Bayesian statistical models allow for complicated probabilistic relationships and constraints on exposure surfaces
- constraints ensure smooth estimated exposure surfaces and borrow strength to estimate in areas with no data
- incorporate local characteristics to do spatial interpolation (spatial downscaling)
- similar specification with two sets of data, although possibly no bias term and no aggregation
- similar model specification with satellite data instead of physical model



MODIS AOT



$PM_{2.5}$ monitors

Uncertainty considerations

- statistical models can account for uncertainty in a probabilistically rigorous fashion
  - (inputs) weight observations based on certainty
  - (inputs) weight parameter values/models based on certainty
  - (outputs) propagate uncertainty through analysis to final estimates
- uncertainty can be estimated based on:
  - quantification of the levels of uncertainty in the observations (e.g., from instrument manufacturers)
  - repeated measurements or measurements at nearby locations or times
  - ground truth against which to internally calibrate (e.g., model output to observations)

Outline

- Model evaluation/assessment
- Calibrating parameters in models and averaging models based on data
- Statistical downscaling (model extension)
- Combining models and data via statistical representations
  - techniques also useful for low resolution reanalysis data or remote sensing data