# Overview of Spatial Statistical Downscaling

Brian Reich

North Carolina State University

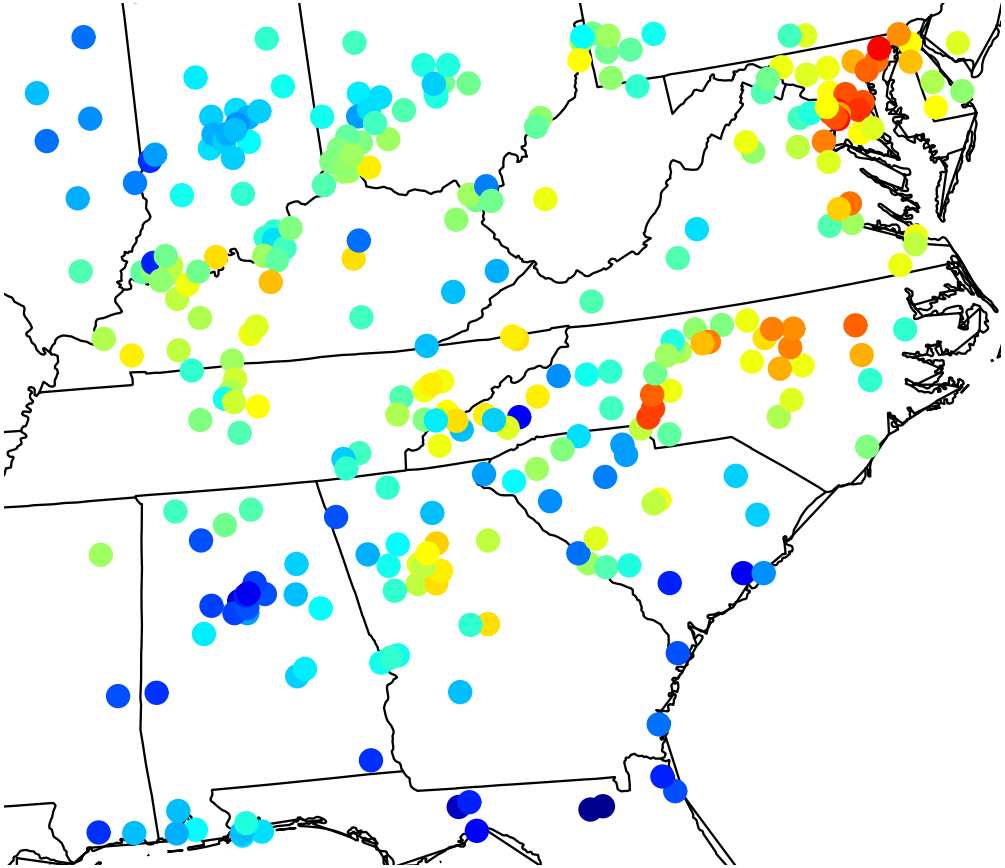
September 28-29, 2017

CAR Post ISEE Workshop
Data/model blending: How the whole can be bigger than the sum of the parts

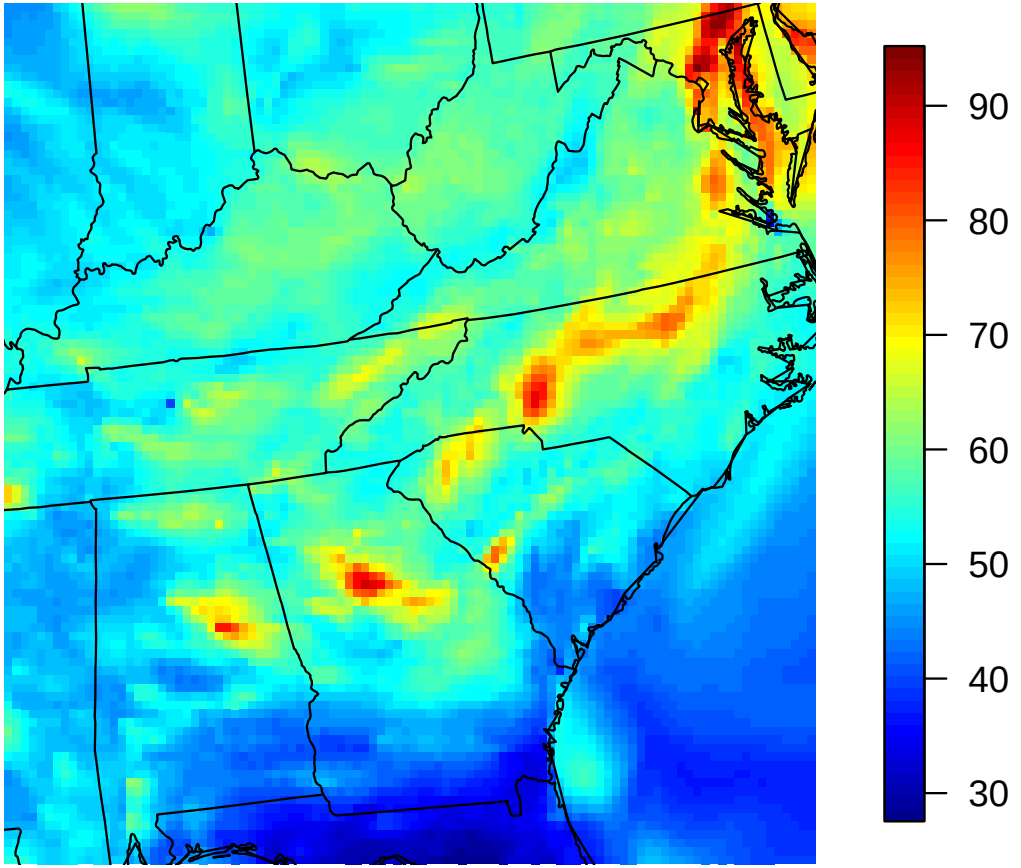
# Statistical downscaling

- Deterministic models (e.g., CTMs) play a key role in environmental epidemiology:
  - Exposure estimates where monitors are unavailable
  - Short-term forecasts
  - Attribution studies
- However, without properly calibrating their output with monitor data the results can be misleading

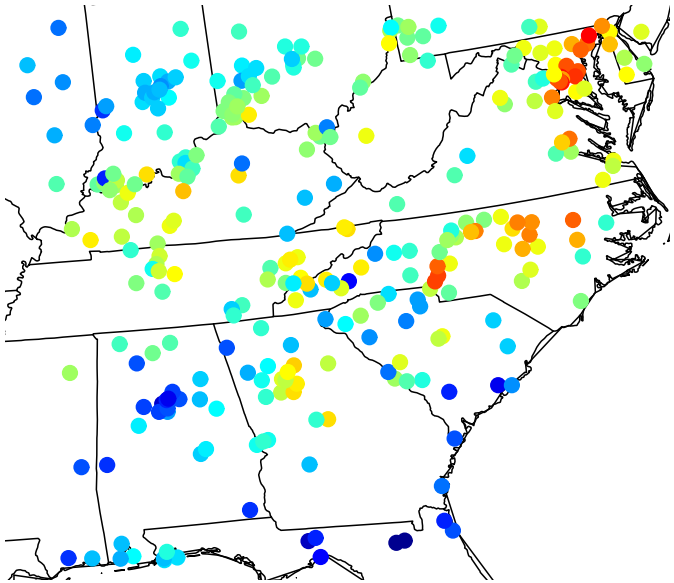
# Monitor data for a one day

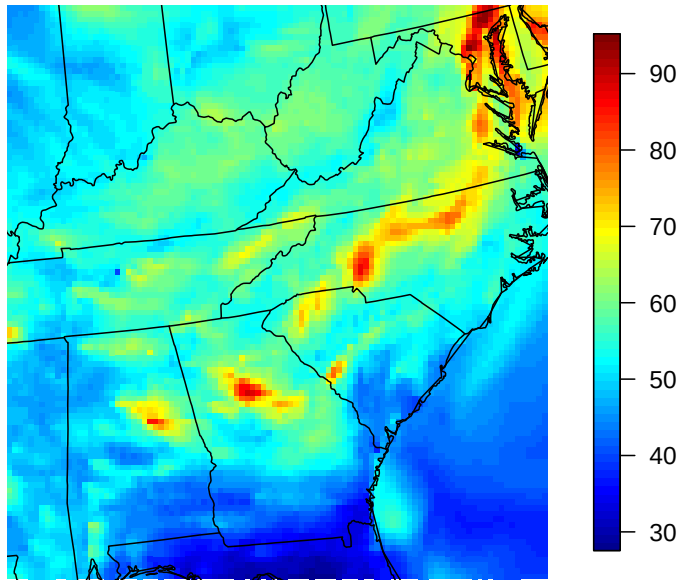


# CMAQ output on that day



# Can we combine monitor (left) and CMAQ (right) data?





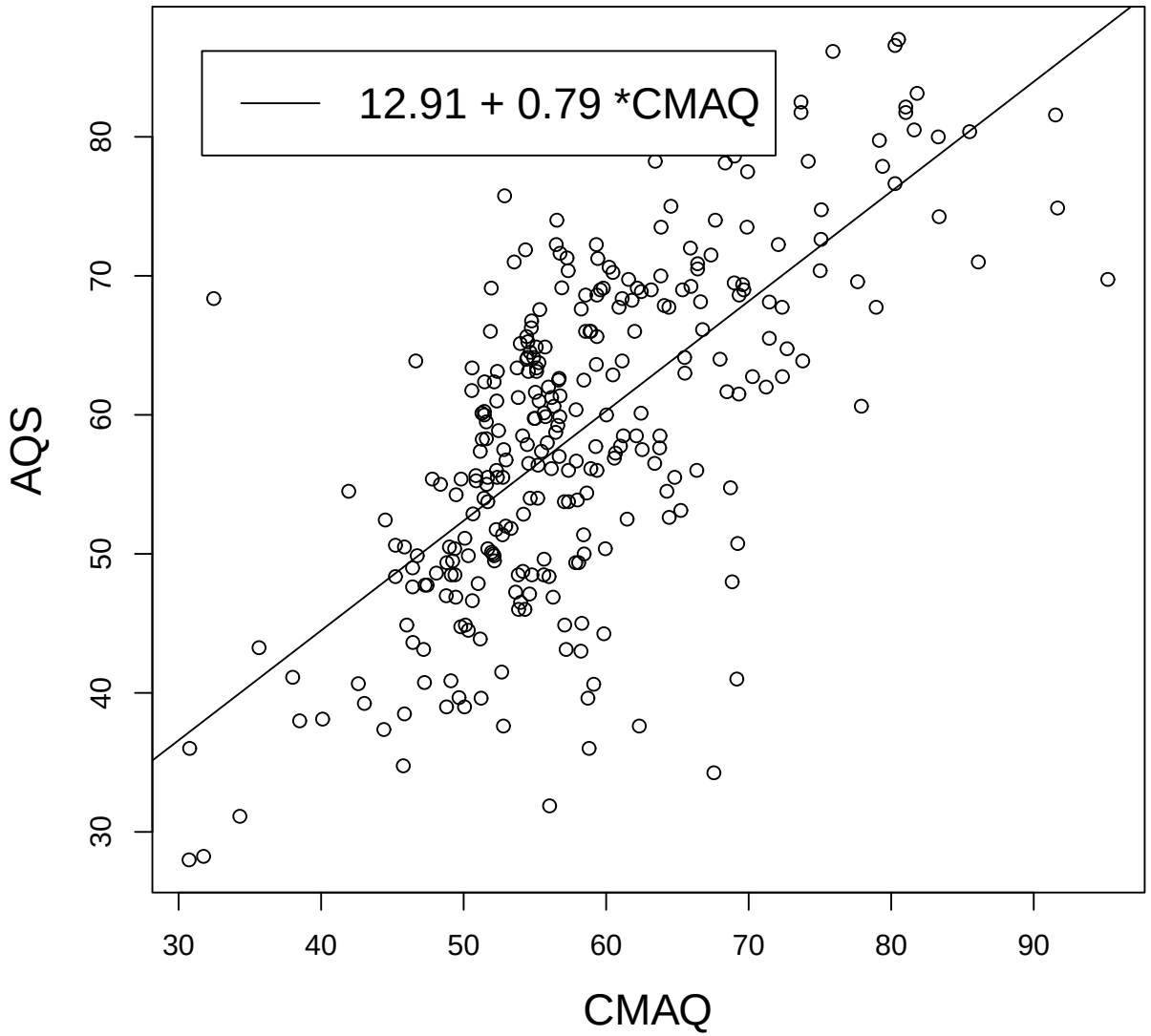
# Simple linear regression (SLR)

- $Y(\mathbf{s})$ is the monitor measurement at spatial location s
- $X(\mathbf{s})$ is the model output for the grid cell containing s
- A common calibration model is

$$Y(\mathbf{s}) = \alpha + X(\mathbf{s})\beta + \varepsilon(\mathbf{s})$$

- The intercept and slope, $\alpha$ and $\beta$, are additive and multiplicative bias terms, respectively.

# CMAQ output (x-axis) versus monitor data (y-axis)



# Sources of bias

- Imprecise inputs
- Numerical approximation
- Spatial resolution
- Change of support problem: monitors measure exposure at a single location, models estimate grid cell averages

# Limitations of SLR

SLR assumes

- the errors $\varepsilon(s)$ are independent - **they are probably spatially-correlated**
- the bias terms are the same for all spatial locations - **they probably vary spatially**
- a linear relationship - **it's probably more complicated**

# Benefits of spatial statistics

- Optimal (Kriging) predictions at locations without data
- We can estimate spatially-varying bias terms
- Parameter estimates are more efficient than SLR
- Standard errors are valid because we appropriately account for spatial correlation

# Fitting a spatial model[1]

$$Y(\mathbf{s}) = \alpha + X(\mathbf{s})\beta + \varepsilon(\mathbf{s})$$

- $Y(\mathbf{s})$ is the response at spatial location s
- $X(\mathbf{s})$ are covariates at s (e.g., model output)
- $\alpha$ and $\beta$ are the regression coefficients, interpreted the same as in non-spatial linear regression
- The Gaussian residuals $\varepsilon(\mathbf{s})$ are spatially correlated

[1] e.g, Berrocal et al (2011), JABES

# Fitting a spatial model

- We model the correlation between two sites as a decreasing function of the distance between them

- The residuals are split into two components

$$\varepsilon(\mathbf{s}) = \theta(\mathbf{s}) + \epsilon(\mathbf{s})$$

- **Nugget**: The pure (uncorrelated) measurement error is

$$\epsilon(\mathbf{s}) \overset{iid}{\sim} \text{Normal}(0, \tau^2)$$

- The spatial errors $\theta(\mathbf{s})$ are correlated

# Fitting a spatial model

- **Partial sill**: The variance of the spatial errors is

$$\text{Var}[\theta(s)] = \sigma^2$$

- **Sill**: The total variance is

$$\text{Var}[\varepsilon(s)] = \sigma^2 + \tau^2$$

- Most analyses assume the correlation between points is:
  - **Stationary**: the same throughout the spatial domain
  - **Isotropic**: depends only on distance between sites

# Fitting a spatial model

- There are many correlation functions (Matern, powered-exponential, spherical, etc.)
- An example is the exponential correlation function

$$\text{Cor}[\theta(\text{s}), \theta(\text{t})] = \exp\left(-\frac{d}{\phi}\right)$$

- Correlation decays exponentially with $d$, the distance between s and t
- **Range**: the parameter $\phi$ controls the range of spatial correlation

# Fitting a spatial model

- The parameters $\alpha$, $\beta$ $\sigma^2$, $\tau^2$ and $\phi$ can be estimated using maximum likelihood estimation
- The `R` package `GeoR` can be used
- There are also Bayesian packages, e.g., `spBayes`
- Estimation can be slow for large datasets because the likelihood involves large matrices

# Spatial prediction

- We use the observed data at the monitors to estimate the model parameters
- Once we have parameter estimates, we can make predictions at other locations
- There are many ways to do this: nearest neighbor, average of observations in a window, etc
- **Kriging** is the optimal method in the sense that it is the Best Linear Unbiased Predictor (BLUP)

# Spatial prediction

- The Kriging prediction at location $s_0$ given the data at $s_1, \ldots, s_n$ is

$$\hat{Y}(s_0) = \mu(s_0) + \sum_{i=1}^{n} w_i [Y(s_i) - \mu(s_i)]$$

- The mean is $\mu(s) = \alpha + \beta X(s)$
- The prediction is a linear combination of the residuals
- The weights $w_i$ are determined by the spatial correlation
- Intuitively, points close to $s_0$ are weighted highest

# Spatial prediction

- Prediction standard deviations have a similar form
- The `R` package `GeoR` performs Kriging
- To make a map, you apply Kriging to a fine grid of points covering the area of interest
- Example: `http://www4.stat.ncsu.edu/~reich/workshop/Ozone_Example.html`

# Spatially-varying coefficients (SVCs)

$$Y(\mathbf{s}) = \alpha(\mathbf{s}) + X(\mathbf{s})\beta(\mathbf{s}) + \varepsilon(\mathbf{s})$$

- $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$ are the spatially-varying coefficients
- The SVCs are assumed to vary smoothly over space
- We need replications over time to estimate these the SVCs
- Bayesian modeling can be used to estimate $\alpha$ and $\beta$ via priors with spatial correlation
- Geographically-weighted regression is a faster option

# Spatially-varying coefficients (SVCs)

Sources of spatial variation in the biases:

- The model was tuned for some areas not others
- The model is missing an input that varies spatially
- Change of support is more problematic in heterogeneous areas

Benefits:

- SVCs can improve spatial prediction
- Studying maps of the SVCs can reveal model deficiencies and lead to refinements

# Advanced topics - distribution matching

- Linear regression calibrates the mean and variance of the model output
- However, it may be important to calibrate other features such as skewness, extreme probabilities, etc
- An approach is to match sample CDFs but this is unstable
- We have used quantile regression[2]

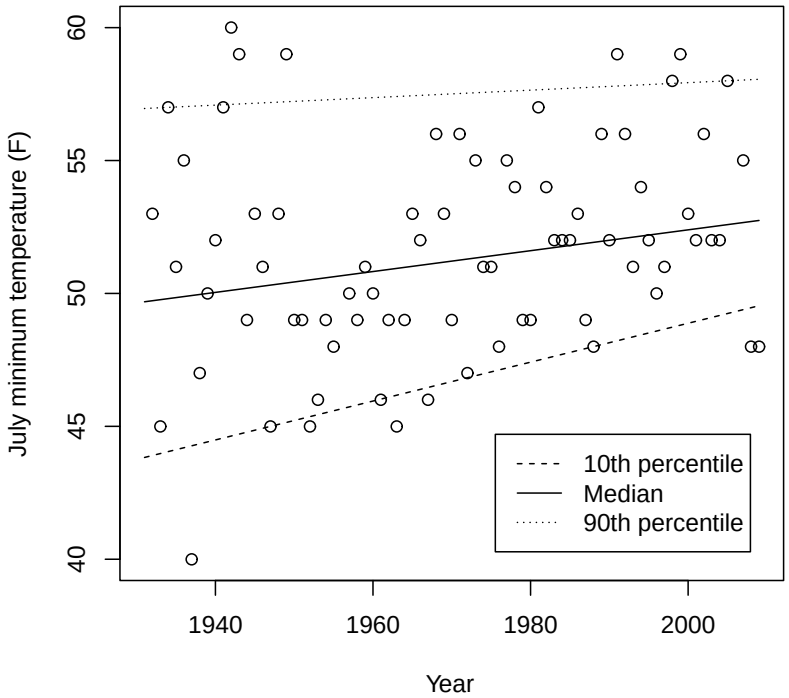
[2] e.g, Koenker (2005)

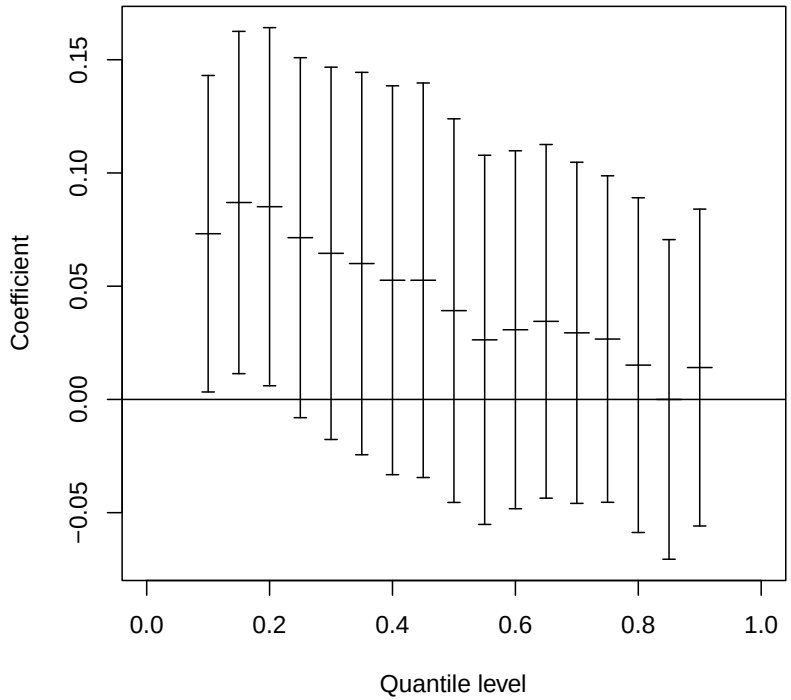
# Advanced topics - distribution matching

- Mean regression: $\mathrm{E}(Y|X) = \alpha + \beta X$
- Let $Q(\tau|X)$ be the $\tau$ quantile of $Y$
  - $Q(0.5|X)$ is the median
  - $Q(0.99|X)$ is the $99^{th}$ percentile
  - $Q(0.75|X) - Q(0.25|X)$ is the IQR
- Quantile regression: $Q(\tau|X) = \alpha(\tau) + \beta(\tau)X$
- The effect of $X$ changes with $\tau$
- We also let the effect change with space

# Quantile calibration

Let $Y_t$ be the value in year $t$.



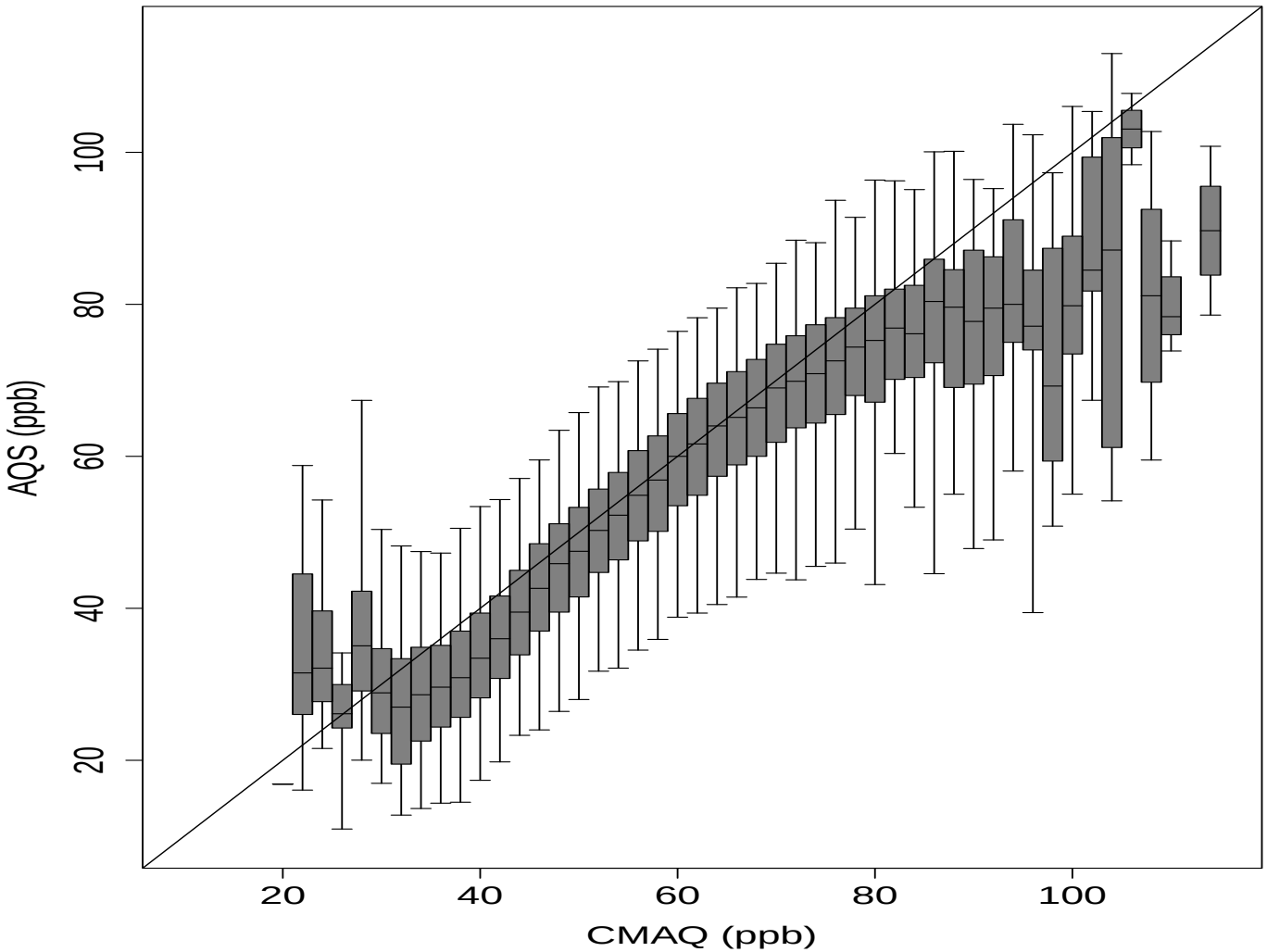

Here the $\tau^{th}$ quantile is modeled as $\beta_0(\tau) + \beta_1(\tau)t$.

# Advanced topics - extreme value analysis (EVA)

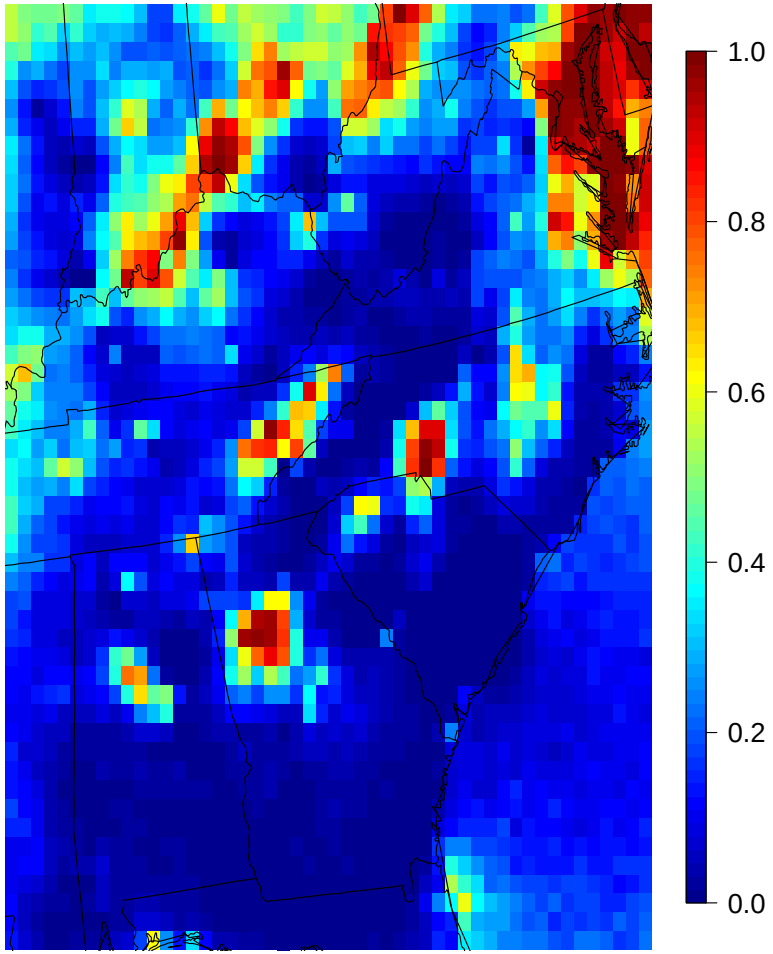
- Often the most important events to model statistically are the extreme events
- For example, we may wish to estimate the 10-year pollution event or test if it is changing over time
- Models are particularly bad at capturing rare events
- We have used EVA to map probabilities of extreme ozone events using model output[3]

[3] Reich et al (2013). AOAS.

# Boxplot of monitor measurements by CMAQ estimates



# Probability of non-compliance* under at 50% reduction in mobile emissions



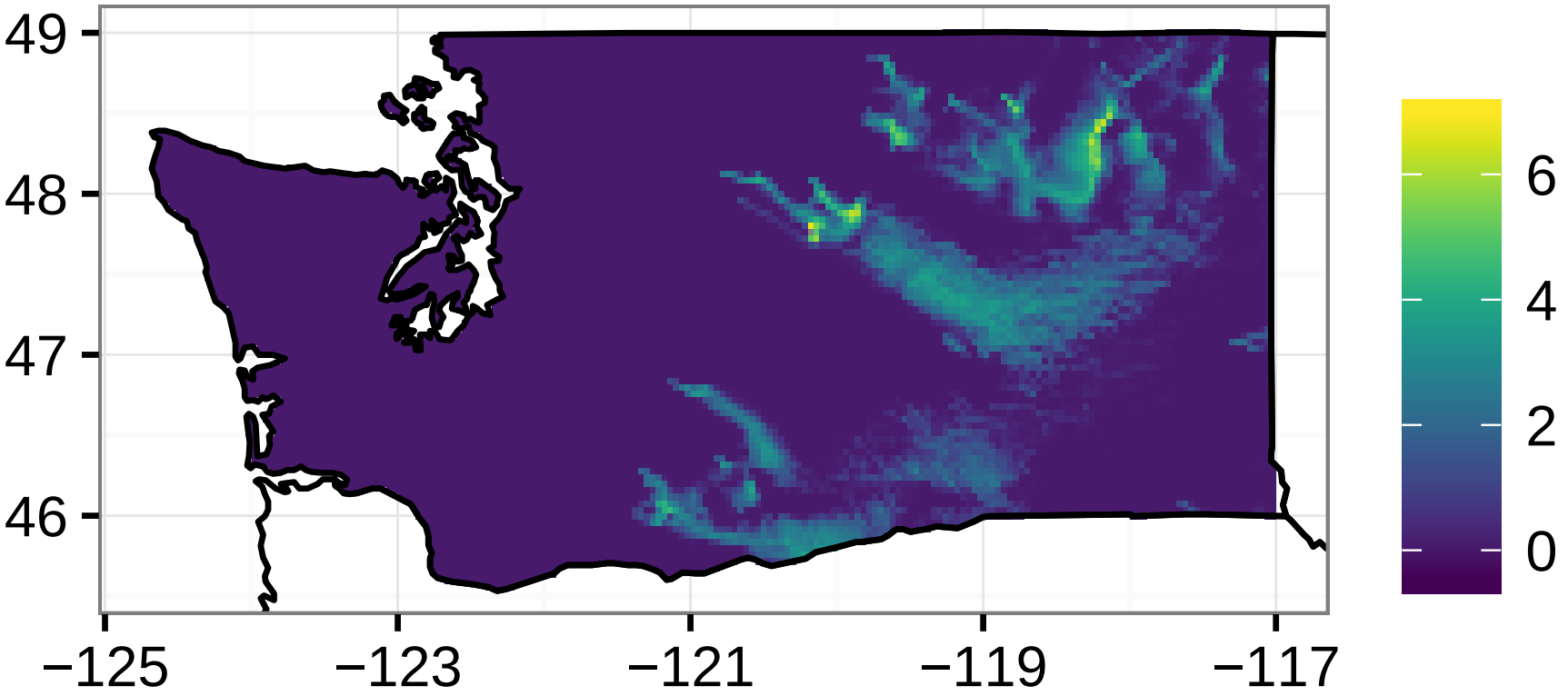

*The $99^{th}$ percentile of daily 8-hour ozone exceeding 70ppb

# Advanced topics - Machine learning

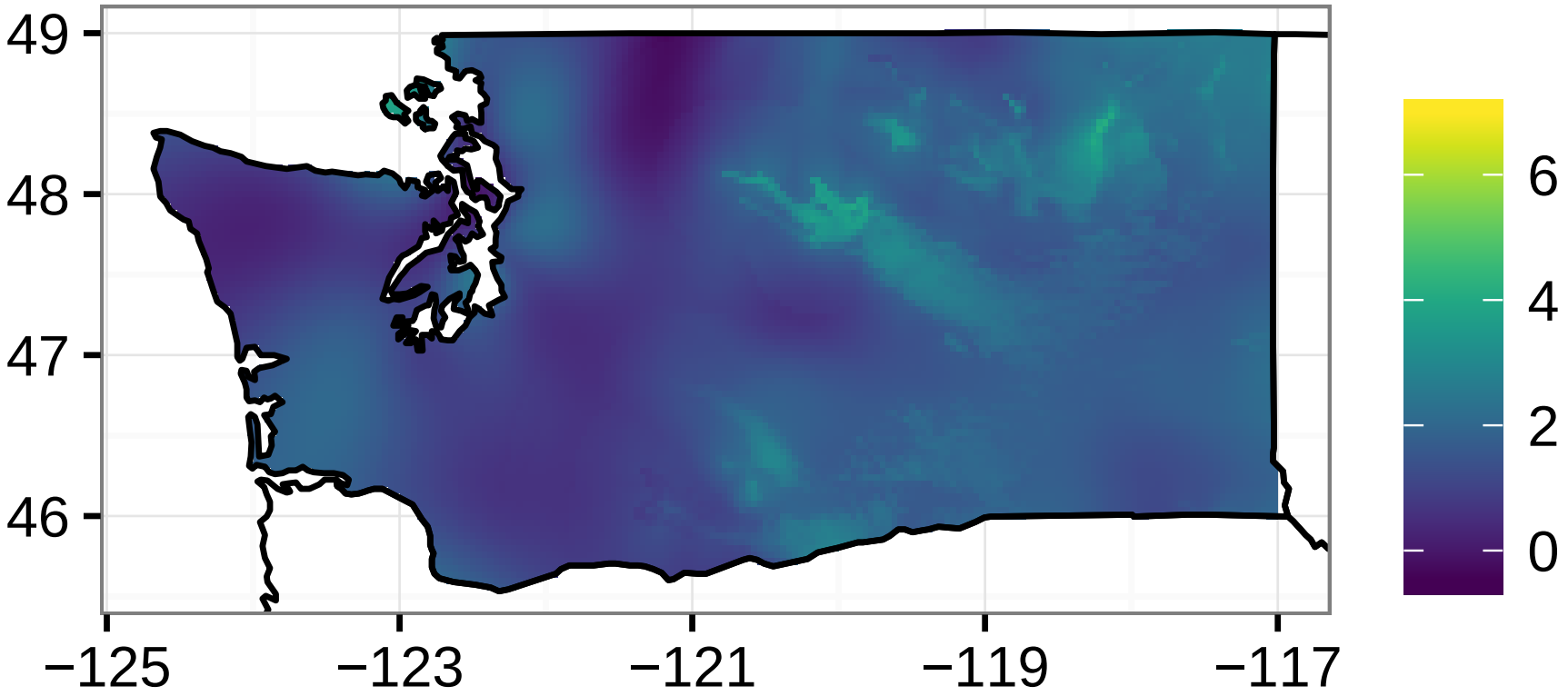
- Machine learning is everywhere these days, and it has begun to permeate statistical downscaling
- The idea is that we assume some super flexible model for the mean of $Y$ given $X$ and let the (big) data figure out this (non-linear) relationship
- Methods include neural networks, random forests, etc.
- We[4] have use generalized additive models that allow for spatially-varying non-linear relationships
- The model uses forecasts at all cells within a radius of the monitor

[4] Wei et al (2017), submitted

# HYSPLIT forecast of forest fire PM2.5



# Calibrated forecast



# Prediction MSE

| Neighbors | SVC | Residuals | Linear | Non-linear |
|---|---|---|---|---|
| No | No | Independent | 0.295 | 0.313 |
| | | Spatial | 0.290 | 0.307 |
| | Yes | Independent | 0.274 | 0.280 |
| | | Spatial | 0.271 | 0.288 |
| Yes | No | Independent | 0.293 | 0.270 |
| | | Spatial | 0.289 | 0.267 |
| | Yes | Independent | 0.255 | 0.248 |
| | | Spatial | 0.250 | **0.243** |