

# Extreme value analysis for evaluating ozone control strategies

Brian J. Reich, NC State

Dan Cooley, Colorado State

Kristen Foley and Sergey Napelenok, US EPA

Ben Shaby, UC-Berkeley

# Spatial quantile regression (QR)

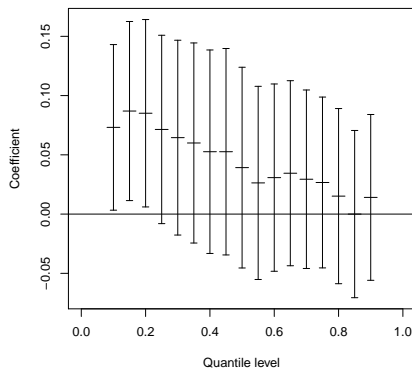
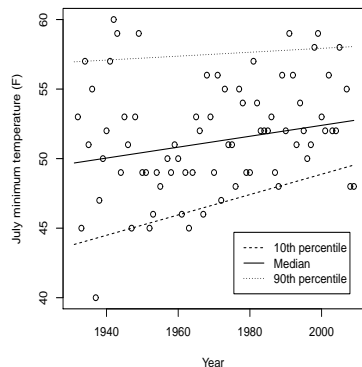
- ▶ Climate change may lead to changes in several aspects of the distribution of climate variables, including changes in the mean, increased variability, and severity of extreme events.
- ▶ We propose using spatiotemporal QR as a flexible and interpretable method for simultaneously detecting changes in several features of the distribution of climate variables.
- ▶ We apply this method to study changes in temperature in the eastern US.
- ▶ Statistical downscaling also requires analysis of several features of the predictive distribution.
- ▶ We use QR to calibrate a deterministic model to capture extreme air pollution events.

# Quantile regression

- ▶ Mean regression:  $E(Y) = X^T \beta$ .
- ▶ Quantile regression:  $Q(\tau) = X^T \beta(\tau)$ , where  $Q(\tau)$  is the  $\tau^{th}$  quantile of  $Y$  so that  $P[Y < Q(\tau)] = \tau \in [0, 1]$ .
- ▶  $\beta(\tau)$  gives the covariate effects on the  $\tau^{th}$  quantile.
- ▶ With  $\tau = 0.5$ , this is median regression.
- ▶ With  $\tau = 0.99$ , this allows us to test for the effects of covariates on the magnitude of extreme values.
- ▶ The covariate will be the year for the climate change application, and the deterministic model output for the downscaling application.

# July minimum temperature data by year

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



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

# Spatiotemporal quantile regression

Let  $Y_t(s)$  be the response in year  $t$  at location  $s$ .

- ▶ We model the  $\tau^{th}$  quantile of  $Y_t(s)$  as

$$Q_t(\tau|s) = \beta_0(\tau|s) + \beta_1(\tau|s)t.$$

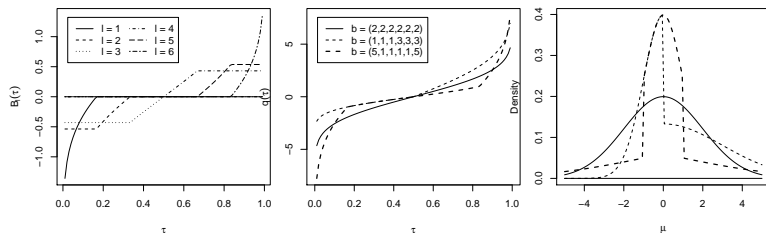
- ▶ All sites and quantiles are modeled simultaneously to borrow strength where appropriate.
- ▶ We perform inference on the evolution of different aspects of the distribution.

## Basis expansion for $\beta_0$ and $\beta_1$

Let  $\beta_j(\tau|s) = \theta_{j0}(s) + \sum_{l=1}^L B_l(\tau)\theta_{jl}(s)$ .

- ▶  $\theta_{j0}(s)$  is a location parameter (the median for this model).
- ▶  $B_l(\tau)$  are known basis functions.
- ▶  $\theta_{jl}(s)$  determine the shape of the quantile function.
- ▶ The  $\theta_{jl}(s)$  vary spatially to allow for a different distribution at each site.

# We use piecewise Gaussian basis functions



- ▶ Left: basis functions  $B_l(\tau)$  with  $L = 6$ .
- ▶ Middle: quantile function  $\beta_0(\tau) = \sum_{l=1}^L B_l(\tau)b_l$ .
- ▶ Right: The corresponding density with  $t = 0$ .

## Properties of these basis function

- ▶ If  $\theta_{jl}(s)$  are the same for all basis functions  $l = 1, \dots, L$ , then  $Y_t(s) \sim N[\theta_{j0}(s), \theta_{j1}(s)^2]$ .
- ▶ In this sense, the model is centered on the heteroskedastic Gaussian spatial model.
- ▶ The model is flexible enough to match any  $L$  quantiles exactly.
- ▶ The tails are like Gaussian tails.

# Spatial model

- ▶ In order for the quantile function to be valid, it must be increasing in  $\tau$  for all  $t$  and  $s$ .
- ▶ For this choice of basis function, this happens if and only if
  - ▶ We restrict  $t$  to a finite interval, e.g.,  $t \in [0, 1]$
  - ▶  $\theta_{0l}(s) > 0$  for all  $l = 1, \dots, L$
  - ▶  $\theta_{0l}(s) + \theta_{1l}(s) > 0$  for all  $l = 1, \dots, L$
- ▶ To ensure the constraints are satisfied, let  $\theta_{jl}^*$  be independent Gaussian processes.
- ▶ We take

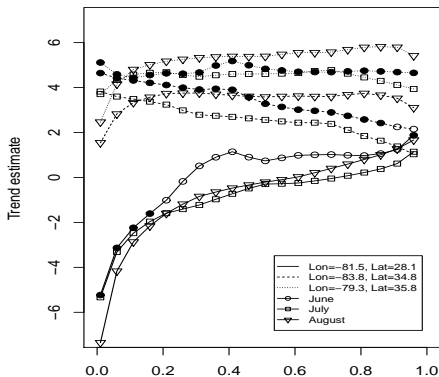
$$\theta_{jl} = \theta_{jl}^* I[\theta_{0l}(s) > 0 \text{ and } \theta_{0l}(s) + \theta_{1l}(s) > 0] + \epsilon.$$

# Residual correlation

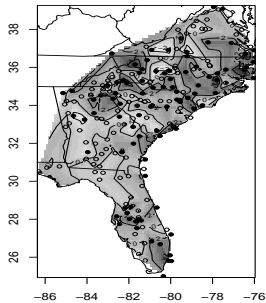
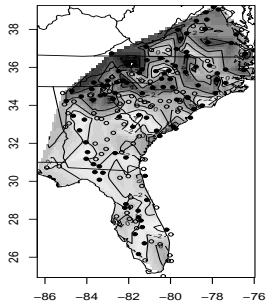
- ▶ Allowing the quantile function to vary spatially accounts for some spatial dependence.
- ▶ However, there remains residual spatial dependence.
- ▶ We account for this with a Gaussian copula.
- ▶ Let  $Q_t(\tau|s)$  be the quantile function of  $Y_t(s)$  and  $\epsilon_t(s)$  be a latent Gaussian process with mean zero and variance one.
- ▶ We take  $Y_t(s) = Q_t[U_t(s)|s]$  where  $U_t(s) = \Phi[\epsilon_t(s)]$ .
- ▶ The likelihood has a fairly simple closed-form.

# Results

- ▶  $Y_t(s)$  is the minimum temperature for July in year  $t$  at location  $s$ .
- ▶ Data are used from 1931-2009 at 191 sites in SE US.

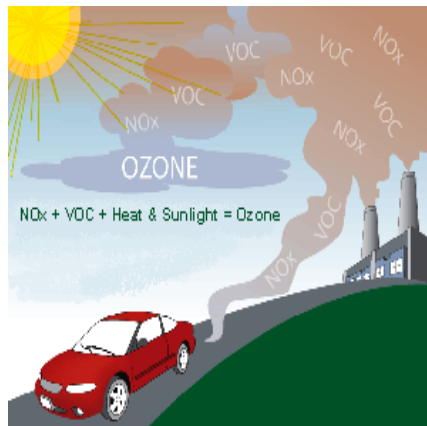


# $\beta_1(\tau|s)$ for $\tau = 0.1, 0.9$



# Changing gears: Tropospheric ozone

- ▶ Tropospheric ozone has been linked with several adverse health effects and is regulated by the EPA.
- ▶ Ozone is primarily a secondary pollutant.



# Evaluating control strategies

- ▶ The EPA is interested in evaluating the effects of control strategies.
- ▶ For example, if emissions from cars is reduced by 50%, how much would this reduce ozone?
- ▶ In particular, they are interested in extremes because ozone is regulated based on the fourth highest day of the year (0.99 quantile).
- ▶ Because ozone is a secondary pollutant, this is virtually impossible to answer without an atmospheric chemistry model.
- ▶ For this (and other) purposes, scientists have developed a numerical model called CMAQ (Community Multiscale Air Quality).

# CMAQ is a deterministic model

- ▶ The inputs include
  - ▶ Meteorology
  - ▶ Emissions of different types of from different sources
  - ▶ Many others...
- ▶ Output
  - ▶ Hourly surface ozone concentration on a 12km by 12km over one summer.
- ▶ For one set of inputs this takes hours/days to run.

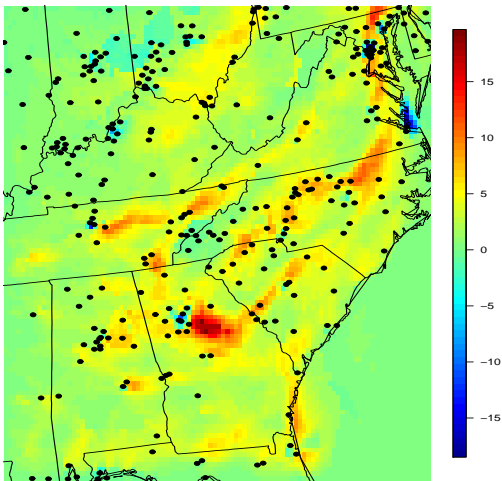
# Reduced-form CMAQ

To overcome the computational difficulty, a second-order approximation has been developed which can be evaluated for new set of inputs virtually immediately:

$$C(t, s | \alpha) = C_0(t, s) + \sum_{j=1}^d S_j^{(1)}(t, s) \alpha_j + \sum_{k \leq j} S_{jk}^{(2)}(t, s) \alpha_j \alpha_k$$

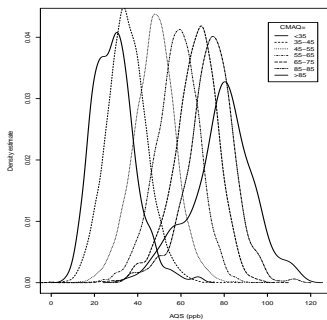
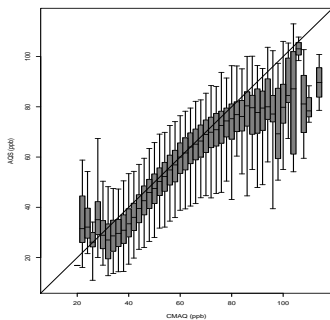
- ▶  $C_0(t, s)$  is the output of a base run using the best guess of the inputs.
- ▶  $\alpha_j$  is the percent change in input  $j$ . We consider perturbing  $d = 6$  inputs: mobile-, point-, and other-source  $\text{NO}_x$ , anthropogenic and biogenic VOCs, and ozone boundary conditions.
- ▶  $S^{(1)}$  and  $S^{(2)}$  are sensitivities to perturbations in the inputs.

# Sensitivity to mobile-source $\text{NO}_x$ for one day, $S^{(1)}(1, s)$



# Data fusion with monitored data

Given the base CMAQ value  $C_0(t, s)$ , there is considerable uncertainty in the monitor data  $Y(t, s)$ .



# Our approach

- ▶ Our approach is to model the conditional distribution of the data  $Y(t, s)$  given CMAQ  $C(t, s|\alpha)$ .
- ▶ Given this conditional model, we can generate new simulated data sets for CMAQ at new values of  $\alpha$ .
- ▶ In particular, we hope to estimate the upper tail of  $Y(t, s)|C(t, s|\alpha)$  well.
- ▶ We also need a good model for the rest of the conditional distribution, since conditioned on large  $C(t, s|\alpha)$ , the median of  $Y(t, s)|C(t, s|\alpha)$  may be “extreme”.
- ▶ We use a mix of flexible quantile regression for the center of the distribution where data are abundant, and extreme value methods for the tails where data are sparse.

# Quantile regression model for downscaling

- ▶ Let  $Q_0(\tau|t, s)$  be the  $\tau^{th}$  quantile of  $Y(t, s)|C(t, s|\alpha)$ .
- ▶ We assume  $Q_0(\tau|t, s) = \beta_0(s) + \beta_1(s)C(t, s|\alpha)$ , where  $\beta_0(s)$  and  $\beta_1(s)$  are modeled as before.
- ▶ This model has Gaussian tails, which may be insufficient for extreme value analysis.

# Quantile regression model for downscaling

We model the tail of the conditional distribution using a Generalized Pareto distribution (GPD).

- ▶ The GPD has quantile function
$$qGP(\tau|\mu, \sigma, \xi) = \mu + \frac{\sigma}{\xi} ([1 - \tau]^{-\xi} - 1).$$
- ▶  $qGP(0|\mu, \sigma, \xi) = \mu$ , therefore  $\mu$  is the lower bound.
- ▶  $\sigma$  is the scale.
- ▶  $\xi$  is the shape.
- ▶ The density is unbounded if  $\xi > 0$ , and has upper bound  $= \mu - \frac{\sigma}{\xi}$  if  $\xi < 0$ .

## Our model for the quantile process

$$Q(\tau|C, s) = \begin{cases} Q_0(\tau|C, s) & \tau \leq T(C) \\ qGP \left[ \frac{\tau - T(C)}{1 - T(C)} \mid, \mu(C, s), \sigma(C, s), \xi(C) \right] & \tau > T(C). \end{cases}$$

- ▶  $C$  is short for  $C(t, s|\alpha)$ .
- ▶  $T(C) \in (0, 1)$  is the quantile level that separates the semiparametric quantile function  $Q_0$  from the parametric GP quantile function.
- ▶ The GPD lower bound is fixed at  $\mu(C, s) = Q_0[T(c)|C, s]$ .
- ▶ The GPD shape and scale depend on  $C$ .

## GPD parameters as a function of CMAQ

- ▶  $\xi(C)$  and  $\log[\sigma(C)]$  are  $M^{\text{th}}$  order polynomials in  $C$ .
- ▶  $T(C) \in [l, u]$  where  $l$  and  $u$  are unknown parameters with  $l \sim \text{Uniform}(0.8, 1.0)$  and  $u|l \sim \text{Uniform}(l, 1.0)$ .
- ▶ The variability of  $T(C)$  within  $(l, u)$  is modeled using the logistic link

$$T(C) = l \frac{\exp[d(C)]}{1 + \exp[d(C)]} + u \frac{1}{1 + \exp[d(C)]}.$$

- ▶ As with the other parameters,  $d(C)$ , and thus  $T$ , varies with  $C$  as a  $M^{\text{th}}$  order polynomial.

# Residual dependence

- ▶ A copula could be used here as well.
- ▶ However, this data set is very large, and so we fit the model in two stages.
- ▶ We first estimate the parameters in  $Q(\tau|t, s)$  assuming the responses are independent.
- ▶ Given estimates of these parameters, we estimate the residual correlation parameters.

# Model comparison

We compare models with quantile scores and Brier scores.

- ▶ The data are split randomly into training and testing sets, each with 13K observations.
- ▶ The quantile rank score for quantile level  $\tau$  is  $2 \{I[y < \hat{q}] - \tau\} (\hat{q} - y)$ :
  - ▶  $y$  is the test set AQS value
  - ▶  $\hat{q}$  is its estimated  $\tau^{th}$  quantile
- ▶ The Brier score for evaluating accuracy of predicting exceedance of threshold  $c$  is  $[e(c) - P(c)]^2$ :
  - ▶  $e(c) = I(y > c)$  is the indicator that the test set AQS value exceeds  $c$
  - ▶  $P(c)$  is the predicted probability of an exceedance.

# Quantile scores

Quantile level	Gaussian				Non-Gaussian (L=4)			
	$M = 1$		$M = 2$		$M = 1$		$M = 2$	
	NoGP	GP	NoGP	GP	NoGP	GP	NoGP	GP
0.75	5.365	5.350	5.327	<b>5.304</b>	5.353	5.353	5.313	5.321
0.95	1.848	1.841	1.836	<b>1.814</b>	1.841	1.840	1.825	1.826
0.99	0.521	0.507	0.518	<b>0.498</b>	0.520	0.505	0.515	0.502
0.995	0.302	0.287	0.299	<b>0.284</b>	0.301	0.287	0.299	0.289

This criteria recommends the model with GP tail, quadratic functions of  $C$  ( $M = 2$ ), and a Gaussian model below the threshold.

# Brier scores (multiplied by 100) for exceedances

Threshold	Gaussian				Non-Gaussian (L=4)			
	$M = 1$		$M = 2$		$M = 1$		$M = 2$	
	NoGP	GP	NoGP	GP	NoGP	GP	NoGP	GP
70	6.179	6.166	6.127	<b>6.105</b>	6.187	6.207	6.124	6.141
75	3.886	3.886	3.788	<b>3.762</b>	3.901	3.940	3.796	3.806
80	2.105	2.108	1.975	<b>1.951</b>	2.122	2.150	1.978	1.973
85	0.985	0.999	0.866	<b>0.852</b>	0.997	1.022	0.859	<b>0.852</b>
90	0.427	0.440	0.350	0.344	0.430	0.445	0.343	<b>0.341</b>
95	0.225	0.229	0.184	<b>0.182</b>	0.226	0.229	<b>0.182</b>	<b>0.182</b>
100	0.098	0.098	0.078	<b>0.077</b>	0.098	0.098	0.079	0.078

This criteria also recommends the model with GP tail, quadratic functions of  $C$  ( $M = 2$ ), and a Gaussian model below the threshold.

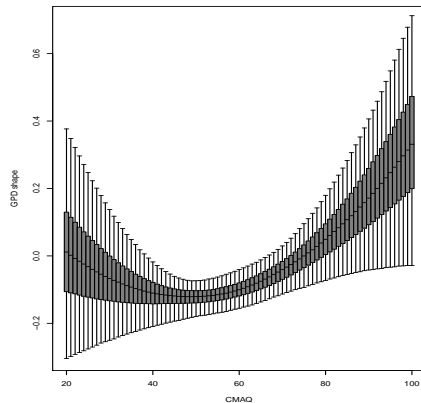
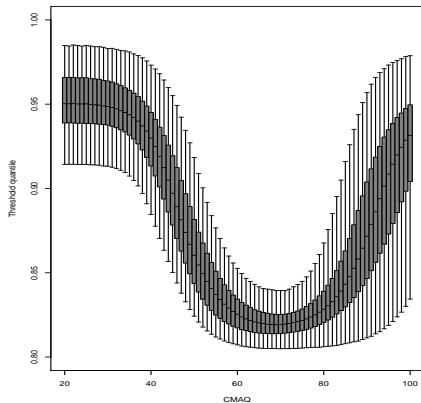
# Evaluating control strategies

- ▶  $\alpha$  corresponds to the adjustment to the initial emissions after comparing CMAQ to monitor data.
- ▶ Control strategies are parameterized by  $\phi$ , which gives the additional reduction in emissions after the adjustment by  $\alpha$ .
- ▶ E.g., a 50% reduction in mobile-source  $\text{NO}_x$  is  $\phi = (-0.5, 0, 0, 0, 0, 0)$ , and we use  $\alpha_j^* = (1 + \alpha_j) * (1 + \phi_j) - 1$  as inputs to the RFM.
- ▶ Given  $\alpha^*$ , we can compute  $C(t, s | \alpha^*)$  over space and time and then generate  $Y(t, s)$  from the conditional distribution of  $Y(t, s) | C(t, s | \alpha^*)$ .
- ▶ To do this, we generate  $z(t, s)$  from an AR(1) at each site, and then set  $U = \Phi(z)$  and  $Y = Q(U | C)$ .

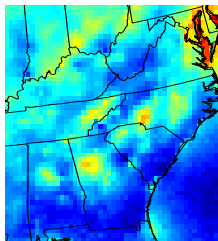
# Evaluating control strategies

- ▶ For each location, we generate 10,000 “ozone seasons” using this approach.
- ▶ Each replication uses a different posterior draw for  $\alpha$  and the parameters in the quantile function.
- ▶ For each season we compute the 4th highest day of the year, which is the value used for regulation.
- ▶ We present the average of the fourth highest day as well as the proportion above 75ppb, the current EPA cutoff.

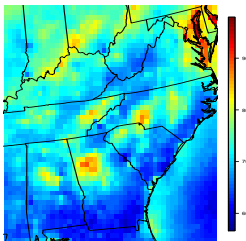
# GP shape, $\xi(C)$ (left), and threshold, $T(C)$ (right)



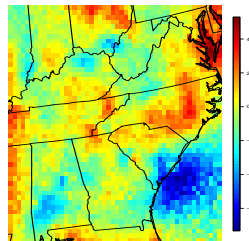
# Base case: average of the 4th highest day



With GP tails

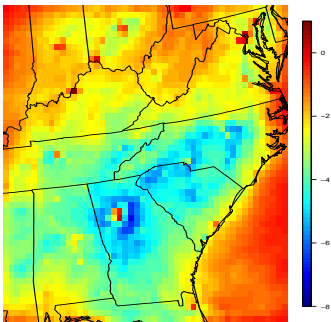


Without GP tails

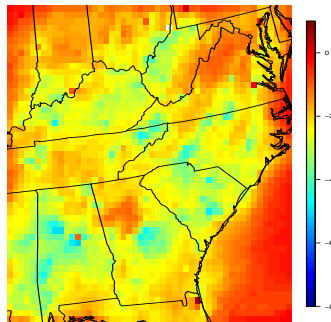


Difference

# Difference in the 4th highest day of the year between control strategies

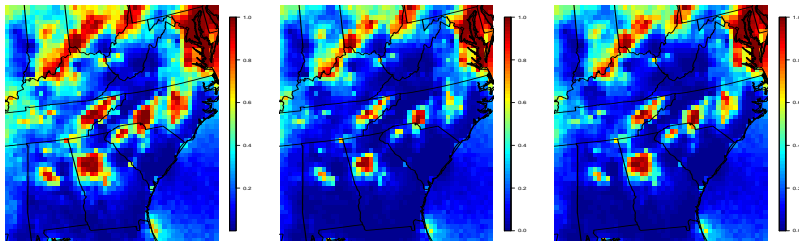


50% reduction in mobile-source  
NO<sub>x</sub> - the base case



50% reduction in point-source  
NO<sub>x</sub> - the base case

# Probability of exceeding the standard



Base case

50% reduction in  
mobile-source NO<sub>x</sub>

50% reduction in  
point-source NO<sub>x</sub>

# Caveats

This method accounts for several sources of uncertainty:

1. Uncertainty in the measurements given model output
2. Uncertainty in the parameters in the statistical model
3. Uncertainty in emissions via  $\alpha$  (partially)

This insufficiently accounts for other sources of uncertainty:

1. Daily variation in emissions and meteorology
2. Spatial variation in emissions calibration

A longer run of CMAQ would help with (1), spatially-varying  $\alpha$  would address (2).