Bayesian spatial quantile regression

Brian J. Reich[a1], Montserrat Fuentes[a], and David B. Dunson[b]

[a] Department of Statistics, North Carolina State University

[b] Department of Statistical Science, Duke University

April 16, 2009

Institute of Statistics Mimeo Series #2624

Abstract

Tropospheric ozone is one of the six criteria pollutants regulated by the US EPA under the Clear Air Act and has been linked with several adverse health effects, including mortality. Due to the strong dependence on weather conditions, ozone may be sensitive to climate change and there is great interest in studying the potential effect of climate change on ozone, and how this change may affect public health. In this paper we develop a Bayesian spatial model to predict ozone under different meteorological conditions, and use this model to study spatial and temporal trends and to forecast ozone concentrations under different climate scenarios. We develop a spatial quantile regression model that does not assume normality and allows the covariates to affect the entire conditional distribution, rather than just the mean. The conditional distribution is allowed to vary from site-to-site and is smoothed with a spatial prior. We apply our model to summer ozone from 1997-2005 in the Eastern US, and use deterministic climate models to project ozone under future climate conditions. Our analysis suggests that holding all other factors fixed, an increase in daily average temperature will lead to the largest increase in ozone in the Industrial Midwest and Northeast.

Key words: Climate change; Ozone; Quantile regression; Semiparametric Bayesian methods; Spatial data.

[1]Corresponding author, email: reich@stat.ncsu.edu.

Bayesian spatial quantile regression

1 Introduction

Beginning in 1970, the U.S. Clean Air Act (CAA) directed the U.S. Environmental Protection Agency (EPA) to consider the best available science on exposure to and effects of several ambient air pollutants, emitted by a wide array of sources. National Ambient Air Quality Standards (NAAQS) were set for pollutants to which the public was widely exposed. Since the inception of NAAQS, EPA has determined that photochemical-oxidant air pollution, formed when specific chemicals in the air react with light and heat, is of sufficient public-health concern to merit establishment of a primary NAAQS. EPA has since 1979 identified ozone, a prominent member of the class of photochemical oxidants, as an indicator for setting the NAAQS and tracking whether areas of the country are in compliance with the standards. To attain the current ozone standard, the 3-year average of the fourth-highest daily maximum 8-hour average ozone concentrations measured at each monitor within an area over each year must not exceed 0.075 ppm (standard effective since May 27, 2008).

Most ozone in the troposphere is not directly emitted to the atmosphere, although there are minor sources of such ozone, including some indoor air cleaners. Rather, it is formed from a complex series of photochemical reactions of the primary precursors: nitrogen oxides (NOx), volatile organic compounds (VOCs), and to a smaller extent other pollutants, such as carbon monoxide (CO). Since the reactions that form ozone are driven by sunlight, ambient ozone concentrations exhibit both diurnal variation (they are typically highest during the afternoon) and marked seasonal variation (they are highest in summer). Ambient con-

centrations are highest during hot, sunny summer episodes characterized by low ventilation (a result of low winds and low vertical mixing).

Due to the strong dependence on weather conditions, ozone levels may be sensitive to climate change (Seinfeld and Pandis, 2006). There is great interest in studying the potential effect of climate change on ozone levels, and how this change may affect public health (Bernard et al. 2001; Haines and Patz, 2004; Knowlton et al., 2004; Bell et al., 2007). In this paper we also study the potential changes in ozone due to climatic change. This type of work is needed to address the impact of climate change on emission control strategies designed to reduce air pollution. Using future numerical climate model forecasts of meteorological conditions, we forecast potential future increases or decreases in ozone levels. In particular, based on current relationships between temperature, cloud cover, wind speed, and ground-level ozone, we predict the percent change in ozone given future temperature and cloud cover levels.

The objective of the paper is to develop an effective statistical model for the daily tropospheric ozone distribution as a function of daily meteorological variables. The daily model is then used to study trends in ozone levels over space and time, and to forecast yearly summaries of ozone under different climate scenarios. We build our model using spatial methods to borrow strength across nearby locations. Several spatial models have been proposed for ozone (Guttorp et al., 1994; Carroll et al., 1997; Meiring et al., 1998; Huang and Hsu, 2004; Huerta et al., 2004; Gilleland and Nychka, 2005; Sahu et al., 2007). These models assume normality for either untransformed ozone or for the square root of ozone. Exploratory analysis suggests that ozone data are non-Gaussian even after a square root transformation. Ozone is often right-skewed (Lee et at., 2006; Zhang and Fan, 2008) in which case Gaussian models

may underestimate the tail probability. Correctly estimating the tail probability is critically important in studying the health effects of ozone exposure, and has policy implications because EPA standards are based on the fourth highest day of the year (approximately the 99^{th} quantile). A further challenge is that the relationship between meteorological predictors and the ozone response can be nonlinear and the meteorological effects are not restricted to the mean. The variance and skewness of the response varies depending on location and meteorological conditions. Recently several methods have been developed for non-Gaussian spatial modeling (Gelfand et al., 2005; Griffin and Steel, 2006; Reich and Fuentes, 2007; Dunson and Park, 2008; An et al., 2008). These methods treat the conditional distribution of the response given the spatial location and the covariates as an unknown quantity to be estimated from the data. We follow this general approach to model the conditional ozone density.

Although these models are quite flexible, one drawback is the difficulty in interpreting the effects of each covariate. For example, many of these models are infinite mixtures, where the spatial location and/or covariates affect the mixture probabilities. In this very general framework, it is difficult to make inference on specific features of the conditional density, for example, whether there is an interaction between cloud cover and temperature, or whether there is a statistically significant time trend in the distribution's upper tail probability. As a compromise between fully-general Bayesian density regression and the usual additive mean regression, we propose a Bayesian spatial quantile regression model. Quantile regression models the distribution's quantiles as additive functions of the predictors. This additive structure permits inference on the effect of individual covariates on the response's quantiles.

There is a vast literature on quantile regression (e.g., Koenker, 2005), mostly from the

frequentist perspective. The standard model-free approach is to estimate the effect of the covariates separately for a few quantile levels by minimizing an objective function. This approach is popular due to computational convenience and theoretical properties. Sousa et al. (2008) applied the usual quantile regression method to ozone data and found it to be superior to multiple linear regression, especially for predicting extreme events. An active area of research is incorporating clustering into the model-free approach (Jung, 1996; Lipsitz et al., 1997; Koenker, 2004; Wang and He 2007; Wang and Fygenson, 2008). Recently, Hallin et al. (2009) propose a quantile regression model for spatial data on a grid. They allow the regression coefficients to vary with space using local regression. This approach, and most other model-free approaches, perform separate analyses for each quantile level of interest. As a result, the quantile estimates can cross, i.e., for a particular combination of covariates the estimated quantile levels are non-increasing, which causes problems for prediction. Several post-hoc methods have been proposed to address this problem (He, 1997; Yu and Jones, 1998; Takeuchi et al., 2006; Dette and Volgushev, 2008) for non-spatial data.

Incorporating spatial correlation may be more natural in a Bayesian setting, which necessarily specifies a likelihood for the data. Model-based Bayesian quantile regression methods for independent (Yu and Moyeed, 2001; Kottas and Gelfand, 2001; Kottas and Krnjajic, 2009; Hjort and Walker, 2009) and clustered (Geraci and Bottai, 2007; Reich et al., 2009) data that focus on a single quantile level have been proposed. Dunson and Taylor (2005) propose a method to simultaneously analyze a finite number of quantile levels for independent data. To our knowledge, we propose the first model-based approach for spatial quantile regression.

Rather than focusing on a single or finite number of quantile levels, our approach is to

specify a flexible semiparametric model for the entire quantile process across all covariates and quantile levels. We assume the quantile function at each quantile level is a linear combination of the covariates and model the quantile functions using a finite number of basis functions with constraints on the basis coefficients to ensure that the quantile function is non-crossing for all covariate values. An advantage of this approach is that we can center the prior for the conditional density on a parametric model, e.g., multiple linear regression with skew-normal errors. Our model is equipped with parameters that control the strength of the parametric prior. Also, the quantile function, and thus the conditional density, is allowed to vary spatially. Spatial priors on the basis coefficients are used to allow the quantile process to vary smoothly across space.

The paper proceeds as follows. Section 2 proposes the spatial quantile regression model. While this model is computationally efficient for moderately-sized data sets, it is not feasible for very large data sets. Therefore Section 2 also includes an approximate model which is able to handle several years of daily data for the entire Eastern US. Section 4 conducts a brief simulation to compare our model with other methods and examine sensitivity to hyperprior choice. In Section 5 we analyze a large spatiotemporal ozone data set. We discuss meteorologically-adjusted spatial and temporal trends for different quantile levels and use the estimated conditional densities to forecast future ozone levels using deterministic climate model output. Section 6 concludes.

2 Bayesian quantile regression for spatiotemporal data

Let y_i be the observed eight-hour maximum ozone for day t_i at location s_i. Our interest lies in estimating the conditional density of y_i as a function of s_i and covariates $\mathrm{X}_i = (X_{i1}, ..., X_{ip})'$. In particular, we would like to study the conditions that lead to extreme ozone days. Extreme events are often summarized with return levels. The n-day return level is the value c_n so that $P(y_i > c_n) = 1/n$. Given our interest in extreme events and return levels, we model y_i's conditional density via its quantile (inverse CDF) function $q(\tau|\mathrm{X}_i, \mathrm{s}_i)$, which is defined so that $P\{y_i < q(\tau|\mathrm{X}_i, \mathrm{s}_i)\} = \tau \in [0, 1]$. We model $q(\tau|\mathrm{X}_i, \mathrm{s}_i)$ as

$$q(\tau|\mathrm{X}_i, \mathrm{s}_i) = \mathrm{X}_i'\boldsymbol{\beta}(\tau, \mathrm{s}_i) \tag{1}$$

where $\boldsymbol{\beta}(\tau, \mathrm{s}_i) = (\beta_1(\tau, \mathrm{s}_i), ..., \beta_p(\tau, \mathrm{s}_i))'$ are the spatially-varying coefficients for the τ^{th} quantile level. Directly modeling the quantile function makes explicit the effect of each covariate on the probability of an extreme value.

Several popular models arise as special cases of Model (1). For example, setting $\beta_j(\tau, \mathrm{s}) \equiv \beta_j$ for all τ, s, and $j > 1$ gives the usual linear regression model with location shifted by $\sum_{j=2}^{p} X_{ij}\beta_j$ and residual density determined by $\beta_1(\tau, \mathrm{s})$. Also, setting $\beta_j(\tau, \mathrm{s}) \equiv \beta_j(\mathrm{s})$ for all τ and $j > 1$ gives the spatially-varying coefficients model (Gelfand et al., 2003) where the effect of X_j on the mean varies across space via the spatial process $\beta_j(\mathrm{s})$. Allowing $\beta_j(\tau, \mathrm{s})$ to vary with s and τ relaxes the assumption that the covariates simply affect the mean response, and gives a density regression model where the covariates are allowed to affect the shape of the response distribution. In particular, the covariates can have different effects on the

center ($\tau = 0.5$) and tails ($\tau \approx 0$ and $\tau \approx 1$) of the density.

2.1 Model for the quantile process

We begin modeling the quantile function by ignoring spatial location and assuming the intercept-only model with $\mathrm{X}_i = 1$. In this case, the quantile function in (1) reduces to $q(\tau) = \beta(\tau)$. The process $\beta(\tau)$ must be constructed so that $q(\tau)$ is nondecreasing in τ. Let

$$\beta(\tau) = \sum_{m=1}^{M} B_m(\tau)\alpha_m, \tag{2}$$

where $B_m(\tau)$ is a known basis function of τ and α_m are unknown coefficients that determine the shape of the quantile function. We use Bernstein basis polynomials

$$B_m(\tau) = \binom{M}{m}\tau^m(1-\tau)^{M-m}. \tag{3}$$

An attractive feature of these basis functions is that if $\alpha_m \geq \alpha_{m-1}$ for all $m > 1$, then $\beta(\tau)$, and thus $q(\tau)$, is an increasing function of τ. This reduces the complicated monotonicity constraint to a sequence of simple constraints $\delta_m = \alpha_m - \alpha_{m-1} \geq 0$, for $m = 2, ..., M$. In addition, Bernstein polynomials bases induce a prior with dense support on the space of continuous monotone functions from $[0, 1] \to \mathcal{R}$ (Chang et al., 2007).

Since the constraints on $\alpha = (\alpha_1, ..., \alpha_M)$ are expressed in terms of difference between adjacent terms, we reparameterize to $\delta_1 = \alpha_1$ and $\delta_m = \alpha_m - \alpha_{m-1}$ for $m = 2, ..., M$. The original basis function coefficients are then $\alpha_m = \sum_{l=1}^{m} \delta_l$. Following Cai and Dunson (2008), we ensure the quantile constraint by introducing latent unconstrained variable δ_m^* and taking

$\delta_1 = \delta_1^*$ and

$$\delta_m = \begin{cases} \delta_m^*, & \delta_m^* \geq 0 \\ 0, & \delta_m^* < 0 \end{cases} \tag{4}$$

for $m > 1$.

The δ_m^* have independent normal priors $\delta_m^* \sim \mathrm{N}(\bar{\delta}_m(\Theta), \sigma^2)$, with unknown hyperparameters Θ. We pick $\bar{\delta}_m(\Theta)$ to center the quantile process on a parametric distribution $f_0(y|\Theta)$, for example, a $\mathrm{N}(\mu_0, \sigma_0)$ random variable with $\Theta = (\mu_0, \sigma_0)$. Letting $q_0(\tau|\Theta)$ be the quantile function of $f_0(y|\Theta)$, the $\bar{\delta}_m(\Theta)$ are then chosen so that

$$q_0(\tau|\Theta) \approx \sum_{m=1}^{M} B_m(\tau)\bar{\alpha}_m(\Theta), \tag{5}$$

where $\bar{\alpha}_m(\Theta) = \sum_{l=1}^{m} \bar{\delta}_l(\Theta)$. The $\bar{\delta}_m(\Theta)$ are chosen to correspond to the following ridge regression estimator:

$$\left(\bar{\delta}_1(\Theta), ..., \bar{\delta}_M(\Theta)\right)' = \arg\min_d \sum_{k=1}^{K} \left(q(\tau_k|\Theta) - \sum_{m=1}^{M} B_m(\tau_k) \left[\sum_{l=1}^{m} d_l \right] \right)^2 + \lambda \sum_{m=1}^{M} d_m^2, \tag{6}$$

where $d_m \geq 0$ for $m > 1$, $\{\tau_1, ..., \tau_K\}$ is a dense grid on (0,1), and λ is a tuning constant for the ridge penalty $\sum_{m=1}^{M} d_m^2$ added to provide numerical stability. For simplicity we fix $\lambda = 1$, which seems to provide a reasonable approximation. As $\sigma \to 0$, the quantile functions this resembles are increasing shrunk towards the parametric quantile function $q_0(\tau|\Theta)$, and the likelihood is similar to $f_0(y|\Theta)$.

2.2 Model for the quantile process with covariates

Adding covariates, the conditional quantile function becomes

$$q(\tau|\mathrm{X}_i) = \mathrm{X}_i\boldsymbol{\beta}(\tau) = \sum_{j=1}^{p} X_{ij}\beta_j(\tau). \tag{7}$$

As in Section 2.1, the quantile curves are modeled using Bernstein basis polynomials

$$\beta_j(\tau) = \sum_{m=1}^{M} B_m(\tau)\alpha_{jm}, \tag{8}$$

where α_{jm} are unknown coefficients. The processes $\beta_j(\tau)$ must be constructed so that $q(\tau|\mathrm{X}_i)$ is nondecreasing in τ for all X_i. Collecting terms with common basis functions gives

$$\mathrm{X}_i'\boldsymbol{\beta}(\tau, \mathrm{s}) = \sum_{m=1}^{M} B_m(\tau)\theta_m(\mathrm{X}_i), \tag{9}$$

where $\theta_m(\mathrm{X}_i) = \sum_{j=1}^{p} X_{ij}\alpha_{jm}$. Therefore, if $\theta_m(\mathrm{X}_i) \geq \theta_{m-1}(\mathrm{X}_i)$ for all $m > 1$, then $\mathrm{X}_i\boldsymbol{\beta}(\tau)$, and thus $q(\tau|\mathrm{X}_i)$, is an increasing function of τ.

To specify our prior for the α_{jm} to ensure monotonicity, we assume that $X_{i1} = 1$ for the intercept and the remaining covariates are suitably scaled so that $X_{ij} \in [0, 1]$ for $j > 1$. Since the constraints are written in terms of the difference between adjacent terms, we reparameterize to $\delta_{j1} = \alpha_{j1}$ and $\delta_{jm} = \alpha_{jm} - \alpha_{jm-1}$ for $j = 2, ..., M$. We ensure the quantile

constraint by introducing latent unconstrained variable $\delta^*_{jm} \sim \mathrm{N}(\bar{\delta}_{jm}(\Theta), \sigma^2_j)$ and taking

$$\delta_{jm} = \begin{cases} \delta^*_{jm}, & \delta^*_{1m} + \sum_{j=2}^{p} I(\delta^*_{jm} < 0)\delta^*_{jm} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Recalling $X_{i1} = 1$ and $X_{ij} \in [0, 1]$ for $j = 2, ..., p$,

$$\theta_m(\mathrm{X}_i) - \theta_{m-1}(\mathrm{X}_i) = \sum_{j=1}^{p} X_{ij}\delta_{jm} \geq \delta_{1m} + \sum_{j=2}^{p} X_{ij} I\left(\delta_{jm} < 0\right) \delta_{jm} \geq 0 \tag{11}$$

for all X_i, giving a valid quantile process. As in Section 2.1 we center the intercept curve on a parametric quantile function $q_0(\Theta)$. The remaining coefficients have $\bar{\delta}_{jm}(\Theta) = 0$ for $j > 1$.

Although this model is quite flexible, we have assumed that the quantile process is a linear function of the covariates, simplifying interpretation. In some applications the linear quantile relationship may be overly-restrictive. In this case, transformations of the original predictors such as interactions or basis functions can be added to give a more flexible model.

2.3 Spatial quantile regression

In this section we let the quantile functions vary spatially by letting

$$\beta_j(\tau, \mathrm{s}) = \sum_{m=1}^{M} B_m(\tau)\alpha_{jm}(\mathrm{s}), \tag{12}$$

where $\alpha_{jm}(\mathrm{s})$ are spatially-varying basis function coefficients. We enforce the monotonicity constraint at each spatial location as in Section 2.2 by introducing latent Gaussian parameters $\delta^*_{jm}(\mathrm{s})$. The latent parameters relate to the basis function coefficients as $\alpha_{jm}(\mathrm{s}) =$

$\sum_{l=1}^{m} \delta_{jl}(\mathrm{s})$ and

$$\delta_{jm}(\mathrm{s}) = \begin{cases} \delta^*_{jm}(\mathrm{s}), & \delta^*_{1m}(\mathrm{s}) + \sum_{j=2}^{p} I(\delta^*_{jm}(\mathrm{s}) < 0)\delta^*_{jm}(\mathrm{s}) \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

The $\delta^*_{jm}(\mathrm{s})$ are independent (over j and m) Gaussian spatial processes with mean $E(\delta^*_{jm}(\mathrm{s})) = \bar{\delta}_{jm}(\Theta)$ and exponential spatial covariance $\mathrm{Cov}(\delta^*_{jm}(\mathrm{s}), \delta^*_{jm}(\mathrm{s}')) = \sigma_j^2 \exp(-||\mathrm{s} - \mathrm{s}'||/\rho_j)$, where σ_j^2 is the variance of $\delta^*_{jm}(\mathrm{s})$ and ρ_j determines the range of the spatial correlation function.

3 Approximate method

Section 2's spatial quantile regression model can be implemented efficiently for moderately-sized data sets. However, it becomes computationally infeasible for Section 5.3's analysis of several years of daily data for the Eastern U.S. To approximate the full Bayesian analysis, we propose a two-stage approach. We first perform separate quantile regression at each site for a grid of quantile levels to obtain estimates of the quantile process and their asymptotic covariance. In a second stage, we analyze these initial estimates using the Bayesian spatial model for the quantile process.

The usual quantile regression estimate (Koenker, 2005) for quantile level τ and spatial location s is

$$\left(\hat{\beta}_1(\tau_k, \mathrm{s}), ..., \hat{\beta}_p(\tau_k, \mathrm{s})\right)' = \arg\min_{\beta} \sum_{\mathrm{s}_i = \mathrm{s},\ y_i > \mathrm{X}_i'\beta} \tau_k |y_i - \mathrm{X}_i'\beta| + \sum_{\mathrm{s}_i = \mathrm{s},\ y_i < \mathrm{X}_i'\beta} (\tau_k - 1)|y_i - \mathrm{X}_i'\beta|. \tag{14}$$

This estimate is easily obtained from the quantreg package in R and is consistent for the

true quantile function and has asymptotic covariance (Koenker, 2005)

$$\text{Cov}\left[\sqrt{n_{\text{S}}}\left(\hat{\beta}_1(\tau_k,\text{s}),...,\hat{\beta}_p(\tau_k,\text{s})\right),\sqrt{n_{\text{S}}}\left(\hat{\beta}_1(\tau_l,\text{s}),...,\hat{\beta}_p(\tau_l,\text{s})\right)\right] = H(\tau_k)^{-1}J(\tau_k,\tau_l)H(\tau_l)^{-1}, \tag{15}$$

where n_{S} is the number of observations at site s, $H(\tau) = \lim_{n_{\text{S}}\to\infty} n_{\text{S}}^{-1}\sum_{i=1}^{n_{\text{S}}} X_i X_i' f_i(\xi_i(\tau))$, $f_i(\xi_i(\tau))$ is the conditional density of y_i evaluated at the τ^{th} quantile level, and $J(\tau_k,\tau_l) = [\tau_k \wedge \tau_l - \tau_k\tau_l]\, n_{\text{S}}^{-1}\sum X_i X_i'$.

Although consistent as the number of observations at a given site goes to ∞, these estimates are not smooth over space or quantile level, and do not ensure a non-crossing quantile function for all X. Therefore we smooth these initial estimates using the spatial model for the quantile process proposed in Section 2. Let $\hat{\boldsymbol{\beta}}(\text{s}_i) = \left[\hat{\beta}_1(\tau_1,\text{s}_i),...,\hat{\beta}_1(\tau_K,\text{s}_i),\hat{\beta}_2(\tau_1,\text{s}_i),...,\hat{\beta}_p(\tau_K,\text{s}_i)\right]'$ and $\text{acov}\left(\hat{\boldsymbol{\beta}}(\text{s}_i)\right) = \Sigma_i$ (with elements defined by (15)). We fit the model

$$\hat{\boldsymbol{\beta}}(\text{s}_i) \sim \text{N}\left(\boldsymbol{\beta}(\text{s}_i),\Sigma_i\right), \tag{16}$$

where the elements of $\boldsymbol{\beta}(\text{s}_i) = [\beta_1(\tau_1,\text{s}_i),...,\beta_1(\tau_K,\text{s}_i),\beta_2(\tau_1,\text{s}_i),...,\beta_p(\tau_1,\text{s}_i)]'$ are functions of Bernstein basis polynomials as in Section 2.3. This approximation provides a dramatic reduction in computational time because the dimension of the response is reduced from the number of observations at each site to the number of quantile levels in the approximation, and the posteriors for the parameters that define $\boldsymbol{\beta}$ are fully-conjugate allowing for Gibbs updates and rapid convergence. Computational details for this model and the full model are given in the Appendix.

4 Simulation study

In this section we analyze simulated data to compare our method with standard quantile regression approaches, and to examine the performance of Section 3's approximate method. For each of the $S = 50$ simulated data sets we generate $n = 20$ spatial locations s_i uniformly on $[0,1]^2$. The $p = 3$ covariates are generated as $X_1 \equiv 1$ and $X_{i2}, X_{i3} \overset{iid}{\sim} \mathrm{U}(0,1)$, independent over space and time. The true quantile function is

$$q(\tau|\mathrm{X}_i, \mathrm{s}_i) = 2s_{i2} + (\tau+1)\Phi^{-1}(\tau) + \left(5s_{i1}\tau^2\right) X_{i3}, \tag{17}$$

which implies that $\beta_1(\tau, \mathrm{s}_i) = 2s_{2i} + (\tau+1)\Phi^{-1}(\tau)$, $\beta_2(\tau, \mathrm{s}_i) = 0$, and $\beta_3(\tau, \mathrm{s}_i) = 5s_{1i}\tau^2$, where Φ is the standard normal distribution function. Figure 1a plots the density corresponding to this quantile function for various covariates and spatial locations. The density is generally right-skewed to mimic ozone data. The second spatial coordinate simply shifts the entire distribution by increasing the intercept function β_1. The first predictor X_2 has no effect on the density because $\beta_2(\tau, \mathrm{s}) = 0$. The second predictor X_3 has little effect on the left tail because $\beta_3(\tau, \mathrm{s})$ is near zero for small τ, but increasing X_3 adds more mass to the density's right tail.

Each data set contains 100 replicates at each spatial location. We first consider the situation where the replicates are independent over space and time. We also generate data with spatially correlated residuals using a Gaussian copula. That is, we generate the data by first generating Z_i as independent (over time) Gaussian processes with mean zero and exponential spatial covariance $\exp\left(-||\mathrm{s} - \mathrm{s}'||/\rho_Z\right)$, and then transform using the marginal

quantile function $y_i = q_o[\Phi(Z_i)] + \sum_{m=1}^{M} B_m[\Phi(Z_i)]\theta_m(\mathrm{X}_i, \mathrm{s}_i)$. We assume spatial range of the residuals is $\rho_Z = 0.5$.

For each simulated data set we fit three Bayesian quantile methods: the full model described in Section 2, the model in Section 2 without spatial modeling (i.e., $\delta^*_{jk} \overset{iid}{\sim} \mathrm{N}(\bar{\delta}_{jk}(\Theta), \sigma_j^2)$), and Section 3's approximate method. For the Bayesian quantile regression models (full and approximate) we use $M = 10$ knots and vague yet proper priors for the hyperparameters that control the prior covariance of the quantile function, $\sigma_j^2 \sim$ InvG(0.1,0.1) and $\rho_j \sim$ Gamma(0.06,0.75). The prior for the ρ_j is selected so the effective range $-\rho_j \log(0.05)$, i.e., the distance at which the spatial correlation equals 0.05, has prior mean 0.25 and prior standard deviation 1. The centering distribution f_0 was taken to be skew-normal with location $\mu_0 \sim \mathrm{N}(0,10^2)$, scale $\sigma_0^2 \sim$ InvGamma(0.1,0.1), and skewness $\psi_0 \sim \mathrm{N}(0,10^2)$. We also compare our methods with usual frequentist estimates in (14), computed using the quantreg package in R. Of course these estimates do not smooth over quantile level or spatial location, so they may not be directly comparable in this highly-structured setting.

For each simulated data set and each method we compute point estimates (posterior means for Bayesian methods) and 90% intervals for $\beta_j(\tau_k, \mathrm{s}_i)$ for $j = 1, 2, 3$, $\tau_k \in \{0.05, 0.10, ..., 0.95\}$, and all spatial locations s_i. We compare methods using mean squared error, coverage probability, and power averaged over space, quantile levels, and simulated data set. Specifically, mean squared error for the j^{th} quantile function is computed as

$$MSE = \frac{1}{SnK} \sum_{sim=1}^{S} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\hat{\beta}_j(\tau_k, s_i)^{(sim)} - \beta_j(\tau_k, s_i) \right)^2, \tag{18}$$

where $\hat{\beta}_j(\tau_k, s_i)^{(sim)}$ is the point estimate for the simulation number sim. Coverage and

power are computed similarly.

Table 1 presents the results. Adding correlation to the residuals has little effect on the performance of any of the models, so we focus our discussion on the simulation without residual correlation. By borrowing strength across quantile level and spatial location all three Bayesian methods provide smaller mean squared error and higher power than the usual quantile regression approach. Figure 1b shows the true quantile curve for β_3 for each spatial location for one representative data set with line width proportional to the first spatial coordinate. The usual estimates in Figure 1c fluctuate greatly across quantile levels compared to the smooth curves produced by the Bayesian spatial quantile regression model in Figure 1d.

The approximate method which smooths the initial estimates from usual quantile regression reduces mean squared error. In fact, in this simulation the approximate method gives smaller mean squared error than the full model without spatial modeling. Therefore, the approximate model appears to provide a computationally efficient means to estimate the true quantile function to be used for predicting future observations. However, the coverage probabilities for the individual covariates using the approximate method sink to as low as 0.60. We repeated the simulation increasing the number of replicates at each site from $n_t = 100$ to $n_t = 1000$ and the coverage probabilities remained unacceptably low. So although the approximate method is useful for prediction, inference regarding the significance of individual predictors should not be made based on this method, even for large data sets.

In this simulation we assumed $M = 10$ basis functions, $\sigma_j^2 \sim \text{InvGamma}(0.1,0.1)$, and the prior standard deviation of the effective range was one. We reran the simulation (without spatially correlated residuals) with $S = 10$ data sets using the full model three times, each

varying one of these assumptions. The alternatives were $M = 25$ basis functions, $\sigma_j \sim$ InvGamma(0.001,0.001), and the prior standard deviation of the effective range set to five. The results of the simulation were fairly robust to these changes; the mean squared error varied from 0.06 to 0.11 for X_1, from 0.07 to 0.12 for X_2, and from 0.18 to 0.26 for X_3, and the average coverage probability was at least 0.86 for all fully Bayes models and all covariates. The prior for the effective range was most influential. Altering this prior affected the posterior of the spatial range, but had only a small effect on the mean squared errors and coverage probabilities.

5 Analysis of Eastern US ozone data

In this section we analyze monitored ozone data from the Eastern US from the summers of 1997-2005 and climate model output from 2041-2045. Section 5.1 describes the data. Sections 5.2 and 5.3 analyze monitored ozone data from 1997-2005, first using the full model and a subset of the data, and then using Section 3's approximate model and the complete data. Sections 5.4 and 5.5 analyze the computer model output.

5.1 Description of the data

Meteorological data were obtained from the National Climate Data Center (NCDC; http://www.ncdc.noaa.gov/oa/ncdc.html). We obtained daily average temperature and daily maximum wind speed for 773 monitors in the Eastern US from the NCDC's Global Summary of the Day Data Base. Daily average cloud cover for 735 locations in the Eastern US was obtained from the NCDC's National Solar Radiation Data Base.

Maximum daily 8-hour average ozone was obtained from the US EPA's Air Explorer Data Base (http://www.epa.gov/airexplorer/index.htm). We analyze daily ozone concentrations measured at 631 locations in the Eastern US during the summers (June-August) of 1997-2005 (470,239 total observations), plotted in Figure 2a. Meteorological and ozone data are not observed at the same locations. Therefore we imputed meteorological variables at the ozone locations using spatial Kriging. Spatial imputation was performed using SAS version 9.1 and the MIXED procedure with spatial exponential covariance function, with covariance parameters allowed to vary by variable and year.

The final source of data is time slice experiments from the North American Regional Climate Change Assessment Program. These data are output from the Geophysical Fluid Dynamics Laboratory's (GFDL) deterministic atmospheric computer model (AM2.1) with 3 hour $\times$ 50 km resolution. We use data from two experiments. The first provides modeled data from 1968 to 2000 using observed sea surface temperature and sea-ice extent. The second provides modeled data for 2038 to 2070 using deviations in sea surface temperature and sea-ice extent from the GFDL's CM2.1 A2 scenario. From these experiments we obtain modeled daily average temperature, maximum wind speed, and average cloud cover fraction. As with the NCDC meteorological data, these data are not on the same scale as the ozone data. Throughout we use the ozone monitoring locations as the spatial unit, and we match modeled climate data with ozone data by extracting climate data from the grid cell containing the ozone monitor.

5.2 Atlanta sub-study

We begin by comparing several models using only data from the 12 stations in the Atlanta area. The continuous variables maximum temperature and wind speed are standardized and transformed to the unit interval by the normal CDF

$$X_j = \Phi\left([temp - m(temp)]/s(temp)\right) \tag{19}$$

where $m(temp)$ and $s(temp)$ are the sample mean and standard deviation of daily maximum temperature over space and time. Cloud cover proportion naturally falls on the unit interval. We also include the year to investigate temporal trends in the quantile process. The year is transformed as $X_j = (year - 1997)/8$. We also include the interaction $X_j X_l$ between temperature and cloud cover and quadratic effects $4(X_j - 0.5)^2$.

We fit the full and approximate model with and without quadratic terms and compare these models with the fully-Gaussian spatial model with spatially-varying coefficients,

$$y(\mathrm{s}, t) = \sum_{j=1}^{p} X_j(\mathrm{s}, t)\beta_j(\mathrm{s}) + \mu(\mathrm{s}, t) + \varepsilon(\mathrm{s}, t), \tag{20}$$

where $\beta_j(\mathrm{s})$ are spatial Gaussian processes with exponential covariance, $\mu(\mathrm{s}, t)$ are independent (over time) spatial Gaussian processes with exponential covariance, and $\varepsilon(\mathrm{s}, t) \overset{iid}{\sim} N(0, \sigma^2)$ is the nugget effect. For both the full and approximate Bayesian quantile regression models, the centering distribution f_0 is taken to be skew-normal with location $\mu_0 \sim$ N(0,10^2), scale $\sigma_0^2 \sim$ InvGamma(0.1,0.1), and skewness $\psi_0 \sim$ N(0,10^2). Figure 2b shows that this distribution is flexible enough to approximate the ozone distributions. However, a

parametric skew-normal spatial model would be insufficient to characterize changes in the right tail of the ozone distribution with covariates and spatial location. We use $M = 10$ knots and priors $\tau_{jm}^2 \sim$ InvGamma(0.1,0.1) and $\rho_j \sim$ Gamma(0.5,0.5), which gives prior 95% intervals (0.00,0.99) and (0.00,0.80) for the correlation between the closest and farthest pairs of points, respectively.

To compare models we randomly removed all observations for 10% of the days ($N =$ 910 total observations), with these test observations labeled $y_1^*, ..., y_N^*$. For each deleted observation we compute the posterior predictive mean $\hat{y}_i^*$ and the posterior predictive 95% equal-tailed intervals. Table 1 gives the root mean squared prediction error $RMSE = \sqrt{\sum_{i=1}^N \left(y_i^* - \hat{y}_i^*\right)^2 / N}$, mean absolute deviation $MAD = \sum_{i=1}^N |y_i^* - \hat{y}_i^*|/N$, and the coverage probabilities and average (over i) width of the prediction intervals.

The coverage probabilities of the 95% intervals are near or above the nominal rate for all models in Table 2. The Gaussian models have the highest $RMSE$ and MAD. The approximate method with quadratic terms minimizes both MSE and MAD. Section 4's simulation study shows that the posterior intervals for the regression coefficients $\boldsymbol{\beta}$ have unacceptable coverage probability. However, for these data the prediction intervals have good coverage probability, justifying their use for prediction.

Figure 3 summarizes the posterior for the full Bayesian spatial quantile regression model with quadratic terms. Panels (a)-(c) plot the quantile curves for one representative location. The main effects are all highly significant, especially for the upper quantile levels. As expected ozone concentration increases with temperature and decreases with cloud cover and wind speed. Figure 3 reveals a complicated relationship between ozone and temperature. The linear temperature effect is near zero for low quantile levels and increases with τ.

Figure 3c plots the data and several fitted quantile curves (τ ranging from 0.05 to 0.95) by year with the transformed meteorological variables fixed at 0.5. All quantile levels decrease from 1997 to 2002; after 2002 the lower quantiles plateau while the upper quantiles continue to decline. Note that in this plot more than 5% of the observations fall above the 95% quantile level. This is the result of plotting the quantile functions without regard to variability in the meterological variables (i.e., fixing them at 0.5). To give a sense of the spatial variability in the quantile curves, Figure 3d plots the posterior mean of the main effect for year for all locations. While there is variability from location to location, all sites show a decreasing trend, especially for upper quantile levels.

5.3 Analysis of Eastern US ozone data

Analyzing data for the entire Eastern US using the full Bayesian spatial quantile regression model is not computationally feasible, so we use only the approximate model. Based on Section 5.2's results, we include quadratic terms for all predictors. Also, we use the same priors as in Section 5.2. The posterior median (95% interval) for the skewness parameter ψ_0 of the skew-normal base distribution is 3.76 (1.12, 10.54), supporting a non-Gaussian analysis. To justify that this model fits well, we randomly removed observations for 2% of the days ($N = 12,468$ observations) and the out-of-sample coverage probability of the 95% prediction intervals was 93.4%.

Figure 4 maps the posterior means of several quantile functions. The two strongest predictors are temperature and cloud cover. The model includes interaction and quadratic terms, so to illustrate the effects, we plot linear combinations of all terms involving these

predictors,

$$X_1\beta_1(\tau,\mathrm{s}) + X_2\beta_2(\tau,\mathrm{s}) + 4(X_1-0.5)^2\beta_3(\tau,\mathrm{s}) + 4(X_2-0.5)^2\beta_4(\tau,\mathrm{s}) + X_1X_2\beta_5(\tau,\mathrm{s}), \quad (21)$$

where X_1 and X_2 are temperature and cloud cover values, respectively, and β_1-β_5 are the corresponding quantile curves. Figure 4a-4d plot the posterior means of linear combinations using X_1 that correspond to 20^oC and 30^oC and X_2 that correspond to 10% and 90% cloud cover.

For both temperature values ozone concentrations are generally higher when cloud cover is low. For clear days with 10% cloud cover temperature has a strong effect on ozone in the north, but only a weak effect in the south. For cloudy days with 90% cloud cover, temperature has less effect overall, but remains significant in the northeast.

The linear time trend for the 95^{th} quantile in Figure 4e is generally decreasing, especially in the south. This agrees with Chan (2009). To compare the rate of decrease in the upper and lower tails, Figure 4f plots the difference between the linear time trend for the 95^{th} and 5^{th} quantile ($\beta_j(0.95,\mathrm{s}) - \beta_j(0.05,\mathrm{s})$). In the Mid-Atlantic (red) the negative trend is stronger in the lower tail than the upper tail; in contrast in Florida the trend is stronger in the upper tail.

5.4 Calibrating computer model output

Before applying our statistical model to forecast ozone levels, we must calibrate the climate model with the observed data. Figure 5 plots the sample quantile function for the observed and modeled temperature and wind speed for all days in 1997-1999 in Georgia. The distribu-

tion of daily average temperature agrees quite well below the median, however the modeled temperature has heavier right tail than the observed temperature. The standard approach to calibration is to simply shift and scale the computer model output by matching the sample mean and standard deviation. However, in light of the differences in the tail of the distributions and our desire to accurately model extreme events, a more sophisticated approach is warranted.

We use nonlinear monotonic regression on the sample quantile functions to calibrate these distributions. Let $\hat{q}_o(\tau_k)$ and $\hat{q}_m(\tau_k)$ be the τ_k^{th} sample quantiles of the observed and modeled 1997-1999 data, respectively. These distributions are standardized by the nonlinear model $E\left[\hat{q}_o(\tau)\right] = g\left[\hat{q}_m(\tau)\right]$ using sample quantiles on the grid $\tau_k \in \{0.01, 0.02, ..., 0.99\}$. We model g using $M = 20$ Bernstein polynomial basis functions with monotonicity constraints as in Section 2 to ensure that g is an increasing function.

Now let X be an output from the 2041-2045 model run. Rather than using X as an input to our forecast, we use the transformed predictor $g(X)$. Figure 5 shows the transformed temperature for Georgia. Model outputs for 2041-2045 with large temperatures are reduced to resolve the discrepancy between observed and modeled 1997-1999 data. The meteorological predictors are transformed separately by state to account for spatial variation in the calibration. Although a fully-Bayesian isotonic regression model to account for uncertainty in this transformation would be desirable, we simply estimate g using restricted least squares and treat the calibration function as fixed in our forecast.

5.5 Projecting ozone levels under different climate scenarios

The additive structure of the quantile regression model gives the effect of each covariate on the maximum daily 8-hour average ozone in closed form. In addition, policy makers are often interested in the effect of covariates on the yearly ozone distribution. For example, in this section we explore the relationship between temperature and yearly median and 95^{th} percentile of ozone. Here we use Section 5.3's estimate of the conditional density of daily ozone to simulate several realizations of the ozone process to forecast yearly summaries under different climate scenarios. These simulations vary temperature, wind speed, and cloud cover and assume all other factors (emissions, land-use, etc.) are fixed. Certainly other factors will change in the future (for example emissions may decline in response to new standards) so these projections are not meant to be realistic predictions. Rather, they are meant to isolate the effect of climate change on future ozone levels.

Two factors contribute to the affect of climate changes on ozone levels at a given location: the magnitude of the climate change and the strength of the association of meteorology and ozone. Figure 4 suggests that the association between temperature and ozone is strongest in the Northeast US. To quantify spatial variability in the effect of temperature increase on yearly summaries, we generate 500 replicates of the ozone process at the data points under different climate scenarios. The first scenario is no change in the meteorological variables. In this case, replicates are generated by simulating the ozone concentration each day at each spatial location from the Section 5.3's conditional daily ozone distribution given the observed meteorological values for that location on that day. For this and all other simulations we fix the year variable to 2005 for all observations to represent the most recent ozone distribution.

The r^{th} replicate at location s and day t, $y^{(r)}(\mathrm{s},t)$, is generated by first drawing $u \sim$U(0,1) and then transforming to

$$y^{(r)}(\mathrm{s},t) = \sum_{j=1}^{p} X_j(\mathrm{s},t)\hat{\beta}_j(u,\mathrm{s}), \tag{22}$$

where $\hat{\beta}_j(u,\mathrm{s})$ is the posterior mean of $\beta_j(u,\mathrm{s})$. For each replication we calculate the yearly summaries $Q_1^{(r)}(\mathrm{s},\tau)$, the τ^{th} quantile of $\{y^{(r)}(\mathrm{s},1),...,y^{(r)}(\mathrm{s},n_t)\}$, and $T_1^{(r)}(\mathrm{s})$, the three-year (2003-2005) average of the fourth-highest daily maximum 8-hour average ozone concentrations.

The second scenario increases the daily average temperature by 2^oC every day at every location and keeps all other variables fixed. Denote $Q_2^{(r)}(\mathrm{s},\tau)$ and $T_2^{(r)}(\mathrm{s})$ as the yearly summaries for replication r from this scenario. Figure 6 plots the mean (over r) of $Q_2^{(r)}(\mathrm{s},\tau) - Q_1^{(r)}(\mathrm{s},\tau)$ for $\tau = 0.5$ and $\tau = 0.95$ to illustrate the effect of a shift in daily temperature holding all other variables fixed. The change in median and 95^{th} quantile of yearly ozone are both the largest in Michigan and Northeast US. The current (as of 2008) EPA ozone standard is that the three-year average of the fourth-highest daily maximum 8-hour average ozone concentrations is less than 0.075 ppm. Figures 6c plots the probability of exceeding the 2008 standard under the current meteorological conditions ($P\{T_1^{(r)}(\mathrm{s}) > 75\}$). Many counties in the industrial midwest and northeast are currently out of compliance, and these regions have the highest probability of exceeding the 2008 ozone standard under our model. These areas are also most susceptible to changes in temperature; the probability of the three-year average exceeding 0.075 ppm increases in these areas under the temperature shift scenario (Figure 6d).

The third scenario uses the calibrated GFDL projected temperature, wind speed, and

cloud cover for 2041-2045. Figure 7 shows the difference in average temperature from the observed 1997-2005 temperatures and calibrated 2041-2045 temperatures. The projected temperature varies spatially, but is generally between 1-4^{o}C. Figure 7d plots the probability of the three-year (2041-2043) average of the fourth-highest daily maximum 8-hour average ozone concentrations is greater than 0.075 ppm under the future climate scenario. The largest increase in this probability is in the industrial midwest where the projected temperature change is the largest and the relationship between temperature and ozone is strong.

6 Discussion

In this paper we propose a Bayesian spatial quantile method for tropospheric ozone. Our model does not assume the response is Gaussian and allows for complicated relationships between the covariates and the response. Working with a subset of data from the Atlanta area we found that temperature, cloud cover, and wind speed were all strongly associated with ozone, and that the effects are stronger in the right tail than the center of the distribution. Working with the entire Eastern US data set we found a decreasing time trend, especially in the South. Applying the model fit under different climate scenarios suggests that the effect of a warmer climate on ozone levels will be strongest in the Industrial Midwest and Northeast, and that a warmer climate will increase the probability of exceeding the EPA ozone standard in these areas.

Our model accounts for spatial variability by modeling the conditional distribution as a spatial process. However, we do not directly account for the spatial correlation of two nearby observations on the same day. A spatial copula (Nelsen, 1999) could be used to

account for this source of correlation while preserving the marginal distribution specified by the quantile function. We experimented with a spatial Gaussian copula and found it dramatically improved prediction of withheld observations when several observations on the same day were observed. However, when all observations on a day were withheld, prediction did not improve substantially. Since our objective is to predict ozone on days with no direct observations, and MCMC convergence and run times are slower using a copula, we elect to present results from the independent model. An efficient way to account for residual correlation is an area of future work.

We used the quantile regression model to calibrate model output with observed values. This calibration matched not only the mean and variance of the two data sources, but also the entire quantile process. An area of future work is to use spatial quantile regression for model evaluation. Model evaluation compares model output and observed data at different spatial locations and/or under different environmental conditions. Comparisons are usually limited to bias and mean squared error. Spatial quantile regression could provide a more comprehensive evaluation, including comparing the frequency of extreme events from each data source.

References

An Q, Wang C, Shterev I, Wang E, Carin L, Dunson D (2009). Hierarchical kernel stick-breaking process for multi-task image analysis. *International Conference on Machine Learning*, to appear.

Bernard SM, Samet JM, Grambsch A, Ebi KL, Romieu I (2001). The potential impacts of climate variability and change on air pollution-related health effects in the United States. *Environmental Health Perspectives*, **109**: S199S209.

Bell ML, McDermott A, Zeger SL, Samet JM, Dominici F (2004). Ozone and short-term mortality in 95 US urban communities, 1987-2000. *Journal of the American Medical Association*, **292**: 2372-2378.

Bell ML, Goldberg R, Hogrefe C, Kinney PL, Knowlton K, Lynn B, Rosenthal J, Rosenzweig C, Patz JA (2007). Climate change, ambient ozone, and health in 50 US cities. *Climatic Change*, **82**: 61–76.

Cai B, Dunson DB (2007). Bayesian multivariate isotonic regression splines: Applications to carcinogenicity studies. *Journal of the American Statistical Association*, **102**: 1158–1171.

Carroll R, Chen R, George E, Li T, Newton H, Schmiediche H, Wang N. (1997). Ozone exposure and population density in Harris County, Texas. *Journal of the American Statistical Association*, **92**: 392-404.

Chan E (2009). Regional ground-level ozone trends in the context of meteorological influences across Canada and the eastern United States from 1997 to 2006. *Journal of Geophysical Research*, **114**: D05301, doi:10.1029/2008JD010090.

Chang I, Chien L, Hsiung CA, Wen C, Wu Y (2007). Shape restricted regression with random Bernstein polynomials. *IMS Lecture Notes –Monograph Series*, **54**: 187–202.

Dette H, Volgushev S (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B*, **70**: 609–627.

Dunson DB, Park JH (2008). Kernel stick-breaking processes. *Biometrika*, **95**: 307–323.

Dunson DB, Taylor JA (2005). Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, **17**, 385–400.

Gelfand AE, Kim HK, Sirmans CF, Banerjee S (2003). Spatial modelling with spatially varying coefficient processes. *Journal of the American Statistical Association*, **98**: 387–396.

Gelfand AE, Kottas A, MacEachern SN (2005). Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing. *Journal of the American Statistical Association*, **100**: 1021–1035.

Gilleland E, Nychka D (2005). Statistical models for monitoring and regulating ground-level ozone. *Environmetrics*, **16**: 535-546.

Geraci M, Bottai M (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, **8**: 140-154.

Griffin JE, Steel MFJ (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, **101**:179–194.

Guttorp P, Meiring W, Sampson PD (1994). A space-time analysis of ground-level ozone data. *Environmetrics*, **5**: 241-254.

Haines A, Patz JA (2004). Health effects of climate change. *Journal of the American Medical Association*, **291**: 99-103.

Hallin M, Lu Z, Yu K (2009). Local Linear Spatial Quantile Regression. *Bernoulli*, to appear.

He X (1997). Quantile curves without crossing. *American Statistician*, **51**: 186191.

Hjort NL, Walker SG (2009). Quantile pyramids for Bayesian nonparametrics. *The Annals of Statistics*, **37**: 105–131.

Huang HC, Hsu NJ (2004). Modeling transport effects on ground-level ozone using a nonstationary space-time model. *Environmetrics*, **15**: 251-268.

Huerta G, Sanso B, Stroud JR (2004). A spatiotemporal model for Mexico City ozone levels. *Applied Statistics*, **53**: 231-248.

Jung S (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, **91**: 251–257.

Knowlton K, Rosenthal JE, Hogrefe C, Lynn B, Gaffin S, Goldberg R, Rosenzweig C, Civerolo K, Ku JY, Kinney PL (2004). Assessing ozone-related health impacts under a changing climate. *Environmental Health Perspectives*, **112**: 1557-1563.

Koenker R (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91**: 74-89.

Koenker R (2005). *Quantile Regression*, Cambridge, U.K.: Cambridge University Press.

Kottas A, Gelfand AE (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, **96**: 1458–1468.

Kottas A, Krnjajić M (2009). Bayesian nonparametric modeling in quantile regression. *Scandinavian Journal of Statistics*, to appear.

Lee CK, Juang LC, Wang CC, Liao YY, Tu CC, Liu UC, Ho DS (2006). Scaling characteristics in ozone concentration time series (OCTS). *Chemosphere*, **62**: 934–946.

Lipsitz SR, Fitzmaurice GM, Molenberghs G, Zhao LP (1997). Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus. *Journal of the Royal Statistical Society, Series C*, **46**: 463-76.

Meiring W, Guttorp P, Sampson PD (1998). Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics*, **5**: 197-222.

Nelsen R (1999). *An introduction to copulas*. New York: Springer-Verlag.

Peacock JL, Marston L, Konstantinou K (2004). Meta-analysis of time-series studies and panel studies of particulate matter (PM) and ozone (O3). World Health Organization, Copenhagen, Denmark.

Reich BJ, Fuentes M (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Annals of Applied Statistics*, **1**: 249–264.

Reich BJ, Bondell HD, Wang H. Flexible Bayesian quantile regression for independent and clustered data. Submitted.

Sahu SK, Gelfand AE, Holland DM (2007). High Resolution Space-Time Ozone Modeling for Assessing Trends. *Journal of the American Statistical Association*, **102**: 1221–1234.

Seinfeld JH, Pandis SN (2006). Atmospheric chemistry and physics: from air pollution to climate change. Wiley, New York, New York.

Sousa SIV, Pires JCM, Martins FG, Pereira MC, Alvim-Ferraz MCM (2008). Potentialities of quantile regression to predict ozone concentrations. *Environmetrics*, **20**: 147–158.

Stieb DM, Judek S, Burnett RT (2003). Meta-analysis of time-series studies of air pollution and mortality: update in relation to the use of generalized additive models. *Journal of the Air and Waste Management Association*, **53**: 258-261.

Takeuchi I, Le QV, Sears TD, Smola AJ (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, **7**: 1231-1264.

Wang H, Fygenson M (2008). Inference for censored quantile regression models in longitudinal studies. *Annals of Statistics*, to appear.

Wang H, He X (2007). Detecting differential expressions in GeneChip microarray studies: a quantile approach. *Journal of American Statistical Association*, **102**: 104–112.

Yu K, Jones MC (1998). Local linear quantile regression. *Journal of American Statistical Association*, **93**: 228237.

Zhang K, Fan W (2008). Forecasting skewed biased stochastic ozone days: analyses, solutions and beyond. *Knowledge and Information Systems*, **14**: 299–326.

Appendix - MCMC details

MCMC sampling is carried out using the software package *R*. Different sampling schemes to update the regression coefficients $\delta^*_{jm}(\mathrm{s})$ are used for the full and approximate models; all other parameters are updated identically for both methods. For Section 2.3's full model, the $\delta^*_{jm}(\mathrm{s})$ are updated individually using Metropolis sampling. This requires computing the likelihood for each observation. This likelihood is approximated by computing $q(\tau_k|\mathrm{X}_i, \mathrm{s}_i)$ on a grid of 100 equally-spaced τ_k from 0 to 1, and taking $p(y_i|\mathrm{X}_i, \boldsymbol{\delta}(\mathrm{s}_i)) \approx 1/[q(\tau_{j+1}|\mathrm{X}_i, \mathrm{s}_i) - q(\tau_j|\mathrm{X}_i, \mathrm{s}_i)]$, where τ_j is the quantile level so that $q(\tau_{j+1}|\mathrm{X}_i, \mathrm{s}_i) \leq y_i < q(\tau_j|\mathrm{X}_i, \mathrm{s}_i)$.

Using Section 3's approximate model, the latent $\delta^*_{jm}(\mathrm{s})$ have conjugate full conditionals and are updated using Gibbs sampling. The full conditionals are given below. Denote the quantile process at location s_i evaluated on the grid of τ in (16) as $\boldsymbol{\beta}(\mathrm{s}_i) = \Omega\boldsymbol{\delta}(\mathrm{s}_i)$, where Ω is the appropriate matrix of basis functions and $\boldsymbol{\delta}(\mathrm{s}_i)$ the vector of $\delta_{jm}(\mathrm{s}_i)$. The full conditional

is a mixture of two truncated normal densities

$$\pi * N_{[-\infty,c]}\left(m_1, s_1\right) + (1-\pi) N_{[c,\infty]}\left(m_2, s_2\right), \tag{23}$$

where $N_A(m, s)$ is the truncated normal density with location m, scale s, and domain A. The first truncated normal density has parameters m_1 and s_1, where $\delta^*_{jm}(\mathrm{s}_i)|\delta^*_{jm}(\mathrm{s}_k), k \neq i \sim N(m_1, s_1)$ is the conditional prior from the Gaussian spatial model. The second has parameters m_2 and s_2, where $s_2^{-2} = 1/s_1^2 + \Omega'_{jm}\Sigma_i^{-1}\Omega_{jm}$, $m_2 = s_2^2\left[m_1/s_1^2 + \Omega'\Sigma_i^{-1}(\hat{\beta}(\mathrm{s}_i) - \Omega_{(jm)}\boldsymbol{\delta}_{(jm)}(\mathrm{s}_i))\right]$, Ω_{jm} is the column of Ω that corresponds to $\delta_{jm}(\mathrm{s}_i)$, $\Omega_{(jm)}$ is the matrix Ω after removing column Ω_{jm}, and $\boldsymbol{\delta}_{(jm)}(\mathrm{s}_i)$ is the vector $\boldsymbol{\delta}(\mathrm{s}_i)$ after removing $\delta_{jm}(\mathrm{s}_i)$.

The probability π and cutpoint c depend on j and m. The first term is unconstrained, so if $m = 1$ then $\pi = 0$ and $c = -\infty$. Terms with $m > 1$ are constrained. For these terms if $j = 1$ then $c = -\sum_{j=2}^{p} I(\delta^*_{jm}(\mathrm{s}) < 0)\delta^*_{jm}(\mathrm{s})$ and

$$\pi = \frac{\Phi(\frac{c-m_1}{s_1})}{\Phi(\frac{c-m_1}{s_1}) + \frac{s_2}{s_1}(1 - \Phi(\frac{c-m_2}{s_2}))\exp\left(-[r_2'\Sigma_i^{-1}r_2 + m_1^2/s_1^2 + m_2^2/s_2^2 - r_1'\Sigma_i^{-1}r_1]/2\right)} \tag{24}$$

where $r_1 = \hat{\beta}(\mathrm{s}_i) - \Omega_{(m)}\boldsymbol{\delta}_{(m)}$, $r_2 = \hat{\beta}(\mathrm{s}_i) - \Omega_{(jm)}\boldsymbol{\delta}_{(jm)}$, and $\Omega_{(m)}$ and $\boldsymbol{\delta}_{(m)}$ have all elements corresponding to $\{\delta_{1m}, ..., \delta_{pm}\}$ removed. Finally, we give the full conditional for terms with $m > 1$ and $j > 1$. For these terms, if $c^* = -\delta^*_{1m}(\mathrm{s}) - \sum_{k>1, k\neq j} I(\delta^*_{km}(\mathrm{s}) < 0)\delta^*_{km}(\mathrm{s}) \geq 0$ then $\pi = 1$ and $c = \infty$, and if $c^* < 0$ then $c = c^*$ and π is given by (24).

For both full and approximate methods, the spatial variances τ_j^2 have conjugate inverse gamma priors and are updated using Gibbs sampling. The spatial ranges ρ_j and centering distribution parameters Θ are updated individually using Metropolis sampling with Gaussian

candidate distributions.

For all analyses we generate 20,000 MCMC samples and discard the first 10,000 as burn-in. Convergence is monitored using trace plots of the deviance and several representative parameters.

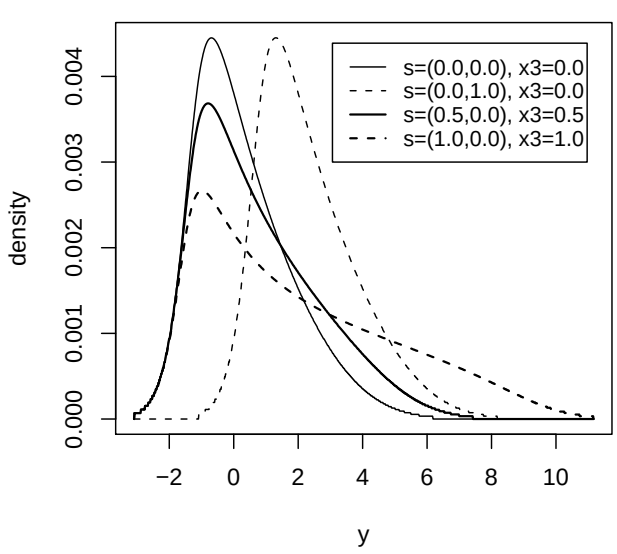
Table 1: Simulation study results. MSE and coverage probabilities are averaged over spatial location, simulated data set, and quantile level. Power is evaluated only for the 95^{th} quantile level for β_3 and is averaged over space and simulated data set.

	Correlated	MSE			Coverage			Power
Method	Residuals	β_1	β_2	β_3	β_1	β_2	β_3	β_3
Quantreg	No	0.51	0.99	1.02	0.89	0.89	0.88	0.35
Bayes - Approx		0.17	0.32	0.42	0.60	0.60	0.61	0.82
Bayes - Full nonspatial		0.31	0.21	0.62	0.82	0.93	0.85	0.57
Bayes - Full spatial		0.13	0.12	0.27	0.86	0.93	0.88	0.79
Quantreg	Yes	0.53	1.02	1.02	0.89	0.88	0.88	0.36
Bayes - Approx		0.18	0.33	0.42	0.59	0.60	0.60	0.83
Bayes - Full nonspatial		0.32	0.20	0.65	0.82	0.92	0.84	0.58
Bayes - Full spatial		0.14	0.10	0.23	0.84	0.93	0.88	0.81

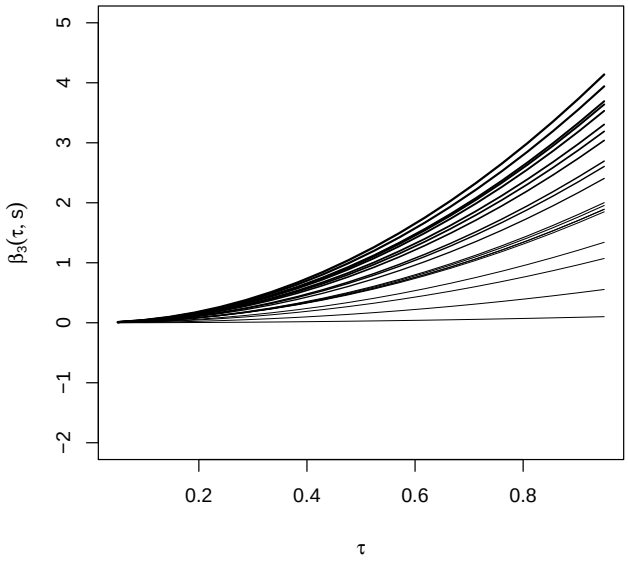
Table 2: Cross-validation results for the Atlanta sub-study.

		Coverage Prob	Average Width		
Model	Covariates	95% interval	95% interval	RMSE	MAD
Gaussian	Linear	0.965	60.4	14.3	9.34
Gaussian	Quadratic	0.969	59.8	14.2	9.07
QR - approx	Linear	0.959	50.4	13.1	7.76
QR - approx	Quadratic	0.947	47.3	12.9	7.62
QR - full	Linear	0.987	70.1	13.2	7.87
QR - full	Quadratic	0.986	69.6	13.0	7.64

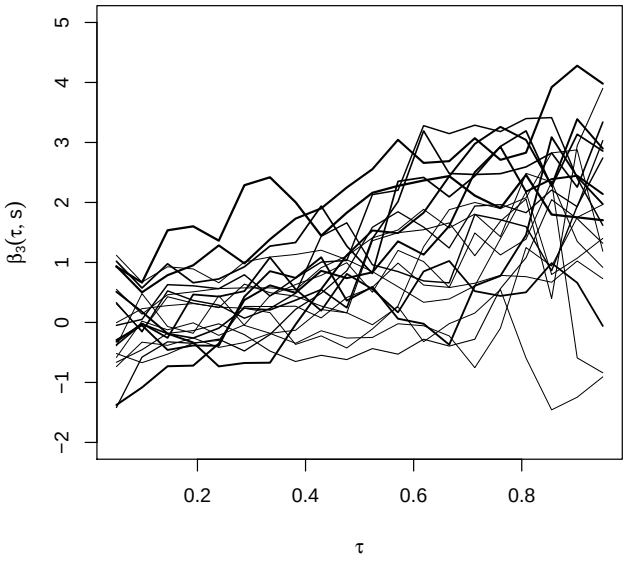
Figure 1: True and estimated quantile curves for the simulation study. Panel (a) gives the true density as a function of space and covariates. Panel (b) plots the true quantile function $\beta_3(\mathrm{s}, \tau)$, Panel (c) plots the usual quantile estimate for one data set, and Panel (d) plots the posterior mean from spatial quantile regression for one data set.



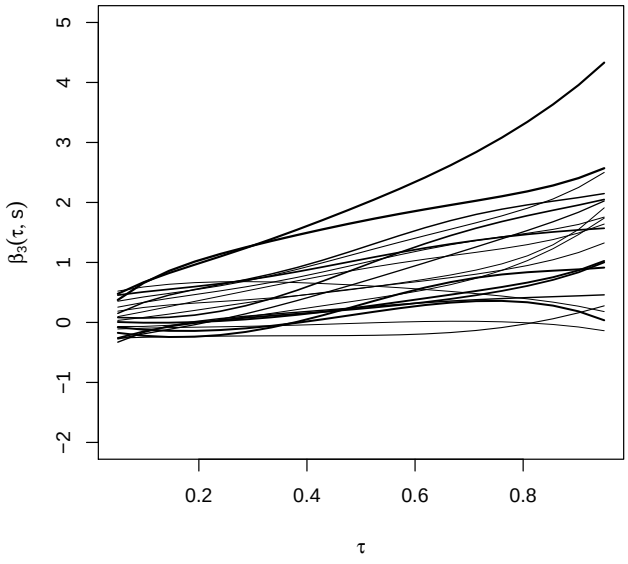
(a) True density



(b) True β_3 by location

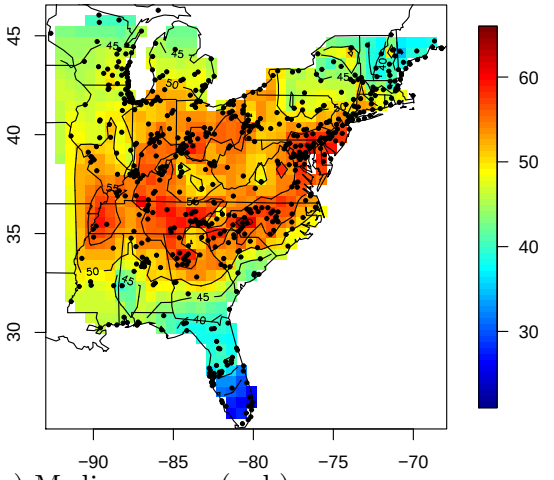


(c) Usual quantile regression estimate

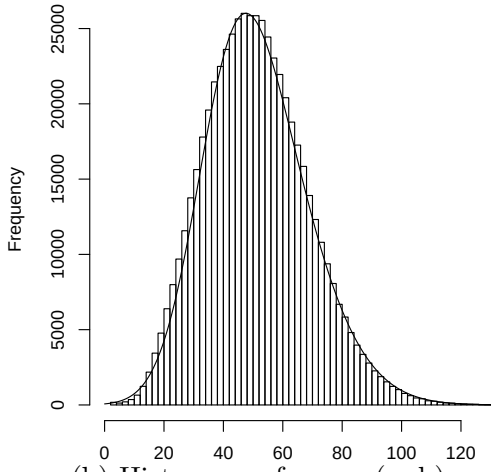


(d) Bayesian spatial quantile regression estimate

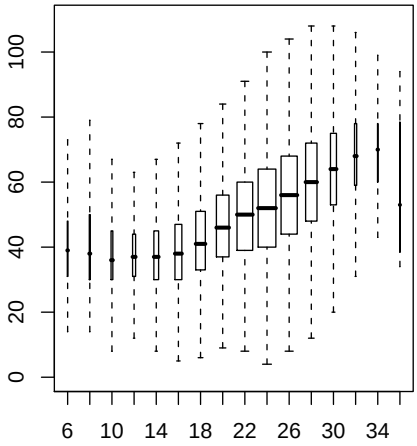
Figure 2: Panel (a) maps the sample median ozone concentration; the points are the 631 monitoring locations. Panel (b) plots the skew-normal(34.5,24.3,1.8) over the histogram of ozone concentrations, pooled over spatial location. Panels (c) and (d) plot ozone by daily average temperature and cloud cover proportion, respectively; the data are pooled over space and the width of the boxplots are proportional to the number of observations in the bins.



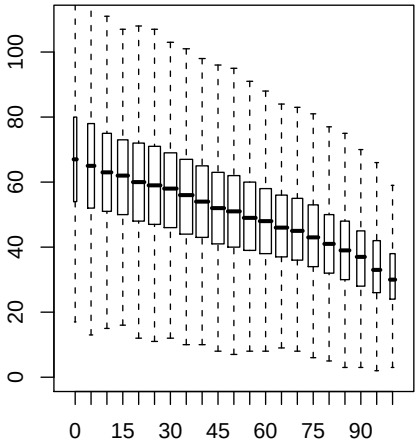
(a) Median ozone (ppb)



(b) Histogram of ozone (ppb)

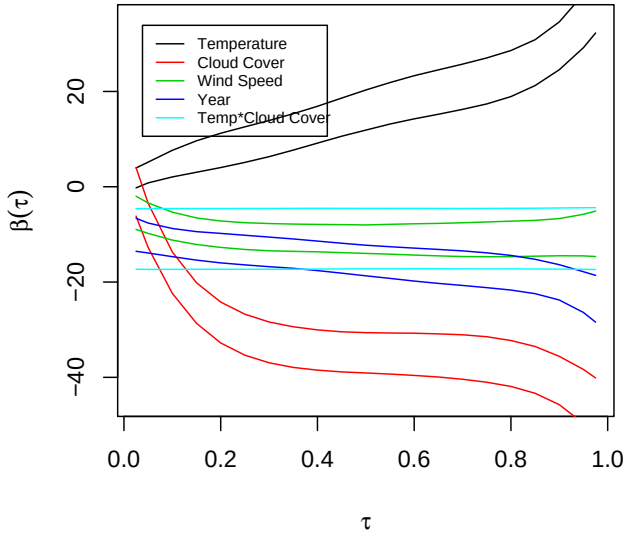


(c) Ozone by temperature (C)

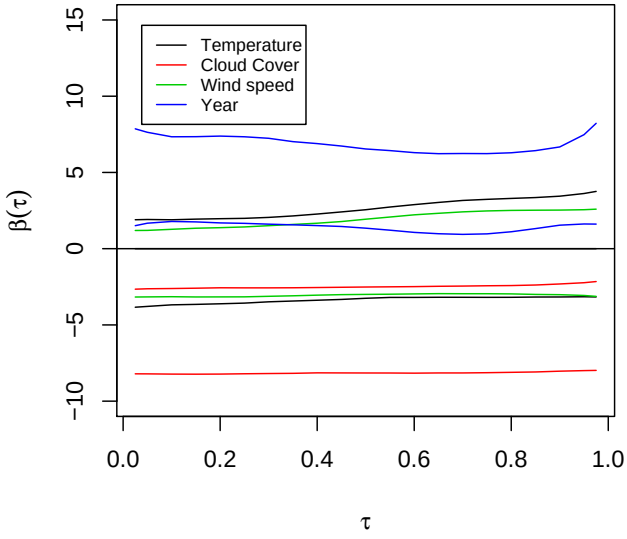


(d) Ozone by cloud cover (%)

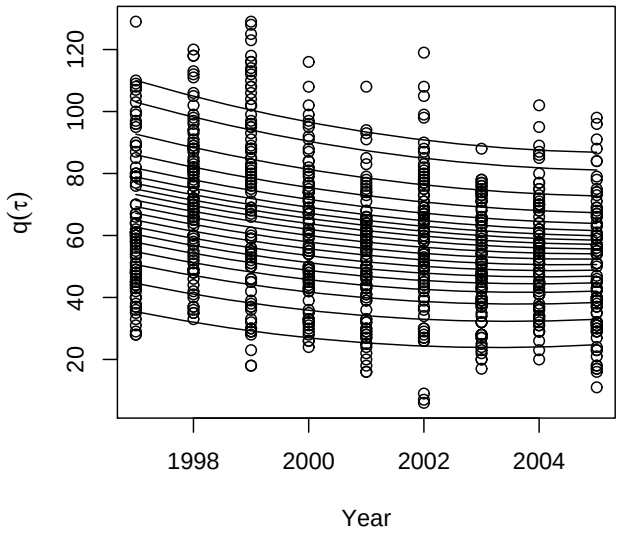
Figure 3: Results for Atlanta sub-study. Panels (a)-(c) plot the results for one location. Panels (a) and (b) give posterior 95% intervals for main effect and second-order quantile curves. Panel (c) plots the data by year, along with the posterior mean quantile curves for several quantile levels ranging from $\tau = 0.05$ to $\tau = 0.95$ with all covariates fixed at 0.5 expect year. Panel (d) plots the posterior mean of the year main effect for each location.



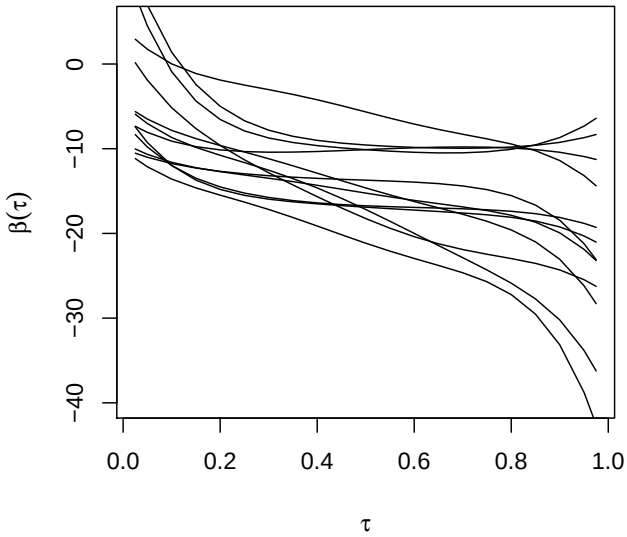
(a) Linear terms



(b) Quadratic terms

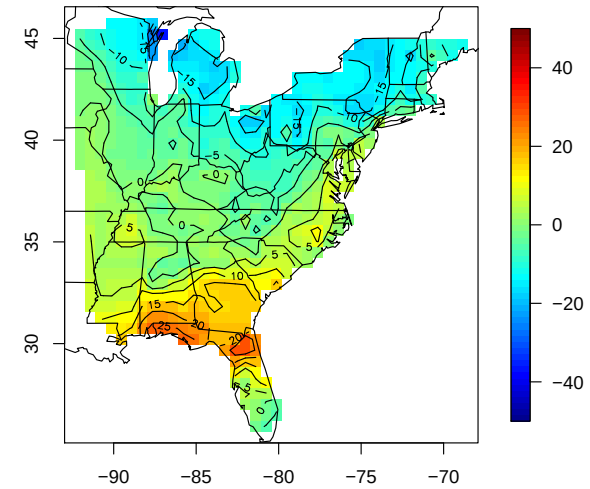


(c) Data and fit by year

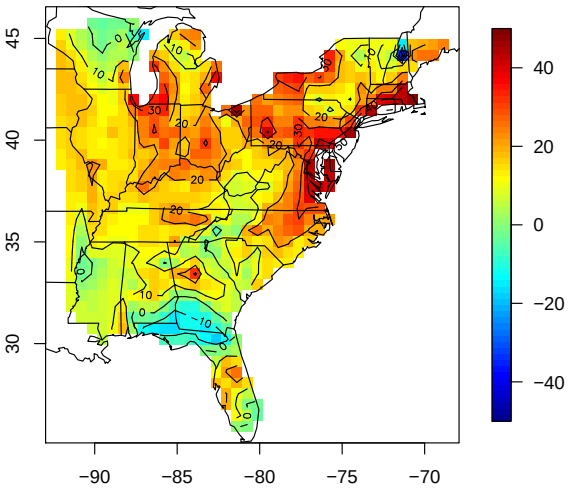


(d) Year main effect by location

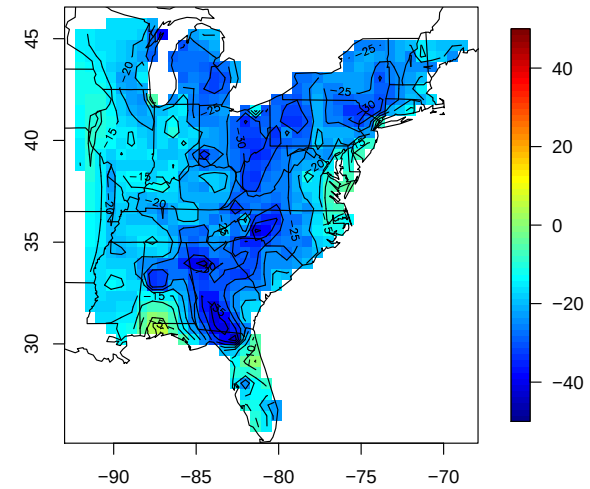
Figure 4: Summary of the posterior mean of $\beta_j(\tau, \mathrm{s})$. Panels (a)-(d) plot linear combinations of $\beta_j(\tau, \mathrm{s})$ as discussed in (21). Panel (f) plots the posterior mean of $\beta_j(0.95, \mathrm{s}) - \beta_j(0.05, \mathrm{s})$. The units are parts per billion in all plots.



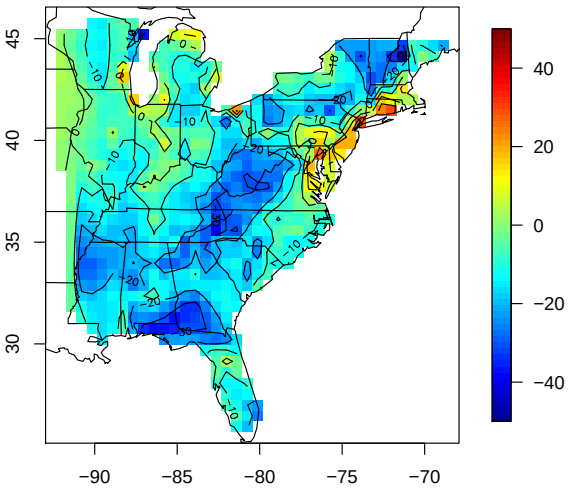
(a) 20^o C, cloud cover 10%, $\tau = 0.95$



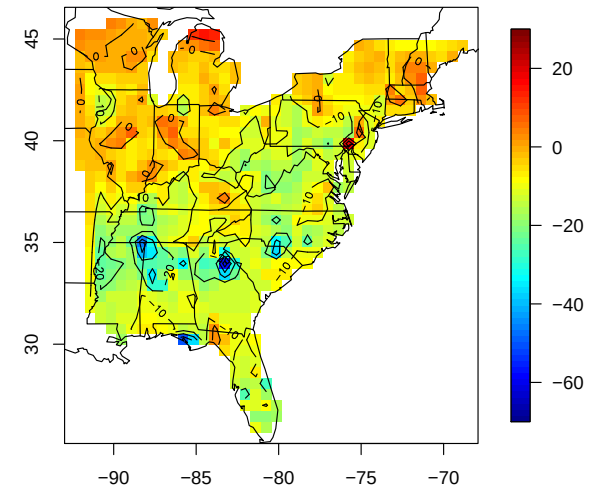
(b) 30^o C, cloud cover 10%, $\tau = 0.95$



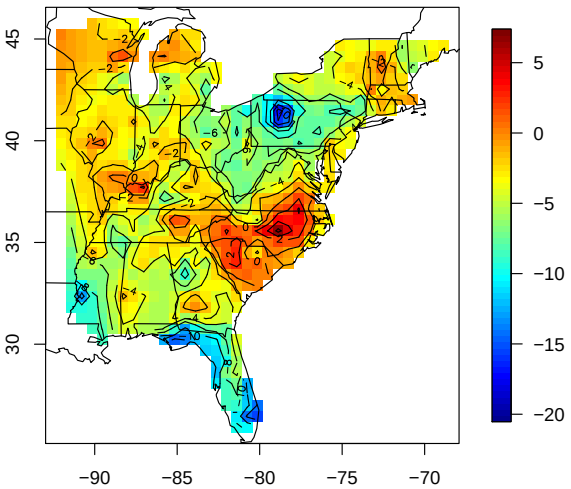
(c) 20^o C, cloud cover 90%, $\tau = 0.95$



(d) 30^o C, cloud cover 90%, $\tau = 0.95$



(e) Linear year, $\tau = 0.95$



(f) Linear year, $\tau = 0.95$ - $\tau = 0.05$

Figure 5: Calibration plots for meteorological data in Georgia.

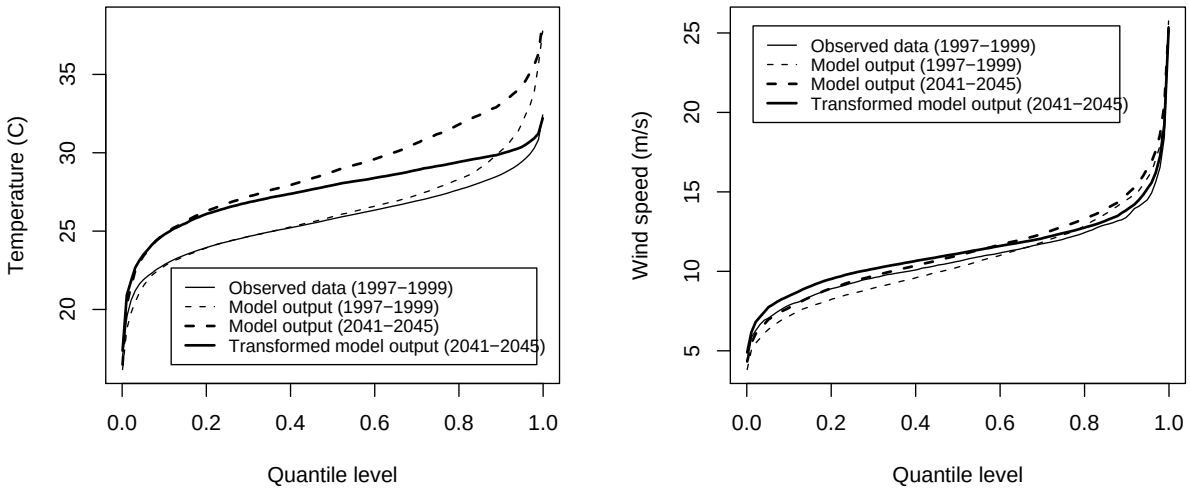
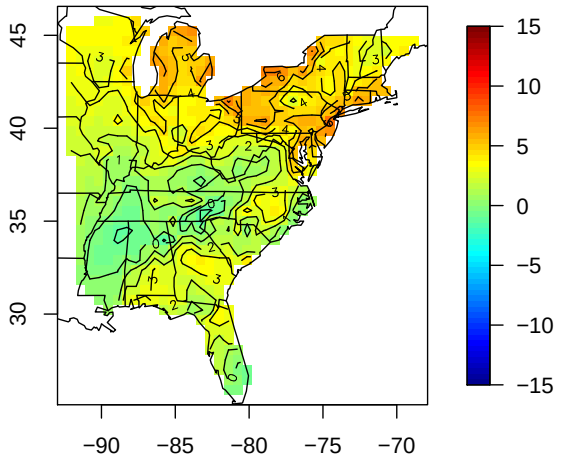
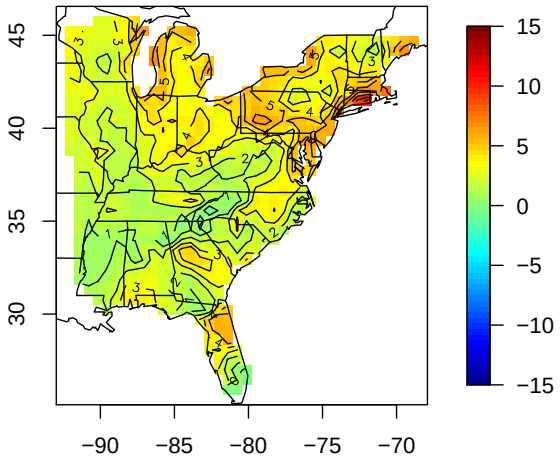


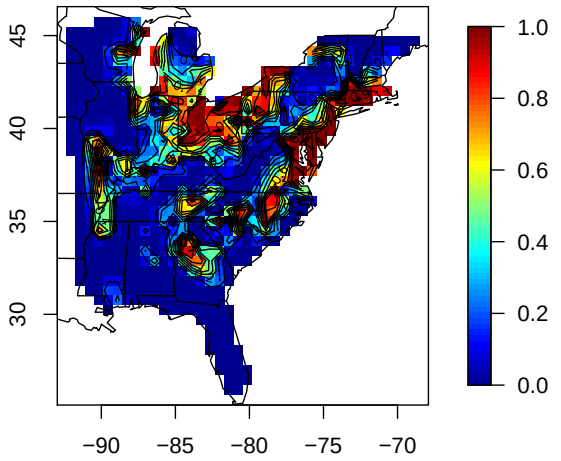
Figure 6: Panels (a) and (b) plot estimates of the change (ppb) in yearly median and 95^{th} quantile due to shifting each daily average temperature by 2^oC (standard errors are less than 3ppb for all sites and quantiles). Panels (c) and (d) give the probabilities that the three-year (2003-2005) average of the fourth-highest daily maximum 8-hour average ozone concentrations exceeds 75 ppb for current temperature and the 2^oC increase in temperature, respectively.



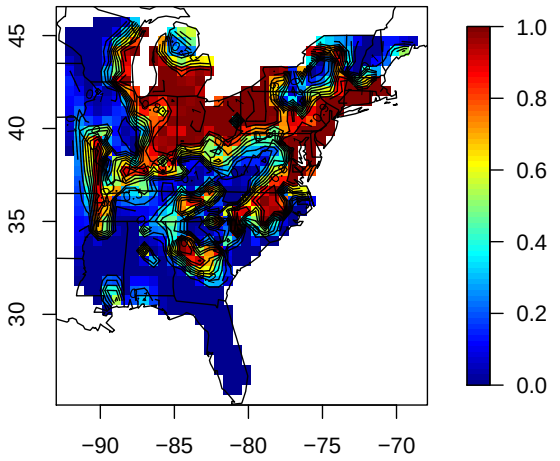
(a) Change in the median



(b) Change in the 95^{th} quantile

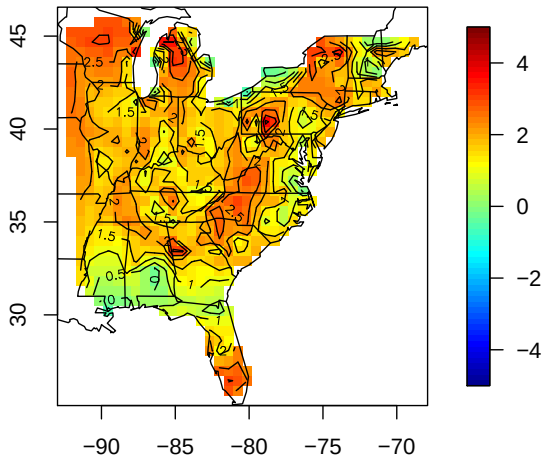


(c) Exceedence probability, current temperature

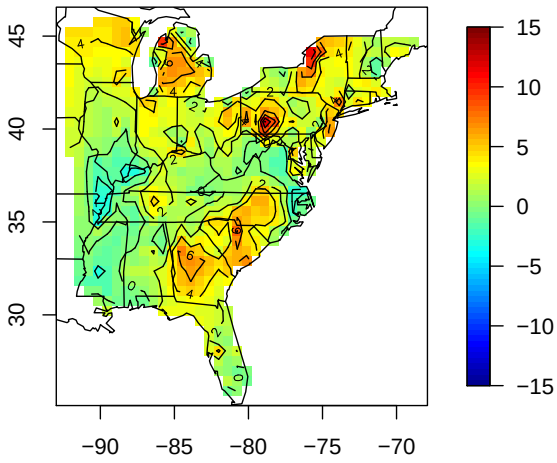


(d) Exceedence probability, increased temperature

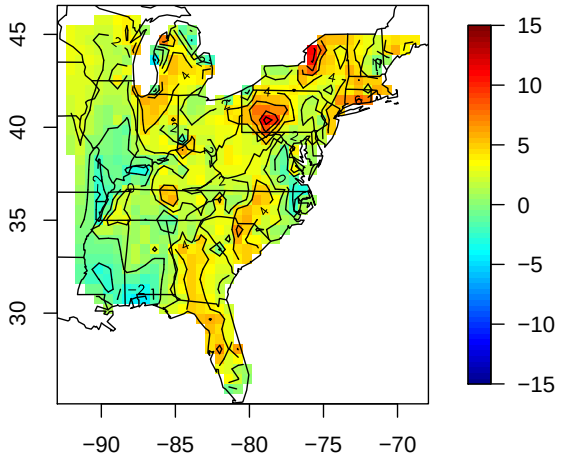
Figure 7: Panel (a) plots the change in average temperature between the calibrated 2041-2045 data and the observed 1997-2005 data. Panels (b) and (c) plot estimates of the change in yearly median and 95^{th} quantile the future and current climate scenarios (standard errors are less than 3ppb for all sites and quantiles). Panel (d) gives the probability that the three-year (2041-2043) average of the fourth-highest daily maximum 8-hour average ozone concentrations exceeds 75 ppb for future climate scenario.



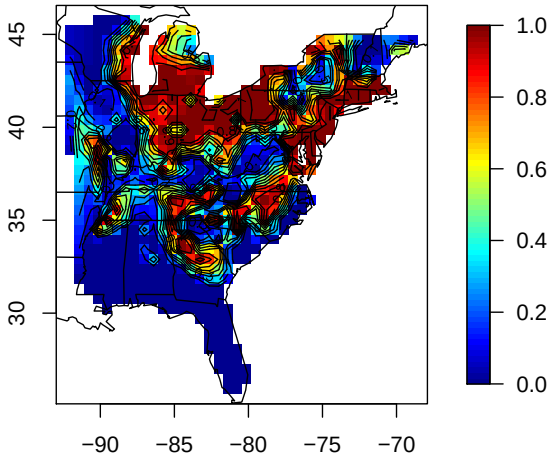
(a) Change in average temperature



(b) Change in the median



(c) Change in the 95^{th} quantile



(d) Exceedence probability, 2041-2045 met