Statistica Sinica **31** (2021), 1167-1187
doi:https://doi.org/10.5705/ss.202019.0002

SMOOTH DENSITY SPATIAL QUANTILE REGRESSION

Halley Brantley, Montserrat Fuentes, Joseph Guinness and Eben Thoma

NC State University, University of Iowa, Cornell University and U.S. Environmental Protection Agency

Abstract: We derive the properties and demonstrate the desirability of a model-based method for estimating the spatially varying effects of covariates on a quantile function. By modeling the quantile function as a combination of I-spline basis functions and Pareto tail distributions, we allow for flexible parametric modeling of the extremes, while preserving the nonparametric flexibility in the center of the distribution. We further establish that the model guarantees the desired degree of differentiability in the density function, and enables us to estimate nonstationary covariance functions that are dependent on the predictors. We use a simulation study to show that the proposed method outperforms other methods in terms of producing efficient estimates of the effects of predictors, particularly in distributions with heavy tails. To illustrate the utility of the model, we apply it to measurements of benzene collected around an oil refinery to determine the effect of an emission source within the refinery on the distribution of the fence line measurements.

Key words and phrases: Conditional density estimation, quantile regression, spatially-varying coefficients.

1. Introduction

Quantile regressions offer an important alternative to the traditional mean regression for problems where the interest lies not in the center of the distribution, but in some other aspect. A large body of literature has developed since the first quantile regression paper was published by Koenker and Bassett (1978) and is reviewed in Koenker (2005). Yu and Moyeed (2001) proposed a form of Bayesian quantile regression employing the asymmetric Laplace distribution (ASL) as the working likelihood, owing to its similarity to the check loss function used by Koenker and Bassett (1978). Both approaches perform separate analyses for each quantile level of interest. When quantiles are estimated separately, there is no guarantee of a valid nondecreasing quantile function. There are several approaches to address this issue. The first is a two-stage method: in the first-stage, the quantiles are fitted separately using one of the above methods. In the second

Corresponding author: Halley Brantley, 9278 Hyland Creek Rd., Bloomington, MN 55437, USA. E-mail: halleybrantley@gmail.com.

stage the estimates are smoothed to ensure monotonicity. This approach has been adopted by numerous authors, including Dette and Volgushev (2008), Neocleous and Portnoy (2008), Chernozhukov, Fernandez-Val and Galichon (2009), Rodrigues and Fan (2017), and Reich, Fuentes and Dunson (2012), who used it as a more computationally efficient Bayesian spatial method. Bondell, Reich and Wang (2010) embed a constraint that ensures monotonicity into the minimization problem, and Cai and Jiang (2015) use prior specifications to ensure the constraints in the Bayesian framework.

The final approach, which we adopt and extend, is to model the entire quantile function jointly using basis functions. This is the approach taken by Reich, Fuentes and Dunson (2012), among others (Reich (2012); Smith et al. (2015)), and is more naturally implemented using a Bayesian framework. Regardless of the approach taken, ensuring monotonicity requires either some form of distributional assumption, or constraints on the quantile regression coefficients and the parameter space of the predictors. Cai and Jiang (2015) demonstrated that when predictors are constrained to be positive, the quantile function is monotonic for every possible predictor value if and only if the basis functions are monotonic. This is the approach taken by Zhou, Fuentes and Davis (2011) and Zhou, Chang and Fuentes (2012), who first proposed the I-spline quantile regression model, the properties of which we derive in this study.

We show how the I-spline model must be constrained in order to guarantee that the resulting density function has a desired number of continuous derivatives. In our numerical studies and in the application to benzene data, we show that this is important to avoid overfitting. We also derive the expectations and covariance of the model, showing that the spatial covariances are nonstationary and depend on the coviarates. The primary difficulty in ensuring differentiability is that, while the center of the distribution is modeled by I-splines, the model has tails that follow a generalized Pareto distribution (GPD). The GPD is used to model the tails because it has been shown to be a natural choice for exceedances over a threshold (Davison and Smith (1990)). Furthermore it provides flexibility owing to the shape parameter, which controls the boundedness and the existence of moments. Zhou, Chang and Fuentes (2012) proposed a two-stage method for estimating parameters. The first stage estimates the GPD shape parameters, which are then fixed in the second stage, which estimates the quantile regression parameters. Here, we assign priors to all parameters and estimate them simultaneously in a Bayesian framework.

As in a mean regression, one way of incorporating a spatial correlation into a quantile regression is to model spatially varying parameters using Gaussian

process priors. Lum and Gelfand (2012) use the ASL for the likelihood, and incorporate a spatial correlation by modeling the error as a function of a Gaussian process and an independent and identically distributed (i.i.d) exponential random variable. For large data sets they propose an asymmetric Laplace predictive process, extending the method introduced by Banerjee et al. (2008). However, the use of the ASL does not allow for a valid posterior inference, because it does not represent the true likelihood of the observations. Yang and He (2015) combined spatial priors with their Bayesian empirical likelihood approach to model the conditional quantiles in the presence of both predictors and spatial correlation. However, their method only allows for effects to be estimated at a small, fixed number of quantile levels. Several other methods of modeling a spatially varying conditional quantile function using basis functions have been proposed (Reich, Fuentes and Dunson (2012); Reich (2012)).

A full description of the I-spline quantile regression model for both independent and spatially correlated data is given in Section 2. Here, we formulate the conditions under which the resulting density has the desired degree of differentiability, and derive the marginal expectations and spatial covariances, which can be nonstationary (Section 3). Our simulation studies demonstrate that ensuring a smooth density can lead to more accurate effect estimates and predictive distributions, as compared with methods that do not ensure differentiability (Section 4). We apply the method to benzene measurements from a petrochemical facility to determine the effects of emission sources on concentrations (Section 5). The final section concludes the paper.

2. Model and Estimation Methods

2.1. Proposed model

We model the quantile function of the stochastic process $Y(s)$ as a linear combination of the predictors:

$$Q(\tau|s,\mathbf{x}(s)) = \beta_0(\tau,s) + \sum_{p=1}^{P} x_p(s)\beta_p(\tau,s), \tag{2.1}$$

where $\mathbf{x}(s) = (x_1(s),\ldots,x_p(s)) \in \mathbb{R}_+^p$ is the vector of predictors observed at location s, $\beta_0(\tau,s)$ is the quantile function at location s when all predictors are zero, and $\beta_p(\tau,s)$ is the effect of predictor p on quantile level τ at location s. We follow the approach of Zhou, Fuentes and Davis (2011), and model $\beta(\tau,s)$ as a linear combination of I-spline basis functions in the center of the distribution.

We denote the mth I-spline basis function evaluated at τ as $I_m(\tau)$, and define the constant basis function $I_0(\tau) = 1$, for all τ. Although I-splines allow for a large degree of flexibility in the center of the distribution, unbounded distributions cannot be estimated using I-splines with a finite number of knots. To solve this issue, we use the quantile function of the GPD to model the relationship between the covariate(s) and the process in the tails of the distribution. The model for $\beta_p(\tau, s)$ can then be expressed as

$$\beta_p(\tau,s) = \begin{cases} \theta_{0,p}(s) - \frac{\sigma_{L,p}(s)}{\alpha_L(s)}\left[\left(\frac{\tau}{\tau_L}\right)^{-\alpha_L(s)} - 1\right] & \tau < \tau_L \\ \sum_{m=0}^{M} \theta_{m,p}(s) I_m(\tau) & \tau_L \le \tau \le \tau_U \\ \left[\sum_{m=0}^{M} \theta_{m,p}(s)\right] + \frac{\sigma_{U,p}(s)}{\alpha_U(s)}\left[\left(\frac{1-\tau}{1-\tau_U}\right)^{-\alpha_U(s)} - 1\right] & \tau > \tau_U, \end{cases} \tag{2.2}$$

where τ_L and τ_U are the thresholds between the tails and the center of the distribution, $\theta_{0,p}$ is the location parameter at the lower tail, and $\theta_{m,p}(s)$ represents the coefficient of the mth I-spline basis function and the pth predictor at location s. I-splines are monotonic polynomials formed by integrating normalized B-splines (Fig. 1) (Ramsay (1988)). They are defined on a sequence of knots $\{\tau_L = \tau_0 = \cdots = \tau_k < \cdots < \tau_{M+1} = \cdots = \tau_{M+1+k} = \tau_U\}$, where k represents the degree of the polynomial, and M is the number of nonconstant basis functions. In both the simulation study and the application, we space the knots evenly between zero and one to maximize the number of observations available to estimate the quantile function between each set of knots.

The GPD has three parameters: the shape parameter α, the scale parameter σ, and a location parameter μ. In our parameterization, the location parameter of the lower tail is equal to $\theta_{0,p}(s)$, and the location parameter of the upper tail is equal to $\sum_{m=0}^{M} \theta_{m,p}(s)$. These ensure that the quantile function is continuous. We denote the shape parameters of the lower and upper tails as $\alpha_L(s)$ and $\alpha_U(s)$, respectively, and the scale parameters as $\sigma_{L,p}(s)$ and $\sigma_{U,p}(s)$, respectively. We require the shape parameter to be constant across predictors in order to ensure that the density in the tails follows a parametric distribution. The scale parameters vary by both predictor and location, and allow the predictors to affect the tails differently. When $\alpha < 0$, the support of the GPD is also bounded above; otherwise, the domain is unbounded above. The case when $\alpha = 0$ is interpreted as the limit when $\alpha \to 0$; that is, $(\sigma_{U,p}/\alpha_U)\left[((1-\tau)/(1-\tau_U))^{-\alpha_U} - 1\right]$ is replaced with $-\sigma_{U,p}\log((1-\tau)/(1-\tau_U))$. The expectation exists if α is less than one, and the variance exists if α is less than $1/2$.

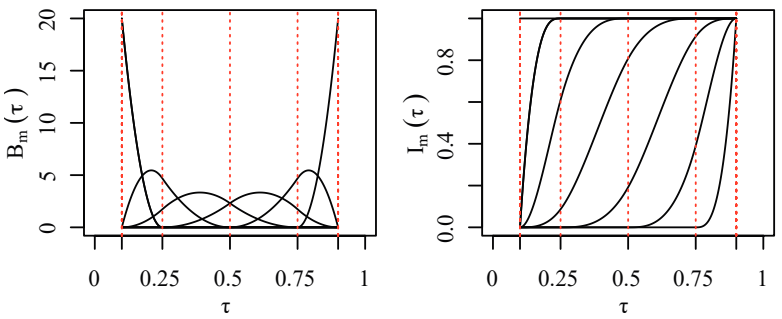

Figure 1. Example set of normalized B-spline (left) and corresponding I-spline (right) basis functions. Dotted vertical lines indicate knot locations.

This model formulation ensures a quantile function that is continuous and differentiable at all but a finite number of points. We can thus exploit the result of Tokdar and Kadane (2012) who demonstrated that a differentiable and invertible quantile function corresponds to the density

$$f(y) = \frac{1}{Q'(Q^{-1}(y))}. \tag{2.3}$$

To ensure the quantile function is monotonic; we introduce latent parameters with Gaussian process priors, $\theta^*_{m,p} \sim \mathcal{GP}(\mu^*_{m,p}, \Sigma^*_{m,p})$, and define $\theta_{0,p}(s) = \theta^*_{0,p}(s)$ and $\theta_{m,p}(s) = \exp(\theta^*_{m,p}(s))$, for $m > 0$. Using this formulation, the resulting $\theta_{m,p}(s)$ is modeled as a log Gaussian process. No constraints are placed on $\theta_{0,p}$, which allows predictors to have a negative effect on the response.

The model formulation has many advantages, including the ability to allow the effect of each predictor to vary by quantile level and by spatial location, while guaranteeing a valid quantile function. It can also accommodate a variety of tail distributions, including both bounded and unbounded tails. Furthermore, we show in Section 3 that we can guarantee the degree of differentiability of the corresponding density function.

Reich (2012) proposed a similar model by constructing the quantile function using parametric Gaussian basis functions. Although these functions allow for straightforward evaluation of the density, they do not guarantee a differentiable quantile function, which results in a noncontinuous density function (Fig. 2). Our simulation and applied data analysis show that constraining the density to be continuous and differentiable can result in better parameter estimates and out-of-sample scores.

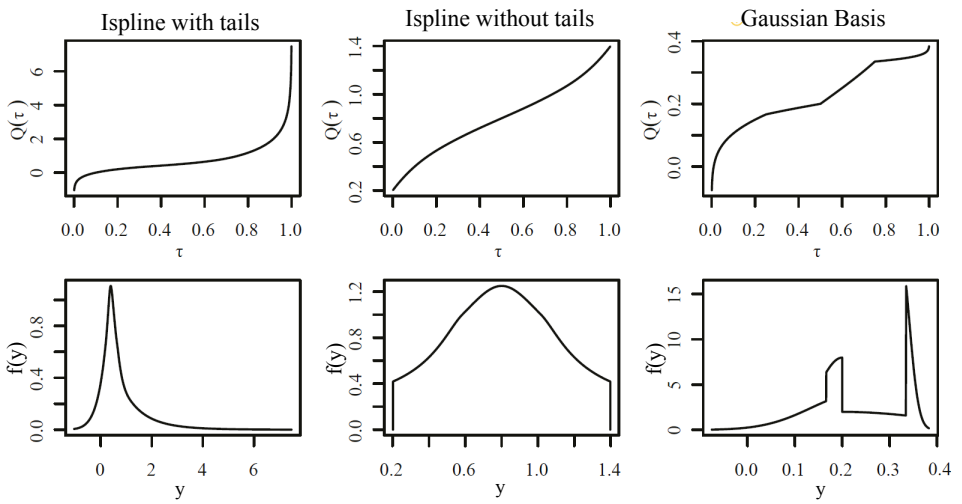

Figure 2. Examples of quantile functions (top row) and the corresponding density functions (bottom row), constructed using different bases.

2.2. Estimation details

We estimate the parameters using MCMC sampling and the R programming language. We calculate the likelihood by inverting the quantile function. Our formulation of the likelihood can be inverted analytically when τ is less than τ_L or greater than τ_U. When $\tau_L < \tau < \tau_U$, the quantile function is a polynomial in τ, and can be inverted either by finding the roots of the cubic polynomial (for I-splines of order three) or by using Halley's root finding algorithm (Hansen and Patrick (1976)). Then the likelihood is calculated using Equation (3.2), as described in the next Section.

To ensure that the degree of smoothness at the thresholds matches the degree of smoothness at the internal knots, we sample $\theta_{1,p}$, $\theta_{M,p}$, α_L, and α_U, and define $\theta_{2,p}$, $\theta_{M-1,p}$, $\sigma_{L,p}$, and $\sigma_{U,p}$ in accordance with Proposition 1 and Theorem 1 (see Section 3). The latent parameters $\theta^*_{m,p}$ are given Gaussian process priors with mean $\mu^*_{m,p}$ and spatial variance $\eta^2_{m,p}$, and the nugget variance is fixed to $\lambda^2_{m,p} = 0.01(\eta^2_{m,p})$. The prior parameters are updated using a Gibbs update from the full conditionals. The latent parameters are updated using the Metropolis–Hastings method with the step size tuned to have an acceptance rate between 0.3 and 0.7.

Latent parameters are also used to sample the tail shape parameters, with $\alpha = \log(\alpha^* - 0.4) - \log(0.4 - \alpha^*)$ and $\alpha^* \sim \mathcal{GP}(\mu_\alpha, \eta^2_\alpha)$. This ensures that the range of α is constrained to $(-0.4, 0.4)$. The upper bound prevents invalid

quantile functions, and the lower bound helps ensure convergence. Conjugate hyperpriors are used in both the simulation study and the application, with $\mu_\alpha \sim N(0,1)$ and $\eta_\alpha^{-2} \sim \text{Gamma}(2, 0.1)$. The prior parameters are updated using a Gibbs sampling step. Theorem 1 is used to update the corresponding θ to ensure the differentiability of the quantile functions. The tail scale parameters, σ_p, are updated using Proposition 1. The knots for the I-splines are evenly spaced between zero and one.

3. Model Properties

3.1. Validity of the quantile function

Assuming an I-spline order $k > 1$, the proposed quantile function is continuous everywhere and is differentiable for all values of $\tau \in (0,1)$, except τ_L and τ_U. Thus, a necessary and sufficient constraint to ensure a valid quantile function is $Q'(\tau) \geq 0$, for all τ at which the derivative exists. For all values of τ such that $\tau_L < \tau < \tau_U$, $Q'(\tau) = \sum_{m=1}^{M} B_m(\tau) \sum_{p=1}^{P} \theta_{m,p} x_p$. By definition, $B_0(\tau) = 0$ for all τ, and $B_m(\tau) \geq 0$ for all m and τ. Without loss of generality, we henceforth assume that the predictors are all nonnegative; that is, $\mathbf{x} \in \mathbb{R}_+^P$. Therefore a sufficient constraint to ensure a valid quantile function is $\theta_{m,p} \geq 0$, for all p and $m > 0$. If $\sigma_{L,p} > 0$ for any p and $\tau \leq \tau_L$, $Q'(\tau) = \sigma_{L,p} x_p \tau^{-\alpha_L - 1} \tau_L^{\alpha_L} > 0$. Similarly, if $\sigma_{U,p} > 0$ for any p, $Q'(\tau) > 0$ when $\tau \geq \tau_U$.

3.2. Continuity and differentiability

In many cases, such as the application described below, it is desirable to ensure that the density is continuous and smooth. Proposition 1 establishes the conditions for the continuity of the density function.

Proposition 1. *Let Y have a quantile function as defined in* (2.1) *and* (2.2)*, with $\sigma_{L,p} > 0$ for at least one p. Then the density of Y is continuous at $Q(\tau_L|\mathbf{x}, \Theta)$, for any $\mathbf{x} \in \mathbb{R}_+^P$, if and only if*

$$\theta_{1,p} = \frac{\sigma_{L,p}}{\tau_L I_1'(\tau_L)}, \tag{3.1}$$

for all p. Similarly, given $\sigma_{U,p} > 0$ for at least one p, the density of Y is continuous at $Q(\tau_U|x, \Theta)$ if and only if

$$\theta_{M,p} = \frac{\sigma_{U,p}}{(1-\tau_U) I_M'(\tau_U)}. \tag{3.2}$$

Having clarified the conditions for a continuous density, which can be viewed

as the zeroth-order differentiability, Theorem 1 proceeds to establish the conditions for qth-order differentiability of the density function of Y.

Theorem 1. *Let Y have a quantile function as defined in* (2.1) *and* (2.2)*, with an I-spline basis order greater than $q+1$, and a density that is continuous and $(q-1)^{th}$ order differentiable at $Q(\tau_L)$. If α_L is constrained such that Eq.* (3.3) *does not result in $\theta_{q+1,p} < 0$, then Y has a density that is qth-order differentiable at $Q(\tau_L)$, for any $\mathbf{x} \in \mathbb{R}_+^P$, if and only if*

$$\theta_{q+1,p} = \frac{1}{I_{q+1}^{(q+1)}(\tau_L)} \left\{ \frac{-\sigma_{L,p}}{\alpha_L \tau_L^{q+1}} (-\alpha_L - q)_{q+1} - \sum_{m=1}^{q} \theta_{m,p} I_m^{(q+1)}(\tau_L) \right\} \tag{3.3}$$

where $I_{q+1}^{(q+1)}(\tau_L)$ is the $(q+1)$th-order derivative of the $(q+1)$th I-spline basis function, $(-\alpha_L - q)_{q+1} = \prod_{j=0}^{q}(-\alpha_L - j)$.

The conditions that guarantee differentiability at τ_U are similar, and are given in the Supplementary Material. Combined with the positivity constraint on θ, these results imply that the shape parameters have an upper bound that is a function of the knot placement. Ensuring a density that is first-order differentiable yields the possible values for α_L being bounded above by $-1-\tau_L(I_1^{(2)}(\tau_L)/I_1'(\tau_L))$. This bound is a function of $I_1'(\tau_L)$ and $I_1^{(2)}(\tau_L)$, which are functions of the first two knot locations. We can still model any tail behavior, provided the outermost knots are placed sufficiently close.

3.3. Expectations and covariance

While our models allow for flexible nonGaussian distributions, sometimes the first two moments are of interest (e.g., for the best linear unbiased prediction). We now elaborate on the various types of covariance structure that can be estimated using the proposed model. We model the covariances of the latent parameters $\theta_{m,p}^*$ using the covariance function C, such that $Cov[\theta_{m,p}^*(s), \theta_{m,p}^*(s')] = \eta_{m,p}^2 C(s,s')$ and $Var[\theta_{m,p}^*(s)] = \eta_{m,p}^2 + \lambda_{m,p}^2$. Consequently, the expectation of $\theta_{m,p}$ can be expressed as $E[\theta_{m,p}] = \mu_{m,p} = \exp[\mu_{m,p}^* + (\eta_{m,p}^2 + \lambda_{m,p}^2)/2]$, and the covariance of $\theta_{m,p}$ is

$$\Sigma_{m,p}(s,s') = \mathrm{Cov}[\theta_{m,p}(s), \theta_{m,p}(s')] = \mu_{m,p}^2(\exp[\eta_{m,p}^2 C(s,s')] - 1). \tag{3.4}$$

In this section, we describe the covariance of the case when $\tau_L = 0$ and $\tau_U = 1$. We elaborate on other cases in the Supplementary Material. Under these conditions, the conditional expectation of $Y(s)|\Theta(s), \mathbf{x}(s)$ is

$$E[Y(s)|\Theta(s), \mathbf{x}(s)] = \int_0^1 Q_Y[\tau|\Theta(s), \mathbf{x}(s)]d\tau = \sum_m \sum_p \theta_{m,p}(s)x_p(s)G_m, \quad (3.5)$$

where $G_m = \int_0^1 I_m(\tau)d\tau$. We further marginalize over the log Gaussian processes $\theta_{m,p}(s)$, with mean $\mu_{m,p}$ and covariance $\Sigma_{m,p}$, to obtain the expectation and covariance of $Y(s)$,

$$E[Y(s)|\mathbf{x}(s)] = \sum_m \sum_p \mu_{m,p}(s)x_p(s)G_m, \quad (3.6)$$

$$\text{Cov}[Y(s), Y(s')|\mathbf{x}(s), \mathbf{x}(s')] = \sum_m \sum_p G_m^2 x_p(s)x_p(s')[\Sigma_{m,p}(s, s')]. \quad (3.7)$$

This simple case shows that the covariance is dependent on the values of the predictors, in addition to the covariance functions of the latent parameters. This dependence on the predictors can result in nonstationary covariances if x_p varies across space, even if $C(s, s')$ is stationary. Other authors have used covariates to construct nonstationary Gaussian processes. See Risser and Calder (2015), and the references therein, for several examples.

4. Simulation Study

Our simulation studies demonstrate the superior efficiency of the proposed I-spline quantile regression method (IQR) using four designs from data-generating models that are not in the proposed model class (Table 1). The designs include cases with both light tails (D1 and D3) and heavy tails (D2 and D4), and with (D3 and D4) and without (D1 and D2) spatial correlation. The designs illustrate the flexibility of the proposed method compared with previously established methods.

For each design, the observed response is indexed as $y_t(s_i)$, where $t \in \{1, \ldots, n\}$ indexes the observations at a given location s_i, with $i \in \{1, \ldots, S\}$. The predictor vector $\mathbf{x}_{1,t}$ is generated by sampling from a uniform random variable in D1 and D2. In D3 and D4, $\mathbf{z}_t$ is generated by sampling from a Gaussian process with mean zero, and an exponential covariance with range one and $\mathbf{x}_{1,t} = \Phi^{-1}(\mathbf{z}_t)$, where $\Phi^{-1}(\tau)$ is the quantile function of the standard normal. The predictor $\mathbf{x}_{2,t}$ is generated by sampling from a uniform random variable in all designs. The observed response is generated by drawing an independent random uniform variable $u_t(s_i)$ and setting:

$$y_t(s_i) = \beta_0(u_t(s_i), s_i) + \beta_1(u_t(s_i), s_i)x_{1,t}(s_i) + \beta_2(u_t(s_i), s_i)x_{2,t}(s_i). \quad (4.1)$$

In all designs, we assume multiple observations are obtained for each location.

Table 1. True parameter functions, by design, used in the simulation study. The location is given as $s = (s_1, s_2)$, $\Phi^{-1}(\tau)$ represents the quantile function of the standard normal evaluated at τ, and Q_{Pareto} represents the quantile function of the Pareto distribution with the given parameters.

	$\beta_0(\tau, s)$	$\beta_1(\tau, s)$	$\beta_2(\tau, s)$
D1	$0.1\Phi^{-1}(\tau)$	0.3τ	$Q_{Pareto}(\tau, \alpha = -0.2 \quad, \mu = 0 \quad, \sigma = 0.1)$
D2	$0.1\Phi^{-1}(\tau)$	0.3τ	$Q_{Pareto}(\tau, \alpha = \quad 0.3 \quad, \mu = 0 \quad, \sigma = 0.3)$
D3	$(0.05 + 0.2s_1s_2)\Phi^{-1}(\tau)$	$0.3e^{s_2} + 0.2\tau$	$Q_{Pareto}(\tau, \alpha = -0.1 \quad, \mu = 0 \quad, \sigma = 0.1)$
D4	$(0.05 + 0.2s_1s_2)\Phi^{-1}(\tau)$	$0.3e^{s_2} + 0.2\tau$	$Q_{Pareto}(\tau, \alpha = \quad 0.4s_1, \mu = 0.3, \sigma = 0.4)$

For each design, we simulate $B = 50$ independent data sets. In D1 and D2, we simulate 1,000 observations per data set, assuming all observations are from a single location, and thus independent. In D3 and D4, we use $S = 16$ locations, evenly spaced on a unit square, and simulate 100 observations per site, for a total of 1,600 observations per data set. For each of the data sets we randomly assign 10% of the data to be used as validation data for the out-of-sample calculations, and use the other 90% as training data.

We compare the estimates from the proposed model (IQR) with those from the model using parametric Gaussian basis functions (GAUS) proposed by Reich (2012), and with the noncrossing quantile regression estimates (NCQR) proposed by Bondell, Reich and Wang (2010). For the IQR and GAUS methods, four basis functions were used, and the estimates of $\beta(\tau, s)$ represent the means of the corresponding posterior samples. For the NCQR method, the estimates of $\beta(\tau, s)$ are obtained by minimizing the check loss function combined with the noncrossing constraint. The GAUS model allows for spatially varying coefficients and spatial correlation, whereas the NCQR method assumes i.i.d. samples.

For our proposed IQR method, we draw 25,000 MCMC samples, discarding the first 5,000 as burn-in, and monitor the convergence using trace plots of the deviance, as well as several representative parameters. The Gaussian process prior parameters $\mu^*_{m,p}$ and $\eta^{-2}_{m,p}$ are given conjugate hyperpriors. Specifically, $\mu^*_{m,p} \sim N(-3, 1)$ and $\eta^{-2}_{m,p} \sim \text{Gamma}(0.5, 0.005)$. The unconstrained parameters $\mu^*_{0,p}$ are given the prior $N(0, 10)$. An exponential covariance function with fixed range 0.5 is used.

For the GAUS method, in the simulation study, we draw 25,000 MCMC samples, discarding the first 5,000 as burn-in. A fixed range of 0.5 for the exponential covariance was also used for this method. In both the simulation study, and the application, all parameters are given conjugate hyperpriors, with $N(0, 100)$ used as the hyperprior for the means of the Gaussian processes. The hyperpriors for the spatial precision are $\text{Gamma}(0.5, 0.005)$.

We index the quantile levels at which the methods are compared by $j \in 1, \ldots, J$. For each quantile level, τ_j, and simulated data set replicate, $b \in \{1, \ldots, B\}$, the estimated coefficients $\widehat{\beta}_p(\tau_j, s_i)$ were compared using the root mean integrated squared error (RMISE). The RMISE for simulated data set b was calculated for a given β_p and sequence $\tau_1, \ldots, \tau_J$:

$$RMISE(\beta_p)^{(b)} = \sqrt{\frac{1}{S}\sum_{i=1}^{S}\sum_{j=1}^{J}\delta_j\left[\widehat{\beta}_p(\tau_j, s_i)^{(b)} - \beta_p(\tau_j, s_i)\right]^2}, \tag{4.2}$$

where $\delta_j = \tau_j - \tau_{j-1}$. The means and standard errors of the RMISEs, as well as the coverage of the 95% confidence (NCQR) or credible (IQR and GAUS) intervals, were then calculated for each method and design (Table 2).

The IQR and GAUS methods both produce density estimates. The NCQR method does not estimate the entire quantile function and, therefore, cannot be used to create a density estimate without substantial additional calculation. To evaluate the predictive densities, we use the log score, which is the logarithm of the predicted density evaluated at the training and validation data. This is a strictly proper scoring rule (Gneiting and Raftery (2007)). We calculate the log score for each observation as the log of the posterior mean of the predictive density evaluated at the observation. The total log score for each data set is calculated as the mean of the log scores for the individual observations. The mean and standard error by the simulation design are calculated using the total log score values of the 50 simulated data sets.

We compare all three methods using $\tau = \{0.05, 0.06, \ldots, 0.94, 0.95\}$. Four nonconstant basis functions per predictor were used in both the IQR and GAUS methods. The results given in Table 2 show that, although the three methods perform similarly for D1 (independent, light tails), the IQR method performs substantially better than the GAUS method does in the heavy-tailed designs (D2 and D4), and substantially better than the NCQR method does in the spatially varying designs (D3 and D4). Compared with the nominal coverage rate of 0.95, the IQR method has good coverage for all of the designs, with the lowest coverage being 0.88 for β_1 in D1. The GAUS method shows poor coverage for D2, and the NCQR method exhibits poor coverage for D3 and D4.

Unlike the NCQR method, both our method and the GAUS method assume parametric forms for the tails, and so can be used to estimate parameter effects on extreme quantiles. We compare the parameter estimates for these two methods evaluated at $\tau = \{0.950, 0.951, \ldots, 0.994, 0.995\}$ in Table 3. Our method outperforms the other methods in all cases, except D1 β_1, which is a linear function of

Table 2. Comparison of fitted $\beta(\tau)$ functions $\tau = (0.05, 0.06, \ldots, 0.94, 0.95)$. COV represents the coverage of the 95% credible interval (IQR and GAUS) or confidence interval (NCQR).

	β_0			β_1			β_2		
	RMISE	SE	COV	RMISE	SE	COV	RMISE	SE	COV
D1									
IQR	0.014	0.001	0.92	0.027	0.001	0.89	0.022	0.002	0.91
GAUS	0.016	0.001	0.92	0.022	0.002	0.93	0.025	0.002	0.93
NCQR	0.017	0.001	0.96	0.025	0.001	0.97	0.026	0.001	0.97
D2									
IQR	0.019	0.001	0.93	0.035	0.002	0.91	0.047	0.002	0.90
GAUS	0.038	0.005	0.83	0.065	0.009	0.83	0.113	0.007	0.76
NCQR	0.025	0.001	0.98	0.045	0.002	0.97	0.051	0.002	0.97
D3									
IQR	0.029	0.001	0.95	0.050	0.002	0.95	0.027	0.001	0.99
GAUS	0.027	0.001	0.97	0.046	0.001	0.97	0.032	0.001	0.98
NCQR	0.050	0.000	0.64	0.201	0.001	0.16	0.026	0.002	0.92
D4									
IQR	0.038	0.001	0.94	0.062	0.002	0.96	0.094	0.004	0.94
GAUS	0.094	0.047	0.94	0.104	0.034	0.95	0.182	0.027	0.93
NCQR	0.054	0.001	0.75	0.197	0.001	0.24	0.112	0.002	0.84

Table 3. Comparison of fitted $\beta(\tau)$ functions $\tau = (0.950, 0.951, \ldots, 0.994, 0.995)$. COV represents the coverage of the 95% credible interval (IQR and GAUS) or confidence interval (NCQR).

	β_0			β_1			β_2		
	RMISE	SE	COV	RMISE	SE	COV	RMISE	SE	COV
D1									
IQR	0.0047	0.0004	0.96	0.0095	0.0010	0.89	0.0072	0.0007	0.94
GAUS	0.0051	0.0005	0.98	0.0089	0.0009	0.90	0.0077	0.0006	0.98
D2									
IQR	0.0139	0.0014	0.95	0.0377	0.0022	0.85	0.0810	0.0051	0.76
GAUS	0.0266	0.0054	0.78	0.0469	0.0109	0.75	0.0913	0.0045	0.57
D3									
IQR	0.0094	0.0003	0.97	0.0124	0.0004	0.95	0.0089	0.0005	0.99
GAUS	0.0119	0.0004	0.95	0.0139	0.0005	0.99	0.0132	0.0005	0.99
D4									
IQR	0.0196	0.0012	0.96	0.0286	0.0027	0.93	0.1424	0.0048	0.90
GAUS	0.0802	0.0476	0.94	0.0666	0.0314	0.97	0.2007	0.0271	0.81

Table 4. Comparison of mean estimated log scores.

	In-sample		Out-of-sample	
	Mean	SE	Mean	SE
D1				
IQR	0.339	0.003	0.315	0.008
GAUS	0.356	0.003	0.322	0.008
D2				
IQR	−0.223	0.004	−0.254	0.017
GAUS	−0.219	0.005	−0.288	0.022
D3				
IQR	0.476	0.003	0.418	0.010
GAUS	0.536	0.003	0.419	0.009
D4				
IQR	−0.191	0.004	−0.238	0.012
GAUS	−0.126	0.006	−0.287	0.022

τ.

The results of the log score comparisons are consistent with the parameter estimates (Table 4). However, the GAUS method consistently produces higher log scores in-sample than the IQR method does. Because the likelihood is not constrained to be continuous in the GAUS method, very large likelihood values can be obtained for the in-sample observations (Fig. 2). In the heavy-tailed designs, the IQR method results in higher out-of-sample log scores.

5. Application

5.1. Data

An amendment to the U.S. National Emission Standards for Hazardous Air Pollutants for petroleum refineries requires the use of two-week time-integrated passive samplers at specified intervals around the facility fence line to establish the levels of benzene in the air (EPA (EPA(2014)). The utility of fence line measurements as a method of controlling emissions is contingent on their distributions being dependent on nearby sources within the facility. To evaluate the efficacy of passive samplers in monitoring benzene emissions from petroleum refineries, researchers from the US EPA Office of Research and Development conducted a yearlong field study in collaboration with Flint Hills Resources in Corpus, Christi, TX (Thoma et al. (2011)). Preliminary analyses found that under consistent wind conditions, downwind concentrations were statistically higher than upwind concentrations (Thoma et al. (2011)). More sophisticated modeling should reveal the contributions of individual sources to the concentrations observed at the

fence line. Modeling these concentrations requires an extra level of complexity, because near-source air pollutant measurements typically exhibit strong spatial correlation, along with nonstationary and nonGaussian distributions, even after a transformation. Both the spatial covariance and the distribution of the pollutant concentrations can vary as a function of wind and emission source location. Accurately modeling the entire distribution and spatial structure of the pollutant concentrations should improve the quality of inferences related to the strengths of the known sources. Additionally, owing to the stochastic nature of the dispersion and variation in the background pollutant concentration levels, the effect of a specific source on the pollutant distribution might not be detected by a mean regression. Of particular concern, both for exposure and compliance evaluation, are the source effects on the upper tail of the distribution, particularly the 95th percentile.

The measurements used in this study were collected between December 3, 2008, and December 2, 2009 around the Flint Hills West Refinery (Thoma et al. (2011)). The samplers were attached to the boundary fence around the facility, approximately 1.5 m above the ground at 15 locations (Fig. 3). In addition, one sampler (633) was deployed at a nearby Texas Commission on Environmental Quality (TCEQ) continuous air monitoring station (CAMS). A total of 406 two-week time-integrated benzene concentration measurements were collected over the course of the year, and were used in the analysis. Hourly temperature, wind speed, and direction were also measured at TCEQ CAMS 633.

The concentrations exhibited both spatial and temporal trends (Fig. 3). In particular, the variance increased dramatically during the summer months. The highest concentrations were observed on the northern edge of the refinery (sites 360, 20, and 50), while the lowest concentrations were observed on the southern edge (sites 250, 633, and 270). The increase in variance can partly be explained by meteorology (Fig. 4). During the summer, the wind blows consistently from the southeast, while during the rest of the year, the wind direction is more evenly distributed.

A visual analysis of the concentrations and wind roses for the hourly measurements at each time period suggests that the concentrations are correlated with a source within the refinery. Two probable emission source locations, $\mathbf{e}_1$, and $\mathbf{e}_2$, were selected using the reported emission inventory. To determine the effect of the emission sources on the distribution of the benzene concentration, we denote the tth observed value of the benzene concentration at site s_i as $y_t(s_i)$, where $i = 1, \ldots, 16$ and $t = 1, \ldots, 26$. Then we model the quantile function of Y using equation (2.1) and (2.2). Our full model includes an intercept and three

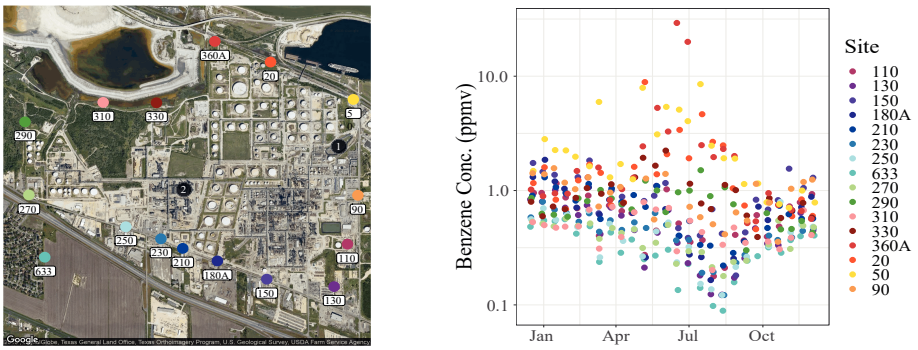

Figure 3. Benzene measurements by time and location. Source locations, $\mathbf{e}_1$ and $\mathbf{e}_2$, are labeled one and two. Points have been jittered slightly to improve visibility.

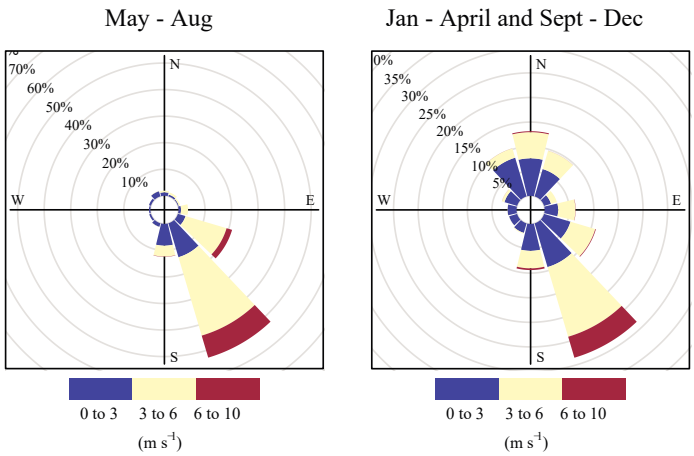

Figure 4. Wind roses for different seasons.

predictors: transport from source 1, transport from source 2, and temperature.

The predictors that represent transport from a source are calculated from the observed hourly wind vectors and relative spatial locations of the source and measurement. The tth observed value of the transport from source 1 to location s_i is defined as

$$x_{1,t}(s_i) = \sum_{h=1}^{336} \left\{ \max \left(\frac{\mathbf{w}_{t,h} \cdot (\mathbf{e}_1 - \mathbf{s}_i)}{||(\mathbf{e}_1 - \mathbf{s}_i)||}, 0 \right) \right\}, \tag{5.1}$$

where $\mathbf{e}_1$ is the location of emission source 1, and $\mathbf{s}_i$ is the measurement location. Each hourly wind vector, $\mathbf{w}_{t,h}$, for the two-week period, with $h = 1, \ldots, 336$, was transformed into the same coordinate system and projected onto the vector from

the source to the measurement $(\mathbf{e}_1 - \mathbf{s}_i)$. Assuming a constant emission source, the resulting scalar quantity represents the amount of pollutant transported from $\mathbf{e}_1$ to $\mathbf{s}_i$, ignoring the effects of vertical dispersion. When the wind is blowing from $\mathbf{s}_i$ toward $\mathbf{e}_1$, transport from $\mathbf{e}_1$ is negative. However, owing to finite, small background concentrations, the integrated benzene concentration remains the same rather than decreasing under these conditions. Therefore, the maximum of the transport from $\mathbf{e}_1$ and zero was taken before taking the sum over h in period t (5.1). The transport from source 2 was calculated similarly.

We use 10-fold cross-validation to determine the most appropriate model for the benzene concentrations. Using each fold as a validation data set, the in-sample and out-of-sample log scores were calculated using both the proposed IQR method and the GAUS method proposed by Reich (2012) for each combination of predictors (Table 5). Four basis functions were used for both methods. The priors are the same as those in the simulation study, except that $\eta_{m,p}^{-2} \sim \text{Gamma}(0.1, 0.1)$ was used for both methods. An exponential covariance function and range of 0.5 were used for both methods. The predictors were transformed to be between zero and one before the models were fitted. The IQR method was run for 25,000 samples, discarding the first 5,000. The GAUS method converged more slowly, and so 35,000 samples were drawn, discarding the first 15,000.

For both methods, the in-sample log score tends to increase with the number of predictors included in the model. All of the models with predictors have higher in-sample and out-of-sample log scores than those of the intercept-only model. Of the models with predictors, the one that included all three produced the largest out-of-sample log score for the IQR method, but the lowest out-of-sample log score for the GAUS method, indicating that adding additional predictors exacerbates the probability of over-fitting using the GAUS method. In all cases, the out-of-sample performance of our method is substantially better than that of the GAUS method.

The model was fit to the entire data set to determine the effects of the two sources and the temperature on the distribution of benzene at the fence line. We plot the coefficients by quantile level and location in Figure 5. We can see that the base distribution does not vary as much by location as the effects of the sources and the temperature do. The effects of the sources on the quantiles of the concentrations range from positive to negative, with the majority of the source effects being positive. The negative effects could be due to these sources not being constant over the course of the year. If wind from a given source corresponded to time points when the source was not emitting, this could result in a negative effect on the concentrations. As can be seen in Figure 6, the effect of source 1 on

Table 5. Estimated log scores for training and validation data by method.

Predictors	In-sample				Out-of-sample			
	IQR		GAUS		IQR		GAUS	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
None	−0.29	0.01	−0.16	0.01	−0.42	0.07	−0.45	0.09
Source 1	0.04	0.01	0.49	0.01	−0.13	0.07	−0.18	0.07
Source 2	0.04	0.01	0.47	0.01	−0.14	0.08	−0.21	0.08
Temperature	0.10	0.01	0.57	0.02	−0.07	0.08	−0.25	0.08
Source 1 + Source 2	0.21	0.01	0.81	0.02	−0.02	0.08	−0.22	0.06
Source 1 + Temp	0.30	0.01	1.00	0.03	0.08	0.08	−0.19	0.09
Source 2 + Temp	0.27	0.01	0.88	0.02	0.06	0.08	−0.24	0.09
All	0.39	0.01	0.95	0.04	0.13	0.09	−0.39	0.08

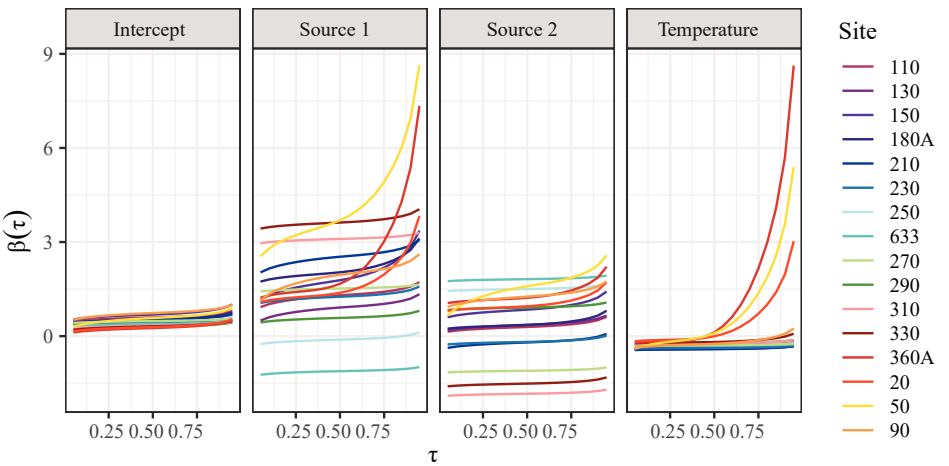

Figure 5. Estimated predictor effect by quantile and location.

the 95th quantile is large and positive for the sites on the northern edge of the refinery, as well as some sites along the southern edge of the refinery. The northern sites were also the locations where the highest concentrations were observed. The effect of source 2 on the 95th quantile was smaller overall, and varied by site, with positive effects observed on the background site and sites on the northern edge of the refinery (Fig. 6). Temperature had a strong positive effect on concentrations on the northern edge of the refinery, indicating the possibility of another emission source during the summer near the northern edge of the refinery that was not accounted for.

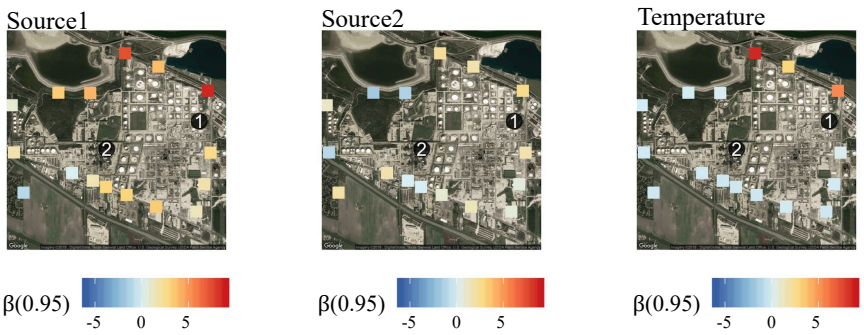

Figure 6. Spatial variation in the effect of the predictors on the 95th quantile of fence-line benzene measurements.

6. Conclusion

We have derived the properties and demonstrated the utility of a spatial quantile regression method that allows for spatially varying coefficients and flexible tail distributions. By modeling the entire quantile function, we exploit the flexibility of the nonparametric basis functions in the center of the distribution, and the constraints of the parametric tails where the data are sparse. We have shown the conditions under which the model guarantees a smooth density function, with the desired degrees of differentiability, and enables the estimation of a nonstationary covariance that is dependent on the predictors. Using both simulations and an application to fence line benzene concentrations, we have demonstrated the utility of ensuring a smooth density function with parametric tails, as well as the flexibility and accuracy of the method compared with previous methods.

Although the model does not currently account for temporal correlation in the response variable, a nonlinear function of time could easily be incorporated as a predictor using the current framework. Additionally, temporal correlation could be accounted for by adjusting the priors of the coefficients or incorporating a copula. A multivariate extension for modeling multiple pollutants simultaneously could also be developed using multivariate spatial priors. Recently, Yang and Tokdar (2017) provided a characterization for noncrossing quantile regressions over convex predictor spaces. It would be interesting to explore extensions of their method to the spatial or spatial-temporal case.

Supplementary Material

The online Supplementary Material provides proofs of Proposition 1 and Theorem 1, as well as various computations.

Disclaimer

The views expressed in this publication are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Acknowledgments

This work was supported by the National Science Foundation grant No. 1613219, and the National Institutes of Health grant No. R01ES027892. The authors would like to thank Oak Ridge Institute for Science and Education for the fellowship funding that also supported this work.

References

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **70**, 825–848.

Bondell, H. D., Reich, B. J. and Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika* **97**, 825–838.

Cai, Y. and Jiang, T. (2015). Estimation of non-crossing quantile regression curves. *Australian & New Zealand Journal of Statistics* **57**, 139–162.

Chernozhukov, V., Fernandez-Val, I. and Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* **96**, 559–575.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)* **52**, 393–442.

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

U.S. Environmental Protection Agency. (2014). *Petroleum Refinery Sector Risk and Technology Review and New Source Performance Standards*. Rule document, 75177–75354.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Hansen, E. and Patrick, M. (1976). A family of root finding methods. *Numerische Mathematik* **27**, 257–269.

Koenker, R. (2005). *Quantile Regression (Econometric Society Monographs; no. 38)*. Cambridge University Press, Cambridge.

Koenker, R. and Bassett Jr., G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

Lum, K. and Gelfand, A. E. (2012). Spatial quantile multiple regression using the asymmetric laplace process. *Bayesian Analysis* **7**, 235–258.

Neocleous, T. and Portnoy, S. (2008). On monotonicity of regression quantile functions. *Statistics & Probability Letters* **78**, 1226–1229.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**, 425–441.

Reich, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **61**, 535–553.

Reich, B. J., Fuentes, M. and Dunson, D. B. (2012). Bayesian spatial quantile regression. *Journal of the American Statistical Association* **106**, 6-20.

Risser, M. and Calder, C. (2015). Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics* **26**, 284–297.

Rodrigues, T. and Fan, Y. (2017). Regression adjustment for noncrossing bayesian quantile regression. *Journal of Computational and Graphical Statistics* **26**, 275–284.

Smith, L. B., Reich, B. J., Herring, A. H., Langlois, P. H., and Fuentes, M. (2015). Multilevel quantile function modeling with application to birth outcomes. *Biometrics* **71**, 508–519.

Thoma, E. D., Miller, M. C, Chung, K. C, Parsons, N. L. and Shine, B. C. (2011). Facility fence-line monitoring using passive samplers. *Journal of the Air & Waste Management Association* **61**, 834–842.

Tokdar, S. T. and Kadane, J. B. (2012). Simultaneous linear quantile regression: A semiparametric bayesian approach. *Bayesian Analysis* **7**, 51–72.

Yang, Y. and He, X. (2015). Quantile regression for spatially correlated data: An empirical likelihood approach. *Statistica Sinica* **25**, 261–274.

Yang, Y. and Tokdar S. (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association* **112**, 1107–1120.

Yu, K. and Moyeed, R.A. (2001). Bayesian quantile regression. *Statistics & Probability Letters* **54**, 437–447.

Zhou, J., Fuentes, M. and Davis, J. (2011). Calibration of numerical model output using nonparametric spatial density functions. *Journal of Agricultural, Biological, and Environmental Statistics* **16**, 531–553.

Zhou, J., Chang, H. and Fuentes, M. (2012). Estimating the health impact of climate change with calibrated climate model output. *Journal of Agricultural, Biological, and Environmental Statistics* **17**, 377–394.

Halley Brantley

9278 Hyland Creek Rd., Bloomington, MN 55437, USA.

E-mail: halleybrantley@gmail.com

Montserrat Fuentes

101 Jessup Hall, Iowa City, Iowa 52242-1316, USA.

E-mail: montserrat-fuentes@uiowa.edu

Joseph Guinness

1178 Comstock Hall, 129 Garden Ave., Ithaca, NY 14853, USA.

E-mail: guinness@cornell.edu

Eben Thoma

109 TW Alexander Drive (M/S E343-02), Research Triangle Park, NC 27711, USA.

E-mail: Thoma.Eben@epa.gov

(Received January 2019; accepted September 2019)