Received: 27 September 2016 | Revised: 5 May 2017 | Accepted: 8 May 2017

DOI: 10.1002/env.2448

RESEARCH ARTICLE

WILEY

Spatially modeling the effects of meteorological drivers of $PM_{2.5}$ in the Eastern United States via a local linear penalized quantile regression estimator

Brook T. Russell[1] | Dewei Wang[2] | Christopher S. McMahan[1]

[1]Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A.

[2]Department of Statistics, University of South Carolina, Columbia, SC 29208, U.S.A.

Correspondence
Christopher S. McMahan, Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A.
Email: mcmaha2@clemson.edu

Funding information
National Institutes of Health, Grant/Award Number: R01 AI121351

Fine particulate matter ($PM_{2.5}$) poses a significant risk to human health, with long-term exposure being linked to conditions such as asthma, chronic bronchitis, lung cancer, and atherosclerosis. In order to improve the current pollution control strategies and to better shape public policy, the development of a more comprehensive understanding of this air pollutant is necessary. To this end, this work attempts to quantify the relationship between certain meteorological drivers and the levels of $PM_{2.5}$. It is expected that the set of important meteorological drivers will vary both spatially and within the conditional distribution of $PM_{2.5}$ levels. To account for these characteristics, a new local linear penalized quantile regression methodology is developed. The proposed estimator uniquely selects the set of important drivers at every spatial location and for each quantile of the conditional distribution of $PM_{2.5}$ levels. The performance of the proposed methodology is illustrated through simulation, and it is then used to determine the association between several meteorological drivers and $PM_{2.5}$ over the Eastern United States. This analysis suggests that the primary drivers throughout much of the Eastern United States tend to differ based on season and geographic location, with similarities existing between "typical" and "high" $PM_{2.5}$ levels.

KEYWORDS

adaptive LASSO, fine particulate matter, local linear quantile regression, meteorological drivers of $PM_{2.5}$

1 | INTRODUCTION

Particulate matter is an air pollutant that is comprised of microscopic particles of compounds, such as metals, soot, sulfates, nitrates, smoke, and dust. The size of the particles is inextricably linked to their potential to pose risk to human health. Of the various types, fine particulate matter ($PM_{2.5}$), characterized by particles less than 2.5 μm, poses the highest degree of risk because it has the propensity to settle deep within the lungs and can pass into the bloodstream. Pope, Ezzati, and Dockery (2009) and Krewski et al. (2009) link long-term exposure to $PM_{2.5}$ to a decrease in human life expectancy in the United States. In particular, it has been conjectured that long-term exposure to $PM_{2.5}$ can lead to medical conditions, such as asthma, chronic bronchitis, lung cancer, and atherosclerosis; for further discussion, see Khafaie, Yajnik, Salvi, and Ojha (2016) and the references therein. Consequently, developing a more comprehensive understanding of the meteorological drivers of $PM_{2.5}$ levels is of great importance with respect to shaping pollution control strategies and public health policies. Further, it has been posited that acute air pollution events are particularly harmful (Porter, Heald, Cooley, & Russell, 2015), and thus, gaining an understanding of the drivers of these types of events is also important.

Routine variability in air pollution levels can often be linked to meteorological conditions. Jacob and Winner (2009) surmise that air quality in general is "strongly dependent" on meteorological variables (e.g., precipitation, temperature, and wind speed) and that understanding this relationship is tantamount to understanding air pollution. Moreover, it is reasonable to believe that the set of meteorological drivers that is associated with air pollution (or more specifically $PM_{2.5}$) varies spatially; that is, for example, in certain regions of the United States, precipitation might be useful with respect to explaining the trends in $PM_{2.5}$, but not in other regions. Several authors have conducted spatial and spatiotemporal analyses of $PM_{2.5}$ over different geographic regions. For example, Smith, Kolenikov, and Cox (2003) summarize the results of a spatiotemporal analysis of $PM_{2.5}$ levels in three Southeastern U.S. states, and Lopez et al. (2015) model air pollution extremes, including $PM_{2.5}$ levels, in the Southwest United States. Both of these studies model air pollution, not the associated meteorological variables. However, others have studied the relationship between meteorological conditions and air pollution; for example, see Jacob and Winner (2009), Tai, Mickley, and Jacob (2010), and Porter et al. (2015). It is important to note that these authors considered either a global relationship between meteorological variables and $PM_{2.5}$ levels or the association at specific locations. For example, Porter et al. (2015) and Tai et al. (2010) use quantile regression and standard linear regression techniques, respectively, to assess the relationship between meteorological variables and $PM_{2.5}$ levels at numerous geographic locations throughout the Continental United States. Porter et al. (2015) perform the analysis at air pollution monitoring station locations, whereas Tai et al. (2010) use points on a coarse grid; however, both analyses suggest that the relationship between $PM_{2.5}$ and its meteorological drivers varies spatially.

Motivated by these studies, the regression methodology developed herein is targeted at spatially modeling the conditional quantiles of $PM_{2.5}$ levels as a function of various meteorological variables over the Eastern United States. In particular, the functional relationship between these variables is allowed to change spatially, and the methodology can be used to uniquely select the set of important meteorological drivers of $PM_{2.5}$ levels at any spatial location, even if data are unavailable at the location of interest. Moreover, this estimation and selection process can be completed uniquely within any of the conditional quantiles of $PM_{2.5}$ levels. All of these goals are accomplished through the development of a new quantile regression methodology. Quantile regression, first proposed by Koenker and Bassett (1978), has become an increasingly popular alternative to standard mean regression techniques. Owing to its popularity and utilitarian nature, many extensions and generalizations of quantile regression have been proposed; for example, Kai and Li (2010) propose a nonparametric robust mean estimator based on composite quantile regression, and Zhu, Huang, and Li (2012) develop a semiparametric quantile regression estimator within the high-dimensional covariate setting. For modeling spatial data, Hallin, Lu, and Yu (2009) propose a local linear estimator, and Sun, Wang, and Fuentes (2016) develop a fused adaptive least absolute shrinkage and selection operator (LASSO) approach within the quantile regression framework. Although both of these methods are substantive contributions to the literature, their scope differs from that of the current work. In particular, these methods neither provide for spatially varying effect estimates nor possess the ability to identify a spatially unique set of meteorological drivers that are related to $PM_{2.5}$ levels.

The proposed regression methodology is developed by extending and generalizing the local linear quantile regression estimator studied in Fan, Hu, and Truong (1994) and Yu and Jones (1998). To allow for spatially varying effect estimates, the proposed approach views each of the regression coefficients associated with the meteorological variables as an unknown surface that varies spatially, thus allowing the relationship between these variables and $PM_{2.5}$ to change across the spatial domain. In other words, the proposed approach could be viewed as a varying coefficient quantile regression model. Other authors have considered such models (Cai & Xu, 2008; Honda, 2004; Kim, 2007; Wang, Zhu, & Zhou, 1998), but these works allow the coefficient to vary in a single dimension; that is, typically the coefficient is allowed to vary in time or with the levels of another covariate. To allow the coefficient to vary spatially, the approach taken here is very akin to the proposal of Chen, Deng, Yang, and Matthews (2012). The primary advantage of the proposed approach over the technique outlined in Chen et al. (2012) is that it employs regularization to obtain a sparse estimator; that is, during the estimation process, regression coefficients associated with insignificant variables are set to be identically equal to zero, thus completing the model fitting and variable selection simultaneously. This is accomplished by adopting and adapting the adaptive LASSO of Zou (2006). Taking advantage of the formulation of the proposed model, a computationally efficient technique for model fitting is developed. Moreover, the asymptotic properties of the proposed estimator are established, and it is shown that the proposed estimator possesses what are commonly referred to as the "oracle properties"; for further discussion, see Fan and Li (2001) and Zou (2006).

The remainder of this article is organized as follows. In Section 2, the proposed methodology is developed, and model fitting strategies are discussed. Section 3 provides the asymptotic properties of the proposed estimator. The performance of the proposed approach is examined through numerical studies in Section 4, and the results of the analysis of the motivating data application are presented in Section 5. Section 6 concludes with a summary discussion. All technical proofs and conditions are provided in the Supporting Information.

2 | METHODOLOGY

The proposed methodology seeks to assess the explanatory capacity of the available covariates (e.g., precipitation, wind speed, and turbulence kinetic energy) for key quantiles of the conditional distribution of $PM_{2.5}$. Based on this goal, adaptations to the usual quantile regression methodology are considered. Specifically, these generalizations allow the association between covariates and the response to vary spatially. In contrast, standard quantile regression techniques (e.g., see Koenker & Bassett, 1978) obtain an estimate of the conditional τth quantile, $\tau \in (0, 1)$, of the response variable (Y) given a vector of covariate values ($\mathbf{x}$), denoted by $Q_Y(\tau, \mathbf{x}, \boldsymbol{\beta}_\tau)$. Herein, it is assumed that $Q_Y(\tau, \mathbf{x}, \boldsymbol{\beta}_\tau) = \mathbf{x}'\boldsymbol{\beta}_\tau$, where $\boldsymbol{\beta}_\tau = (\beta_{0_\tau}, \ldots, \beta_{p_\tau})'$ denotes a $(p+1)$-dimensional vector of regression coefficients, with β_{0_τ} being the usual intercept parameter. Note that a primary strength of quantile regression is that it is capable of estimating different types of associations at different quantiles of interest; that is, it is not necessary for $\boldsymbol{\beta}_\tau = \boldsymbol{\beta}_{\tau'}$, for $\tau \neq \tau'$. Under this parametric framework, estimating $\boldsymbol{\beta}_\tau$ is tantamount to estimating the entire conditional τth quantile function, for all values of the covariate, and this estimator can be obtained as

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{s=1}^{S}\sum_{t=1}^{T_s} \rho_\tau\left(Y_{st} - \mathbf{x}_{st}'\boldsymbol{\beta}_\tau\right),$$

where $\rho_\tau(z) = z\{\tau - 1(z < 0)\}$ is the usual "check loss" function, Y_{st} denotes the observed response (i.e., $PM_{2.5}$ level) at the sth location at the tth time point, and $\mathbf{x}_{st}$ is the corresponding vector of covariates, for $t = 1, \ldots, T_s$ and $s = 1, \ldots, S$. It is worthwhile to point out that $\widehat{\boldsymbol{\beta}}_\tau$ can be viewed as a "global" estimator, because it does not spatially vary; that is, this estimator is the same for all locations within the spatial domain.

To acknowledge that the relationship between the response and covariates may differ geographically, one could fit a quantile regression model at each spatial location individually; that is, one could obtain the location-specific estimator of $\boldsymbol{\beta}_\tau$ as

$$\widehat{\boldsymbol{\beta}}_\tau^S = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{t=1}^{T_s} \rho_\tau\left(Y_{st} - \mathbf{x}_{st}'\boldsymbol{\beta}_\tau\right),$$

for $s = 1, \ldots, S$. Consequently, an estimator of the location-specific conditional τth quantile function is obtained as $Q_Y(\tau, \mathbf{x}, \widehat{\boldsymbol{\beta}}_\tau^S) = \mathbf{x}'\widehat{\boldsymbol{\beta}}_\tau^S$, for $s = 1, \ldots, S$. In general, this approach would allow one to detect different relationships between the response variable and covariates at different geographic regions, but it does not allow for the interpolation of this relationship to regions where data are not available. Note that this approach is similar to the methodology employed by Porter et al. (2015).

To allow for such an interpolation, the proposed methodology views each of the regression coefficients as an unknown surface; that is, for a geographic location $\boldsymbol{l}$, it is assumed that $\beta_{j_\tau} := \beta_{j_\tau}(\boldsymbol{l})$, for $j = 0, \ldots, p$, where $\boldsymbol{l} = (l_0, l_1)'$ denotes a 2-dimensional vector of spatial coordinates, for example, latitude and longitude. Thus, define $\boldsymbol{\beta}_\tau(\boldsymbol{l}) = \{\beta_{0_\tau}(\boldsymbol{l}), \ldots, \beta_{p_\tau}(\boldsymbol{l})\}'$, and let $\boldsymbol{l}_s = (l_{s_0}, l_{s_1})'$ denote the spatial coordinates of the sth location. The primary goal is to estimate $\boldsymbol{\beta}_\tau(\boldsymbol{l})$ at any location $\boldsymbol{l}^*$ of interest, whether or not $\boldsymbol{l}^*$ corresponds to a location in the observed data. To accomplish this task, it is assumed that the available data $\{(Y_{st}, \mathbf{x}_{st}, \boldsymbol{l}_s) : t = 1, \ldots, T_s; s = 1, \ldots, S\}$ are independent realization arising from a joint model that possesses the following property:

$$\mathbf{x}'\boldsymbol{\beta}_\tau(\boldsymbol{l}) = \underset{a\in\mathbb{R}}{\operatorname{argmin}} E\left\{\rho_\tau(Y - a)|\mathbf{x}, \boldsymbol{l}\right\}.$$

This is equivalent to assuming that

$$Y = \mathbf{x}'\boldsymbol{\beta}_\tau(\boldsymbol{l}) + \epsilon_\tau,$$

where $P(\epsilon_\tau \leq 0|\mathbf{x}, \boldsymbol{l}) = \tau$.

In order to develop an estimator of $\beta_{j_\tau}(\boldsymbol{l}^*)$, the proposed approach makes use of the first-order Taylor series expansion of $\beta_{j_\tau}(\boldsymbol{l})$, about $\boldsymbol{l}^*$, given by

$$\beta_{j_\tau}(\boldsymbol{l}) \approx \theta_{j\tau}^{0*} + \theta_{j\tau}^{1*}\left(l_0 - l_0^*\right) + \theta_{j\tau}^{2*}\left(l_1 - l_1^*\right), \tag{1}$$

where $\theta_{j\tau}^{0*} = \beta_{j_\tau}(\boldsymbol{l}^*)$, $\theta_{j\tau}^{1*} = \partial\beta_{j_\tau}(\boldsymbol{l})/\partial l_0|_{\boldsymbol{l}=\boldsymbol{l}^*}$, and $\theta_{j\tau}^{2*} = \partial\beta_{j_\tau}(\boldsymbol{l})/\partial l_1|_{\boldsymbol{l}=\boldsymbol{l}^*}$. It is assumed that all necessary derivatives exist; that is, it is assumed that $\beta_{j_\tau}(\cdot)$, for $j = 0, \ldots, p$, is continuously differentiable. For notational convenience, define $\boldsymbol{\theta}_\tau^{0*} = (\theta_{0\tau}^{0*}, \ldots, \theta_{p\tau}^{0*})'$, $\boldsymbol{\theta}_\tau^{1*} = (\theta_{0\tau}^{1*}, \ldots, \theta_{p\tau}^{1*})'$, $\boldsymbol{\theta}_\tau^{2*} = (\theta_{0\tau}^{2*}, \ldots, \theta_{p\tau}^{2*})'$, $\boldsymbol{\theta}_\tau^* = (\boldsymbol{\theta}_\tau^{0*\prime}, \boldsymbol{\theta}_\tau^{1*\prime}, \boldsymbol{\theta}_\tau^{2*\prime},)'$, and $\mathbf{x}_{st}^* = \{\mathbf{x}_{st}', (l_{s_0} - l_0^*)\mathbf{x}_{st}', (l_{s_1} - l_1^*)\mathbf{x}_{st}'\}'$. Inspired by local polynomial regression techniques (e.g., see Fan & Gijbels, 1996; Fan et al., 1994), an estimator of $\boldsymbol{\theta}_\tau^*$ is

$$\hat{\boldsymbol{\theta}}_\tau^*(h) = \underset{\boldsymbol{\theta}_\tau^*\in\mathbb{R}^{3p+3}}{\operatorname{argmin}} \sum_{s=1}^{S}\sum_{t=1}^{T_s} \rho_\tau\left(Y_{st} - \mathbf{x}_{st}^{*'}\boldsymbol{\theta}_\tau^*\right) K\left(\frac{\left\|\boldsymbol{l}_s - \boldsymbol{l}^*\right\|_2}{h}\right), \tag{2}$$

where $\widehat{\boldsymbol{\theta}}^{*}_{\tau}(h) = \{\widehat{\boldsymbol{\theta}}^{0*\prime}_{\tau}(h), \widehat{\boldsymbol{\theta}}^{1*\prime}_{\tau}(h), \widehat{\boldsymbol{\theta}}^{2*\prime}_{\tau}(h)\}'$, $K(\cdot)$ is a symmetric kernel function (e.g., biweight, Epanechnikov, and Gaussian), h is the bandwidth parameter, and $\|\cdot\|_2$ is the usual Euclidean norm. Note that, from Equation 2, an estimator of $\boldsymbol{\beta}_{\tau}(\boldsymbol{l}^{*})$ is given by $\widehat{\boldsymbol{\beta}}_{\tau}(\boldsymbol{l}^{*}) = \widehat{\boldsymbol{\theta}}^{0*\prime}_{\tau}(h)$. In general, the approximation suggested in Equation 1 is good for $\boldsymbol{l}$ within a neighborhood of $\boldsymbol{l}^{*}$. This fact is acknowledged through the use of $K(\cdot)$; that is, the kernel function downweights the influence of observations that are spatially "far" from $\boldsymbol{l}^{*}$. Conceptually, the smoothing parameter h reflects what is meant by "far," that is to say larger values of h equate to larger neighborhoods of influence, and vice versa. As one might expect, different values of h inherently lead to different estimates of $\boldsymbol{\beta}_{\tau}(\boldsymbol{l}^{*})$. Note that the methodology outlined above is very akin to the technique presented in Chen et al. (2012), and both provide estimates of $\boldsymbol{\beta}_{\tau}(\boldsymbol{l})$ that spatially vary. Additionally, the first-order approximation provided in Equation 1 could easily be extended to a higher order approximation, but this generalization is not explored for two primary reasons: first, through numerical studies and the motivating data application the first-order approximation appears to be sufficient in realistic scenarios and second, so that the proposed methodology could be succinctly presented.

A primary goal of an analysis of this form is to identify the regions where each of the covariates are significantly related to the response; that is, to perform model selection spatially. That is to say, it is expected that some covariates will be useful in explaining the response variable in some geographical regions while not being useful in others. To allow and account for these effects, the methodology described above is further extended and recast in the penalized regression context. Motivated by the works of Tibshirani (1996), Zou (2006), and Wu and Liu (2009), the following sparse estimator is considered:

$$\widetilde{\boldsymbol{\theta}}^{*}_{\tau}(h,\lambda) = \underset{\boldsymbol{\theta}^{*}_{\tau}\in\mathbb{R}^{3p+3}}{\operatorname{argmin}} \sum_{s=1}^{S}\sum_{t=1}^{T_s} \rho_{\tau}\left(Y_{st} - \mathbf{x}^{*\prime}_{st}\boldsymbol{\theta}^{*}_{\tau}\right) K\left(\frac{\|\boldsymbol{l}_s - \boldsymbol{l}^{*}\|_2}{h}\right) + \lambda \sum_{k=0}^{2}\sum_{j=0}^{p} \left|\theta^{k*}_{j\tau}\right| \Big/ \left|\hat{\theta}^{k*}_{j\tau}(h)\right|, \tag{3}$$

where λ is a penalty parameter and $\hat{\theta}^{k*}_{j\tau}(h)$ is an initial estimate of $\theta^{k*}_{j\tau}$ obtained from Equation 2. This approach yields $\widetilde{\boldsymbol{\beta}}_{\tau}(\boldsymbol{l}^{*}) = \widetilde{\boldsymbol{\theta}}^{0*}_{\tau}(h,\lambda)$, which is a sparse estimator (i.e., some coefficients are set to be identically equal to zero) of $\boldsymbol{\beta}_{\tau}(\boldsymbol{l}^{*})$, where $\widetilde{\boldsymbol{\theta}}^{*}_{\tau}(h,\lambda) = \{\widetilde{\boldsymbol{\theta}}^{0*\prime}_{\tau}(h,\lambda), \widetilde{\boldsymbol{\theta}}^{1*\prime}_{\tau}(h,\lambda), \widetilde{\boldsymbol{\theta}}^{2*\prime}_{\tau}(h,\lambda)\}'$. The sparsity of the estimator is due to the utilization of the adaptive LASSO penalty by the proposed modeling framework. Further, as with the estimator obtained from Equation 2, the proposed sparse estimator is inherently dependent on the bandwidth parameter h and the penalty parameter λ. In fact, the sparsity of the estimator is directly controlled by λ, with large values of λ promoting a more sparse solution and vice versa. Given their influence, a method of determining the tuning parameters h and λ is presented and evaluated in Section 2.2.

Note that the sparse estimator proposed in Equation 3 can be used to select covariates related to the τth quantile of the response variable at a particular geographic location $\boldsymbol{l}^{*}$. Moreover, the effect size, direction, and significance associated with each of the covariates are allowed to change from location to location. Consequently, given the scope of the proposed work, it is desirable to identify regions of significance for each of the covariates. To this end, let $\mathcal{S}$ denote the entire spatial region of interest, and for the jth covariate define the region $\mathcal{I}_{j\tau} = \{\boldsymbol{l} \in \mathcal{S} : \beta_{j\tau 0}(\boldsymbol{l}) \neq 0\}$, where $\beta_{j\tau 0}(\boldsymbol{l})$ is the true value of $\beta_{j\tau}(\boldsymbol{l})$, for all $\boldsymbol{l}$; that is, $\mathcal{I}_{j\tau}$ is the region of $\mathcal{S}$ on which the jth covariate is truly related to the τth quantile of the response variable. Note that the region described by $\mathcal{I}_{j\tau}$, for all j, represents an uncountable set and is therefore impossible to identify exactly. In order to provide an approximation to these regions, a fixed grid consisting of M points is selected within $\mathcal{S}$; denote these points as $\boldsymbol{l}^{*}_{m}$, for $m = 1, \ldots, M$. Let $I_{j\tau} = \{\boldsymbol{l}^{*}_{m} : \boldsymbol{l}^{*}_{m} \in \mathcal{I}_{j\tau}\}$, and note that if the grid is selected to be large enough, then $I_{j\tau}$ is a natural fine approximation of $\mathcal{I}_{j\tau}$. An estimator of $I_{j\tau}$ can be constructed via $\tilde{I}_{j\tau} = \{\boldsymbol{l}^{*}_{m} : \tilde{\beta}_{j\tau}(\boldsymbol{l}^{*}_{m}) \neq 0\}$, where $\tilde{\beta}_{j\tau}(\boldsymbol{l}^{*}_{m})$ is the estimator resulting from the proposed approach.

2.1 | Model fitting strategy

In this section, data augmentation techniques that can be used to efficiently obtain the estimators described in Equations 2 and 3 are presented. First, define the transformed response and covariate vector as $Z_{st} = w_s Y_{st}$ and $\mathbf{u}_{st} = w_s \mathbf{x}^{*}_{st}$, where $w_s = K(h^{-1}\|\boldsymbol{l}_s - \boldsymbol{l}^{*}\|_2)$. Based on this transformed data, the estimator described in Equation 2 can be equivalently expressed as

$$\widehat{\boldsymbol{\theta}}^{*}_{\tau}(h) = \underset{\boldsymbol{\theta}^{*}_{\tau}\in\mathbb{R}^{3p+3}}{\operatorname{argmin}} \sum_{s=1}^{S}\sum_{t=1}^{T_s} \rho_{\tau}\left(Z_{st} - \mathbf{u}'_{st}\boldsymbol{\theta}^{*}_{\tau}\right). \tag{4}$$

Note that the estimator resulting from the minimization problem described in Equation 4 is identically equal to the standard quantile regression estimator (about the τth quantile) obtained from treating Z_{st} as the response variable and $\mathbf{u}_{st}$ as the covariate vector. Consequently, this optimization step can be carried out using existing software packages designed to fit quantile regression models, for example, `quantreg` in R (for further details, see Koenker, 2015).

In order to fit the penalized model, it is first noted that $a|\phi| = \rho_\tau(a\phi) + \rho_\tau(-a\phi)$, for all $a > 0$ and $\phi \in \mathbb{R}$. Thus, the terms in the penalty of Equation 3 can be expressed as

$$\begin{aligned}\lambda \left|\theta_{j\tau}^{k*}\right| / \left|\hat{\theta}_{j\tau}^{k*}(h)\right| &= \rho_\tau\left(\lambda\theta_{j\tau}^{k*} / \left|\hat{\theta}_{j\tau}^{k*}(h)\right|\right) + \rho_\tau\left(-\lambda\theta_{j\tau}^{k*} / \left|\hat{\theta}_{j\tau}^{k*}(h)\right|\right) \\ &= \rho_\tau\left(\dot{Z}_{jk1} - \dot{\mathbf{u}}_{jk1}'\boldsymbol{\theta}_\tau^*\right) + \rho_\tau\left(\dot{Z}_{jk2} - \dot{\mathbf{u}}_{jk2}'\boldsymbol{\theta}_\tau^*\right),\end{aligned}$$

for $k = 0, 1, 2$ and $j = 0, \ldots, p$, where $\dot{Z}_{jk1} = \dot{Z}_{jk2} = 0$, $\dot{\mathbf{u}}_{jk1} = -\dot{\mathbf{u}}_{jk2}$, and $\dot{\mathbf{u}}_{jk1}$ is a vector containing all zeros with the exception of the element corresponding to $\theta_{j\tau}^{k*}$ that takes value $-\lambda/|\hat{\theta}_{j\tau}^{k*}(h)|$. Consequently, the penalized estimator depicted in Equation 3 can be fit after introducing appropriately structured synthetic data. In particular, define the synthetic response variable $\dot{Z}_r = 0$, for $r = 1, \ldots, R = 6p + 6$, and the corresponding covariate vector $\dot{\mathbf{u}}_r$, where $\dot{\mathbf{u}}_r$ is the rth row of the matrix $\dot{\mathbf{U}} = [\text{diag}\{\lambda/|\widehat{\boldsymbol{\theta}}_\tau^*(h)|\}, \text{diag}\{-\lambda/|\widehat{\boldsymbol{\theta}}_\tau^*(h)|\}]'$. Constructing synthetic data in this fashion allows one to impose the penalty in Equation 3 as a part of an unpenalized problem. That is, based on the transformed and synthetic data, the estimator described in Equation 3 can be equivalently obtained via

$$\widetilde{\boldsymbol{\theta}}_\tau^*(h, \lambda) = \underset{\boldsymbol{\theta}_\tau^* \in \mathbb{R}^{3p+3}}{\text{argmin}} \sum_{s=1}^{S}\sum_{t=1}^{T_s} \rho_\tau\left(Z_{st} - \mathbf{u}_{st}'\boldsymbol{\theta}_\tau^*\right) + \sum_{r=1}^{R} \rho_\tau\left(\dot{Z}_r - \dot{\mathbf{u}}_r'\boldsymbol{\theta}_\tau^*\right). \tag{5}$$

It should be emphasized that, after adding the synthetic data to the observed data, the minimization problem described in Equation 5 can easily be solved using standard numerical routines used to fit quantile regression models.

2.2 | Tuning parameter selection

In order to determine appropriate values for the tuning parameters h and λ, an iterative leave-one-out cross-validation scheme is suggested. Similar proposals have been made in Li (1984), Rice (1984), and Zou and Li (2008). The difference between the proposed scheme and standard leave-one-out cross-validation procedures is that rather than leaving a single observation out, as is specified by the latter approach, the proposed scheme omits all observations associated with a particular location. Thus, for a given value of h, define $\widehat{\boldsymbol{\beta}}_\tau^s(\boldsymbol{l}_s)$ to be the estimator of $\boldsymbol{\beta}_\tau(\boldsymbol{l}_s)$ resulting from Equation 2 after removing the data associated with sth location. The proposed leave-one-out cross-validation score used to select h is given by

$$CV_1(h) = \sum_{s=1}^{S}\sum_{t=1}^{T_s} \rho_\tau\left\{Y_{st} - \mathbf{x}_{st}'\widehat{\boldsymbol{\beta}}_\tau^s(\boldsymbol{l}_s)\right\}. \tag{6}$$

It is then suggested that the smoothing parameter h be chosen to minimize Equation 6, and its value is denoted by $\hat{h}$. Once this step is accomplished, let $\widetilde{\boldsymbol{\beta}}_\tau^s(\boldsymbol{l}_s)$ be the estimator of $\boldsymbol{\beta}_\tau(\boldsymbol{l}_s)$ resulting from Equation 3 after removing the data associated with sth location, for a given value of λ with the smoothing parameter being set to be $\hat{h}$. The proposed leave-one-out cross-validation score used to select λ is given by

$$CV(\hat{h}, \lambda) = S^{-1}\sum_{s=1}^{S}\sum_{t=1}^{T_s} \rho_\tau\left\{Y_{st} - \mathbf{x}_{st}'\widetilde{\boldsymbol{\beta}}_\tau^s(\boldsymbol{l}_s)\right\}. \tag{7}$$

Using these cross-validation scores, it is suggested that one implement the one-standard error rule of Hastie, Tibshirani, and Friedman (2009) to select the penalty parameter. That is, the penalty parameter is selected to be the largest value of λ satisfying

$$CV(\hat{h}, \lambda) \leq CV(\hat{h}, \lambda^*) + \frac{SD\left\{CV(\hat{h}, \lambda^*)\right\}}{\sqrt{S}},$$

where λ^* is the penalty parameter value that minimizes $CV(\hat{h}, \lambda)$ and $SD\{CV(\hat{h}, \lambda^*)\}$ is the sample standard deviation of the cross-validation scores computed at the S locations using λ^*. Note that, utilizing the computationally efficient model fitting strategies discussed in Section 2.1, one can easily minimize Equations 6 and 7 using standard grid search techniques over a grid of potential values for h and λ, respectively. Note that, when conducting this process, it is generally advisable, for both the selection of h and λ, to plot the cross-validation scores versus the tuning parameter of interest to ensure that a reasonable range of values have been considered.

It is worthwhile to note that the performance of the proposed methodology is inherently tied to the selection of both h and λ. That is, misspecifying either of these tuning parameters can be deleterious to the performance of the proposed regression methodology. Through simulation, it has been ascertained that the aforementioned process of selecting h and λ is reliable; for further details, see Section 4. Further, given the computationally efficient model fitting strategy outlined in Section 2.1, it is conjectured that the proposed approach could be used to handle extremely large data sets, but the computational time required to complete the entire process (whether for small or larger data sets) is highly dependent on several features. In particular, the computational time highly depends on the number of tuning parameter configurations under consideration, the number of spatial units available in the data, and the number of spatial units at which one desires to obtain an estimate.

3 | THEORETICAL RESULTS

In this section, three theoretical properties of the estimator in Equation 3 are discussed: consistency in variable selection, asymptotic consistency, and asymptotic normality. The combination of these three characteristics provides that the proposed estimator possesses what are commonly referred to as the "oracle properties." That is to say, asymptotically, the proposed estimator will correctly identify the collection of covariates that are truly related to the response variable, and the estimator incurs no asymptotic bias that is attributable to the penalization process.

To establish these results, the asymptotic properties of the proposed estimator are first studied at an arbitrary geographic location $\boldsymbol{l}^*$. At this location, let $\boldsymbol{\theta}^*_{\tau 0}$ denote the true value of $\boldsymbol{\theta}^*_{\tau}$, where $\boldsymbol{\theta}^*_{\tau 0} = (\boldsymbol{\theta}^{0*\prime}_{\tau 0}, \boldsymbol{\theta}^{1*\prime}_{\tau 0}, \boldsymbol{\theta}^{2*\prime}_{\tau 0})'$ and $\boldsymbol{\theta}^{k*}_{\tau 0} = (\theta^{k*}_{0\tau 0}, \ldots, \theta^{k*}_{p\tau 0})$, for $k = 0, 1, 2$. Based on $\boldsymbol{\theta}^*_{\tau 0}$, define the collection of indices given by $\mathcal{A}^* = \{j : \theta^{0*}_{j\tau 0} \neq 0\}$. Note that, for every $j \in \mathcal{A}^*$, one has that $\theta^{0*}_{j\tau 0} \neq 0$ (i.e., the true value of $\beta_{j_\tau}(\boldsymbol{l}^*)$ is nonzero), which is equivalent to saying that the τth quantile of the response variable is truly related to the jth covariate at the geographic location $\boldsymbol{l}^*$. Moreover, for every $j \notin \mathcal{A}^*$, one has that $\theta^{0*}_{j\tau 0} = 0$, thereby indicating that the jth covariate is not related to the τth quantile of the response variable at the geographic location $\boldsymbol{l}^*$. Similarly, define the set of indices $\widetilde{\mathcal{A}}^*_\lambda = \{j : \widetilde{\boldsymbol{\theta}}^{0*}_{j\tau}(h, \lambda) \neq 0\}$, with respect to the proposed estimator. This set of indices identifies the collection of covariates selected by the proposed estimator as being related to the τth quantile of the response variable at location $\boldsymbol{l}^*$. Thus, the property referred to as consistency in variable selection can be succinctly stated as $\lim_{N\to\infty} P(\widetilde{\mathcal{A}}^*_\lambda = \mathcal{A}^*) = 1$, where $N = \sum_{s=1}^S T_s$; that is, as the sample size tends to infinity the proposed estimator will identify the collection of covariates that are truly related to the τth quantile of the response with probability approaching unity. A formal statement of this result and the asymptotic properties of the proposed estimator is now provided.

Theorem 1. *Under conditions 1–4 provided in Appendix S1, if* $\lambda \to \infty$ *and* $\lambda(Nh^4)^{-1/2} \to 0$ *as* $N = \sum_{s=1}^S T_s \to \infty$, *the following results hold:*

1. *consistency in variable selection:* $\lim_{N\to\infty} P(\widetilde{\mathcal{A}}^*_\lambda = \mathcal{A}^*) = 1$
2. *asymptotic consistency:* $\widetilde{\boldsymbol{\theta}}^{0*}_{\tau\mathcal{A}^*}(h, \lambda) \xrightarrow{p} \boldsymbol{\theta}^{0*}_{\tau 0\mathcal{A}^*}$.
3. *asymptotic normality:* $\sqrt{Nh^2}\{\widetilde{\boldsymbol{\theta}}^{0*}_{\tau\mathcal{A}^*}(h, \lambda) - \boldsymbol{\theta}^{0*}_{\tau 0\mathcal{A}^*} - \mathcal{B}_{\mathcal{A}^*}(\boldsymbol{l}^*)\} \xrightarrow{d} N(\mathbf{0}, \Sigma)$,

where $\boldsymbol{a}_{\mathcal{A}}$ *denotes the subvector of* $\boldsymbol{a}$ *corresponding to the index set* $\mathcal{A}$.

A proof of this result is provided in Appendix S1, along with closed-form expressions for the asymptotic bias $\mathcal{B}_{\mathcal{A}^*}(\boldsymbol{l}^*)$ and covariance matrix Σ.

The statement of Theorem 1 warrants several comments. First, unlike many classical regression methodologies, the proposed approach is not reliant on the aforementioned asymptotic properties to perform variable selection; that is, because the proposed method results in a sparse estimator, it does not make use of asymptotic inference to perform variable selection. The asymptotic normality of the proposed estimator is established solely for completeness. Second, the aforementioned result guarantees, under the stated regularity conditions, that the proposed estimator is asymptotically consistent, and that the approach possesses the consistency in variable selection property; that is, in the limit, the estimator in Equation 3 will not only select the truly significant covariates, but it will also precisely estimate the associated effects. Last, this result holds at one particular geographic location, that is, at $\boldsymbol{l}^*$. Given the scope of this work, extending this result to the entire spatial region is of interest. That is, based on the estimator $\tilde{I}_{j\tau}$ proposed in Section 2, it would be desirable to have that the $P(\tilde{I}_{j\tau} = I_{j\tau}, \forall j)$ goes to unity as the sample sizes tend to infinity, and the following corollary provides this result.

Corollary 1. *Under the conditions of Theorem 1, one has that* $P(\tilde{I}_{j\tau} = I_{j\tau}, \forall j)$ *converges to 1 as N goes to infinity.*

A proof of this result is provided in Appendix S1.

4 | SIMULATION STUDY

In order to illustrate the finite sample performance of the proposed methodology, the following simulation is conducted. This study aims to evaluate two characteristics of the proposed approach: estimation accuracy and its ability to perform variable selection at locations in a spatial domain. To this end, this study considers $T = 100$ observations at each of $S = 100$ locations. These sample sizes were chosen to be significantly smaller than the sample sizes that are available in the motivating data application. The spatial locations $\boldsymbol{l}_s$ are chosen to be an equally spaced grid on $[-3, 3] \times [-3, 3]$, as depicted in Figure 1. Each of the data points $(Y_{st}, \mathbf{x}_{st})$, for $s = 1, \ldots, S$ and $t = 1, \ldots, T$, is generated according to the following model:

$$Y_{st} = \mathbf{x}'_{st}\gamma(\boldsymbol{l}_s) + \epsilon_{st}, \tag{8}$$

where $\mathbf{x}_{st} \overset{iid}{\sim} N(\mathbf{0}, I_4)$, I_4 is a 4×4 identity matrix, $\gamma(\boldsymbol{l}_s) = \{f_1(\boldsymbol{l}_s), \ldots, f_4(\boldsymbol{l}_s)\}'$, and ϵ_{st} are iid random variables generated from a rescaled t-distribution with three degrees of freedom, where rescaling provides a standard deviation of 0.1. For $j = 1, \ldots, 4$, the functions $f_j(\cdot)$ are the sum of truncated bivariate Gaussian density functions, truncated to create regions of significance/insignificance, that is, regions where the regression coefficient surfaces are identically equal to zero. Contour plots of $f_j(\cdot)$, for $j = 1, \ldots, 4$, over the region of interest are provided in the first column of Figure 1, and the first column of Figure 2 provides a depiction of where $f_j(\boldsymbol{l}) \neq 0$; that is, these figures depict I_{jr}, for $j = 1, \ldots, 4$. Note that the choices of these functions provide for a broad range of scenarios: the true regression parameter surface changing sign spatially (f_1 and f_3), a very small signal relative to the noise level (f_2), small areas of insignificance between areas of significance (f_1 and f_3), and large areas of insignificance (f_2 and f_4).

Several comments about the simulation design are warranted. First, the study considers four covariates, each of which are generated according to a standard normal distribution. This emulates the process of standardizing covariates that is common in the penalized regression literature. Second, the effect sizes given by $f_j(\cdot)$ range from approximately -1.0 to 1.0 for three of the covariates and -0.20 to 0.20 for one of the covariates. Thus, this specification leads to a broad spectrum of signal-to-noise ratios when one considers the standard deviation of the error terms (i.e., 0.10). Last, by generating data in this fashion, one has that for all τ, $\beta_{j\tau}(\boldsymbol{l}) = f_j(\boldsymbol{l})$, for $j = 1, \ldots, 4$.

The aforementioned process is used to generate $B = 1,000$ independent data sets, which are analyzed using the methodology outlined in Section 2. The analysis of each data set is performed at two quantiles, that is, at $\tau = 0.50$ and $\tau = 0.95$. These two separate analyses illustrate the characteristics of the proposed approach when used to estimate the central tendency and the tails of the conditional distribution of the response. The leave-one-out cross-validation technique described in Section 2.2 is utilized to identify the smoothing and penalty parameters h and λ, respectively, for each of the 1,000 data sets. In order to graphically depict the resulting estimators, the regression coefficients are estimated at $M = 10,000$ locations $\boldsymbol{l}^*_m$ throughout the spatial region of interest. The spatial locations are taken to be a 100×100 grid of equally spaced points on $[-3, 3] \times [-3, 3]$. For a given h and λ, the corresponding leave-one-out cross-validation value took approximately one minute, on average, to compute. After selecting h and λ, the computing time necessary to perform each spatial interpolation was less than 0.5 seconds.

The second and third columns of Figure 1 provide contour plots of the sample median of the $B = 1,000$ estimates of $\beta_{j\tau}(\boldsymbol{l}^*_m)$ for $\tau = 0.50$ and $\tau = 0.95$, respectively, at every considered value of $\boldsymbol{l}^*_m$. Note that the true surface that is being estimated is depicted in the contour plots in the first column of Figure 1. This figure illustrates that the proposed methodology can accurately estimate $\beta_{j\tau}(\boldsymbol{l})$, for $j = 1, \ldots, 4$, across a spatial domain, for both the central tendencies (i.e., when $\tau = 0.50$) and the extremes (i.e., when $\tau = 0.95$) of the response. One will note that a minor loss in accuracy is observed when $\tau = 0.95$, but this is expected because the estimator is attempting to estimate the tails of the conditional distribution of the response. With that being said, the proposed approach is still able to effectively estimate the general spatial trends of the regression coefficient surfaces when $\tau = 0.95$. The second and third columns of Figure 2 provide a spatial depiction of the proportion of times that the proposed estimator is nonzero for $\tau = 0.50$ and $\tau = 0.95$, respectively, at every considered spatial location $\boldsymbol{l}^*_m$. The first column of Figure 2 depicts the regions of true significance/insignificance. From this figure, it can be seen that the proposed methodology accurately identifies the regions on which the covariates are truly related to the response, for all considered values of τ. In order to assess variability, Web Figure 1 provides contour plots of the sample standard deviation of the $B = 1,000$ estimates of $\beta_{j\tau}(\boldsymbol{l}^*_m)$, for $\tau = 0.50$ and $\tau = 0.95$, at every considered value of $\boldsymbol{l}^*_m$. In summary, the results of this simulation study indicate that the proposed approach is capable of accurately quantifying the relationship between a set of covariates and the response at multiple quantiles across a spatial domain. Moreover, the methodology developed in Section 2 is capable of accurately identifying spatial regions of significance/insignificance.

Several additional simulation studies were conducted in order to evaluate the performance of the proposed approach in other settings that are commonly encountered in spatial analyses. First, a study (results not shown) considering normal errors

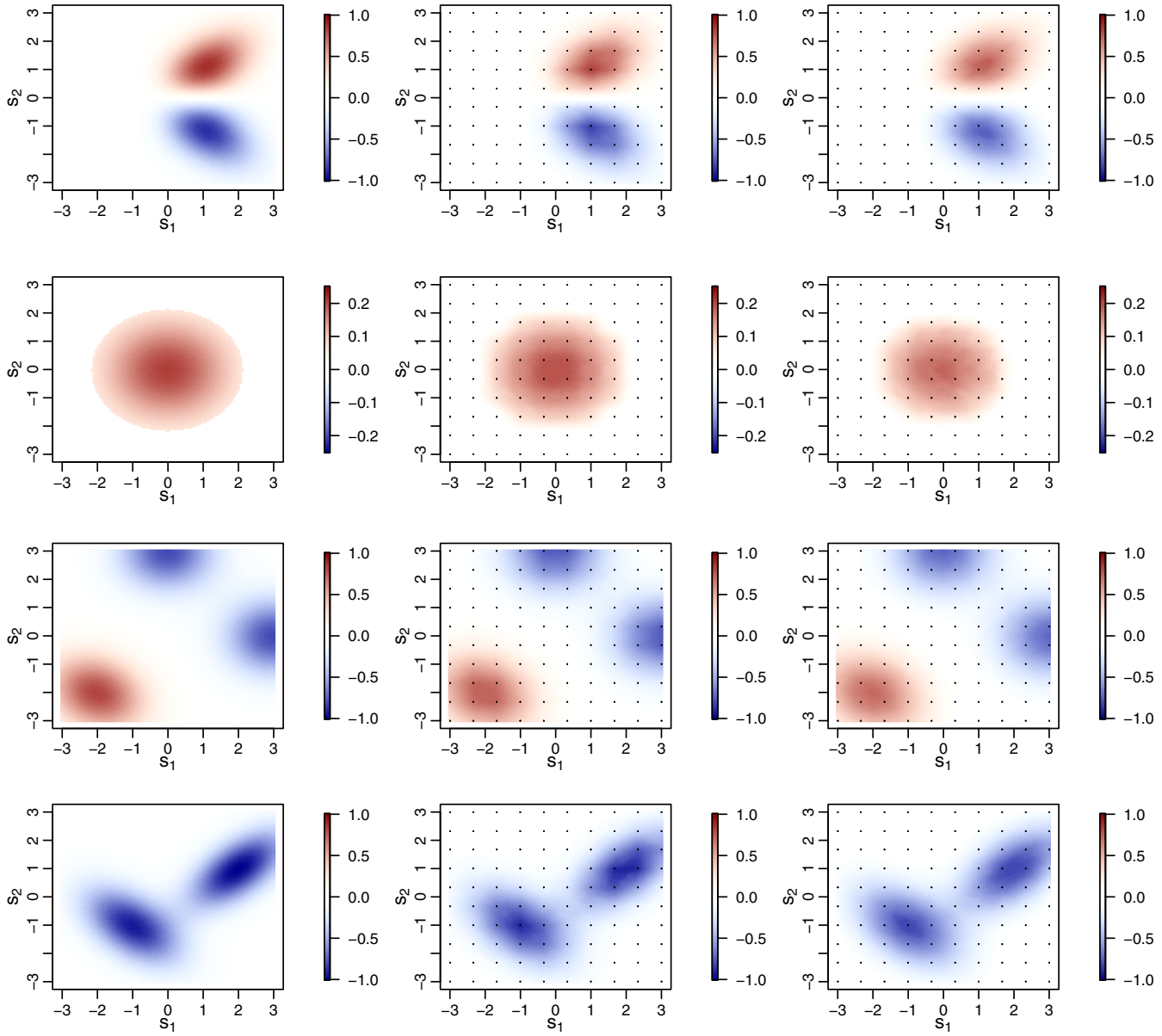

FIGURE 1 Plots of $\boldsymbol{\beta}_j(\boldsymbol{l}_m^*)$ (left panels), the pointwise medians of the estimates when $\tau = 0.50$ (middle panels), and the pointwise medians of the estimates when $\tau = 0.95$ (right panels)

was also performed and provided practically identical results to those discussed above. Second, as environmental covariates could potentially be correlated with one another, a simulation study considering correlated covariates was conducted. Third, in spatial analyses, the errors could exhibit spatial correlation, and as such a simulation investigating the performance of the proposed methodology was conducted to assess the impact of this characteristic. Last, a study considering both spatial correlation and correlated covariates was performed. The details and a summary of the results of these additional studies are provided in Appendix S2. Through all of these additional studies, no appreciable differences were found with the conclusions drawn above; that is, these additional studies again indicate that the proposed approach is capable of accurately quantifying the relationship between a set of covariates and the response at multiple quantiles across a spatial domain, as well as being able to identify spatial regions of significance/insignificance. Further, one should note that a primary strength of the proposed methodology is that it is capable of estimating different types of spatial associations at the different quantiles of the conditional distribution of the response given the covariates; that is, the proposed technique can be used to estimate $\beta_{j\tau}(\boldsymbol{l})$, for $\tau \in \{\tau_1, \tau_2\}$, even when $\beta_{j\tau_1}(\boldsymbol{l}) \neq \beta_{j\tau_2}(\boldsymbol{l})$. For ease of exposition, this particular feature is not illustrated through the study design discussed above, but is demonstrated through the results obtained from the motivating example.

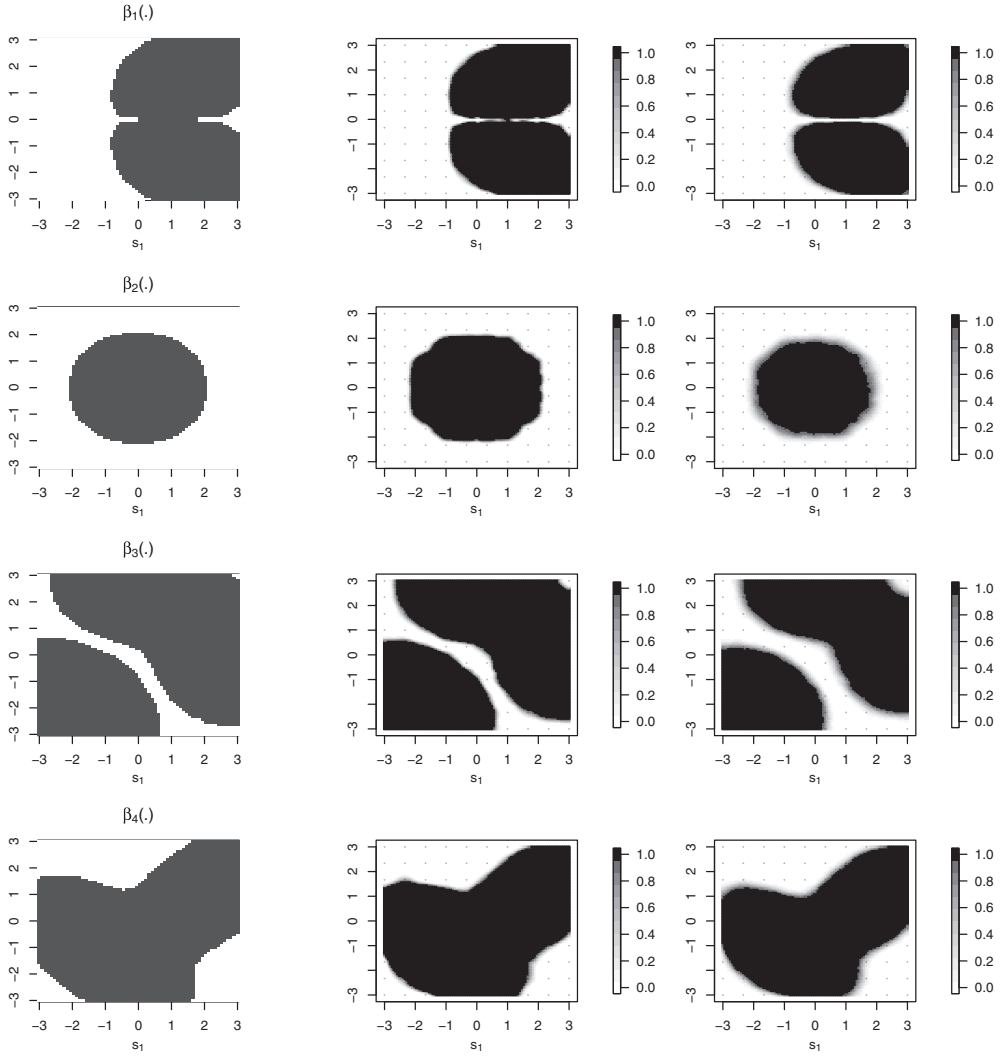

FIGURE 2 The region of significance for $\boldsymbol{\beta}_j(t_m^*)$ (left panels), the pointwise proportion of nonzero estimates when $\tau = 0.50$ (middle panels), and the pointwise proportion of nonzero estimates when $\tau = 0.95$ (right panels). Note that the white and black regions in the left panels depict regions where $\boldsymbol{\beta}_j(t_m^*)$ is zero and nonzero, respectively

5 | SPATIALLY MODELING THE METEOROLOGICAL DRIVERS OF $PM_{2.5}$

Attention is now turned to modeling the meteorological drivers of $PM_{2.5}$ over the Eastern United States.

5.1 | Data and study area

The study region considered in this analysis roughly corresponds to the Eastern Time Zone of the United States. The response variable of interest is daily average $PM_{2.5}$ levels recorded at 174 Environmental Protection Agency (EPA) stations, with consistent data records, within this region. Figure 3 provides a spatial map that depicts the location of each of these stations. The data used in this analysis were collected between the years 2010 and 2014. Further, as it is believed that the drivers of $PM_{2.5}$ may differ by season, the data are divided into four seasons. The analysis presented here focuses on the summer and winter seasons, with summer defined to be the months of June–August, and winter being the months of December–February.

The meteorological variables for this analysis are obtained from the North American Regional Reanalysis (NARR) and consist of the 12 covariates given in Table 1. The process of selecting these covariates is driven by information gained from other similar studies; for example, see Jacob and Winner (2009) and Porter et al. (2015). Note that the precipitation indicator variable takes the value 1 if any of the corresponding day's NARR categorical rain readings (presence/absence of precipitation) takes the value of 1. Further, lower tropospheric stability represents the difference between the potential temperature at the surface and

FIGURE 3 Plot of the locations of the EPA stations used in the $PM_{2.5}$ analysis

TABLE 1 The 12 considered meteorological variables that were obtained from the NARR

Variable	Abbreviation	Comment
Precipitation	Precip	Daily presence/absence
Night air temperature	Night Temp	Nighttime average
Day air temperature	Day Temp	Daytime average
Night planetary boundary layer height	Night HPBL	Nighttime average
Day planetary boundary layer height	Day HPBL	Daytime average
Relative humidity	RH	Daytime average
Lifted index	LFTX	Daytime average
Lower Tropospheric Stability	LTS	Average of previous 24 hr
Wind speed	Wnd Spd	Average of previous 48 hr
Turbulence kinetic energy	TKE	Same day average
Downward shortwave radiative flux	DSWRF	Average of previous day
Percent cloud cover	TCDC	Average of previous day

the potential temperature at 700 hPa (Klein and Hartmann, 1993; Porter et al., 2015). In order to be able to compare estimated coefficients throughout the spatial domain, all variables are standardized.

5.2 | Spatial analysis

The aim of this analysis is to improve the level of understanding regarding the spatial relationship between $PM_{2.5}$ and the 12 meteorological variables presented in Table 1, throughout the study region. Moreover, it is desired to assess this relationship throughout different seasons (i.e., summer and winter) and at different quantiles of the conditional distribution of $PM_{2.5}$ levels ($\tau = 0.50$ and $\tau = 0.95$). To accomplish this task, the proposed model is fit to the available data (within season) at a grid of 1,783 points covering the Eastern United States. The strategy discussed in Section 2.2 is utilized to determine the tuning

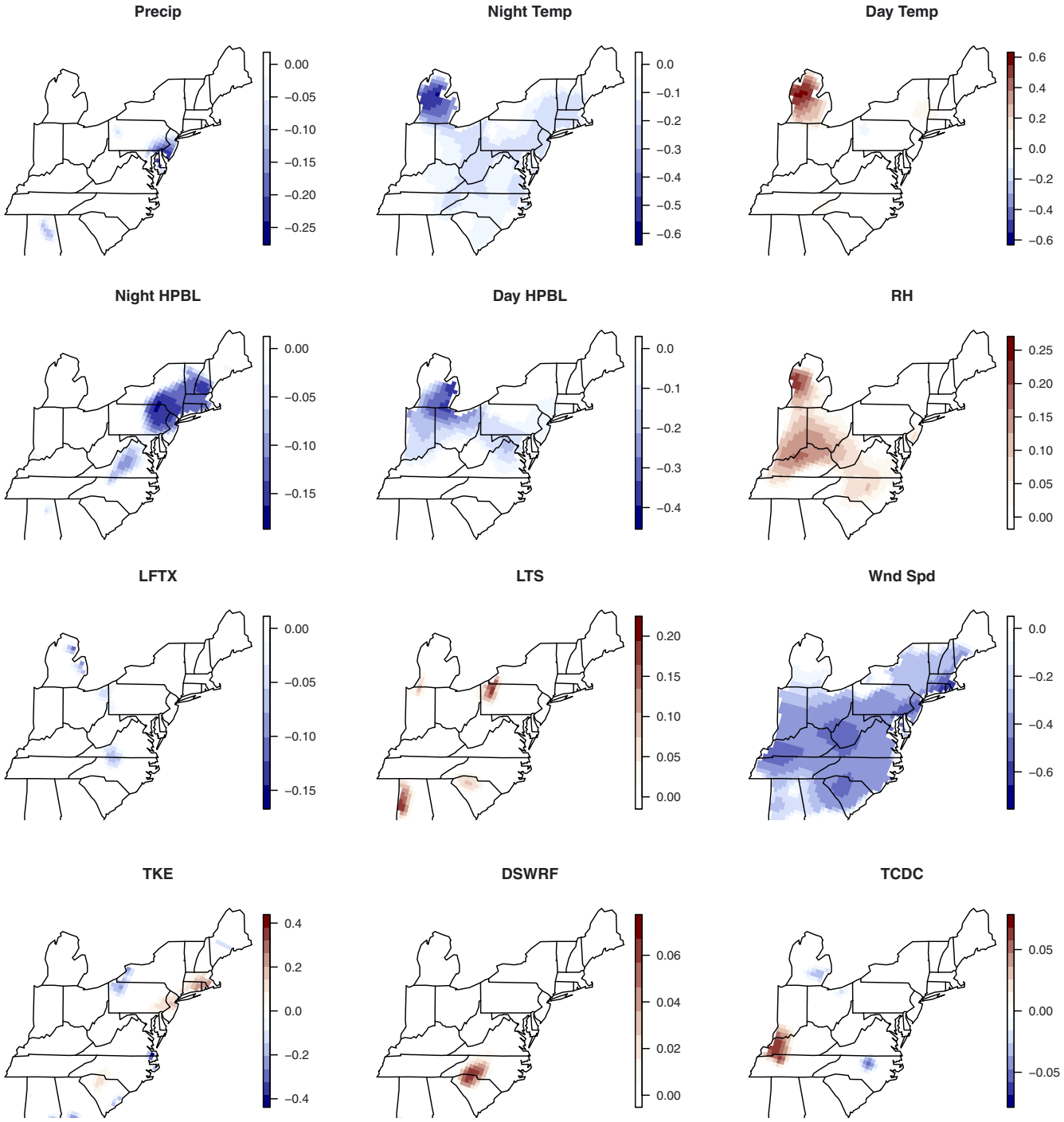

FIGURE 4 Results of the spatial analysis of the winter $PM_{2.5}$ data: Presented results include the estimates of $\beta_{j\tau}(\boldsymbol{l}_m^*)$ for each of the considered meteorological drivers, when $\tau = 0.50$. Note that regions of insignificance are depicted in white

parameters h and λ. For each season and quantile, Figures 4 –7 present the estimated regression coefficient surfaces for all 12 variables, and Web Figures 2 –5 provide the same results but on a common scale so that one can examine relative importance. In particular, these figures summarize the model fits for the different seasons and for the different considered quantiles, in addition to providing regions of significance/insignificance for each of the considered meteorological variables.

5.2.1 | Meteorological drivers of $PM_{2.5}$ in the winter

Through the results presented in Figures 4 and 5, it appears that wind speed is the primary driver over most of the region during the winter, and as expected is negatively related to $PM_{2.5}$ levels at both quantiles of interest. Interestingly, wind speed seems to

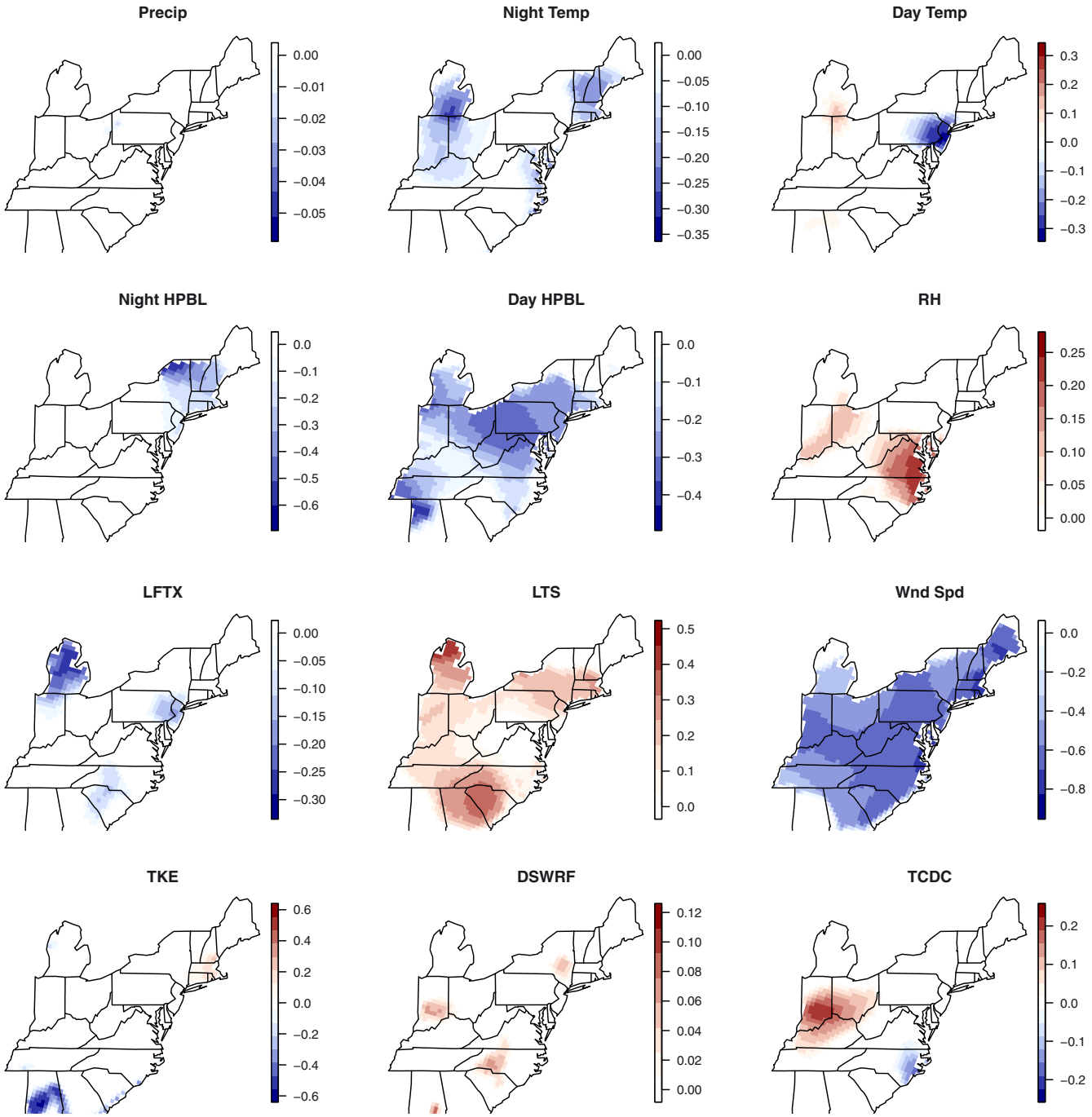

FIGURE 5 Results of the spatial analysis of the winter $PM_{2.5}$ data: Presented results include the estimates of $\beta_{j\tau}(l_m^*)$ for each of the considered meteorological drivers, when $\tau = 0.95$. Note that regions of insignificance are depicted in white

play a larger role in describing $PM_{2.5}$ levels at the 0.95 quantile in the winter, as the magnitude of its coefficient appears to be larger at this quantile. The height of the planetary boundary layer is also an important variable throughout much of the study region, suggesting that inversions may play an important role in the winter. Planetary boundary layer height (HPBL) seems to be most important in the Northern part of the region for $\tau = 0.50$. At this quantile, nighttime HPBL is most important in the far Northeast, whereas daytime HPBL is most important in the upper Midwest. For $\tau = 0.95$, daytime height of the planetary boundary layer is also important in Southern portions of the study area.

Air pollution is commonly associated with warmer air temperatures, but this analysis finds that nighttime air temperature has a negative relationship with $PM_{2.5}$ at both quantiles during the winter. This negative relationship between surface-level air temperature and $PM_{2.5}$ could be consistent with the importance of inversions. Lower tropospheric stability is found to be

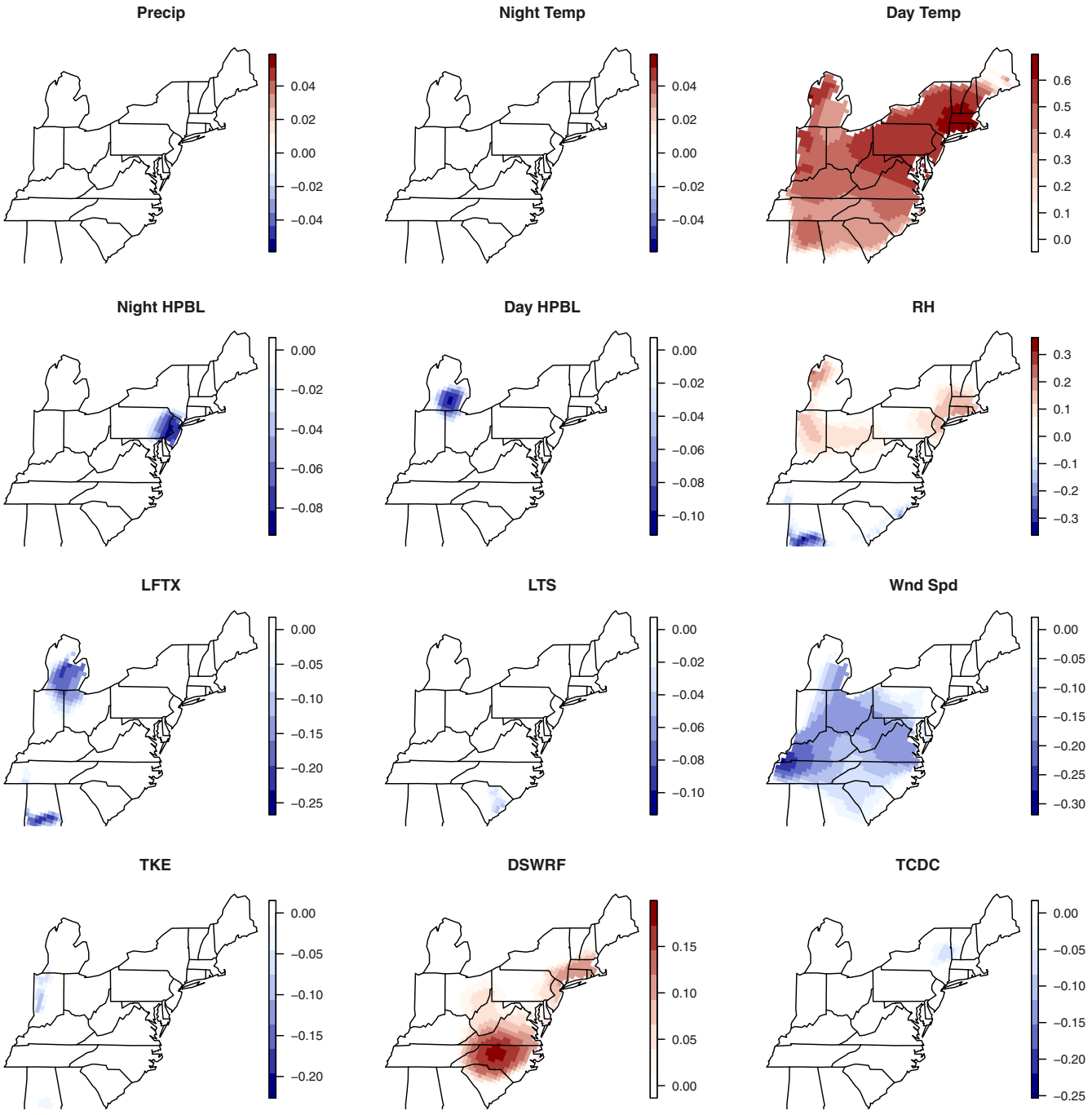

FIGURE 6 Results of the spatial analysis of the summer $PM_{2.5}$ data: Presented results include the estimates of $\beta_{j\tau}(\boldsymbol{l}_m^*)$ for each of the considered meteorological drivers, when $\tau = 0.50$. Note that regions of insignificance are depicted in white

positively related to $PM_{2.5}$ levels throughout much of the study region for $\tau = 0.95$, but does not look to be significant for $\tau = 0.50$. Relative humidity's association tends to be positive in large portions of the region at both quantiles. In much of the region, cloud cover does not seem to be strongly related to $PM_{2.5}$ levels for $\tau = 0.50$, but seems to have more importance in parts of the study region for $\tau = 0.95$.

5.2.2 | Meteorological drivers of $PM_{2.5}$ in the summer

Through the results presented in Figures 6 and 7, air temperature appears to be the most significant meteorological driver for describing median (i.e., when $\tau = 0.50$) $PM_{2.5}$ throughout the Eastern United States. Although nighttime air temperature is found to be negatively related to $PM_{2.5}$ in the winter, daytime air temperature is found to be positively related to $PM_{2.5}$ during the

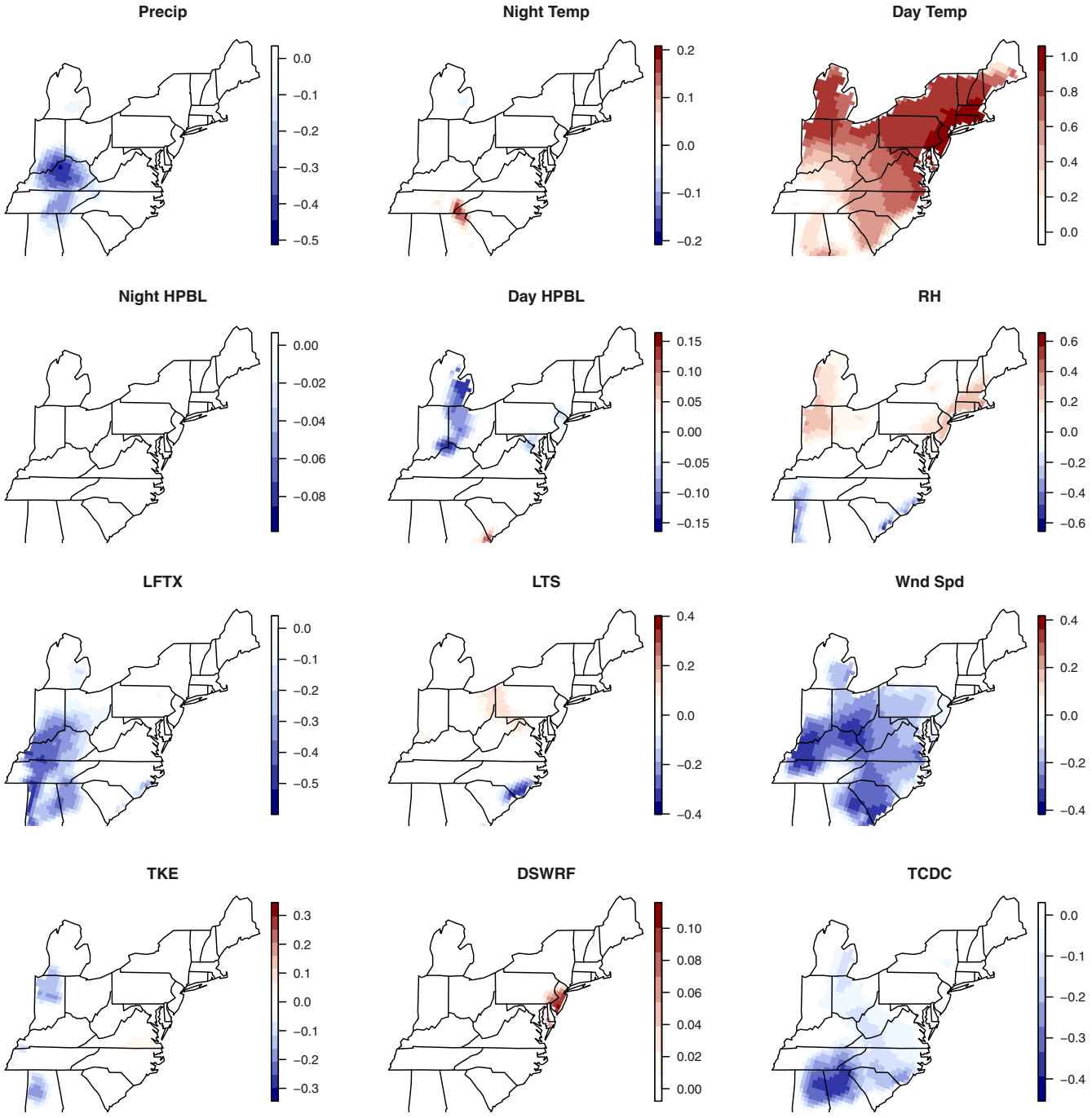

FIGURE 7 Results of the spatial analysis of the summer $PM_{2.5}$ data: Presented results include the estimates of $\beta_{jt}(l_m^*)$ for each of the considered meteorological drivers, when $\tau = 0.95$. Note that regions of insignificance are depicted in white

summer, especially in the Northeast. Not unexpectedly, wind speed seems to be a secondary driver and is negatively related to $PM_{2.5}$ throughout most of the region. Relative humidity seems to be positively related to $PM_{2.5}$ at both quantiles in the Northern portion of the region, but looks to have a negative relationship in Southern portions of the region.

At the 0.50 quantile of $PM_{2.5}$ levels, the proposed modeling approach tends to select downward shortwave radiative flux on the previous day in the Carolinas, but tends to select cloud cover over that region at the 0.95 quantile. Interestingly, precipitation and lifted index looks to be important over Kentucky and Tennessee for $\tau = 0.95$, but not for $\tau = 0.50$.

5.2.3 | Summary discussion of analysis

It is worthwhile to point out that throughout the region of interest, during both the winter and summer and at the different considered quantiles, the association between $PM_{2.5}$ levels and the meteorological covariates vary spatially. In particular, the

magnitude of the estimated effect associated with each of the meteorological variables changes with the geographic area, with the propensity to even change direction; that is, in some regions, variables are positively related with $PM_{2.5}$ levels, and in others, they possess a negative relationship. Moreover, the proposed approach finds that, in some regions of the study area, meteorological drivers are significantly related, while they are insignificant in other areas. These findings are possibly attributable to the variable composition of $PM_{2.5}$ (Jacob & Winner, 2009); that is, the composition of $PM_{2.5}$ tends to vary spatially, and as a consequence, the set of significant meteorological drivers should as well. Last, it is possible that $PM_{2.5}$ levels are spatially correlated, but given the results of the numerical studies presented in Section 4, it is believed that this effect (if present) would not unduly influence the results of this analysis.

The primary goal of this work is not to model $PM_{2.5}$ levels, but rather to spatially model the effects of meteorological drivers on different quantiles of $PM_{2.5}$ levels. The work of Russell, Cooley, Porter, and Heald (2016) is similar in spirit, in the sense that these authors spatially model the meteorological drivers' effects on air pollution, but they take a drastically different approach, and focus on ground-level ozone extremes. The results of the analysis presented in Porter et al. (2015) are interesting to compare and contrast with the results presented above. In particular, Porter et al. (2015) performs variable selection at a large number of U.S. locations individually, using standard quantile regression models. This complimentary analysis found that air temperature is the main driver of $PM_{2.5}$ during the summer, with wind speed and lifted index also being important, throughout the study region considered in this work. Also coinciding with the findings presented above, Porter et al. (2015) found that the height of the planetary boundary layer is a primary driver throughout the Eastern United States during the winter, with turbulence kinetic energy and relative humidity also being important. It is worthwhile to point out that any differences between these two analyses are likely attributable to the differing variable selection strategies, and the fact that the two analyses consider slightly different sets of meteorological variables.

6 | DISCUSSION AND CONCLUSION

In this work, a local linear quantile regression methodology is developed for the purposes of estimating the spatial relationship between a set of covariates and the conditional quantiles of a response variable. In particular, at any spatial location within the region of interest, the proposed methodology can be used to address two main issues, that is, parameter estimation and variable selection, and these are accomplished uniquely at every spatial location. In this sense, the proposed modeling procedure is quite different compared to many existing spatial quantile regression models, because it makes use of an adaptive LASSO penalty to perform model selection. The theoretical properties of the proposed estimator have been established, and the finite sample characteristics are illustrated through simulation. Further, the proposed methodology is used to spatially model the effects of meteorological drivers for different quantiles of the conditional distribution of $PM_{2.5}$ levels throughout the Eastern United States.

There are several topics for future research pertaining to this proposal that could be undertaken. First, and foremost, the development of techniques that could be implemented to conduct model validation would be of key interest. Second, developing an approach that would allow the tuning parameters to vary spatially could also help with the performance of the proposal, especially in areas where the effect size is relatively small. Third, efforts to extend the theoretical results presented in Section 3 could be made to allow for spatial and/or temporal correlation. This could likely be accomplished by adapting the techniques outlined in Wu (2007). Last, generalizing the methodology to allow the effect estimates to vary in time could also be a reasonable pursuit.

ACKNOWLEDGEMENTS

The authors wish to thank the Editor, the Associate Editor, and two anonymous referees for their helpful comments on an earlier version of this article. Clemson University is acknowledged for its generous allotment of computing time on the Palmetto cluster. Christopher S. McMahan was partially supported by Grant R01 AI121351 from the National Institutes of Health.

REFERENCES

Cai, Z., & Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, *103*(484), 1595–1608.

Chen, V. Y., Deng, W., Yang, T., & Matthews, S. A. (2012). Geographically weighted quantile regression (GWQR): An application to US mortality data. *Geographical Analysis*, *44*(2), 134–150.

Fan, J., & Gijbels, I. (1996). *Local polynomial modeling and its applications*. London: Chapman & Hall.

Fan, J., Hu, T. C., & Truong, Y. K. (1994). Robust nonparametric function estimation. *The Scandinavian Journal of Statistics*, *21*, 433–446.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Hallin, M., Lu, Z., & Yu, K. (2009). Local linear spatial quantile regression. *Bernoulli*, *15*(3), 659–686.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York: Springer-Verlag.

Honda, T. (2004). Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inference*, *121*(1), 113–125.

Jacob, D. J., & Winner, D. A. (2009). Effect of climate change on air quality. *Atmospheric Environment*, *43*(1), 51–63.

Kai, B., & Li, R. (2010). Local composite quantile regression smoothing: An efficient and safe alternative to local polynomial regression. *Journal of the Royal Statistical Society: Series B*, *72*(1), 49–69.

Khafaie, M. A., Yajnik, C. S., Salvi, S. S., & Ojha, A. (2016). Critical review of air pollution health effects with special concern on respiratory health. *Journal of Air Pollution and Health*, *1*(2), 123–136.

Kim, M. (2007). Quantile regression with varying coefficients. *The Annals of Statistics*, *35*(1), 92–108.

Klein, S. A., & Hartmann, D. L. (1993). The seasonal cycle of low stratiform clouds. *Journal of Climate*, *6*(8), 1587–1606.

Koenker, R. (2015). quantreg: Quantile Regression. R package version 5.19. Retrieved from http://CRAN.R-project.org/package=quantreg

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

Krewski, D., Jerrett, M., Burnett, R. T., Ma, R., Hughes, E., Shi, Y., … Calle, E. E. (2009). *Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality* (Number 140). Boston, MA: Health Effects Institute.

Li, K. (1984). Consistency for cross-validated nearest neighbor estimates in nonparametric regression. *The Annals of Statistics*, *12*(1), 230–240.

Lopez, D. H., Rabbani, M. R., Crosbie, E., Raman, A., Arellano, A. F., & Sorooshian, A. (2015). Frequency and character of extreme aerosol events in the Southwestern United States: A case study analysis in Arizona. *Atmosphere*, *7*(1), 1.

Pope, C. A., III, Ezzati, M., & Dockery, D. W. (2009). Fine-particulate air pollution and life expectancy in the United States. *New England Journal of Medicine*, *360*(4), 376–386.

Porter, W., Heald, C., Cooley, D., & Russell, B. (2015). Investigating the observed sensitivities of air quality extremes to meteorological drivers via quantile regression. *Atmospheric Chemistry and Physics Discussions*, *15*(10), 14075–14109.

Rice, J. (1984). Bandwidth selection for nonparametric regression. *The Annals of Statistics*, *12*(4), 1215–1230.

Russell, B., Cooley, D., Porter, W., & Heald, C. (2016). Modeling the spatial behavior of the meteorological drivers' effects on extreme ozone. *Environmetrics*, *27*(6), 334–344.

Smith, R. L., Kolenikov, S., & Cox, L. H. (2003). Spatiotemporal modeling of $PM_{2.5}$ data with missing values. *Journal of Geophysical Research: Atmospheres*, *108*(D24), 9004.

Sun, Y., Wang, H., & Fuentes, M. (2016). Fused adaptive lasso for spatial and temporal quantile function estimation. *Technometrics*, *58*(1), 127–137.

Tai, A. P., Mickley, L. J., & Jacob, D. J. (2010). Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: Implications for the sensitivity of $PM_{2.5}$ to climate change. *Atmospheric Environment*, *44*(32), 3976–3984.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288.

Wang, H., Zhu, Z., & Zhou, J. (1998). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, *37*(6), 3841–3866.

Wu, W. (2007). M-estimation of linear models with dependent errors. *The Annals of Statistics*, *35*(2), 495–521.

Wu, Y., & Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, *19*(1), 801–817.

Yu, K., & Jones, M. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, *93*(441), 228–237.

Zhu, S., Huang, M., & Li, R. (2012). Semiparametric quantile regression with high-dimensional covariates. *Statistica Sinica*, *22*(1), 1379–1401.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, L., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, *36*(4), 1509–1533.

How to cite this article: Russell BT, Wang D, McMahan CS. Spatially modeling the effects of meteorological drivers of $PM_{2.5}$ in the Eastern United States via a local linear penalized quantile regression estimator. *Environmetrics*. 2017;28:e2448. https://doi.org/10.1002/env.2448