The Annals of Applied Statistics
2019, Vol. 13, No. 4, 2158–2188
https://doi.org/10.1214/19-AOAS1276

PREDICTION OF SMALL AREA QUANTILES FOR THE CONSERVATION EFFECTS ASSESSMENT PROJECT USING A MIXED EFFECTS QUANTILE REGRESSION MODEL[1]

BY EMILY BERG AND DANHYANG LEE

Iowa State University and University of Alabama

Quantiles of the distributions of several measures of erosion are important parameters in the Conservation Effects Assessment Project, a survey intended to quantify soil and nutrient loss on crop fields. Because sample sizes for domains of interest are too small to support reliable direct estimators, model based methods are needed. Quantile regression is appealing for CEAP because finding a single family of parametric models that adequately describes the distributions of all variables is difficult and small area quantiles are parameters of interest. We construct empirical Bayes predictors and bootstrap mean squared error estimators based on the linearly interpolated generalized Pareto distribution (LIGPD). We apply the procedures to predict county-level quantiles for four types of erosion in Wisconsin and validate the procedures through simulation.

1. Introduction. Agricultural production is associated with water and wind erosion. The Natural Resources Conservation Service (NRCS) of the United States Department of Agriculture (USDA) assists farmers with implementation of conservation practices intended to mitigate erosion. With the partial aim of assessing the impact of provisions in the 2002 farm bill that increased funding for conservation programs, the USDA initiated the Conservation Effects Assessment Project (CEAP). The first national CEAP survey, conducted from 2003–2006, was followed by four regional surveys in 2011–2014, and data processing for the 2015–2016 national CEAP survey is ongoing.

The estimation domains for the 2003–2006 CEAP survey are 12 major watersheds (USDA/NRCS (2012)), regions of land in which water flows into relatively large water bodies. For instance, the Upper Mississippi River Basin, which has a CEAP sample size of approximately 3703 units, covers much of Wisconsin, Minnesota and Iowa. Estimates for geographic regions, such as counties, that are smaller than the twelve major watersheds can help direct conservation policies, inform farmers' decisions and provide a more detailed understanding of erosion. Because sample sizes for counties intersecting the Upper Mississippi River Basin

Received November 2017; revised April 2019.

[1]Supported in part by the United States Department of Agriculture (USDA NRCS CESU Agreement 68-7482-13-529) and by the the National Science Foundation (MMS-1733572).

Key words and phrases. Quantile regression, empirical Bayes, parametric bootstrap, erosion, environmental monitoring.

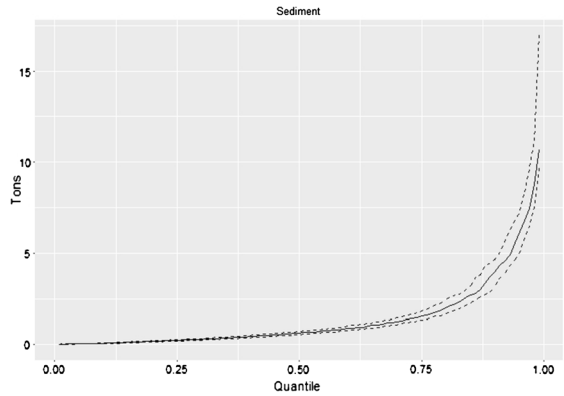

FIG. 1. *Direct estimates of quantiles of sediment loss for Wisconsin with* 95% *confidence intervals.*

typically range from 2 to 40, and some are less than 2, model based small area estimation methods (Rao and Molina (2015)) are needed.

CEAP publications have estimates of the quantiles of the distributions of several measures of erosion, including surface water runoff, sheet and rill erosion, sediment, and the annual change in soil organic carbon. Figure 1, modeled after similar plots in USDA/NRCS (2012), shows direct estimates of quantiles of sediment loss for Wisconsin, along with upper and lower 95% confidence interval limits calculated using the Woodruff (1952) method. For CEAP variables, such as sediment, with skewed distributions, the median is a more interpretable estimator of a typical value than the mean. Estimates of quartiles and extreme quantiles give information on the distribution of erosion in the study domain. Estimates of multiple quantile levels are useful for assessing the efficacy of different conservation strategies and for evaluating interactions between agriculture and the environment more generally. As explained in Goebel and Kellogg (2002), "The most unusual situations are often the most harmful relative to environmental factors; these are in the *tails* of the statistical distributions of...variates and will be lost or averaged out if only aggregate or *representative* values are used. This is an important consideration when analyzing agri-environmental issues with any type of modeling." Our objective is to construct estimates of quantiles of the distributions of several measures of erosion for Wisconsin counties and provide appropriate measures of uncertainty.

Use of quantile regression to construct small area predictors of quantiles for CEAP is appealing because quantile regression ties the estimation procedure to the parameters of interest. To construct small area predictors of quantiles for CEAP, we apply the mixed effects version of the linearly interpolated generalized Pareto distribution (LIGPD) defined in Jang and Wang (2015). To introduce the LIGPD, let y_{ij} denote the variable of interest for unit j in area i, where $i = 1, \ldots, D$, and $j = 1, \ldots, N_i$. Let $\boldsymbol{b}_i \in \mathbb{R}^{p_1}$ be an area random effect with density function

$f_b(\boldsymbol{b}_i \mid \boldsymbol{\Sigma}_b)$ such that $E[(\boldsymbol{b}_i, \boldsymbol{b}_i\boldsymbol{b}_i')] = (\mathbf{0}, \boldsymbol{\Sigma}_b)$. The centerpiece of the LIGPD is the mixed effects quantile regression model defined by

$$q_{ij}(\tau) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau) + \boldsymbol{z}_{ij}'\boldsymbol{b}_i, \quad i = 1, \ldots, D,\ j = 1, \ldots, N_i, \tag{1}$$

where $P(y_{ij} \le q_{ij}(\tau) \mid \boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = \tau$, $y_{ij} \perp y_{ik} \mid \boldsymbol{b}_i$ for $j \ne k$, y_{ij} has an absolutely continuous distribution, $\boldsymbol{x}_{ij} \in \mathbb{R}^{p_2}$ and $\boldsymbol{z}_{ij} \in \mathbb{R}^{p_1}$ are vectors of fixed covariates, and

$$\boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau) \le \boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau + \delta) \tag{2}$$

for $\delta \ge 0$. Because $\boldsymbol{b}_i$ does not depend on τ, $q_{ij}(\tau)$ is nondecreasing in τ for every (i, j). The LIGPD uses a generalized Pareto distribution to approximate the distribution of y_{ij} for quantiles below or above specified lower and upper bounds, as we explain precisely in Section 2. The objective is to predict functions of the distribution of $\{y_{ij} : j = 1, \ldots, N_i\}$, principally finite population quantiles.

An alternative to quantile regression, the empirical Bayes prediction (EBP) method of Molina and Rao (2010) provides a fully parametric approach to prediction of nonlinear small area parameters, such as quantiles. A seminal parametric model for small area estimation is the linear mixed effects model with normally distributed random components (Battese, Harter and Fuller (1988)). Extensions to more complex parametric forms, such as generalized linear mixed models or models with spatial or temporal dependence structures, are reviewed in Pfeffermann (2013), Rao and Molina (2015), and Jiang and Lahiri (2006). Diallo and Rao (2018) apply the EBP approach for a situation in which the random terms have skew normal distributions. Molina, Nandram and Rao (2014) use hierarchical Bayes instead of empirical Bayes to predict nonlinear small area parameters, assuming a satisfactory parametric form is specified. In CEAP, quantile estimates are desired for several measures of water and wind erosion for subdivisions of the United States. As we demonstrate in Section 3, finding a single family of parametric models that adequately describes all distributions of interest is difficult.

Quantile regression (Koenker (2005)) offers a unified framework that can accommodate diverse distributional forms. Chambers and Tzavidis (2006) use M-quantile regression for small area estimation, focusing on means and medians. Chen and Liu (2017, 2012) use empirical likelihood to estimate a quantile regression model for small area prediction, where each small area has a different tilting parameter in the density ratio model. Estimation of a different tilting parameter for each small area is undesirable for CEAP because the county sample size can be less than two. Weidenhammer et al. (2016) develop small area predictors based on the mixed effects version of the asymmetric Laplace distribution introduced by Geraci and Bottai (2007, 2014). Similar to the hierarchical models traditionally used for small area estimation, the Geraci and Bottai (2007, 2014) model, described in Appendix A.5, has a set of fixed parameters that relates the quantile of the distribution of interest to a set of covariates, and random parameters describe variation in this

relationship across the areas. Because the asymmetric Laplace distribution specifies a separate model for each quantile level, the estimates of the quantiles can decrease as τ increases and can be unstable in the tails of the distribution. The limitations of the asymmetric Laplace distribution are important for small area prediction because an estimate of the full distribution is required, rather than an estimate for an individual quantile level.

The LIGPD approximation for the model (1) supports a computationally feasible small area prediction procedure such that the estimated quantile function for any population element is nondecreasing, estimates for the tails are stable, and empirical Bayes prediction and bootstrap mean squared error (MSE) estimation are possible. A further benefit of the LIGPD is that the model makes fewer distributional assumptions than the asymmetric Laplace distribution and is therefore more broadly applicable. Jang and Wang (2015) use Bayesian methods for inference and focus on estimation of the quantile regression coefficients. We emphasize prediction, rather than parameter estimation, and develop a computationally simple frequentist procedure.

An alternative to the Jang and Wang (2015) procedure is the approach of Reich, Fuentes and Dunson (2011). We pursue the Jang and Wang (2015) procedure for small area estimation instead of Reich, Fuentes and Dunson (2011) because the estimation procedure of Reich, Fuentes and Dunson (2011) uses asymptotic distributions that may be inappropriate if the number of sampled units in an area is small. Further, the quantile function in Reich, Fuentes and Dunson (2011) is a nonlinear transformation of the random effect. In contrast, the random effect in the model (1) enters in a linear fashion and therefore has a straightforward interpretation.

We develop a small area estimation procedure based on the LIGPD of Jang and Wang (2015) with the aim of obtaining county level estimators of the quantiles of erosion measurements for CEAP that are more reliable than direct estimators. In Section 2, we present the estimation procedure. In Section 3, we apply the LIGPD estimation procedure to data from CEAP. In Section 4, we validate the estimation procedure through simulations. In Section 5, we summarize and discuss areas for future work.

2. LIGPD model and estimation procedures.

Assume that the population satisfies the model (1) and y_{ij} is observed for a sample of n_i elements for each area i. As is common in small area estimation, assume that $(\boldsymbol{x}'_{ij}, \boldsymbol{z}'_{ij})'$ is known for all N_i elements in the population for area i.

2.1. *LIGPD approximation and Bayes predictor.*

Define a sequence of quantile levels by $\tau_k = k(K+1)^{-1}$ for $k = 1, \ldots, K$, where $K \to \infty$ as $D \to \infty$. The LIGPD approximation for the density function of y_{ij} given $\boldsymbol{b}_i$ corresponding to

model (1) is defined by

$$
\begin{aligned}
&f_Y(y \mid \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i, \boldsymbol{\theta}) \\
&\quad = I\big[y < q_{ij}(\tau_1)\big]\tau_1 f_\ell(y \mid \rho_\ell, \xi_\ell) \\
&\qquad + I\big[y \geq q_{ij}(\tau_K)\big](1-\tau_K) f_u(y \mid \rho_u, \xi_u) \\
&\qquad + \sum_{k=1}^{K-1} I\big[q_{ij}(\tau_k) \leq y < q_{ij}(\tau_{k+1})\big] \frac{\tau_{k+1} - \tau_k}{q_{ij}(\tau_{k+1}) - q_{ij}(\tau_k)},
\end{aligned} \tag{3}
$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}_K', \operatorname{vech}(\boldsymbol{\Sigma}_b)', \rho_\ell, \xi_\ell, \rho_u, \xi_u)'$ is the vector of fixed parameters to be estimated, $\boldsymbol{\beta}_K = (\boldsymbol{\beta}(\tau_1)', \ldots, \boldsymbol{\beta}(\tau_K)')'$, $\operatorname{vech}(\cdot)$ denotes vector half, $I[\cdot]$ is the indicator function that is equal to 1 if the argument is true and zero otherwise, and $f_s(y \mid \rho_s, \xi_s)$ for $s = \ell, u$ are densities of generalized Pareto distributions defined as follows. Letting $u_{ij} = 0.5(\boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau_K) + \boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau_{K-1}))$ and $\ell_{ij} = 0.5(\boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau_1) + \boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau_2))$, $f_u(y \mid \rho_u, \xi_u) = (1-\tau_K)^{-1}\{1 - 0.5(\tau_{K-1} + \tau_K)\}g(y - u_{ij} \mid \rho_u, \xi_u)$, and $f_\ell(y \mid \rho_\ell, \xi_\ell) = \tau_1^{-1} 0.5(\tau_1 + \tau_2) g(-y + \ell_{ij} \mid \rho_\ell, \xi_\ell)$, where

$$
g(y \mid \rho_s, \xi_s) = \begin{cases} \rho_s^{-1}(1 + \xi_s y/\rho_s)^{-(1+1/\xi_s)} & \xi_s \neq 0, \\ \rho_s^{-1} \exp(-y/\rho_s) & \xi_s = 0, \end{cases} \tag{4}
$$

for $s = \ell, u$ with $y > 0$ for $\xi_s \geq 0$, and $0 \leq y < -\rho_s/\xi_s$ for $\xi_s < 0$. In the interest of brevity, we refer the reader to Jang and Wang (2015) for further discussion and motivation of the form (3).

The Bayes predictor of $q_{ij}(\tau)$ for squared error loss corresponding to the approximate density function (3) and the model (1) is

$$
q_{ij}^B(\tau) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}(\tau) + \boldsymbol{z}_{ij}' E[\boldsymbol{b}_i \mid \boldsymbol{y}_i; \boldsymbol{\theta}], \tag{5}
$$

where

$$
E[\boldsymbol{b}_i \mid \boldsymbol{y}_i; \boldsymbol{\theta}] = \frac{\int_{\mathbb{R}^{p_1}} \prod_{j=1}^{n_i} \boldsymbol{b}_i f_Y(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i, \boldsymbol{\theta}) f_b(\boldsymbol{b}_i \mid \boldsymbol{\Sigma}_b)\, d\boldsymbol{b}_i}{\int_{\mathbb{R}^{p_1}} \prod_{j=1}^{n_i} f_Y(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i, \boldsymbol{\theta}) f_b(\boldsymbol{b}_i \mid \boldsymbol{\Sigma}_b)\, d\boldsymbol{b}_i}, \tag{6}
$$

and $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$. If the area has no sampled units, then the conditional density of $\boldsymbol{b}_i$ is $f_b(\boldsymbol{b}_i; \boldsymbol{\Sigma}_b)$, and the conditional mean is $\boldsymbol{0}$. The predictor (6) is a function of $\boldsymbol{\theta}$; in Section 2.2, we define an estimator of $\boldsymbol{\theta}$.

2.2. *Parameter estimation for the LIGPD.* We define an iterative procedure that we call the simplified EM algorithm to estimate $\boldsymbol{\beta}_K$ and $\boldsymbol{\Sigma}_b$. The iteration alternates between calculation of conditional moments and optimization but is not a full EM algorithm. The optimization step minimizes Koenker's check function (Koenker (2005)) defined by $\rho_\tau(u) = u(\tau - I[u < 0])$, a standard optimization criterion for quantiles because for Z with absolutely continuous distribution function $F_Z(z)$, $\operatorname{argmin}_a E[\rho_\tau(Z - a)] = F_Z^{-1}(\tau)$. Let $\hat{\boldsymbol{\theta}}^{(0)}$ denote the vector of initial estimators of $\boldsymbol{\theta}$, where the procedure to obtain the initial values, defined in

Appendix A.1, uses moment type methods to estimate $\boldsymbol{\Sigma}_b$. For $m = 1, 2, \ldots, M$, alternate between the following steps.

1. Define the updated estimator of $\boldsymbol{\Sigma}_b$ by

$$\hat{\boldsymbol{\Sigma}}_b^{(m)} = (D - p_2)^{-1} \sum_{i=1}^{D} E\big[\boldsymbol{b}_i \boldsymbol{b}_i' \mid \boldsymbol{y}_i; \hat{\boldsymbol{\theta}}^{(m-1)}\big]. \tag{7}$$

Define a predictor of $\boldsymbol{b}_i$ in the mth step by

$$\hat{\boldsymbol{b}}_i^{(m)} = E\big[\boldsymbol{b}_i \mid \boldsymbol{y}_i; \hat{\boldsymbol{\theta}}^{(m-1)}\big].$$

Appendix A.2 defines the numerical approximations to the integrals defining the expectations for univariate $\boldsymbol{b}_i$.

2. We use the method of Koenker and Ng (2005) to update the estimator of $\boldsymbol{\beta}_K$ to maintain the monotonicity restriction (2). First, define

$$\hat{\boldsymbol{\beta}}^{(m)}(\tau_{\lfloor 0.5(K+1)\rfloor}) = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{D} \sum_{j=1}^{n_i} \rho_{\tau_{\lfloor 0.5(K+1)\rfloor}}(y_{ij} - \boldsymbol{z}_{ij}' \hat{\boldsymbol{b}}_i^{(m)} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}), \tag{8}$$

where $\lfloor 0.5(K+1)\rfloor$ is the integer part of $0.5(K+1)$. For $k = \lfloor 0.5(K+1)\rfloor + 1, \ldots, K$, define

$$\hat{\boldsymbol{\beta}}^{(m)}(\tau_k) = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{D} \sum_{j=1}^{n_i} \rho_{\tau_k}(y_{ij} - \boldsymbol{z}_{ij}' \hat{\boldsymbol{b}}_i^{(m)} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}) \tag{9}$$

subject to the restriction that $\boldsymbol{x}_{ij}' \hat{\boldsymbol{\beta}}^{(m)}(\tau_k) \geq \boldsymbol{x}_{ij}' \hat{\boldsymbol{\beta}}^{(m)}(\tau_{k-1})$ for $j = 1, \ldots, N_i$ and $i = 1, \ldots, D$. Then, for $k = \lfloor 0.5(K+1)\rfloor - 1, \ldots, 1$, define $\hat{\boldsymbol{\beta}}^{(m)}(\tau_k)$ as in (9) subject to the restriction that $-\boldsymbol{x}_{ij}' \hat{\boldsymbol{\beta}}^{(m)}(\tau_k) \geq -\boldsymbol{x}_{ij}' \hat{\boldsymbol{\beta}}^{(m)}(\tau_{k+1})$ for $j = 1, \ldots, N_i$ and $i = 1, \ldots, D$. We implement the constrained optimization method of Koenker and Ng (2005) using the method `fn` in the `R` function `rq`.

3. We use the method of Jang and Wang (2015) to estimate ρ_s and ξ_s for $s = \ell, u$. Specifically,

$$\begin{aligned}
\hat{\rho}_\ell^{(m)} &= 0.5(\tau_1 + \tau_2) \sum_{i=1}^{D} \sum_{j=1}^{n_i} \frac{\hat{q}_{ij}^{(m)}(\tau_2) - \hat{q}_{ij}^{(m)}(\tau_1)}{n(\tau_2 - \tau_1)}, \\
\hat{\rho}_u^{(m)} &= \big[1 - 0.5(\tau_K + \tau_{K-1})\big] \sum_{i=1}^{D} \sum_{j=1}^{n_i} \frac{\hat{q}_{ij}^{(m)}(\tau_K) - \hat{q}_{ij}^{(m)}(\tau_{K-1})}{n(\tau_K - \tau_{K-1})},
\end{aligned} \tag{10}$$

$\hat{q}_{ij}^{(m)}(\tau_k) = \boldsymbol{x}_{ij}' \hat{\boldsymbol{\beta}}^{(m)}(\tau_k) + \boldsymbol{z}_{ij}' \hat{\boldsymbol{b}}_i^{(m)}$, and $n = \sum_{i=1}^{D} n_i$. Holding $\hat{\rho}_\ell^{(m)}$ and $\hat{\rho}_u^{(m)}$ fixed, the estimator of ξ_s is the maximum likelihood estimator using only $\{y_{ij} < \hat{\ell}_{ij}^{(m)}\}$

for $s = \ell$ and $\{y_{ij} > \hat{u}_{ij}^{(m)}\}$ for $s = u$, where $\hat{\ell}_{ij}^{(m)} = 0.5(\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}^{(m)}(\tau_1) + \boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}^{(m)}(\tau_2))$ and $\hat{u}_{ij}^{(m)} = 0.5(\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}^{(m)}(\tau_K) + \boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}^{(m)}(\tau_{K-1}))$. Precisely,

$$\hat{\xi}_{\ell}^{(m)} = \operatorname*{argmax}_{\xi} \prod_{\{(ij):y_{ij} < \hat{\ell}_{ij}^{(m)}\}} g\bigl(-(y_{ij} - \hat{\ell}_{ij}^{(m)})\bigr) \mid \hat{\rho}_{\ell}^{(m)}, \xi) \tag{11}$$

and

$$\hat{\xi}_{u}^{(m)} = \operatorname*{argmax}_{\xi} \prod_{\{(ij):y_{ij} > \hat{u}_{ij}^{(m)}\}} g\bigl(y_{ij} - \hat{u}_{ij}^{(m)} \mid \hat{\rho}_{u}^{(m)}, \xi\bigr). \tag{12}$$

Let $\hat{\boldsymbol{\theta}} = ((\hat{\boldsymbol{\beta}}_K)', \operatorname{vech}(\hat{\boldsymbol{\Sigma}}_b)', \hat{\rho}_{\ell}, \hat{\xi}_{\ell}, \hat{\rho}_u, \hat{\xi}_u)'$ denote the estimator of $\boldsymbol{\theta}$ obtained in the final step of the iteration.

REMARK 1. An algorithm more similar to a full EM algorithm would replace the optimization in (8–9) with

$$\hat{\boldsymbol{\beta}}^{(m)}(\tau_k) = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{D} E\left[\sum_{j=1}^{n_i} \rho_{\tau_k}\bigl(y_{ij} - \boldsymbol{z}_{ij}'\boldsymbol{b}_i - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\bigr) \,\middle|\, \boldsymbol{y}_i, \hat{\boldsymbol{\theta}}^{(m-1)}\right]. \tag{13}$$

In simulations discussed in Berg and Lee (2019), we find that the increase in computational time to implement (13) is not justified by an important decrease in prediction MSE. In the interest of computational speed, we prefer the simplified EM algorithm outlined in steps 1–3 above.

REMARK 2. In the data analysis and the simulations of Sections 3 and 4, respectively, we use $\tau_k = k(K+1)^{-1}$ with $K = 99$. Alternatively, one can use the number of unique quantile levels as determined by Portnoy (1991) in the model with $\boldsymbol{b}_i$ as fixed effects. (Operationally, specify `tau = -1` in the `R` function `rq`.) We prefer $\tau_k = k(K+1)^{-1}$ because using evenly spaced quantile levels simplifies predictors of small area parameters, as we explain in Section 2.3 below. The choice $\tau_k = k(K+1)^{-1}$ also satisfies the condition of Feng, Chen and He (2015) that $\tau_k - \tau_{k-1} = O(K^{-1})$.

2.3. *Small area parameters, predictors and mean squared error estimators.* The primary objective is to use the LIGPD approximation (3) for the model (1) to predict functions of the distribution of $\{y_{ij} : j = 1, \ldots, N_i\}$. To predict small area parameters, we create an approximation for the estimated distribution of $\{y_{ij} : j = 1, \ldots, N_i\}$ for area i. We define an estimate of the predictor (5) for each element of the population by

$$\hat{q}_{ij}(\tau) = \boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}(\tau) + \boldsymbol{z}_{ij}'\hat{\boldsymbol{b}}_i, \tag{14}$$

where $\hat{\boldsymbol{b}}_i = E[\boldsymbol{b}_i \mid \boldsymbol{y}_i; \hat{\boldsymbol{\theta}}]$. We evaluate (14) at the grid defined by $\{\tau_1, \ldots, \tau_K\}$. The $\{\hat{q}_{ij}(\tau_k) : k = 1, \ldots, K;\ j = 1, \ldots, N_i\}$ is an approximation for the distribution of $\{y_{ij} : j = 1, \ldots, N_i\}$. We use this approximation for the distribution to define small area parameters.

Define the τth population quantile for area i by

$$q_i(\tau) = \inf\bigl\{t : F_{y_i}(t) \geq \tau\bigr\}, \tag{15}$$

where $F_{y_i}(t) = \int_{\Omega_{(\boldsymbol{x}_i, z_i)}} P(y \leq t \mid \boldsymbol{x}, z, \boldsymbol{b}_i)\, dF_{(\boldsymbol{x}_i, z_i)}(\boldsymbol{x}, z)$, $F_{(\boldsymbol{x}_i, z_i)}(\boldsymbol{x}, z)$ is the cumulative distribution function of $(\boldsymbol{x}'_{ij}, z'_{ij})'$ for the population of $(\boldsymbol{x}'_{ij}, z'_{ij})'$ in area i, and $\boldsymbol{\Omega}_{(\boldsymbol{x}_i, z_i)}$ is the sample space for $(\boldsymbol{x}'_{ij}, z'_{ij})'$. The $(\boldsymbol{x}'_{ij}, z'_{ij})'$ are known for $j = 1, \ldots, N_i$. Thus, $F_{(\boldsymbol{x}_i, z_i)}(\boldsymbol{x}, z)$ is the step function with steps at $\{(\boldsymbol{x}'_{ij}, z'_{ij})' : j = 1, \ldots, N_i\}$. Then,

$$\begin{aligned} F_{y_i}(t) &= \frac{1}{N_i} \sum_{j=1}^{N_i} P(y_{ij} \leq t \mid \boldsymbol{x}_{ij}, z_{ij}, \boldsymbol{b}_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \int_0^1 I\bigl[q_{ij}(\tau) \leq t\bigr]\, d\tau \\ &\approx \frac{1}{N_i} \sum_{j=1}^{N_i} \sum_{k=1}^{K-1} I\bigl[q_{ij}(\tau_k) \leq t\bigr](\tau_{k+1} - \tau_k) \\ &\approx \frac{1}{N_i K} \sum_{j=1}^{N_i} \sum_{k=1}^{K} I\bigl[q_{ij}(\tau_k) \leq t\bigr], \end{aligned} \tag{16}$$

where the first approximation is a Riemann approximation to the integral, and the second approximation holds for $\tau_k = (K+1)^{-1}k$ with large K. The definition of the parameter in (15) and the approximation for the CDF in (16) motivate a predictor, $\hat{q}_i^{(0)}(\tau) = \min\{\hat{q}_{ij}(\tau_k) : \hat{F}_{y_i}(\hat{q}_{ij}(\tau_k)) \geq \tau;\ j = 1, \ldots, N_i; k = 1, \ldots, K\}$, where $\hat{F}_{y_i}(t) = (N_i K)^{-1} \sum_{j=1}^{N_i} \sum_{k=1}^{K} I[\hat{q}_{ij}(\tau_k) \leq t]$. Rather than use $\hat{q}_i^{(0)}(\tau)$, we use the default sample quantile in the `R` function `quantile` defined as

$$\hat{q}_i(\tau) = \bigl[1 - \bigl(\tau(N_i K) + m - h\bigr)\bigr]\hat{q}_{i(h)} + \bigl(\tau(N_i K) + m - h\bigr)\hat{q}_{i(h+1)}, \tag{17}$$

with $m = 1 - \tau$ and $h = \lfloor \tau N_i K + m \rfloor$ (Hyndman and Fan (1996)). The predictor $\hat{q}_i(\tau)$ is nearly identical to $\hat{q}_i^{(0)}(\tau)$ for large enough N_i and K.

While estimation of quantiles is our focus, one can use $\{\hat{q}_{ij}(\tau_k) : k = 1, \ldots, K;\ j = 1, \ldots, N_i\}$ to estimate other population parameters, such as the area mean defined by

$$\mu_i = N_i^{-1} \sum_{j=1}^{N_i} \int_0^1 q_{ij}(\tau)\, d\tau \approx \frac{1}{N_i K} \sum_{j=1}^{N_i} \sum_{k=1}^{K} q_{ij}(\tau_k), \tag{18}$$

where the justification for the approximation (18) is similar to (16). Define a predictor of μ_i by

$$\hat{\mu}_i = \frac{1}{N_i K} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \hat{q}_{ij}(\tau_k). \tag{19}$$

To define a bootstrap MSE estimator, repeat the following for $t = 1, \ldots, T$.

1. Generate $\boldsymbol{b}_i^{*(t)} \sim f_b(\boldsymbol{b}_i, \hat{\boldsymbol{\Sigma}}_b)$, and define $q_{ij}^{*(t)}(\tau_k) = \boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}(\tau_k) + \boldsymbol{z}_{ij}'\boldsymbol{b}_i^{*(t)}$ for $k = 1, \ldots, K$.

2. To generate a bootstrap population, generate $u_{ij}^{*(t)} \overset{\text{iid}}{\sim} \text{Unif}(0,1)$ for $i = 1, \ldots, D$, and $j = 1, \ldots, N_i$. Define $y_{ij}^{*(t)} = y_{ij}^*(\hat{\boldsymbol{\theta}}, \boldsymbol{b}_i^{*(t)}, u_{ij}^{*(t)})$ by

$$y_{ij}^{*(t)} = \begin{cases} \Bigg[q_{ij}^{*(t)}(\tau_{k_{ij}^{*(t)}}) \\ \quad + \big(u_{ij}^{*(t)} - \tau_{k_{ij}^{*(t)}}\big)\Bigg(\dfrac{q_{ij}^{*(t)}(\tau_{k_{ij}^{*(t)}+1}) - q_{ij}^{*(t)}(\tau_{k_{ij}^{*(t)}})}{\tau_{k_{ij}^{*(t)}+1} - \tau_{k_{ij}^{*(t)}}}\Bigg)\Bigg]\delta_{1ij}^{*(t)}, \\ \big[-G_\ell^{-1}(\tilde{u}_{\ell,ij}^{*(t)}; \hat{\rho}_\ell, \hat{\xi}_\ell) + 0.5\big(q_{ij}^{*(t)}(\tau_1) + q_{ij}^{*(t)}(\tau_2)\big)\big]\delta_{2ij}^{*(t)}, \\ \big[G_u^{-1}(\tilde{u}_{u,ij}^{*(t)}; \hat{\rho}_u, \hat{\xi}_u) + 0.5\big(q_{ij}^{*(t)}(\tau_{K-1}) + q_{ij}^{*(t)}(\tau_K)\big)\big]\delta_{3ij}^{*(t)}, \end{cases} \tag{20}$$

where $k_{ij}^{*(t)} = \max\{k : \tau_k \le u_{ij}^{*(t)}\}$, $\delta_{gij}^{*(t)} = I[u_{ij}^{*(t)} \in A_g]$, $A_1 = (0.5(\tau_1 + \tau_2), 0.5(\tau_{K-1} + \tau_K))$, $A_2 = (0, 0.5(\tau_1 + \tau_2)]$, $A_3 = [0.5(\tau_{K-1} + \tau_K), 1)$, $\tilde{u}_{\ell,ij}^{*(t)} = u_{ij}^{*(t)}/(0.5(\tau_1 + \tau_2))$, $\tilde{u}_{u,ij}^{*(t)} = u_{ij}^{*(t)} - 0.5(\tau_{K-1} + \tau_K)/(1 - 0.5(\tau_{K-1} + \tau_K))$, and $G_s(y; \rho_s, \xi_s) = \int_{-\infty}^{y} g(a; \rho_s, \xi_s)\, da$ for $s = \ell, u$. The procedure (20) simulates from the model (1) using linear interpolation for the step function with steps at τ_k for $k = 1, \ldots, K$ and using the inverse of the estimate of the generalized Pareto cumulative distribution function for extreme quantiles. Use $\{y_{ij}^{*(t)} : j = 1, \ldots, N_i\}$ to construct the bootstrap version of the population parameters. Specifically, $q_i^{*(t)}(\tau) = [1 - (\tau(N_i) + m - h)]q_{i(h)}^{*(t)} + (\tau(N_i) + m - h)q_{i(h+1)}^{*(t)}$, where $q_{i(h)}^{*(t)}$ is the hth order statistic of $\{y_{ij}^{*(t)} : j = 1, \ldots, N_i\}$, $m = 1 - \tau$, and $h = \lfloor \tau N_i + m \rfloor$.

3. Define a bootstrap sample by $\boldsymbol{y}_s^{*(t)} = \{y_{ij}^{*(t)} : (i,j) \in \mathcal{S}\}$, where $\mathcal{S}$ denotes the original sample. Use $\boldsymbol{y}_s^{*(t)}$ to obtain a parameter estimator $\hat{\boldsymbol{\theta}}^{*(t)}$ and predictors of the quantiles $\{\hat{q}_{ij}^{*(t)}(\tau_k) : i = 1, \ldots, D;\ j = 1, \ldots, N_i;\ k = 1, \ldots, K\}$. Define

$$\hat{q}_i^{*(t)}(\tau) = \big[1 - \big(\tau(N_i K) + m - h\big)\big]\hat{q}_{(h)}^{*(t)} + \big(\tau(N_i K) + m - h\big)\hat{q}_{(h+1)}^{*(t)}, \tag{21}$$

where $\hat{q}_{(h)}^{*(t)}$ is the hth order statistic of $\{\hat{q}_{ij}^{*(t)}(\tau_k) : k = 1, \ldots, K;\ j = 1, \ldots, N_i\}$ with h and m defined as for (17). Likewise, define $\hat{\mu}_i^{*(t)} = (N_i K)^{-1} \times$

$\sum_{j=1}^{N_i}\sum_{k=1}^{K}\hat{q}_{ij}^{*(t)}(\tau_k)$. We simplify the estimation procedure of Section 2.2 to obtain $\hat{q}_{ij}^{*(t)}(\tau_k)$. Rather than estimate the quantile regression coefficients sequentially to enforce the monotonicity constraint, as in (8)–(9), we simultaneously minimize Koenker's check function for all quantile levels and then sort the estimates of the quantiles to obtain a nondecreasing quantile function (Chernozhukov, Fernández-Val and Galichon (2009)) for element (i, j). We describe the sorting operation in Appendix A.1.

Define the bootstrap MSE estimator for $\hat{q}_i(\tau)$ and $\hat{\mu}_i$, respectively, by

$$\hat{\mathrm{MSE}}_i(\tau) = \frac{1}{T}\sum_{t=1}^{T}\big(\hat{q}_i^{*(t)}(\tau) - q_i^{*(t)}(\tau)\big)^2, \tag{22}$$

and $\hat{\mathrm{MSE}}_i(\mu) = T^{-1}\sum_{t=1}^{T}(\hat{\mu}_i^{*(t)} - \mu_i^{*(t)})^2$. We define a prediction interval with nominal coverage $(1-\alpha)100\%$ by

$$\big[L_i(\tau,\alpha), U_i(\tau,\alpha)\big] = \big[\hat{q}_i(\tau) + \Phi_{\alpha/2}^{-1}\sqrt{\hat{\mathrm{MSE}}_i(\tau)}, \hat{q}_i(\tau) - \Phi_{\alpha/2}^{-1}\sqrt{\hat{\mathrm{MSE}}_i(\tau)}\big], \tag{23}$$

where $\Phi_{\alpha/2}^{-1}$ is the $\alpha/2$ quantile of the standard normal distribution. We estimate the covariance matrix of $\hat{\boldsymbol{\beta}}(\tau)$ by

$$\hat{\boldsymbol{V}}\big(\hat{\boldsymbol{\beta}}(\tau)\big) = \frac{1}{T}\sum_{t=1}^{T}\big(\hat{\boldsymbol{\beta}}(\tau)^{*(t)} - \bar{\boldsymbol{\beta}}^{(\cdot)}\big)\big(\hat{\boldsymbol{\beta}}(\tau)^{*(t)} - \bar{\boldsymbol{\beta}}^{(\cdot)}\big)', \tag{24}$$

where $\bar{\boldsymbol{\beta}}^{(\cdot)} = T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{\beta}}(\tau)^{*(t)}$, and $\hat{\boldsymbol{\beta}}(\tau)^{*(t)}$ is the estimate of $\boldsymbol{\beta}(\tau)$ based on bootstrap sample $\boldsymbol{y}_s^{*(t)}$.

2.4. *Transformations*. Because an estimate based on the linear quantile regression model can be negative, one may choose to transform the observations if the support of the variable of interest is positive. For the CEAP application, we consider the class of transformations in Geraci and Jones (2015) and conclude that the log transformation is adequate. Let $\tilde{y}_{ij}$ be the original observation, and let $y_{ij} = \log(\tilde{y}_{ij} + \Delta)$, where Δ is specified. Assume y_{ij} satisfies the model (1). Let $\tilde{q}_{ij}(\tau)$ satisfy $P(\tilde{y}_{ij} \le \tilde{q}_{ij}(\tau) \mid \boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = \tau$. By monotonicity of the log transformation, the population quantile of interest is $\tilde{q}_i(\tau) = \exp(q_i(\tau))$, where $q_i(\tau)$ is the population quantile for the transformed y_{ij} define in (15). We define a predictor of $\tilde{q}_i(\tau)$ and corresponding confidence interval that exploit the invariance of the quantile to monotone transformations. We define the predictor of $\tilde{q}_i(\tau)$ by $\hat{\tilde{q}}_i(\tau) = \exp(\hat{q}_i(\tau)) - \Delta$, where $\hat{q}_i(\tau)$ is defined for y_{ij} as in (17). We define a prediction interval with nominal coverage $100(1-\alpha)\%$ by

$$\big[\exp\big(L_i(\tau,\alpha)\big) - \Delta, \exp\big(U_i(\tau,\alpha)\big) - \Delta\big], \tag{25}$$

TABLE 1
CEAP response variables used for analysis. All variables are annual averages for the field

Variable	Definition	Units
Runoff	Annual surface runoff based on daily rainfall	Inches
RUSLE2	Sheet and rill erosion for the cropped area of the field	Tons
Sediment	Edge of field sediment loss	Tons
CDiff	Annual change in soil organic carbon (January–December)	Tons

where $L_i(\tau, \alpha)$ and $U_i(\tau, \alpha)$ are defined in (23). We also define a bootstrap MSE estimator of $\hat{\tilde{q}}_i(\tau)$ by

$$\hat{\mathrm{MSE}}_i(\tau) = \frac{1}{T}\sum_{t=1}^{T}\left(\exp\left(\hat{q}_i^{*(t)}(\tau)\right) - \exp\left(q_i^{*(t)}(\tau)\right)\right)^2. \tag{26}$$

3. Application to CEAP data. Measures of erosion in the 2003–2006 CEAP survey result from processing administrative and survey data through a computer model called the Agricultural Policy Environmental Extender (APEX). The APEX model produces measures of total erosion as well as losses for specific nutrients, nitrogen and phosphorus. We consider APEX output variables (y) that are not nutrient specific, as summarized in Table 1.

3.1. *Transformations of response variables for CEAP modeling.* With the exception of CDiff, the response variables have nonnegative support. We use the log transformation, a member of the class defined in Geraci and Jones (2015). The log transformation has substantive support because equations defining erosion in the APEX model involve multiplication of input variables related to the model covariates defined in Section 3.3. We obtain data driven support for the log transformation using the procedure described in Appendix A.4. For Runoff, RUSLE2, and Sediment, we let $\tilde{y}_{ij}$ be the CEAP response variable for crop field j in county i, and we let $y_{ij} = \log(\tilde{y}_{ij} + 0.0005)$. We select $\Delta = 0.0005$ because 0.001 is the smallest possible positive value for an APEX model output. Three RUSLE2 values that equal zero are judged to be errors because of inconsistencies with NRI data and are therefore removed from the analysis. For CDiff, the support is the real line, and $\tilde{y}_{ij} = y_{ij}$.

3.2. *Population and samples for CEAP models.* The population of interest consists of area in cropland between 2003 and 2006 in Wisconsin. The CEAP sample is a subset of a larger survey called the National Resources Inventory (NRI) (Nusser and Goebel (1997)). For this analysis, we define the target population to be the collection of NRI locations that are classified as cropland for at least one year between 2003 and 2006. Because the covariates are known for the full population,

an extension to prediction for the full population is possible. The CEAP sample is approximately an 11% sample of the NRI. We exclude CEAP data collected in 2006 because of complications associated with the 2006 CEAP survey documented in Goebel (2009). The CEAP sample sizes for Wisconsin counties range from 0 to 27, the 25th percentile of the sample sizes is 5, the 75th percentile is 14, and the median county sample size is 9. We obtain predictors for 69 out of the 72 counties in Wisconsin, where we omit three counties that have no NRI points classified as cropland in the time frame of interest. Out of the 69 eligible counties, 61 have collected data for CEAP.

3.3. *Auxiliary variables for CEAP models*. Ideal auxiliary variables are inputs to the APEX model that are known for the full population of cropland in Wisconsin. The inputs to the APEX model relate to weather, soil properties, crop managements and conservation practices (Williams and Izaurralde (2006)). In CEAP, the data for weather and soil properties are from administrative sources that contain information for the full population of cropland of interest, while the information on crop managements and conservation practices is from survey data, unknown for the full population. Because our model assumes that the auxiliary variables are known for the full population, we consider covariates related to weather and soils. Table 2 describes the auxiliary variables.

We consider weather variables related to precipitation and temperature. The auxiliary variable related to temperature (TEMP) is the average of the maximum and minimum temperature for a county recorded for July 2004 by the Centers for Disease Control and Prevention in the United States (CDC (undated)). The rainfall factor, RFACT, is the sum of rainfall erosion index units and an additional factor to account for runoff due to snow melt and irrigation (USDA (2015)).

The auxiliary variables related to soils are obtained from the NRCS Soil Survey, a census of soils in the United States. We determined which soils variables to include as potential auxiliary variables through consultation with a soil scientist

TABLE 2
Auxiliary variables for CEAP models

Variable	Description
TEMP	Average of min. and max. county level 2004 July temperatures
RFACT	$\log(\text{RFACT} + 0.001)$, where RFACT is the USLE rainfall factor
KFACT	$\log(\text{KWFACT} + 0.001)$, where KWFACT is a soil erodibility index
SLOPE-R	Difference in elevation divided by distance between two locations
SLOPELENUSLE	Slope length
LSLOG	$\log(\text{SLOPE-R} + 0.001) + \log(\text{SLOPELENUSLE} + 0.001)$
HYDGRP	Values 1, 2, 3, 4 from low (1) to high (4) runoff potential
OM	Percent of organic matter in the soil
SAND	Percent sand in the soil

and research into the equations defining the response variables in the APEX model (Williams and Izaurralde (2006)). The variable KFACT quantifies the vulnerability of the soil to erosion. The variable LSLOG combines slope steepness (SLOPE-R) with slope length, the distance from the top of a hill to location where the gradient is judged flat. Hydrologic group categories are used to form an ordered categorical variable HYDGRP, which is treated as continuous in the model. The percent of organic matter (OM) and percent sand (SAND) in the soil are included specifically for CDiff.

The covariates included in the model for each response variable are presented in Table 3 in Section 3.5, where a blank space indicates that a covariate is not included in the model for that response variable. The loglinear form for positive response variables is motivated by the Universal Soil Loss Equation (Wischmeier and Smith (1978)), a model for sheet and rill erosion as a product of KFACT, RFACT, a slope length/steepness factor, and factors representing crop managements and conservation practices. We exclude RFACT from the model for RUSLE2 because $|\hat{\beta}(\tau_k)|[\mathrm{SE}(\hat{\beta}(\tau_k))]^{-1} < 1$ for $\tau_k = 0.25, 0.5, 0.75$, where $\hat{\beta}(\tau_k)$ and $\mathrm{SE}(\hat{\beta}(\tau_k))$ are the estimates and bootstrap standard errors, respectively, for the regression coefficient for log(RFACT) in a model for RUSLE2 that contains log(RFACT) in addition to the covariates in Table 3. Before fitting any models, we standardize the covariates to have mean 0 and standard deviation 1 for the full NRI. In the model, $\boldsymbol{x}_{ij} = (1, \boldsymbol{x}'_{1,ij})'$, where $\boldsymbol{x}_{1,ij}$ is the vector of standardized covariates for unit j in county i.

3.4. *Parametric models.* As an exploratory step, we consider a lognormal model for Runoff, RUSLE2, and Sediment, and a linear mixed effects model for CDiff. The model is defined by

$$y_{ij} = \gamma_0 + \boldsymbol{x}'_{1,ij}\boldsymbol{\gamma}_1 + \alpha_i + w_{ij}, \tag{27}$$

where $w_{ij} \sim \mathrm{N}(0, \sigma_w^2)$ and $\alpha_i \sim \mathrm{N}(0, \sigma_a^2)$. For RUSLE2 and Sediment, the model (27) is a lognormal model (Berg and Chandra (2014)) because $y_{ij} = \log(\tilde{y}_{ij} + 0.0005)$. For CDiff, no transformation is used, and the model (27) is a linear mixed effects model (Battese, Harter and Fuller (1988)). The EBP of Molina and Rao (2010) provides a mechanism for small area prediction based on (27). To diagnose the fit of the model (27), we define a conditional residual by

$$r_{ij,lm} = \frac{y_{ij} - (\hat{\gamma}_0 + \boldsymbol{x}'_{1,ij}\hat{\boldsymbol{\gamma}}_1 + \hat{\alpha}_i)}{\hat{\sigma}_e}, \tag{28}$$

where $\hat{\alpha}_i$ is an EBLUP of α_i for the linear model with REML estimators of regression coefficients and variances. Ignoring parameter estimation, the residuals $r_{ij,lm}$ would have a standard normal distribution if the linear mixed effects model holds.

The normal probability plots of the residuals $r_{ij,lm}$ in Figure 2 show heavier left tails than a normal distribution for the logarithms of RUSLE2 and Sediment and

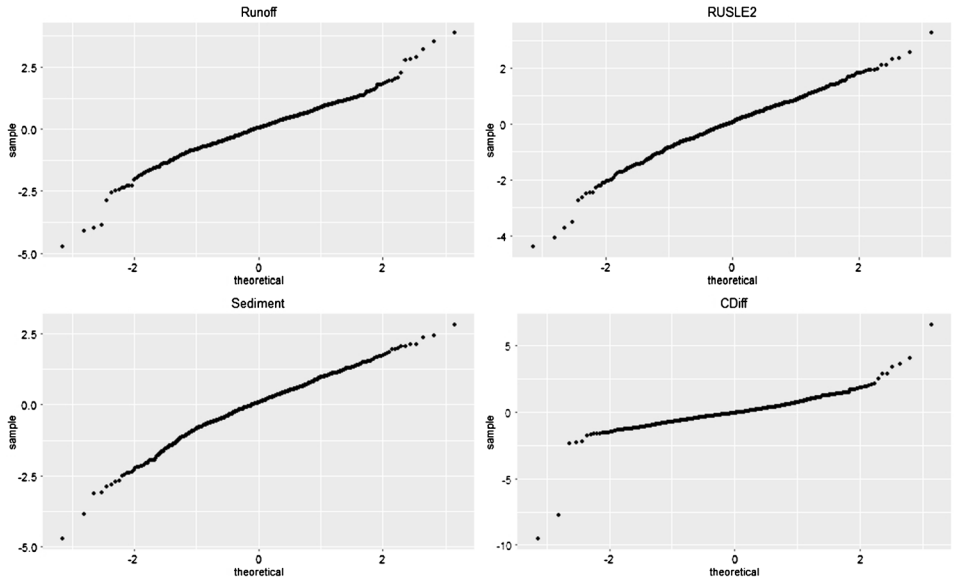

FIG. 2. *Normal quantile-quantile plots of residuals. Lognormal models are fit for Runoff, RUSLE2 and Sediment. The linear mixed effects model with normally distributed errors is used for CDiff.*

show both heavy lower and upper tails for CDiff and the logarithm of Runoff. The p-values of Shapiro–Wilk tests for normality of the residuals are less than 10^{-6}. Berg and Chandra (2014) show that the lognormal model provides an adequate fit to the RUSLE2 data for Iowa, a state that is relatively homogeneous with respect to agricultural production. While the lognormal is adequate for certain variables in homogeneous regions, Figure 2 and the corresponding Shapiro–Wilk p-values indicate that the lognormal model is not flexible enough to describe the distributions for the full range of variables and geographic domains of interest in CEAP. An analysis of a generalized linear mixed model based on a gamma distribution, described in Berg and Lee (2019), leads to a similar conclusion. While the gamma model appears adequate for RUSLE2 (Shapiro–Wilk p-value 0.1), the gamma model is inconsistent with the data for Runoff and Sediment (Shapiro–Wilk p-values < 0.01). These exploratory analyses illustrate the difficulty in obtaining an adequate parametric form for the distributions of all CEAP response variables of interest.

3.5. *Quantile regression models for CEAP data.* In an effort to obtain a unified approach that will adequately describe the distributions of multiple CEAP variables, we apply the LIGPD of Section 2. We define the model by $P(y_{ij} \le q_{ij}(\tau) \mid b_i, \boldsymbol{x}_{ij}) = \tau$, where $q_{ij}(\tau) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}(\tau) + b_i$, $b_i \sim \mathrm{N}(0, \sigma_b^2)$, and the covariates for each response variable are presented in Table 3. We use $K = 99$, terminate the iterative estimation procedure with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(2)}$, and use $T = 100$ bootstrap samples.

TABLE 3

*Estimates of $\boldsymbol{\beta}(\tau)$ and bootstrap standard errors (24) for $\tau = 0.25, 0.50$ and 0.75 for Runoff, RUSLE2, Sediment and CDiff. *SLOPE is LSLOG for Runoff, RUSLE2 and Sediment, and SLOPE is SLOPE-R for CDiff. Covariates defined in Table 2*

		Est. (SE)			
x_{ij}	τ	Runoff	RUSLE2	Sediment	CDiff
Intercept	0.25	1.323 (0.009)	−1.892 (0.048)	−1.467 (0.070)	−0.038 (0.006)
Intercept	0.50	1.401 (0.007)	−1.352 (0.043)	−0.712 (0.053)	0.010 (0.004)
Intercept	0.75	1.481 (0.007)	−0.808 (0.045)	0.014 (0.065)	0.058 (0.005)
TEMP	0.25	0.008 (0.016)	−0.034 (0.048)	−0.277 (0.100)	−0.013 (0.004)
TEMP	0.50	−0.001 (0.012)	−0.090 (0.048)	−0.218 (0.087)	−0.014 (0.004)
TEMP	0.75	−0.015 (0.015)	−0.042 (0.044)	−0.182 (0.124)	−0.013 (0.005)
HYDGRP	0.25	0.126 (0.010)	0.254 (0.066)	0.312 (0.067)	0.003 (0.005)
HYDGRP	0.50	0.132 (0.007)	0.242 (0.046)	0.274 (0.064)	0.003 (0.004)
HYDGRP	0.75	0.136 (0.007)	0.157 (0.049)	0.234 (0.061)	0.008 (0.005)
RFACT	0.25	0.024 (0.015)		0.292 (0.120)	
RFACT	0.50	0.037 (0.011)		0.270 (0.094)	
RFACT	0.75	0.046 (0.013)		0.341 (0.104)	
SLOPE*	0.25		0.615 (0.057)	0.524 (0.071)	0.022 (0.005)
SLOPE	0.50		0.636 (0.055)	0.513 (0.067)	0.029 (0.005)
SLOPE	0.75		0.489 (0.044)	0.626 (0.074)	0.037 (0.006)
KWFACT	0.25		0.159 (0.083)	0.336 (0.080)	
KWFACT	0.50		0.100 (0.046)	0.286 (0.067)	
KWFACT	0.75		0.131 (0.047)	0.276 (0.054)	
OM	0.25				0.099 (0.034)
OM	0.50				0.132 (0.017)
OM	0.75				0.147 (0.015)
SAND	0.25				0.028 (0.004)
SAND	0.50				0.026 (0.004)
SAND	0.75				0.020 (0.004)

Recall that y_{ij} is the logarithm of RUSLE2, Runoff and Sediment, as explained in Section 3.1.

Table 3 contains estimates of the quantile regression coefficients and corresponding standard errors for $\tau = 0.25, 0.5$ and 0.75, where the standard errors are the square roots of the diagonal elements of (24). The positive signs of the estimated coefficients for rainfall and soils variables in the models for Runoff, RUSLE2 and Sediment are consistent with the definitions of these APEX output variables. The positive signs of the estimated coefficients for OM and SAND in the model for CDiff are also consistent with the theory that soils with more organic matter have more potential for carbon loss and that carbon stores in sandier soils are more susceptible to the effects of agricultural production.

To check for spatial structure in the model random effects, we apply Moran's I statistic (using the `R` function `Moran.I`) to the $\{\hat{b}_i : i = 1, \ldots, D\}$, where the

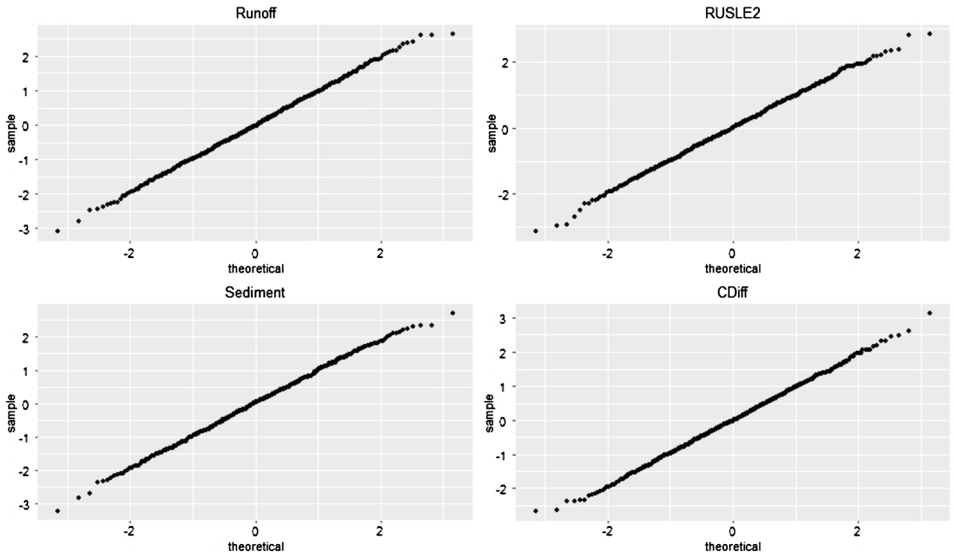

FIG. 3. *Normal probability plots of residuals r_{ij} based on the LIGPD quantile regression model.*

weights for Moran's I statistic are based on an adjacency matrix in which two counties are considered neighbors if they share a border. After incorporating the average temperature (TEMP) as a covariate in the model, Moran I p-values for for CDiff and Sediment are 0.08 and 0.85, respectively. In analyses not presented here, we find that incorporating agricultural statistics districts (groups of counties) in the model, removes spatial dependence for Runoff and RUSLE2. Here, we present the more parsimonious models for Runoff and RUSLE2 for consistency across the four variables.

To assess the plausibility of the LIGPD model assumptions, we define a conditional residual by $r_{ij} = \Phi^{-1}(\hat{F}_{yij}(y_{ij}))$, where Φ^{-1} is the quantile function of a standard normal distribution, and $\hat{F}_{yij}$, defined in Appendix A.3, estimates the approximate cumulative distribution function corresponding to the LIGPD. Ignoring parameter estimation, the residuals r_{ij} would have a standard normal distribution. The normal probability plots of r_{ij} in Figure 3 and Shapiro–Wilk p-values (0.98 0.95 0.86 and 0.99 for Runoff, RUSLE2, Sediment and CDiff, respectively) support the LIGPD.

3.6. *Small area predictors for CEAP data.* In this section, we demonstrate how the LIGPD enables us to attain the benefits of estimating quantiles at the county level using the CEAP data discussed below Figure 1 of the Introduction. We first consider all counties in Wisconsin and then focus on a single county of historical importance. We conclude this section with a comparison of the efficiency of the estimates based on the LIGPD to the efficiency of the direct estimates for Wisconsin counties.

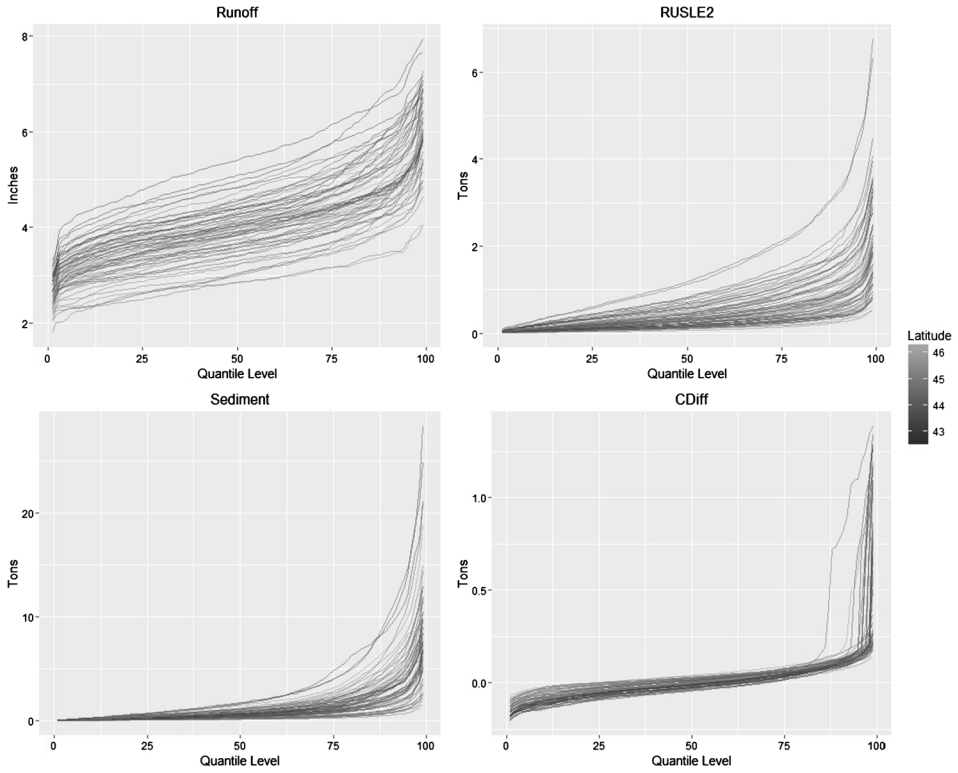

FIG. 4. *Estimates of quantiles for* 61 *counties in Wisconsin with at least one sampled NRI point.*

Figure 4 shows the estimates of the quantiles for all counties in Wisconsin that contain at least one sampled NRI point. The extent of the variation across counties largely reflects the size of the estimate of σ_b^2 (0.016, 0.426, 0.379 and 0.00018 for Runoff, RUSLE2, Sediment and CDiff, respectively). The darkness of the line relates to the latitude of a centroid of the county. For CDiff, northern counties tend to have higher values of CDiff than southern counties, which is consistent with the negative slope of the estimated coefficient for TEMP in the model for CDiff.

As noted in Goebel and Kellogg (2002), areas of extremely high erosion are of substantive interest. For CDiff, the jagged line with the highest estimated quantiles for $\tau_k > 0.9$ corresponds to Marquette County. The sample size for Marquette County is relatively small (only 5), and Marquette County has the fourth largest variance of the sample mean for CDiff in the state. Because of the small sample size and high variance for CDiff, the small area predictors of the upper quantiles for Marquette County are driven largely by the auxiliary variables, namely the percent organic matter (OM). The mean of OM for the NRI is larger for Marquette County than for any of the other counties in Wisconsin. Simultaneously, the estimate of the regression coefficient associated with OM tends to increase with the quantile level, as $\hat{\beta}(\tau_k)$ for OM is 0.058, 0.0716, 0.0962, 0.1195 and 0.1491,

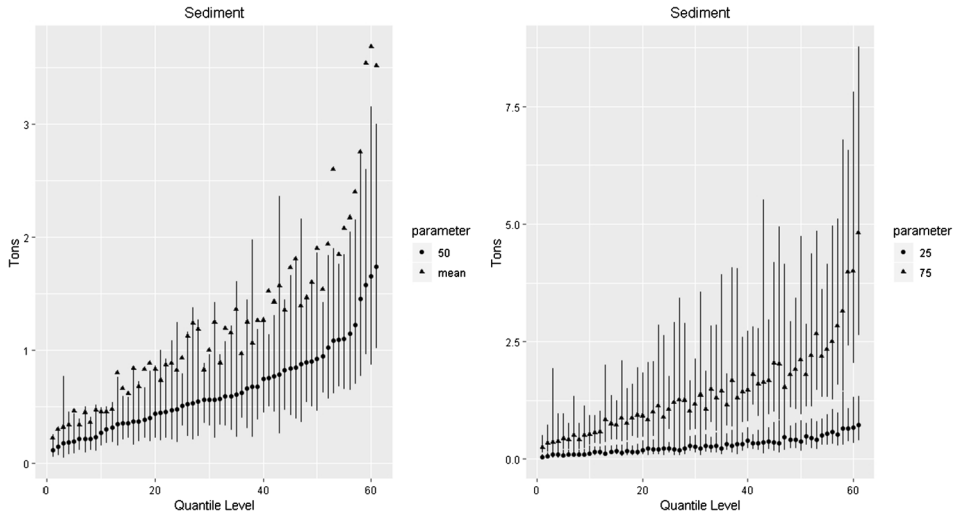

FIG. 5. *Left panel*: *County level predictors and confidence intervals for median sediment loss*, *with corresponding estimates of the mean*. *Right panel*: *County level predictors and confidence intervals for the* 25 *percentile and the* 75 *percentile of sediment loss*.

respectively, for $\tau_k = 0.75, 0.80, 0.85, 0.90$ and 0.95. The two counties with the highest RUSLE2 erosion estimates (Sheboygan County and Manitowac County) have relatively large sample sizes (16 for Manitowac County and 15 for Sheboygan County) and have the second and third largest sample medians for RUSLE2. (The county with the largest median has a sample size of one.) While the covariates are responsible for the extreme predictors of quantiles for CDiff, the observed RUSLE2 explains the extreme predictors of quantiles for RUSLE2. This contrast illustrates the value of small area estimation in using both auxiliary information and collected response variables to gain a more complete picture of the concept under study.

Figure 4 illuminates the skewed nature of the estimated distributions, which suggests that the median might be preferable to the mean as a measure of central tendency. Figure 5 contains predictions and confidence intervals for quartiles and the median as well as estimates of the means for sediment erosion. To conserve space, analogous plots for Runoff, RUSLE2 and CDiff are deferred to Berg and Lee (2019). As shown in the left panel of Figure 5, the estimated means exceed the upper endpoints of the corresponding 95% prediction intervals for the medians. The estimates of the quartiles in the right panel of Figure 5 provide information on the variation of erosion in Wisconsin counties. The confidence intervals for the 25 percentiles and the 75 percentiles are typically disjoint. The confidence interval width and the estimated interquartile range increases with the estimated median erosion, a reflection of the mean-variance relationship in the original data. The confidence interval widths for the 75 percentiles are undesirably wide because the

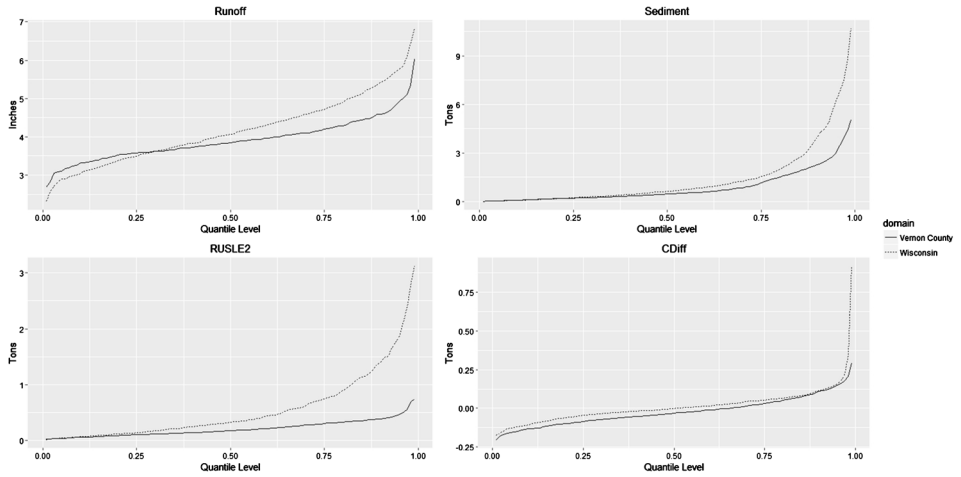

FIG. 6. *Dashed line*: *predicted quantiles for Vernon County based on the LIGPD*. *Solid line*: *state-level direct estimates of quantiles*.

derivative of the curve defined by the quantile estimates for sediment in Figure 4 is close to zero at the 75 percentile and because the sample sizes for the counties are small.

The first watershed conservation project in the history of NRCS took place in Coon Creek watershed, located in Vernon County, in the early 1930s. We examine the predicted quantiles for CEAP variables for Vernon County, where the sample size is 6. Figure 6 contains model based predictors of quantiles for 99 quantile levels for Vernon County and corresponding direct estimates for Wisconsin. The direct estimator is $\hat{q}_{i,D}(\tau) = \min\{y : \hat{F}_{i,D}(y) \geq \tau\}$, where $\hat{F}_{i,D}(y) = n_i^{-1} \sum_{j=1}^{n_i} I[y_{ij} \leq y]$, and the Woodruff (1952) method provides a corresponding confidence interval (`method = "constant"` and `interval.type = "Wald"` in the `R svyquantile` function).

Table 4 contains nominal 95% confidence intervals for $q_i(\tau)$ for $\tau = 0.25, 0.5, 0.75$ for Vernon County calculated as in (23) for CDiff and as in (25) for Runoff, RUSLE2 and Sediment. For the 75 percentile of Runoff and for the 75 percentile and median of RUSLE2, the confidence intervals for Vernon County are disjoint from the state-level intervals, with Vernon County estimates below the corresponding state level esitmates. For Sediment and CDiff, the estimates for Vernon County are close to the state-level estimates.

We compare the average widths of 95% confidence intervals and average estimated root mean squared errors (RMSE) for the LIGPD predictors and the county level direct estimators in Table 5 for counties with at least two sampled units. The standard error for a direct estimator is calculated using the bootstrap implemented in the `boot` method in the `R` function `summary.rq`. For the LIGPD predictors, the confidence intervals and estimated RMSEs are defined in (25) and (26) for

TABLE 4

Estimates and limits of 95% *confidence intervals for* $q_i(\tau)$ $(\tau = 0.25, 0.5, 0.75)$ *for four response variables* (y). *The direct estimator is used for Wisconsin. The LIGPD is used for Vernon County*

		Vernon County			Wisconsin		
y	τ	Lower	Estimate	Upper	Lower	Estimate	Upper
Runoff	0.25	3.239	3.577	3.951	3.421	3.493	3.584
Runoff	0.50	3.489	3.848	4.244	3.985	4.061	4.183
Runoff	0.75	3.813	4.196	4.619	4.630	4.718	4.841
RUSLE2	0.25	0.060	0.099	0.164	0.118	0.135	0.157
RUSLE2	0.50	0.110	0.172	0.269	0.291	0.321	0.351
RUSLE2	0.75	0.184	0.301	0.495	0.672	0.743	0.819
Sediment	0.25	0.090	0.212	0.496	0.200	0.228	0.282
Sediment	0.5	0.201	0.465	1.080	0.557	0.636	0.723
Sediment	0.75	0.454	1.138	2.856	1.349	1.543	1.827
CDiff	0.25	−0.123	−0.085	−0.047	−0.056	−0.046	−0.038
CDiff	0.50	−0.067	−0.032	0.004	−0.009	−0.002	0.005
CDiff	0.75	−0.016	0.030	0.077	0.047	0.054	0.063

Runoff, RUSLE2 and Sediment, and the confidence intervals and RMSEs are defined in (23) and (22) for CDiff. On average, the RMSE and confidence interval widths are smaller for the LIGPD predictors than for the direct estimators. Although we focus on counties with collected data for CEAP, a further benefit of the LIGPD is the ability to obtain estimates for counties where the sample size is zero.

4. Simulations. We compare the LIGPD predictor to alternatives through simulation and evaluate the properties of the MSE estimator. One alternative is based on the asymmetric Laplace distribution (ALD), developed in Geraci and

TABLE 5

Average widths of nominal 95% *confidence intervals and average RMSE for LIGPD predictors and direct county level estimators*

		LIGPD Predictors			Direct Estimators		
Variable	Criterion	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
Runoff	Width	0.697	0.751	0.900	0.866	1.279	1.486
RUSLE2	Width	0.220	0.397	0.722	0.281	0.674	1.032
Sediment	Width	0.446	0.913	2.208	0.694	2.059	3.809
CDiff	Width	0.064	0.062	0.075	0.100	0.148	0.218
Runoff	RMSE	0.180	0.194	0.232	0.283	0.334	0.407
RUSLE2	RMSE	0.064	0.113	0.200	0.102	0.179	0.271
Sediment	RMSE	0.117	0.246	0.591	0.257	0.548	1.021
CDiff	RMSE	0.016	0.016	0.019	0.248	0.530	0.987

Bottai (2007, 2014) and used for small area estimation in Weidenhammer et al. (2016). The ALD predictor that we consider is similar to that of Weidenhammer et al. (2016) and is defined in Appendix A.5. A procedure based on a fully parametric model may have optimality properties under the assumptions of the specified parametric form. For example, Diallo and Rao (2018) consider a model in which both the area random effects and the unit level errors have skew-normal distributions. As a representative of a fully parametric approach, we consider the EBP of Molina and Rao (2010) for a linear mixed effects model with normally distributed random components and constant variances. Specifically, the model underlying the normal empirical Bayes predictor (NEB) is

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + \eta_{ij}, \tag{29}$$

where $(v_i, \eta_{ij})' \sim \mathrm{N}[\mathbf{0}, \mathrm{diag}(\sigma_v^2, \sigma_\eta^2)]$. To compute the NEB predictor, we generate for $r = 1, \ldots, 100$,

$$y_{ij}^{(r)} \sim \mathrm{N}\big(\hat\beta_0 + \hat\beta_1 x_{ij} + \hat v_i, \hat\gamma_i \hat\sigma_e^2 n_i^{-1} + \hat\sigma_\eta^2\big), \tag{30}$$

where $\hat v_i = \hat\gamma_i(\bar y_{n_i} - \bar{\boldsymbol{x}}_{n_i}'\hat{\boldsymbol{\beta}})$, $\bar{\boldsymbol{x}}_{n_i} = (1, n_i^{-1}\sum_{j=1}^{n_i} x_{ij})'$, $\hat{\boldsymbol{\beta}} = (\hat\beta_0, \hat\beta_1)'$, $\hat\gamma_i = \hat\sigma_v^2(\hat\sigma_v^2 + \hat\sigma_\eta^2 n_i^{-1})^{-1}$, and REML is used to estimate the model parameters. The distribution (30) is an estimate of the conditional distribution of y_{ij} given the observed data, evaluated at the REML estimates. We then define the NEB predictors of the small area parameters as in (17), with $\{y_{ij}^{(r)} : r = 1, \ldots, 100;\ j = 1, \ldots, N_i\}$ in place of $\{\hat q_{ij}(\tau_k) : k = 1, \ldots, K;\ j = 1, \ldots, N_i\}$. The third predictor is the sample quantile for an area (obtained with the default `type = 7` in the `R quantile` function). For the LIGPD procedure, we use $K = 99$ to partition (0,1) into 100 evenly spaced intervals, terminate the iterative estimation procedure with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(2)}$ and use $T = 100$ bootstrap samples.

4.1. *Comparison of distributions.* We first consider a simulation model,

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + e_{ij}, \tag{31}$$

where $\beta_0 = -1.5$, $\beta_1 = 0.5$, and we consider two distributions for each of e_{ij} and b_i defined as follows. We simulate b_i from normal and Laplace distributions with mean zero and variance equal to 0.5. To represent the left skew in the residuals based on the linear mixed effects model applied to the log of RUSLE2 and Sediment, we consider $e_{ij} \sim \mathrm{SN}(\xi, \omega, \alpha)$, where $\xi = 1.26$, $\omega = 1.61$ and $\alpha = -5$. The notation $X \sim \mathrm{SN}(\xi, \omega, \alpha)$ means that X has skew-normal density function defined by

$$f_X(x; \xi, \omega, \alpha) = \frac{2}{\omega\sqrt{2\pi}} \exp\big(-(x-\xi)^2/(2\omega^2)\big) \int_{-\infty}^{\alpha(x-\xi)/\omega} \frac{1}{\sqrt{2\pi}} \exp\big(-t^2/2\big)\, dt.$$

We also consider a skewed and heteroskedastic distribution for e_{ij}, where $e_{ij} = (1 + 0.1x_{ij})(e_{ij}^* - 2)/2$ and $e_{ij}^* \sim \chi^2_{(2)}$. In each Monte Carlo (MC) sample, $x_{ij} \sim$

N(0, 1). The simulation parameters are based on a linear mixed effects model with the logarithm of the CEAP variable RUSLE2 as the response, standardized natural log of the slope as the covariate and normally distributed county effects. To roughly represent the sample sizes for the CEAP data, we generate $D = 60$ areas with $(N_i, n_i) = (143, 5)$ for 20 areas, $(N_i, n_i) = (286, 10)$ for 20 areas and $(N_i, n_i) = (571, 20)$ for 20 areas. The MC sample size for each simulation is 200. The τ population quantile is defined as the τ sample quantile of $\{y_{ij} : j = 1, \ldots, N_i\}$ obtained using the default in the `R` function `quantile`.

Table 6 contains the average MC MSE and MC bias of the alternative predictors, where the average is across areas of the same sample size. The comparison of the MSEs of the direct estimator to the MSEs of the model based predictors demonstrates the improvement in efficiency due to the use of models and auxiliary information. The MC MSE of the LIGPD is less than or equal to the MC MSE of the other predictors. With the exception of the estimator of $q_i(0.9)$ under the $\chi^2_{(2)}$ error distribution, the squared MC bias of the LIGPD is less than 10% of the MC MSE. For a given distribution and parameter, the MC MSE of the LIGPD decreases as the area sample size increases.

Table 7 contains average MC relative biases of the bootstrap MSE estimators and empirical coverage of normal theory 95% prediction intervals. The MSE estimator $\hat{\text{MSE}}_i(\tau)$ is defined in (22), and the prediction interval is defined in (23). The bootstrap sample size is $T = 100$ for these simulations. The MC relative bias of the bootstrap MSE estimator for area i is defined as $\text{RB}_i = [\text{MSE}_{\text{MC}}\{\hat{q}_i(\tau)\}]^{-1}(E_{\text{MC}}[\hat{\text{MSE}}_i(\tau)] - \text{MSE}_{\text{MC}}\{\hat{q}_i(\tau)\})$, where $E_{\text{MC}}[\hat{\text{MSE}}_i(\tau)]$ is the MC mean of the MSE estimator (22), and $\text{MSE}_{\text{MC}}\{\hat{q}_i(\tau)\}$ is the MC MSE of a predictor $\hat{q}_i(\tau)$. In Table 7, the MC relative biases and empirical coverages are averages across areas of the same sample size. We conjecture that a substantial part of the bias of the bootstrap MSE estimator for the $\chi^2_{(2)}$ error distribution occurs because the estimation procedure used in step 3 of the bootstrap does not use constrained optimization to estimate the quantile regression coefficients. For the $\chi^2_{(2)}$ error distribution with normally distributed b_i, the average ratio of the MC MSE of predictors of quantiles based on constrained optimization to the MC MSE of the corresponding predictors based on the sorting algorithm is approximately 0.9. Regardless of the approximations, the empirical coverages of normal theory 95% confidence intervals are between 92% and 96%.

4.2. *Analysis of transformed data.* We consider a simulation to represent the transformation used for Runoff, RUSLE2 and Sediment. Let $\tilde{y}_{ij}$ denote the original observations, and let $y_{ij} = \log(\tilde{y}_{ij})$ satisfy the model (31) with normally distribution b_i and skew-normal e_{ij}. The skew-normal distribution for e_{ij} is used to represent the skewness in the residuals for RUSLE2 and Sediment from the lognormal model. For the NEB predictor, we use (17) with $\{\exp(y_{ij}^{(r)}) : r = 1, \ldots, 100;\ j = 1, \ldots, N_i\}$ in place of $\{\hat{q}_{ij}(\tau_k) : k = 1, \ldots, K;\ j = 1, \ldots, N_i\}$. We define an ALD

TABLE 6

Average MC MSE and MC bias, where the average is across areas with the same sample size, and the data are generated as in (29)

			Normal b_i						Laplace b_i					
			$n_i = 5$		$n_i = 10$		$n_i = 20$		$n_i = 5$		$n_i = 10$		$n_i = 20$	
e_{ij}	Method	τ	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
$\chi^2_{(2)}$	LIGPD	0.1	0.046	0.004	0.017	−0.004	0.008	−0.002	0.044	−0.005	0.017	−0.001	0.009	0.002
$\chi^2_{(2)}$	ALD	0.1	0.150	0.045	0.091	0.072	0.050	0.069	0.164	0.045	0.095	0.065	0.059	0.074
$\chi^2_{(2)}$	NEB	0.1	0.290	−0.379	0.204	−0.346	0.162	−0.341	0.283	−0.381	0.204	−0.351	0.155	−0.330
$\chi^2_{(2)}$	Dir.	0.1	0.187	0.250	0.091	0.137	0.042	0.069	0.192	0.252	0.093	0.139	0.043	0.073
$\chi^2_{(2)}$	LIGPD	0.25	0.045	0.005	0.017	−0.006	0.008	−0.005	0.044	−0.003	0.016	−0.001	0.009	0.000
$\chi^2_{(2)}$	ALD	0.25	0.163	0.126	0.107	0.147	0.064	0.140	0.177	0.126	0.111	0.143	0.071	0.149
$\chi^2_{(2)}$	NEB	0.25	0.142	−0.038	0.080	−0.017	0.043	−0.022	0.135	−0.041	0.077	−0.024	0.043	−0.013
$\chi^2_{(2)}$	Dir.	0.25	0.190	0.154	0.078	0.078	0.037	0.033	0.195	0.158	0.082	0.081	0.039	0.040
$\chi^2_{(2)}$	LIGPD	0.5	0.049	−0.007	0.019	−0.018	0.009	−0.019	0.046	−0.013	0.018	−0.014	0.010	−0.012
$\chi^2_{(2)}$	ALD	0.5	0.200	0.223	0.151	0.256	0.111	0.256	0.215	0.245	0.159	0.265	0.126	0.277
$\chi^2_{(2)}$	NEB	0.5	0.207	0.243	0.148	0.251	0.100	0.235	0.194	0.239	0.139	0.241	0.106	0.244
$\chi^2_{(2)}$	Dir.	0.5	0.261	0.061	0.116	0.049	0.058	0.013	0.255	0.063	0.118	0.037	0.059	0.019
$\chi^2_{(2)}$	LIGPD	0.75	0.063	−0.024	0.029	−0.037	0.016	−0.039	0.062	−0.032	0.029	−0.035	0.016	−0.034
$\chi^2_{(2)}$	ALD	0.75	0.257	0.298	0.220	0.354	0.194	0.380	0.302	0.375	0.260	0.408	0.236	0.428
$\chi^2_{(2)}$	NEB	0.75	0.257	0.307	0.184	0.302	0.127	0.277	0.244	0.300	0.174	0.290	0.133	0.284
$\chi^2_{(2)}$	Dir.	0.75	0.529	−0.123	0.275	−0.022	0.141	−0.032	0.502	−0.123	0.281	−0.041	0.145	−0.022
$\chi^2_{(2)}$	LIGPD	0.9	0.112	−0.034	0.058	−0.054	0.036	−0.067	0.115	−0.040	0.060	−0.057	0.036	−0.061
$\chi^2_{(2)}$	ALD	0.9	0.290	0.262	0.226	0.315	0.187	0.340	0.366	0.395	0.297	0.417	0.261	0.433

TABLE 6
(*Continued*)

			Normal b_i						Laplace b_i					
			$n_i = 5$		$n_i = 10$		$n_i = 20$		$n_i = 5$		$n_i = 10$		$n_i = 20$	
e_{ij}	Method	τ	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
$\chi^2_{(2)}$	NEB	0.9	0.207	0.078	0.119	0.060	0.064	0.018	0.201	0.073	0.115	0.043	0.067	0.026
$\chi^2_{(2)}$	Dir.	0.9	1.076	−0.435	0.668	−0.244	0.392	−0.171	1.051	−0.429	0.682	−0.275	0.408	−0.154
SN	LIGPD	0.1	0.145	0.033	0.091	0.062	0.055	0.058	0.149	0.048	0.088	0.061	0.054	0.067
SN	ALD	0.1	0.209	−0.140	0.134	−0.148	0.095	−0.181	0.231	−0.199	0.152	−0.215	0.108	−0.219
SN	NEB	0.1	0.167	−0.016	0.100	0.029	0.058	0.039	0.178	0.012	0.101	0.034	0.058	0.053
SN	Dir.	0.1	0.756	0.442	0.471	0.288	0.268	0.142	0.817	0.479	0.462	0.275	0.254	0.166
SN	LIGPD	0.25	0.124	0.026	0.076	0.043	0.044	0.040	0.128	0.030	0.073	0.039	0.042	0.039
SN	ALD	0.25	0.190	−0.175	0.132	−0.193	0.102	−0.224	0.218	−0.218	0.150	−0.238	0.111	−0.247
SN	NEB	0.25	0.165	−0.123	0.099	−0.097	0.058	−0.092	0.170	−0.104	0.098	−0.093	0.055	−0.085
SN	Dir.	0.25	0.504	0.175	0.258	0.090	0.143	0.037	0.521	0.191	0.256	0.089	0.140	0.042
SN	LIGPD	0.5	0.114	0.007	0.069	0.018	0.041	0.015	0.115	0.009	0.067	0.020	0.037	0.017
SN	ALD	0.5	0.160	−0.125	0.104	−0.132	0.070	−0.147	0.181	−0.138	0.112	−0.147	0.070	−0.154
SN	NEB	0.5	0.159	−0.133	0.100	−0.125	0.064	−0.128	0.163	−0.126	0.100	−0.123	0.061	−0.127
SN	Dir.	0.5	0.340	−0.046	0.166	−0.020	0.092	−0.015	0.344	−0.042	0.170	−0.017	0.088	−0.012
SN	LIGPD	0.75	0.113	−0.016	0.068	−0.006	0.039	−0.009	0.112	−0.014	0.066	−0.007	0.036	−0.011
SN	ALD	0.75	0.148	−0.056	0.089	−0.052	0.051	−0.058	0.170	−0.046	0.094	−0.053	0.050	−0.057
SN	NEB	0.75	0.139	−0.017	0.084	−0.022	0.047	−0.033	0.146	−0.010	0.084	−0.024	0.045	−0.037
SN	Dir.	0.75	0.326	−0.196	0.151	−0.095	0.075	−0.053	0.331	−0.203	0.144	−0.092	0.073	−0.053
SN	LIGPD	0.9	0.116	−0.027	0.069	−0.023	0.041	−0.027	0.116	−0.029	0.067	−0.025	0.038	−0.030
SN	ALD	0.9	0.151	−0.004	0.089	−0.008	0.050	−0.015	0.171	0.011	0.092	−0.001	0.048	−0.006
SN	NEB	0.9	0.171	0.173	0.107	0.153	0.066	0.134	0.179	0.174	0.106	0.148	0.061	0.128
SN	Dir.	0.9	0.363	−0.343	0.186	−0.194	0.092	−0.109	0.364	−0.354	0.188	−0.199	0.089	−0.103

TABLE 7
Relative bias (%) of bootstrap MSE estimator (22) and empirical coverage of normal theory 95% prediction intervals

			Relative Bias (%)			Coverage		
Parameter	e_{ij}	b_i	$n_i = 5$	$n_i = 10$	$n_i = 20$	$n_i = 5$	$n_i = 10$	$n_i = 20$
0.25	$\chi^2_{(2)}$	Normal	27.190	17.502	10.313	0.962	0.958	0.955
0.50	$\chi^2_{(2)}$	Normal	22.006	14.210	6.868	0.961	0.960	0.951
0.75	$\chi^2_{(2)}$	Normal	14.093	8.108	4.162	0.959	0.952	0.950
0.25	$\chi^2_{(2)}$	Laplace	23.226	11.436	16.026	0.962	0.953	0.955
0.50	$\chi^2_{(2)}$	Laplace	19.905	9.783	8.909	0.959	0.955	0.952
0.75	$\chi^2_{(2)}$	Laplace	12.464	3.440	6.216	0.957	0.953	0.951
0.25	SN	Normal	−3.140	−5.168	−14.502	0.944	0.946	0.926
0.50	SN	Normal	−0.884	−7.227	−15.615	0.943	0.940	0.928
0.75	SN	Normal	−0.938	−5.090	−13.824	0.942	0.939	0.928
0.25	SN	Laplace	−2.935	−5.943	−13.983	0.945	0.943	0.926
0.50	SN	Laplace	−2.789	−8.426	−15.790	0.942	0.942	0.927
0.75	SN	Laplace	−1.823	−7.425	−13.749	0.939	0.935	0.926

predictor for the transformed model in Appendix A.5. For the LIGPD, we define the MSE estimator and confidence interval as in (26) and (25), respectively. Table 8 summarizes the properties of the MSE estimator for the LIGPD and alternative predictors of the quantiles of $\tilde{y}_{ij}$ for an MC sample size of 400. The LIGPD has smaller MC MSE than the alternative predictors. The confidence interval coverages are omitted because they are identical to the coverages in Table 7 by construction.

5. Discussion. The LIGPD approximation for the mixed effects quantile regression model (1) with area random effects provides a viable approach to small area prediction. Because the model makes few assumptions about the distribution of the error terms, use of the LIGPD has potential to unify the analysis of multiple response variables with diverse distributional properties. In simulations designed to represent the CEAP data, predictors of small area quantiles based on the LIGPD have smaller MSEs than predictors of corresponding quantiles based on parametric models. The efficiency gain of the LIGPD relative to the NEB and ALD predictors is greatest when the error distribution is far from normal, the number of areas is large, and the area sample size is small. The bootstrap MSE estimator leads to confidence intervals with average coverage within 2–3% of the nominal level. In the application to the Conservation Effects Assessment Project, the LIGPD-based small area predictors have smaller estimated RMSEs than the direct estimators, on average. An analysis of residuals indicates that the LIGPD is appropriate for

a wider range of CEAP variables than the lognormal distribution or the gamma distribution.

The benefits of estimating quantiles of the distribution discussed in the Introduction are realized in the CEAP data analysis. Because the distributions of CEAP response variables are skewed and have outliers, the median is preferable to the mean as a measure of center. Estimates of quartiles and extreme quantile levels, which are important in practice (Goebel and Kellogg (2002)), reflect both collected survey data and auxiliary information.

This study suggests future work related to the application and the methodology. In this application, we treat the NRI survey as a population. Constructing predictors for the full population is an area for future work. We consider a computationally simple frequentist procedure; however, a Bayesian analysis is a possible alternative direction. Extensions of the LIGPD approach to incorporate multivariate response variables or spatio-temporal dependence structures are other areas for methodological development. Ongoing research involves refinements in the context of an informative sample design.

APPENDIX

A.1. Initial estimators. We define an initial estimator of $\boldsymbol{\beta}(0.5)$ and $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_D')'$ by

$$(\hat{\boldsymbol{\beta}}^{(0)}(0.5), \hat{\boldsymbol{b}}^{(0)}) = \underset{\boldsymbol{\beta},\boldsymbol{b}}{\operatorname{argmin}} \sum_{i=1}^{D} \sum_{j=1}^{n_i} \rho_{0.5}(y_{ij} - \boldsymbol{x}_{ij}'\boldsymbol{\beta} - \boldsymbol{z}_{ij}'\boldsymbol{b}_i), \tag{32}$$

where $-\sum_{i=1}^{D-1} \hat{\boldsymbol{b}}_i^{(0)} = \hat{\boldsymbol{b}}_D^{(0)}$ for identifiability because we assume $\boldsymbol{x}_{ij}$ contains an intercept. If $\boldsymbol{x}_{ij}$ contains any covariates (other than the intercept) that are in the

TABLE 8

Summary of simulation results with $\tilde{y}_{ij} = \exp(y_{ij})$, *where* y_{ij} *is generated as in* (31). $100 \times \hat{\text{MSE}}$ *is MC mean of MSE estimator* (26)

			MC MSE × 100				MC Bias × 100			
τ	n_i	$100 \times \hat{\text{MSE}}$	LIGPD	NEB	ALD	Dir.	LIGPD	NEB	ALD	Dir.
0.25	5	0.462	0.418	0.556	0.767	4.032	0.187	2.276	3.59	−6.849
0.25	10	0.255	0.246	0.349	0.552	1.37	−0.232	1.869	3.59	−3.446
0.25	20	0.15	0.182	0.203	0.406	0.679	−0.518	1.455	3.444	−1.804
0.50	5	2.017	1.887	2.652	2.989	6.819	1.023	5.365	6.09	−4.16
0.50	10	1.113	1.104	1.729	1.969	3.189	0.283	4.975	5.852	−2.186
0.50	20	0.64	0.806	1.101	1.335	1.593	−0.201	4.425	5.4	−1.158
0.75	5	7.299	7.008	8.249	9.958	15.09	3.356	3.779	8.38	4.261
0.75	10	4.065	4.091	5.268	6.193	8.663	2.127	3.785	7.23	1.379
0.75	20	2.353	3.019	3.063	3.721	4.353	1.346	3.216	5.79	0.815

column space of z_{ij}, then one option is to replace x_{ij} in (32) with $\tilde{x}_{ij}$, where $\tilde{x}_{ij}$ contains the intercept and the set of covariates that are not in the column space of z_{ij}. Let $\hat{V}_1(\hat{b}_1^{(0)}), \ldots, \hat{V}_{D-1}(\hat{b}_{D-1}^{(0)})$ be estimates of the variance of the asymptotic distribution of $(\hat{b}_1^{(0)}, \ldots, \hat{b}_{D-1}^{(0)})$. The asymptotic covariance matrix of the initial estimators is defined in Berg and Lee (2019) and estimated with the option `se = "ker"` in the `R` function `summary.rq`. To define an initial estimator of Σ_b, define the area-level Fay–Herriot model,

$$\hat{b}_i^{(0)} = b_i + a_i, \tag{33}$$

where a_i has a distribution with mean $\mathbf{0}$ and variance $\hat{V}_i\{\hat{b}_i^{(0)}\}$, and b_i has a distribution with mean $\mathbf{0}$ and variance Σ_b for $i = 1, \ldots, D-1$. For univariate b_i, the initial estimate of Σ_b, denoted by $\hat{\Sigma}_b^{(0)}$, is obtained by applying the estimation procedure of Wang, Fuller and Qu (2008) to the area level model (33). The preliminary estimate of $\beta(\tau_k)$ for $k = 1, \ldots, K$ is defined by

$$\hat{\beta}^{(0)}(\tau_k) = \operatorname*{argmin}_{\beta} \sum_{i=1}^{D} \sum_{j=1}^{n_i} \rho_{\tau_k}\big(y_{ij} - x_{ij}'\beta - z_{ij}'\hat{b}_i^{(0)}\big). \tag{34}$$

We sort $\{x_{ij}'\hat{\beta}^{(0)}(\tau_k) : k = 1, \ldots, K\}$ for every (i, j) to obtain a nondecreasing quantile function (Chernozhukov, Fernández-Val and Galichon (2009)). The estimate $\hat{q}_{ij}^{(0)}(\tau_k)$ is the k order statistic of $\{x_{ij}'\hat{\beta}^{(0)}(\tau_k) + z_{ij}'\hat{b}_i^{(0)} : k = 1, \ldots, K\}$. Given the initial estimates of the quantile function, we use the procedure in Step 3 of Section 2.2 to obtain estimates $\hat{\rho}_s^{(0)}$ and $\hat{\xi}_s^{(0)}$ for $s = \ell, u$.

A.2. Details of numerical integration procedure. Let b_i be univariate and $b_i \sim f_b(b_i; \sigma_b^2)$. For $m = 0, \ldots, M-1$, let $t_r = F_b^{-1}(r/(R+1) \mid \hat{\sigma}_b^{2(m)})$ for $r = 1, \ldots, R$, where $F_b(\cdot \mid \hat{\sigma}_b^{2(m)})$ is the estimate of the cumulative distribution function of b_i evaluated at the parameter estimate obtained in step m. Let $h_i(b)$ denote the function to integrate. Let ℓ and u denote the lower and upper limits of the integral, and let $r_\ell = \min\{r : t_r \geq \ell\}$ and $r_u = \max\{r : t_r \leq u\}$. The approximation for the integral that we use is

$$\int_\ell^u h_i(b)\,db \approx \sum_{r=r_\ell}^{r_u - 1} (t_{r+1} - t_r)\left(\frac{h_i(t_r) + h_i(t_{r+1})}{2}\right).$$

In this work, we take $R + 1 = 1000$.

A.3. Calculation of residuals for quantile regression model. For $y_{ij} < 0.5(\tau_1 + \tau_2)$, define

$$\hat{F}_{yij}(y_{ij}) = -G_\ell(-y_{ij} + \hat{\ell}_{ij})0.5(\tau_1 + \tau_2) + 0.5(\tau_1 + \tau_2).$$

For $y_{ij} > 0.5(\tau_K + \tau_{K-1})$, define

$$\hat{F}_{yij}(y_{ij}) = G_u(y_{ij} - \hat{u}_{ij})\big[1 - 0.5(\tau_K + \tau_{K-1})\big] + 0.5(\tau_K + \tau_{K-1}).$$

For $y_{ij} \in (0.5(\tau_1 + \tau_2), 0.5(\tau_K + \tau_{K-1}))$, define

$$\hat{F}_{yij}(y_{ij}) = \tau_{k_{ij}-1} + \big(y_{ij} - \hat{q}_{ij}(\tau_{k_{ij}-1})\big)\bigg(\frac{\tau_{k_{ij}} - \tau_{k_{ij}-1}}{\hat{q}_{ij}(\tau_{k_{ij}}) - \hat{q}_{ij}(\tau_{k_{ij}-1})}\bigg),$$

where $\tau_{k_{ij}} = \min\{\tau_k : \hat{q}_{ij}(\tau_k) \geq y_{ij}, k = 1, \ldots, K\}$.

A.4. Selection of the transformation. For nonnegative $\tilde{y}_{ij}$, we consider a subset of the class of transformations defined in Geraci and Jones (2015) by

$$h(y, \lambda) = \begin{cases} \dfrac{1}{2\lambda}\bigg((y + \Delta)^{\lambda} - \dfrac{1}{(y + \Delta)^{\lambda}}\bigg) & \text{if } \lambda \neq 0, \\ \log(y + \Delta) & \lambda = 0, \end{cases}$$

where Δ is specified. We use a procedure to estimate λ that differs from the procedure of Geraci and Jones (2015) because we require a single λ for all quantile levels. We define a preliminary estimator of λ by $\tilde{\lambda} = \operatorname{argmin}_{\lambda} \sum_{i=1}^{D} \sum_{j=1}^{n_i} \rho_{0.5}(h(\tilde{y}_{ij}, \lambda) - \boldsymbol{x}'_{ij}\tilde{\boldsymbol{\beta}}(0.5))$, where $\tilde{\boldsymbol{\beta}}(0.5)$ is the minimizer of $\sum_{i=1}^{D} \sum_{j=1}^{n_i} \rho_{0.5}(\tilde{y}_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta})$. We obtain an initial estimator of $\boldsymbol{\theta}$ using the procedure defined in Appendix A.1 with $h(\tilde{y}_{ij}, \tilde{\lambda}) = y_{ij}$ as the observations. We define $\hat{\lambda}$ such that $L_p(\hat{\lambda}) = \max\{L_p(k/10) : k = 0, \ldots, 9\}$, where $L_p(\lambda) = \sum_{i=1}^{D} \log(\int_{\mathbb{R}^{p_1}} \prod_{j=1}^{n_i} f_Y(h(\tilde{y}_{ij}, \lambda) \mid \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i, \hat{\boldsymbol{\theta}}^{(0)}) f_b(\boldsymbol{b}_i \mid \hat{\boldsymbol{\Sigma}}_b^{(0)})\, d\boldsymbol{b}_i)$, and $f_Y(\cdot \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{b}, \boldsymbol{\theta})$ is defined in (3). For the CEAP data, $\hat{\lambda} = 0$ for Runoff, RUSLE2 and Sediment with $\Delta = 0.0005$. Limited simulations using the log transformation indicate that the profile likelihood procedure is capable of correctly selecting $\lambda = 0$. Further study of the profile likelihood procedure for estimating the transformation parameter is a potential area for future investigation. Because the log transformation is also justified on the basis of the loglinear form of the Universal Soil Loss Equation, we treat $\lambda = 0$ as fixed for the analysis.

A.5. Asymmetric Laplace distribution. For specified $\tau \in (0, 1)$, a variable $z \sim \mathrm{ALD}(\mu_\tau, \sigma_\tau)$ if z has density function

$$f_Z(z \mid \mu_\tau, \sigma_\tau) = \sigma_\tau^{-1} \exp\bigg\{-\rho_\tau\bigg[\frac{z - \mu_\tau}{\sigma_\tau}\bigg]\bigg\}. \tag{35}$$

The value of μ_τ that maximizes the likelihood based on (35) minimizes Koenker's check function. Geraci and Bottai (2007, 2014) define a model by

$$\begin{aligned} y_{ij} \mid \alpha_i(\tau) &\sim \mathrm{ALD}\big(q_{ij}(\tau), \sigma^2(\tau)\big), \\ q_{ij}(\tau) &= \boldsymbol{x}'_{ij}\boldsymbol{\beta}(\tau) + \alpha_i(\tau) \end{aligned} \tag{36}$$

and $\alpha_i(\tau) \sim \mathrm{N}(0, \sigma_\alpha^2(\tau))$. Because we consider multiple quantile levels, we index the model (36) by τ. In the model (36), $\alpha_i(\tau_1) \perp \alpha_i(\tau_2)$ for $\tau_1 \neq \tau_2$. The `R` function `lqmm` (Geraci and Bottai, 2007, 2014) uses maximum likelihood to obtain estimators $\hat{\boldsymbol{\beta}}(\tau)$ and $\hat{\sigma}_\alpha^2(\tau)$ and uses a predictor of $\hat{\alpha}_i(\tau)$ with the form of a best estimated linear predictor. When using `lqmm`, we specify `nK = 30` quadrature points and use "normal" and "robust" types for normal and Laplace b_i, respectively. A predictor of $q_{ij}(\tau_k)$ is $\check{q}_{ij}(\tau_k) = \boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}(\tau_k) + \hat{\alpha}_i(\tau_k)$.

Weidenhammer et al. (2016) use the Geraci and Bottai (2007, 2014) model for small area prediction. In the simulations of Section 4, we define small area predictors based on $\{\check{q}_{ij}(\tau_k) : j = 1, \ldots, N_i; k = 1, \ldots, K\}$ as in (17) and (19) with $\tau_k = k(K+1)^{-1}$ with $K = 99$. For the transformed model, we replace $\check{q}_{ij}(\tau_k)$ with $\exp(\check{q}_{ij}(\tau_k)) - \Delta$. Weidenhammer et al. (2016) use a Monte Carlo procedure to define the small area predictors, and in simulations presented Berg and Lee (2019), the prediction MSE using simulation is essentially the same as the prediction MSE from (17).

SUPPLEMENTARY MATERIAL

Supplement to "Small area estimation for the conservation effects assessment project using a mixed effects quantile regression model" (DOI: 10.1214/19-AOAS1276SUPP; .pdf). We provide the link to the Github repository with code, the covariance matrix used for the initial estimators, a comparison to an iterative procedure similar to a full EM algorithm, a description of the mixed effects gamma model applied to the data, and versions of Figure 5 for Runoff, RUSLE2 and CDiff.

REFERENCES

BATTESE, G. E., HARTER, R. and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.

BERG, E. and CHANDRA, H. (2014). Small area prediction for a unit-level lognormal model. *Comput. Statist. Data Anal.* **78** 159–175. MR3212164

BERG, E. and LEE, D. (2019). Supplement to "Prediction of small area quantiles for the conservation effects assessment project using a mixed effects quantile regression model." DOI:10.1214/19-AOAS1276SUPP.

CDC. https://wonder.cdc.gov/nasa-nldas.html.

CHAMBERS, R. and TZAVIDIS, N. (2006). *M*-Quantile models for small area estimation. *Biometrika* **93** 255–268. MR2278081

CHEN, J. and LIU, Y. (2012). Small area estimation under density ratio model. In *JSM Proceedings* 5162–5173. Amer. Statist. Assoc., Alexandria, VA.

CHEN, J. and LIU, Y. (2017). Small area quantile estimation. Available at arXiv:1705.10063.

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and GALICHON, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* **96** 559–575. MR2538757

DIALLO, M. S. and RAO, J. N. K. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scand. J. Stat.* **45** 1092–1116. MR3884901

FENG, Y., CHEN, Y. and HE, X. (2015). Bayesian quantile regression with approximate likelihood. *Bernoulli* **21** 832–850. MR3338648

GERACI, M. and BOTTAI, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8** 140–154.

GERACI, M. and BOTTAI, M. (2014). Linear quantile mixed models. *Stat. Comput.* **24** 461–479. MR3192268

GERACI, M. and JONES, M. C. (2015). Improved transformation-based quantile regression. *Canad. J. Statist.* **43** 118–132. MR3324431

GOEBEL, J. J. (2009). Statistical methodology for the NRI-CEAP cropland survey. USDA/NRCS.

GOEBEL, J. J. and KELLOGG, R. L. (2002). Using survey data and modeling to assist the development of agri-environmental policy. In *Conference on Agricultural and Environmental Statistical Applications in Rome* 695–705. National Statistical Institute of Italy, Rome.

HYNDMAN, R. J. and FAN, Y. (1996). Sample quantiles in statistical packages. *Amer. Statist.* **50** 361–365.

JANG, W. and WANG, H. J. (2015). A semiparametric Bayesian approach for joint-quantile regression with clustered data. *Comput. Statist. Data Anal.* **84** 99–115. MR3292800

JIANG, J. and LAHIRI, P. (2006). Mixed model prediction and small area estimation. *TEST* **15** 1–96. MR2252522

KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

KOENKER, R. and NG, P. (2005). Inequality constrained quantile regression. *Sankhyā* **67** 418–440. MR2208897

MOLINA, I., NANDRAM, B. and RAO, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *Ann. Appl. Stat.* **8** 852–885. MR3262537

MOLINA, I. and RAO, J. N. K. (2010). Small area estimation of poverty indicators. *Canad. J. Statist.* **38** 369–385. MR2730115

NUSSER, S. M. and GOEBEL, J. J. (1997). The national resources inventory: A long-term multi-resource monitoring programme. *Environ. Ecol. Stat.* **4** 181–204.

PFEFFERMANN, D. (2013). New important developments in small area estimation. *Statist. Sci.* **28** 40–68. MR3075338

PORTNOY, S. (1991). Asymptotic behavior of the number of regression quantile breakpoints. *SIAM J. Sci. Statist. Comput.* **12** 867–883. MR1102413

RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. *Wiley Series in Survey Methodology*. Wiley, Hoboken, NJ. MR3380626

REICH, B. J., FUENTES, M. and DUNSON, D. B. (2011). Bayesian spatial quantile regression. *J. Amer. Statist. Assoc.* **106** 6–20. MR2816698

USDA (2015). Summary report: 2012 National Resources Inventory, Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology, Iowa State Univ., Ames, IA.

USDA/NRCS (2012). Assessment of the effects of conservation practices on cultivated cropland in the upper Mississippi river basin. Conservation Effects Assessment Project. U.S. Dept. Agriculture. Available at https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1042093.pdf.

WANG, J., FULLER, W. A. and QU, Y. (2008). Small area estimation under a restriction. *Surv. Methodol.* **34**, 29–36.

WEIDENHAMMER, B., SCHMID, T., SALVATI, N. and TZAVIDIS, N. (2016). A unit-level quantile nested error regression model for domain prediction with continuous and discrete outcomes. Economics Discussion Paper, School of Business and Economics, Freie Univ., Berlin.

WILLIAMS, J. R. and IZAURRALDE, R. C. (2006). The APEX model. In *Watershed Models* (V. P. Singh and D. K. Frevert, eds.) 437–482. CRC Press, Boca Raton, FL.

WISCHMEIER, W. H. and SMITH, D. D. (1978). Predicting rainfall erosion losses—a guide to conservation planning. U.S. Dept. Agriculture Handbook No. 537.

WOODRUFF, R. S. (1952). Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc.* **47** 635–646. MR0050845

DEPARTMENT OF STATISTICS
IOWA STATE UNIVERSITY
AMES, IOWA 50011
USA
E-MAIL: emilyb@iastate.edu