Random Forest Prediction Intervals for Spatially Dependent Data

Yoon Bae Jun*
Department of Statistics
Iowa State University
Ames, IA 50011
yjun@iastate.edu

Abstract

Random forest is a popular machine learning technique that is often used on spatial data for prediction. However, in practice, the spatial dependence of a response variable has typically been ignored in constructing prediction intervals with random forest algorithms, which may result in either low accuracy or low efficiency in such applications. We propose a generalized version of out-of-bag guided random forest prediction intervals, which is well-adapted for spatially dependent data. We use dependency-adjusted regression tree (DART) node-splitting and a novel non-parametric kernel out-of-bag estimator to estimate the underlying conditional prediction error distribution. Theoretical results on the asymptotic consistency of our approach are obtained. Empirical simulation studies and the analysis of global earthquake data indicate that our proposed prediction interval provides good coverage, and is generally more efficient than existing approaches when observations are spatially dependent.

1 Introduction

Random forest is one of the most popular machine learning techniques for predicting a quantitative response. In practice, it is often essential to assess the reliability of the prediction and quantify the prediction uncertainty [5, 4, 28]. One common way to provide information about the prediction uncertainty is to construct a prediction interval that contains an unobserved response value with a specified coverage rate. The literature on prediction intervals for random forests began with the development of quantile regression forests by Meinshausen *et al.* [21]. Zhang *et al.* [28] proposed to use the empirical quantiles of out-of-bag prediction errors to generate prediction intervals. Roy and Larocque [22] compared several variations of prediction interval methods for random forests. Recently, Lu and Hardin [18] introduced a weighted out-of-bag estimator for conditional prediction error distribution functions. More broadly, conformal inference [25, 26] offers a generic way of obtaining prediction intervals that can be applied to virtually any estimator of regression functions [15, 14, 16, 20]. The existing literature, however, usually ignores spatial dependence among observations. The current approaches considering mixed effect modelling or spatial dependency in random forests focus on point prediction [7, 13, 10, 12]. They do not provide rigorous theoretical justification or consider prediction intervals.

The determination of prediction intervals for spatially dependent data remains an unresolved challenge, even though random forests are widely used for prediction with geospatial data. The lack of attention on accounting for spatial correlation may have an adverse impact on both point prediction and uncertainty quantification. In spatial statistics, we usually consider a mixed model regression

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

framework $Y(s) = m(X(s)) + \epsilon(s)$, $\epsilon(s) = w(s) + e(s)$, $s \in \mathcal{D} \subset \mathbb{R}^d$, where s is the point location in a d-dimensional spatial domain $\mathcal{D}$, Y is a univariate stochastic process of interest, m represents a real-valued function of a p-dimensional spatially indexed covariate X, w is a zero-mean Gaussian process that allows for spatial dependence, and e is the (non-spatial) measurement error process [2]. In this regard, there is a well-established theory on kriging, providing an estimate of the conditional expectation of Y given observed data (Y_{obs}, X_{obs}) at an unobserved location s_0, i.e. $\mathbb{E}(Y(s_0)|Y_{obs}, X_{obs}) = m(X(s_0)) + h^T\Gamma^{-1}(Y_{obs} - m(X_{obs}))$, where $h = Cov(\epsilon(s_0), \epsilon)$, and $\Gamma = Cov(\epsilon)$. In this paper, we will show that random forests that incorporate a well-specified spatial covariance structure can provide not only accurate point prediction but also efficient prediction intervals, with performance guarantees we establish via theory. We propose a generalized out-of-bag weighted density estimation of a random forest prediction error distribution for spatially dependent data. We adopt the dependency-adjusted regression tree (DART) node-splitting criteria proposed by Saha *et al.* [23] and combine it with a novel non-parametric kernel estimator for irregularly spaced sampling sites [19].

The paper is organized in the following way. Section 2 reviews the spatial adjustment of a standard random forest algorithm, literature on random forest prediction interval estimation, and a novel extension of the out-of-bag random forest prediction interval. In Section 3 we use simulation to demonstrate that our proposed prediction interval is typically more efficient than the other existing approaches. Section 4 shows applications to global earthquake data. Finally, Section 5 summarizes our work and discusses possible future studies.

2 Methods

2.1 Random Forest Point Prediction for Spatial Data

We begin with the standard random forest [5, 4, 17] (Algorithm 1). Consider a dataset consisting of a response variable Y, covariates X from a domain $\mathcal{X}$, and spatial coordinates s from a domain $\mathcal{S}$. Random forest requires specification of the number of trees (n_{tree}), the number of possible directions for splitting at each node of each tree (m_{try}), and the number of observations in each node at or below which the node is not split ($nodesize$). We follow the algorithm by Liaw and Wiener [17] which uses the classification and regression tree (CART) criterion [5] during tree constructions.

Algorithm 1: Standard Random Forest point prediction

Input:
Data $\mathcal{D}_{1:n} = [Y_{1:n}, \mathbf{Z}_{1:n}]$, where $Y_{1:n} = (y_1, \cdots, y_n)^T$, $\mathbf{Z}_{1:n} = (\mathbf{z}_1, \cdots, \mathbf{z}_n)^T$, $\mathbf{z}_i = (\mathbf{x}_i^T, \mathbf{s}_i^T)^T; \mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T \in \mathcal{X}; \mathbf{s}_i = (s_{i1}, \cdots, s_{id})^T \in \mathcal{S}; i = 1, \cdots, n$
Decide n_{tree}, m_{try}, and $nodesize$
Randomness $\mathcal{R}^{(1)}$: Resampling probability distribution (with replacement) of size n from $\{1, \cdots, n\}$
Randomness $\mathcal{R}^{(2)}_{\mathcal{X}\times\mathcal{S}}$: Subsampling probability distribution of size m_{try} from $\{1, \cdots, p+d\}$
Prediction point : $\mathbf{z}_0 = (\mathbf{x}_0^T, \mathbf{s}_0^T)^T$
Procedure:
for $b = 1 : n_{tree}$ **do**
  Independently sample R_b using $\mathcal{R}^{(1)}$, where R_b assigns each original observation index to a resampled one in the b^{th} tree
  $\mathcal{D}_b^* \leftarrow \mathbf{P}_b\mathcal{D}_{1:n}$, where $\mathbf{P}_b = [\mathbf{e}_{R_b[1]}, \cdots, \mathbf{e}_{R_b[n]}]^T$ denotes the b^{th} resampling matrix, and $\mathbf{e}_l$ denotes a vector of length n with 1 in the l^{th} position and 0 in every other position.
  $\hat{m}_b(\mathbf{z}_0) \leftarrow$ random tree prediction using CART criterion under m_{try} variables sampled under $\mathcal{R}^{(2)}_{\mathcal{X}\times\mathcal{S}}$ for each split, until the number of observations in each cell is either at or below $nodesize$.
end
Output: $\hat{m}(\mathbf{z}_0; \mathcal{D}_{1:n}) = \frac{1}{n_{tree}} \sum_{b=1}^{n_{tree}} \hat{m}_b(\mathbf{z}_0)$

For a given tree, let B be a list of observation indexes in a generic node, and let $N(B)$ be the length of the list. Let $M_{try}(B)$ be the subset of size m_{try} drawn from $\{1, \cdots, p+d\}$ for node B using $\mathcal{R}^{(2)}_{\mathcal{X}\times\mathcal{S}}$. A cut in B is a pair (j, k), where j is some value (dimension) from $M_{try}(B)$ and k is the

position of the cut along the j-th coordinate. Let C_B be the set of all such possible cuts in B, and let x_{ij} be the i-th observation of the j-th covariate. For any $(j,k) \in C_B$, the CART-split criterion takes the form

$$v_n^{CART}(j,k) = \frac{1}{N(B)} \sum_{i \in B} \left(Y_i - \bar{Y}_B\right)^2 - \frac{1}{N(B)} \sum_{i \in B} \left(Y_i - \bar{Y}_{B^{(L)}} \mathbb{I}_{x_{ij} < k} - \bar{Y}_{B^{(R)}} \mathbb{I}_{x_{ij} \geq k}\right)^2$$

where $B^{(L)} = \{\mathbf{i} \in B : \mathbf{x}_{ij} < k\}$, $B^{(R)} = \{\mathbf{i} \in B : \mathbf{x}_{ij} \geq k\}$, and $\bar{Y}_B$ (resp. $\bar{Y}_{B^{(L)}}$, $\bar{Y}_{B^{(R)}}$) is the sample average of Y_i belonging to B (resp. $B^{(L)}$, $B^{(R)}$). For node B, the best cut (j_n^*, k_n^*) is selected by maximizing $v_n^{CART}(j,k)$ over C_B, that is,

$$(j_n^*, k_n^*) = \underset{(j,k) \in C_B}{\operatorname{argmax}} v_n^{CART}(j,k)$$

CART split criterion and subsequent assignment of the node representatives are both based only on members within the parent node. However, geo-referenced data units could be distant in the covariate domain while being close in the spatial domain. That is, members outside of a parent node could be strongly correlated. Thus, the local optimization of node-splitting is not enough for spatially dependent data. Global GLS-style quadratic loss and Dependency-Adjusted-Regression-Tree (DART) split criterion [23] can be used to address spatial dependence.

Consider a regression tree grown up to the list of leaf node indexes $\{B_1, B_2, \cdots, B_\Psi\}$. The node representatives $\hat{\beta}^{(0)} = (\hat{\beta}_1, \cdots, \hat{\beta}_\Psi)^T = (\bar{Y}_{B_1}, \cdots, \bar{Y}_{B_\Psi})^T$ are the corresponding means. After the split (j,k) of the node, say B_Ψ without loss of generality, suppose that the new set of nodes is $\{B_1, B_2, \cdots, B_{\Psi-1}, B_\Psi^{(L)}, B_\Psi^{(R)}\}$ and the corresponding representatives $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_{\Psi-1}, \hat{\beta}_\Psi^{(L)}, \hat{\beta}_\Psi^{(R)})^T = (\bar{Y}_{B_1}, \cdots, \bar{Y}_{B_{\Psi-1}}, \bar{Y}_{B_\Psi^{(L)}}, \bar{Y}_{B_\Psi^{(R)}})^T$ are to be updated. Let $\mathbf{\Phi}^{(0)} = [(\phi_{i\psi})]$, $\phi_{i\psi} = \mathcal{I}(i \in B_\psi)$, be the $n \times \Psi$ membership matrix before the split, and $\mathbf{\Phi}$ denote the $n \times (\Psi + 1)$ membership matrix after a split of B_Ψ using (j,k). When the training data observations are independent, and for $b = 1, \cdots, B$, the optimization can be written in the following way:

$$(j^*, k^*, \hat{\beta}) = \underset{j,k;\beta \in R^{\Psi+1}}{\operatorname{argmax}} \left[\frac{1}{n}\left[\|Y_b - \mathbf{\Phi}^{(0)}\hat{\beta}^{(0)}\|_2^2 - \|Y_b - \mathbf{\Phi}\beta\|_2^2\right]\right]$$

where $Y_b = \mathbf{P}_b Y = \left[Y_{R_b[1]}, \cdots, Y_{R_b[n]}\right]^T$. This seeks to a maximizer of an OLS-style CART split criterion:

$$v_n^{CART}(j,k) = \frac{1}{n}\left[\|Y_b - \mathbf{\Phi}^{(0)}\hat{\beta}_{OLS}(\mathbf{\Phi}^{(0)})\|_2^2 - \|Y_b - \mathbf{\Phi}\hat{\beta}_{OLS}(\mathbf{\Phi})\|_2^2\right]$$

where $\hat{\beta}_{OLS}(\mathbf{\Phi}) = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T Y_b$, for $b = 1, \cdots, B$.

Under spatial dependence, we can simply replace the squared error loss $\|Y - \mathbf{\Phi}\beta\|_2^2$ with a quadratic loss $(Y - \mathbf{\Phi}\beta)^T\mathbf{\Gamma}^{-1}(Y - \mathbf{\Phi}\beta)$, where $\mathbf{\Gamma}$ denotes the covariance matrix of the marginalized response. Saha *et al.*(2021)[23] proposed a GLS-style DART split criterion

$$\begin{aligned}\mathbf{v}_{n,\mathbf{Q}_b}^{DART}(j,k) = \frac{1}{n}\Bigg[&\left(Y_b - \mathbf{\Phi}^{(0)}\hat{\beta}_{GLS}(\mathbf{\Phi}^{(0)})\right)^T \mathbf{Q}_b \left(Y_b - \mathbf{\Phi}^{(0)}\hat{\beta}_{GLS}(\mathbf{\Phi}^{(0)})\right) \\ &- \left(Y_b - \mathbf{\Phi}\hat{\beta}_{GLS}(\mathbf{\Phi})\right)^T \mathbf{Q}_b \left(Y_b - \mathbf{\Phi}\hat{\beta}_{GLS}(\mathbf{\Phi})\right)\Bigg]\end{aligned}$$

where $\hat{\beta}_{GLS}(\mathbf{\Phi}) = \left(\mathbf{\Phi}^T\mathbf{Q}_b\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{Q}_b Y_b$, for $b = 1, \cdots, B$. Algorithm 2 is the GLS-style Random Forest regression [23]. We need to specify a working correlation matrix Γ before the tree construction. Then we should evaluate its Cholesky factor $\mathbf{\Gamma}^{-1/2}$, which will be used to construct working precision matrix $\mathbf{Q}$ during the tree construction.

Algorithm 2: The GLS style of Random Forest point prediction

Procedure:
Evaluate $\mathbf{\Gamma}^{-1/2}$
for $b = 1 : n_{tree}$ **do**
Independently sample R_b using $\mathcal{R}^{(1)}$
$\mathcal{D}_b^* \leftarrow \mathbf{P_b}\mathcal{D}_{1:n}$, where $\mathbf{P}_b = [\mathbf{e}_{R_b[1]}, \cdots, \mathbf{e}_{R_b[n]}]^T$ denotes the b^{th} resampling matrix, and $\mathbf{e}_l$ denotes a vector of length n with 1 in the l^{th} position and 0 in every other position.
$\mathbf{Q}_b = \mathbf{\Gamma}^{-T/2}\mathbf{P}_b^T\mathbf{P}_b\mathbf{\Gamma}^{-1/2}$
$\hat{m}(\mathbf{z}_0; \mathbf{Q}_b) \leftarrow$ Saha et al.(2021)'s RF-GLS algorithm using DART criterion with m_{try} variables sampled under $\mathcal{R}_{\mathcal{X}}^{(2)}$, until the number of observations in each cell is either at or below $nodesize$.
end
Output: $\hat{m}(\mathbf{z}_0; \mathcal{D}_{1:n}) = \frac{1}{n_{tree}} \sum_{b=1}^{n_{tree}} \hat{m}(\mathbf{z}_0; \mathbf{Q}_b)$

2.2 Random Forest Prediction Intervals

Let us denote $l_b(\mathbf{x}, \mathbf{s})$ as the terminal node corresponding to the subspace containing $\mathbf{x}$ with a location $\mathbf{s}$ in tree b. We define the cohabitatnt relationship between a geo-referenced training observation and an arbitrary geo-referenced predicted point if and only if they have the same terminal node. In other words, $\mathbf{z}_i = (\mathbf{x}_i^T, \mathbf{s}_i^T)^T$ is a cohabitant of $\mathbf{z} = (\mathbf{x}^T, \mathbf{s}^T)^T$ if $l_b(\mathbf{X}_i, \mathbf{s}_i) = l_b(\mathbf{x}, \mathbf{s})$. Note that $l_b(\mathbf{x}, \mathbf{s}) \equiv l_b(\mathbf{x})$ if a random forest tree construction does not depend on a location $\mathbf{s}$. In addition, we can define $n_{ib} := \sum_{j=1}^{n} \mathbf{1}(R_b[j] = i)$ as the number of i^{th} observation counts included in the b^{th} bootstrapped sample. Remark that $\{n_{ib} = 0\}$ is equivalent to the statement that the i^{th} observation is not included in the b^{th} bootstrapped sample, which means the i^{th} observation is classified into an "out-of-bag" sample in the b^{th} tree.

Meinshausen [21] suggested that a conditional quantile within the cohabitant group provides a consistent estimator to the conditional distribution of Y given $\mathbf{x}$. This consistency is guaranteed regardless of spatial dependence, but the problem is that this type of prediction interval is not sufficiently efficient in that it usually provides unnecessarily wide finite-sample prediction intervals.

Like other bootstrap aggregating methods, random forests uses subsampling with replacement to create in-bag sample and out-of-bag sample for each tree. Recall that Zhang *et al.* [28] proposed to use out-of-bag errors for constructing prediction intervals. This is based on the idea that the relationship between training observations and the test observations is analogous to in-bag observations and out-of-bag samples when the number of training observations and the number of trees is large. Zhang *et al.* [28] consider the out-of-bag prediction of the i^{th} training unit by employing a weighted average of in-bag responses within the subset of trees for which the i^{th} observation is out-of-bag. Following Zhang *et al.* [28], consider the conditional prediction error distribution estimator $\hat{F}_E(e|\mathbf{x}, \mathbf{s}) := \frac{1}{n} \sum_{i=1}^{n} 1(Y_i - \hat{\varphi}^{(i)}(\mathbf{x}_i, \mathbf{s}_i) \leq e)$, where

$$\hat{\varphi}^{(i)}(\mathbf{x}_i, \mathbf{s}_i) := \frac{1}{\sum_{b=1}^{B} \mathbf{1}(n_{ib} = 0)} \sum_{b:\mathbf{z}_i \notin \mathcal{D}_b^*} \sum_{j=1}^{n} w_{j,b}(\mathbf{x}_i, \mathbf{s}_i) Y_j$$

is the out-of-bag predictor of the i^{th} training unit, and the weights

$$w_{j,b}(\mathbf{x}, \mathbf{s}) := \frac{n_{jb}\mathbf{1}(\mathbf{z}_j \in l_b(\mathbf{x}, \mathbf{s}))}{\sum_{h=1}^{n} n_{hb}\mathbf{1}(\mathbf{z}_h \in l_b(\mathbf{x}, \mathbf{s}))}$$

are corresponding to cohabitation of $\mathbf{z}_j$ with $\mathbf{x}_i$. Lu and Hardin[18] further consider that the out-of-bag prediction errors of training observations which are more frequently out-of-bag cohabitants of a given test observation make better proxies. Followed by Lu and Hardin[18], we can define out-of-bag weight with a spatial argument as

$$v_{i,b}(\mathbf{x}, \mathbf{s}) := \frac{\mathbf{1}(n_{ib} = 0 \text{ and } \mathbf{z}_i \in l_b(\mathbf{x}, \mathbf{s}))}{\sum_{h=1}^{n} \sum_{b=1}^{B} \mathbf{1}(n_{hb} = 0 \text{ and } \mathbf{z}_h \in l_b(\mathbf{x}, \mathbf{s}))}$$

This results in the weighted out-of-bag predictor $\hat{F}_E(e|\mathbf{x},\mathbf{s}) := \sum_{i=1}^{n} v_i(\mathbf{z}) 1(Y_i - \hat{\varphi}^{(i)}(\mathbf{x}_i,\mathbf{s}) \leq e)$, where $v_i(\mathbf{z}) := \sum_{b=1}^{B} v_{i,b}(\mathbf{x},\mathbf{s})$. Our goal is to broaden the scope of the existing idea by taking into account of spatial correlation of a response variable.

On the other hand, conformal prediction can be considered as a noticeable solution. Split Conformal prediction approaches define a certain non-conformity measure and compute a plausibility measure for each candidate values [15, 14, 16]. Remark that Mao *et al.* [20] proposed a (smoothed) local spatial conformal prediction (LSCP) to make asymptotically valid conditional prediction interval for spatially dependent data. We will compare our proposed estimator with those approaches in Section 3.

2.3 Random Forest Prediction Interval for Spatial Data

We propose an out-of-bag generalized kernel-based random forest prediction interval (OOBGK) algorithm as follows:

Algorithm 3: Generalized out-of-bag-weighted Random Forest Prediction Interval

Procedure:
Evaluate $\mathbf{\Gamma}^{-1/2}$
for $b = 1 : n_{tree}$ **do**
  Independently sample R_b using $\mathcal{R}^{(1)}$
  $\mathcal{D}_b^* \leftarrow \mathbf{P_b}\mathcal{D}_{1:n}$, where $\mathbf{P}_b = [\mathbf{e}_{R_b[1]}, \cdots, \mathbf{e}_{R_b[n]}]^T$ denotes the b^{th} resampling matrix, and $\mathbf{e}_l$ denotes a vector of length n with 1 in the l^{th} position and 0 in every other position.
  $\mathbf{Q}_b = \mathbf{\Gamma}^{-T/2}\mathbf{P}_b^T\mathbf{P}_b\mathbf{\Gamma}^{-1/2}$
  $\hat{m}(\mathbf{z}_0;\mathbf{Q}_b) \leftarrow$ Saha et al.(2021)'s RF-GLS algorithm using DART criterion with m_{try} variables sampled under $\mathcal{R}^{(2)}_{\mathcal{X}\times\mathcal{S}}$, until the number of observations in each cell is either at or below $nodesize$.
end
$\hat{m}_{OOB}(\mathbf{z}_i;\mathbf{Q}_b) \leftarrow$ out-of-bag predictor of the i^{th} training unit, $i = 1, \cdots, n$
$v_{i,b}(\mathbf{z}_0) \leftarrow$ out-of-bag weight (Lu and Hardin 2021)
$\hat{m}(\mathbf{z}_0) = \frac{1}{n}\sum_{b=1}^{n_{tree}} \hat{m}(\mathbf{z}_0;\mathbf{Q}_b)$
$\hat{e}_i = Y_i - \sum_{b=1}^{n_{tree}} v_{ib}(\mathbf{z}_0)\hat{m}_{OOB}(\mathbf{z}_i;\mathbf{Q}_b)$, for $i = 1, \cdots, n$
$\hat{H}(\epsilon|\mathbf{z}_0,\mathcal{D}_{1:n}) := \sum_{i=1}^{n} v_i(\mathbf{z}_0)\int_{-\infty}^{\epsilon}\mathbb{K}_h(\hat{e}_i - e)de$, where $v_i(\mathbf{z}_0) = \sum_{b=1}^{n_{tree}} v_{ib}(\mathbf{z}_0)$
$\hat{Q}_\alpha(\mathbf{z}_0;\mathcal{D}_{1:n}) := \inf\{e : \hat{H}(\epsilon|\mathbf{z}_0,\mathcal{D}_{1:n}) \geq \alpha\}$
Return $\left(\hat{m}(\mathbf{z}_0;\mathcal{D}_{1:n}) + \hat{Q}_{\alpha/2}(\mathbf{z}_0;\mathcal{D}_{1:n}), \hat{m}(\mathbf{z}_0;\mathcal{D}_{1:n}) + \hat{Q}_{1-\alpha/2}(\mathbf{z}_0;\mathcal{D}_{1:n})\right)$

The following theorem confirms the consistency of the OOBGK estimator to the true conditional prediction error distribution, under conditions for the error process (Assumption 1,4,5), the working precision (Assumption 2,4), sampling sites (Assumption 6), kernel functions (Assumption 7), bandwidths (Assumption 8,9), the standard regularity conditions on random forest (Assumption 10,11,12,13), and the out-of-bag weight (Assumption 14) (See the Appendix A.1 for details)

Theorem 2.1 *Let* $\hat{F}_n(\epsilon|\mathbf{z}) = \sum_{i=1}^{n} v_i(\mathbf{z})\int_{-\infty}^{\epsilon} K_h(Y_i - \hat{m}^{(i)}_{\hat{\Gamma}}(\mathbf{z}_i) - e)de$. *Under assumptions 1-14,*

$$\sup_{\epsilon\in\mathbb{R}} |\hat{F}_n(\epsilon|\mathbf{z}) - F_E(\epsilon|\mathbf{z})| \xrightarrow{p} 0$$

as $n \to \infty$, *pointwise for every* $\mathbf{z} = (\mathbf{x}^T, \mathbf{s}^T)^T \in \mathcal{X}\times\mathcal{S} \subset \mathbb{R}^p \times \mathbb{R}^d$.

Following the suggestion of Saha *et al.* [23], we run the first pass of the standard random forest [5, 17] on the data to get a preliminary estimate of m, and use it to obtain the residuals and to estimate parameters of Γ by a maximum likelihood approach. We also use the nearest neighbor Gaussion Process (NNGP) [8] sparse Cholesky factor $\tilde{\Gamma}^{-1/2}$ to acquire computational benefit without loss of consistency of the proposed estimator. We follow suggestions of the previous literature to determine the number of tree n_{tree} and the smallest number of the final nodes $nodesize$. Any functions can be used, but we use $K_h(\xi) = h^{-1}K(\xi/h)$ with a normalized Gaussian kernel function K and a bandwidth h decreasing by n (i.e. $h = h_n \to 0$ as $n \to \infty$). We select a pair of the kernel bandwidth

h and the number of features investigated for each node split m_{try} that minimize the average interval score [11], defined as $S_\alpha(I_l, I_u, y_{test}) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left[(I_u - I_l) + \frac{2}{\alpha}(I_l - y_i) + \frac{2}{\alpha}(y_i - I_u)\right]$, where $[I_l, I_u]$ represents the $100(1-\alpha)\%$ prediction interval and $y_{test} = (y_1, \cdots, y_{n_{test}})$ are the test observations. We denote AIS90 of this metric using $\alpha = 0.1$.

3 Empirical Results

We consider the following models and the simulation design factors:

Equations :

$$y_i = m(x_i) + w(s_i) + e_i \quad \tau^2 \in \{2.4, 0.6\} \tag{1}$$

$$w : \text{Gaussian Process (exponential)} \quad \sigma^2 \in \{0.6, 2.4\} \quad \phi = 1. \tag{2}$$

$$\text{Mean functions : } m(x_i) = \begin{cases} x_{i1} + x_{i2} & \text{LINEAR} \\ 5\sin(\pi x_{i1}/3) & \text{SINUSOIDAL} \\ 5(\mathbb{I}(x_{i1} > 0) + x_{i2}) & \text{STEP} \\ 10\sin(\pi x_{i1} x_2) + 20(x_{i3} - 0.5)^2 + 10x_{i4} + 5x_{i5} & \text{FRIEDMAN} \end{cases}$$

$$\text{Distribution of errors : } e_i \sim \begin{cases} N(0, \tau^2) & \text{HOMO} \\ \tau t(3)/\sqrt{3} & \text{HEAVY} \\ \tau\left(\frac{1}{2} + \frac{|m(z)|}{2E|m(z)|}\right) & \text{HETERO} \end{cases}$$

Distribution of predictors: $\mathbf{X} = \{X_1, \cdots, X_{20}\} \sim Unif[-3.5, 3.5]^{20}$

The full-factorial design results in $2 \times 4 \times 3 = 24$ different simulation scenarios. For each of the scenarios, we construct $n_{sim} = 100$ repeated simulations, using $n_{train} = 200$ training observations and $n_{test} = 50$ test observations. We set the arbitrary tuning parameters as $n_{tree} = 0.25 n_{train}$, and $nodesize = 20$. For the bandwidth selection, we use a greedy search with a candidate set of $m_{try} \in \{12, 13, \cdots, 20\}$, and $h \in h_0 \cdot \{1.0, 1.5, \cdots, 4.5, 5.0\}$, where h_0 is the default smoothing bandwidth by Sheather *et al* [24]. We mainly consider the case where nominal error rate is equal to $\alpha = 0.10$ in this section, but also examine the different set of nominal levels for all different scenarios, respectively. (Appendix A.2).

In our simulations, the underlying mean function $E(Y|z, S) = m(X, S)$ is classified in either (1) linear; (2) sinusoidal; (3) step; and (4) Friedman. w is assumed to be the smooth spatial random effect, which is generated as an exponential Gaussian process with spatial variance σ^2 and spatial correlation decay $\phi = 1$; and ϵ is the i.i.d random noise with variance τ^2, which is also called the nugget in spatial literature. The choice of (σ^2, τ^2) are decided to satisfy $\sigma^2 + \tau^2 = 3.0$ in order to fix the total marginal variance of Y and control the spatial proportions to the total variation. The distribution of errors are assumed to be either (1) homoscedastic (HOMO); (2) heavy-tailed (HEAVY); and (3) heteroscadestic (HETERO). We do not fix the predictor values for training cases over simulations. We do not make test cases having same predictor values as training cases.

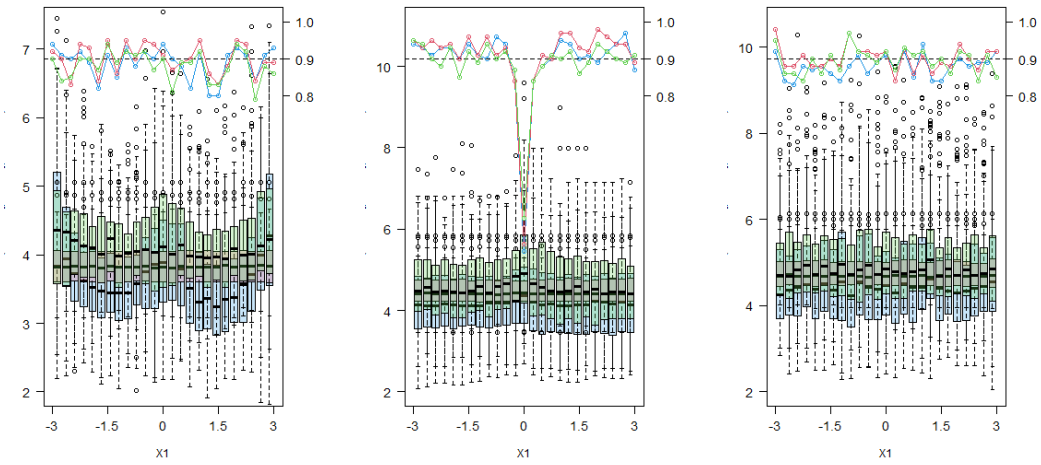
Table 1 shows generally good marginal performance of the proposed approach (OOBGK) under spatial dependence of data. First of all, OOBGK provides the lowest length of prediction interval for all the scenarios. Also, OOBGK usually shows relatively low average interval score. OOBGK only shows a little bit higher AIS90 than LSCP and OOBW shows, but the difference is negligible compared to the performances from the other approaches. Moreover, OOBGK achieves comparable level of empirical coverage rate with the other candidates. QRF provides high coverage probabilities but also high interval length. SC and OOB provides good coverage probabilities, but tends to give higher interval length than OOBGK.

Figure 1 implies that all the methods are affected considerably and often adversely when the target points are either the extreme or the boundary points of Y. For example, we observed "double-smile" shaped pattern of the boxplots across the test locations in the first panel, one peak at the middle in the second panel, wiggly shaped pattern in the third panel. These patterns are implied by the shape of underlying sinusoidal, step, and Friedman mean function, respectively. We can also observe that the estimated conditional coverage rate at the jump point is far below the nominal

Table 1: MARGINAL PERFORMANCES UNDER DOMINANT SPATIAL ERROR Average coverage rates of 90% prediction intervals, widths, and interval score across 100 simulations constructed by Quantile Regression Forests (QRF), split conformal prediction (SC), the unweighted out-of-bag method (OOB), Local Spatial Conformal Prediction (LSCP), the weighted out-of-bag method (OOBW), and the generalized out-of-bag kernel method (OOBGK). Bold quantities represents the case showing the lowest values in interval length or average interval score, respectively, among the candidates.

	LINEAR			STEP			FRIEDMAN		
	CPR	LEN	AIS90	CPR	LEN	AIS90	CPR	LEN	AIS90
QRF	0.96	5.73	6.14	0.97	6.85	7.21	0.93	6.32	7.19
SC	0.93	4.99	5.73	0.91	6.76	8.20	0.91	5.92	7.32
OOB	0.91	4.41	5.30	0.90	4.85	6.52	0.90	5.05	6.35
LSCP	0.91	4.11	4.89	0.90	4.49	6.11	0.90	4.77	5.95
OOBW	0.90	4.29	5.35	0.87	4.63	6.70	0.90	4.98	6.27
OOBGK	0.91	**3.86**	**4.73**	0.88	**4.24**	**6.04**	0.88	**4.64**	**5.93**
	SINUSOIDAL(HOMO)			SINUSOIDAL(HEAVY)			SINUSOIDAL(HETERO)		
	CPR	LEN	AIS90	CPR	LEN	AIS90	CPR	LEN	AIS90
QRF	0.93	5.04	5.77	0.91	5.36	6.93	0.91	7.22	8.82
SC	0.91	4.98	6.05	0.90	5.25	6.97	0.91	7.26	9.06
OOB	0.90	4.18	5.25	0.90	4.47	6.31	0.89	6.44	8.47
LSCP	0.90	3.86	4.79	0.90	4.24	6.08	0.89	6.31	**8.33**
OOBW	0.89	4.12	5.31	0.89	4.50	6.54	0.90	6.52	**8.33**
OOBGK	0.89	**3.74**	**4.78**	0.89	**4.07**	**6.03**	0.86	**5.85**	8.35

Figure 1: CONDITIONAL PERFORMANCES UNDER DOMINANT SPATIAL ERROR Boxplots of estimated conditional prediction interval lengths and lines of estimated conditional conditional coverage probabilities over the all test observations, constructed by our method (OOBGK; lightblue), the out-of-bag weighted method (OOBW; lightgreen), and the local spatial conformal prediction method (LSCP; lightred). Each panel represents sinusoidal (left), step (middle), and Friedman (right) mean function. Left-axis represents interval lengths, right-axis represents coverage rates, and bottom-axis represents the first column of covariates ($X1$). Black horizontal dash line represents the 90% nominal coverage rate.



rate not only for OOBGK, but also for LSCP and OOBW. Nevertheless, our proposed one shows relatively great performances in that it shows relatively low interval length without significant loss of coverage probabilities. OOBW shows similar performances to the OOBGK but gives generally higher prediction interval length. LSCP shows relatively uniform interval length over the test locations compared to OOB-guided methods, but often overestimates the prediction interval length on average for all the scenarios.

4 Real Data Application

We show an example of real data application using the publicly available data[2] provided by the National Earthquake Information Center (NEIC). Our goal here is to validate the magnitude of earthquake in year 2016 using the records of the magnitude, year, month, date, hour, latitude, longitude, depth, and various sources of every earthquake reported magnitude 5.5 or higher from year 1965 to 2015. For the data pre-processing step, we decide to use only the type of "Earthquake", so we removed 4 cases of "Explosion", 175 cases of "Nuclear Explosion, and 1 case of "Rock Burst". We only use the completely observed variables among the original dataset, and removed some invalid observations (e.g. missing dates, occurrence month is greater than 12). Before model fitting, the continuous variable (e.g. Depth) is standardized, and the categorical variables (e.g. Date, Time, and Sources of the earthquake) are transformed into dummy variables. After all, we reconstruct a completely observed dataset with $23,256$ observations and 214 predictor variables.

We set the nominal error rate 0.1. We define $1,000$ randomly chosen samples out of the observations reported before 2016 as the training set, and 469 samples of year 2016 as the test set. We set the number of trees $n_{tree} = 250$, the size of each final node $nodesize = 100$, the size of predictor set for node-splitting $m_{try} = 168$, the LSCP smoothing bandwidth $\eta = 25$, and the OOBGK kernel bandwidth $h = 2.5h_0$, which are based on our simulation studies and some related sensitive analyses. This work was conducted in `R` version 4.1.1 on a machine with Intel(R) Core(TM) I7-7700 CPU @ 3.60 GHz with 16.0GB RAM. The total running time at this experiment was 8.84 minutes. In particular, OOBGK took 7.36 minute, OOBW took 0.44 minutes, and LSCP took 1.05 minutes. I used the custom NatGeoStyle[3] for visualization in Figure 2.

Figure 2 shows the promising validation performance of OOBGK approach. The reported earthquake events are represented as a circle. The epicenter locations are plotted at the center of each circle, and the magnitude of an earthquake are represented as the size of radius. That is, The larger the radius of a circle, the greater the magnitude reported. We evaluate the true-positive case not only the point predicted value is above 5.5, but also the lower bound of interval prediction is higher than 5.5. Despite of the low TPRs in general, OOBGK shows the highest TPR among the candidates. Moreover, OOBGK gives the lowest prediction interval length and the lowest AIS90 as well.

5 Conclusion

We design an out-of-bag guided random forest prediction interval, called OOBGK, for spatially dependent data under mild conditions. Using the dependency adjusted random forest regression framework [23], we construct a novel nonparametric kernel estimator to provide an uncertainty quantified measure resulting from the out-of-bag error guided conditional prediction error distribution. We prove its theoretical consistency to the true conditional prediction error distribution, and verify that the OOBGK offers generally good marginal and conditional performance in finite-sample practice. As well as the context of prediction interval, we can also naturally broaden our scope to either conditional bias or conditional mean squared prediction error similarly to Lu and Hardin [18]. All codes, data, and other relevant assets are included in the URL "https://github.com/junpeea/SpRFPI". Also, we are developing an `R` package entitled `SpRFPI` to make public users facilitate our methodology according to their own research purposes.

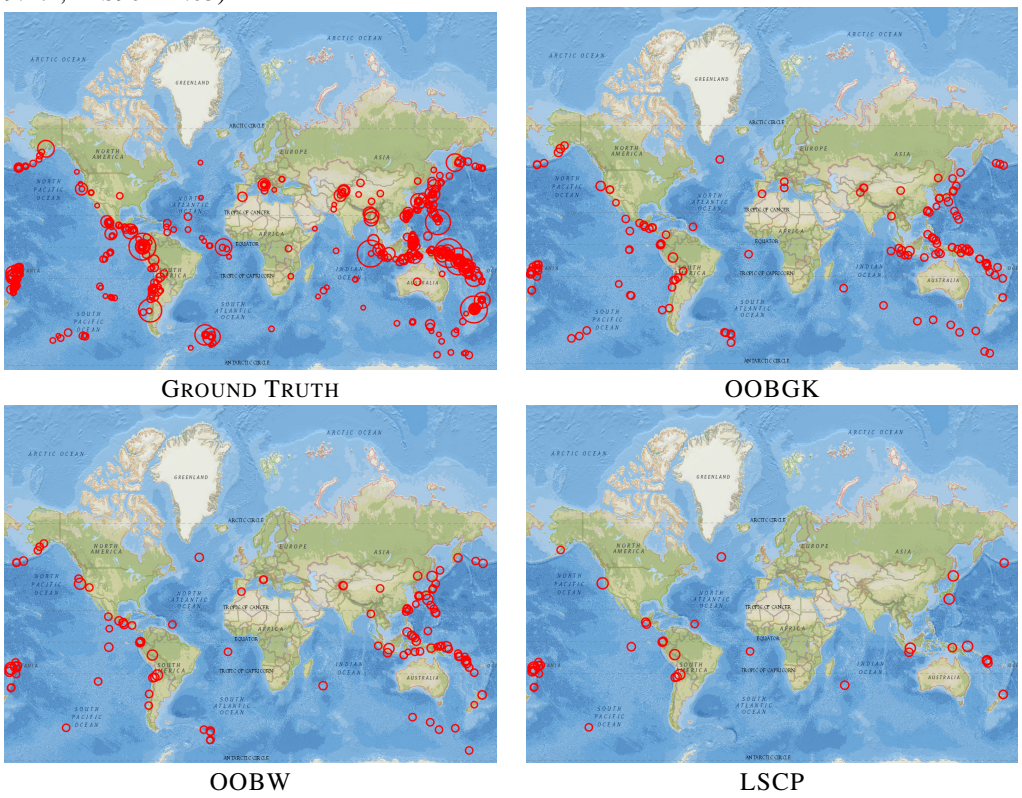
One appealing characteristic of OOBGK is that it does not need to split data into training and validation set out of available observations before applying random forest procedure. This property helps to make fully use of training dataset. We can find both empirical evidences and theoretical verification of the main consistency of the estimator for sufficiently large sample size. On the other hand, split conformal approaches need this kind of data split, so that the training power could be decreased since the calibration set would not be used in the tree construction. Naive conformal prediction may not need such data split, but it usually requires high amount of computational burden [16]. Furthermore, the conformal methods require the assumption of exchangeability, but often it is failed under spatial dependence.

Our main argument and interpretations are consistent with the different set of nominal error rates for the given training sample sizes (A.2Table 4). However, they may differ depending on the spatial

[2]See https://www.kaggle.com/datasets/usgs/earthquake-database

[3]Copyright (c) 2013 Leaflet Providers contributors All rights reserved.

Figure 2: Visualization of the ground truth/model-predicted earthquakes of magnitude greater than 5.5 in year 2016. Each circle represents the reported earthquake events, of which center corresponds to the epicenter locations, and of which radius represents the magnitude. The top-right panel result is from OOBGK approach (TPR = 33.7%, AIS90 = 1.76). The bottom-left panel result is from OOBW approach (TPR = 28.6%, AIS90 = 1.85). The bottom-right panel result is from LSCP approach (TPR = 9.2%, AIS90 = 2.03)



GROUND TRUTH OOBGK

OOBW LSCP

dependence of the variable of interest or the case where the model assumptions do not hold. Under a few spatial dependency (A.2Table 2), OOBGK may not provide the best average interval score, but we can still argue that OOBGK shows relarively low length of prediction interval as well as comparable level of average interval score among the candidates in general. Next, OOBGK requires homoscadesticity assumption of measurement error variance. We observe that OOBGK tends to achieve nominal level regardless of the specified nominal error rate if this assumption holds, but it tends to overestimate the miscoverage rate otherwise. (Table 1, A.2Figure 3).

We also note that the selection of kernel bandwidth h is more sensitive to performance of prediction interval than the other random forest tuning parameters such as the number of tree generation (n_{tree}), and the number of predictors for node-splitting (m_{try}) (A.2 Figure 4). We highly recommend to apply cross validation to check validity of your own choice of h in your studies. Finally, the sensitiveness to working precision specification could be only problematic when the underlying correlation is not sufficiently smooth as shown in A.2Table 3. We wish to extend our work to deal with nonstationary processes and more general spatial sampling designs for the future works.

Acknowledgement

This research was partially supported by AWD-021392-00001: HDR TRIPODS: D4 (Dependable Data-Driven Discovery) 10/01/2019 (version 5)

References

[1] Susan Athey, Julie Tibshirani, and Stefan Wager. "Generalized random forests". In: *The Annals of Statistics* 47.2 (2019), pp. 1148–1178.

[2] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

[3] Gérard Biau. "Analysis of a random forests model". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 1063–1095.

[4] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *Test* 25.2 (2016), pp. 197–227.

[5] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[6] Abhirup Datta et al. "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets". In: *Journal of the American Statistical Association* 111.514 (2016), pp. 800–812.

[7] Ibrahim Fayad et al. "Regional scale rain-forest height mapping using regression-kriging of spaceborne and airborne LiDAR data: Application on French Guiana". In: *Remote Sensing* 8.3 (2016), p. 240.

[8] Andrew O Finley et al. "Efficient algorithms for Bayesian nearest neighbor Gaussian processes". In: *Journal of Computational and Graphical Statistics* 28.2 (2019), pp. 401–414.

[9] Rina Friedberg et al. "Local linear forests". In: *Journal of Computational and Graphical Statistics* 30.2 (2020), pp. 503–517.

[10] Stefanos Georganos et al. "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling". In: *Geocarto International* 36.2 (2021), pp. 121–136.

[11] Tilmann Gneiting and Adrian E Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477 (2007), pp. 359–378.

[12] Ahlem Hajjem, François Bellavance, and Denis Larocque. "Mixed-effects random forest for clustered data". In: *Journal of Statistical Computation and Simulation* 84.6 (2014), pp. 1313–1328.

[13] Tomislav Hengl et al. "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables". In: *PeerJ* 6 (2018), e5518.

[14] Ulf Johansson et al. "Regression conformal prediction with random forests". In: *Machine learning* 97.1 (2014), pp. 155–176.

[15] Jing Lei and Larry Wasserman. "Distribution-free prediction bands for non-parametric regression". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 71–96.

[16] Jing Lei et al. "Distribution-free predictive inference for regression". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1094–1111.

[17] Andy Liaw, Matthew Wiener, et al. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.

[18] Benjamin Lu and Johanna Hardin. "A Unified Framework for Random Forest Prediction Error Estimation." In: *J. Mach. Learn. Res.* 22 (2021), pp. 8–1.

[19] Zudi Lu and Dag Tjøstheim. "Nonparametric estimation of probability density functions for irregularly observed spatial data". In: *Journal of the American Statistical Association* 109.508 (2014), pp. 1546–1564.

[20] Huiying Mao, Ryan Martin, and Brian Reich. "Valid model-free spatial prediction". In: *arXiv preprint arXiv:2006.15640* (2020).

[21] Nicolai Meinshausen and Greg Ridgeway. "Quantile regression forests." In: *Journal of Machine Learning Research* 7.6 (2006).

[22] Marie-Hélène Roy and Denis Larocque. "Prediction intervals with random forests". In: *Statistical Methods in Medical Research* 29.1 (2020), pp. 205–229.

[23] Arkajyoti Saha, Sumanta Basu, and Abhirup Datta. "Random forests for spatially dependent data". In: *Journal of the American Statistical Association* (2021), pp. 1–19.

[24] Simon J Sheather and Michael C Jones. "A reliable data-based bandwidth selection method for kernel density estimation". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3 (1991), pp. 683–690.

[25] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[26] Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. "On-line predictive linear regression". In: *The Annals of Statistics* (2009), pp. 1566–1590.

[27] Stefan Wager and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1228–1242.

[28] Haozhe Zhang et al. "Random forest prediction intervals". In: *The American Statistician* (2019).

A Appendix

A.1 Theoretical Results

In this section, we illustrate the list of required assumptions. The complete proof and the relevant lemmas for our main theorem will be also provided.

Assumption 1 MIXING CONDITION *We assume $Y_i = m(X_i) + \epsilon_i$ where the error process $\{\epsilon_i\}$ is a stationary, absolutely regular (β-mixing) process (Bradley 2005) with finite $(2+\delta)$th moment for some $\delta > 0$.*

Assumption 1 focuses on absolutely regular or β-mixing processes, which enable us to extend uniform law of large numbers from independent process to this dependent process under moderate restriction on the class of functions under consideration. No additional assumptions is required on the decay rate of the β-mixing coefficients.

Note that Assumption 1 implies that the error process $\{\epsilon_i\}$ also guarantees the α-mixing property that there exists a function φ such that $\varphi(t) \downarrow 0$ as $t \to \infty$, and a function $\psi : \mathcal{N}^2 \to \mathbb{R}^+$ symmetric and increasing in each of its two arguments, such that

$$\begin{aligned}\alpha(\mathcal{B}(\mathcal{S}'), \mathcal{B}(\mathcal{S}'')) &:= \sup\{|P(AB) - P(A)P(B)|, A \in \mathcal{B}(\mathcal{S}'), B \in \mathcal{B}(\mathcal{S}'')\} \\ &\leq \psi(Card(\mathcal{S}'), Card(\mathcal{S}''))\varphi(d(\mathcal{S}', \mathcal{S}'')),\end{aligned}$$

for any $\mathcal{S}', \mathcal{S}'' \subset \mathbb{R}^2$. The function φ moreover is such that

$$\lim_{m\to\infty} m^{\gamma} \sum_{j=m}^{\infty} j^2 \{\varphi(j)\}^{\kappa/(2+\kappa)} = 0$$

for some constant $\gamma > \max\{1, 2\kappa/(2+\kappa)\}$ and some $\kappa > 0$.

Assumption 2 REGULARITY OF THE WORKING PRECISION MATRIX *The working precision matrix $Q = \Gamma^{-1}$ admits a regular and sparse lower-triangular Cholesky factor $\Gamma^{-1/2}$ such that*

$$\Gamma^{-1/2} = \begin{pmatrix} L_{q\times q} & 0 & 0 & \cdots & \cdots \\ \rho^T_{1\times(q+1)} & & 0 & \cdots & \cdots \\ 0 & \rho^T_{1\times(q+1)} & & 0 & \cdots \\ \vdots & \ddots & & & \vdots \\ \cdots & 0 & 0 & \rho^T_{1\times(q+1)} & \end{pmatrix}$$

where $\rho = (\rho_q, \rho_{q-1}, \cdots, \rho_0)^T \in \mathbb{R}^{q+1}$ for some fixed $q \in \mathbb{N}$, and L is a fixed lower-triangular $q \times q$ matrix.

Assumption 2 requires the cholesky factor of the precision matrix to be sparse and regular. For spatial data, an exponential covariance family on a one-dimensional grid satisfies this. Other covariances like the Matérn family (except the exponential covariance) do not generally satisfy this assumption. However, NNGP covariance matrices satisfy this and are able to be used as excellent approximation to the full GP covariance matrices[6]. We can always use an approximate working covariance matrix arising from NNGP to the true covariance of the process to satisfy this assumption.

Assumption 3 DIAGONAL DOMINANCE OF THE WORKING PRECISION MATRIX *Q is diagonally dominant satisfying $Q_{ii} - \sum_{j\neq i} |Q_{ij}| > \xi$, for all i, for some constant $\xi > 0$.*

Diagonal dominance (Assumption 3) implies the smallest eigenvalue of Q is bounded away from zero as $n \to \infty$ which is needed to ensure stability of the GLS estimate. Note that under Assumption 2, checking that the first $(q+1)$ rows of Q are diagonally dominant is enough to verify Assumption 3.

Assumption 4 TAIL BEHAVIOR OF THE ERROR DISTRIBUTION

(a) There exist $\{\xi_n\}_{n\geq 1}$ *such that*

$$\xi_n \to \infty, \frac{t_n(\log n)\xi_n^4}{n} \to 0, \text{ and}$$
$$\mathbb{E}\left[(\max_i \epsilon_i^2)\mathbb{1}(\max_i\{\epsilon_i^2 > \xi_i^2\})\right] \to 0 \text{ as } n \to \infty$$

(b) There exist constant $C_\pi > 0$ *and* $n_0 \in \mathbb{N}^*$ *such that with probability* $1-\pi$*, for* $\forall n > n_0$*,*

$$\max_i |\epsilon_i| \leq C_\pi \sqrt{\log n}$$

(c) Let $\mathcal{I}_n \subseteq \{1, 2, \cdots, n\}$ *with* $|\mathcal{I}_n| := a_n$ *and* $a_n \to \infty$ *as* $n \to \infty$*. Then* $\frac{1}{a_n}|\sum_{i\in\mathcal{I}_n}\epsilon_i| > \delta$ *with probability at most* $C\exp(-ca_n)$*, and* $\frac{1}{n}|\sum_i \epsilon_i^2| > \sigma_0^2$ *with probability at most* $C\exp(-cn)$ *for any* $\delta > 0$*, and some constants* $c, C, \sigma_0^2 > 0$*.*

For Gaussian errors, ξ_n needs to be $\mathcal{O}(\log n)^2$ which makes the scaling condition in Assumption 4(a) as $\frac{t_n(\log n)^9}{n} \to 0$. This is the same scaling used in Scornet et al. (2015) for Gaussian errors and using the entire sample. In general, the choice of ξ_n will depend on the error distribution. Assumption 4(a), 4(b), and 4(c) will all be satisfied by sub-Gaussian errors.

Assumption 5 ADDITIVE MODEL ON THE COORDINATES *The true mean function* $m(x(s)))$ *is additive on the coordinates* s_d*, that is,* $m(x(s)) = \sum_{d=1}^D m_d(x_d(s))$*, where each component* m_d *is continuous.*

Assumption 6 SAMPLING SITES *The observations are positioned at* $\{s_i, i = 1, 2, \cdots, n\} \subset \mathbb{R}^d$*, which are defined under domain-expanding infill asymptotics, where*

$$\delta_n := \max_{1\leq j\leq n} \delta_{j,n} \to 0,$$
$$\text{with } \delta_{j,n} := \min\{\|s_i - s_j\| : 1 \leq i \leq n, i \neq j\}$$

that is, the distance between neighboring observations all tends to zero, as $n \to \infty$*, and*

$$\Delta_n := \min_{1\leq j\leq n} \Delta_{j,n} \to \infty,$$
$$\text{with } \Delta_{j,n} := \max\{\|s_i - s_j\| : 1 \leq i \leq n, i \neq j\}$$

that is, the domain at each location is expanding to ∞*, as* $n \to \infty$*, where* $\|\cdot\|$ *denotes the Euclidean norm in* $\mathcal{R}^2$*. We suppose* $\min_{1\leq j\leq n} \delta_{j,n}/\delta_n \geq c_1 > 0$*, and* $\max_{i\leq j\leq n} \Delta_{j,n}/\Delta_n \leq C_1 < \infty$*, for all* n*. Also, there exists a continuous sampling intensity function* f_S *defined on* $\mathbb{R}^d$ *such that*

(a) for any measurable set $A \subset \mathbb{R}^d$*,* $N^{-1}\sum_{i=1}^N I(s_i \in A) \to \int_A f_S(s)ds$ *as* $N \to \infty$

(b) $f_S(s)$ *is bounded and has second derivatives which are continuous on* $\mathbb{R}^d$*.*

Assumption 7 KERNEL FUNCTIONS *The kernel function* $K(\cdot)$ *satisfies that* $\int K(u)du = 1$*,* $\int uK(u)du = 0$*, and* $\mu_{K,2} := \int u^2K(u)du < \infty$*,* $\nu_K := \int K^2(u)du < \infty$*.*

Assumption 8 BANDWIDTHS I *As* $n \to \infty$*, (a)* $h_n \to 0$*; (b)* $nh_n \to \infty$*; and (c)* $\delta_n^{-2(1+2/\gamma)}h_n^{2(\gamma-2\kappa/(2+\kappa)\gamma} \to 0$*.*

Assumption 9 BANDWIDTHS II *Let* $c_N = \{\delta_n^2 h_n^{\kappa/(2+\kappa)}\}^{-1/\gamma}$*, which tends to* ∞ *as* $n \to \infty$*. (a)* $\limsup_{n\to\infty} nh_n^2 > 0$*,* $nh_n^5 = \mathcal{O}(1)$*; (b)* $\liminf_{m\to\infty} m^{4+3\gamma}\sum_{t=m}^\infty t^2[\varphi(t)]^{\kappa/(2+\kappa)} < \infty$*; (c)* $n\psi(1,n)\varphi(c_n) \to 0$*, as* $n \to \infty$*, where* ψ *and* φ *are defined in Assumption 1.*

Assumption 10 *The density of X is positive and bounded in a domain $\mathcal{X} \subset \mathbb{R}^p$.*

For notational convenience, we further assume that Covariate X has the uniform distribution over $[0,1]^p$ during the proof, but the results can be generalized into Assumption 10 as well.

Assumption 11 *Let $k_b(l) := \sum_{h=1}^{n} n_{hb} \mathbb{1}(\mathbf{z}_h \in l_b(\mathbf{x}, \mathbf{s}))$ denote the number of units from its bootstrap sample $\mathcal{D}_n^*$ in its terminal node l containing x given a tree b.*

- *(a) The proportion of observations from $\mathcal{D}_n^*$ in any given node, relative all observations from $\mathcal{D}_n^*$, is decreasing in n, that is, $\max_{l,b} k_b(l) = o(n)$*
- *(b) The minimum number of observations from $\mathcal{D}_n^*$ in a node is increasing in n, that is, $1/\min_{l,b} k_b(l) = o(n)$*
- *(c) The probability that variable $m \in \{1, \cdots, p\}$ is chosen for a given split point is bounded from below for every node by a positive constant.*
- *(d) When a node is split, the proportion of observations belong to $\mathcal{D}_n^*$ in the original node that fall into each of the resulting sub-nodes is bounded from below by a positive constant.*

The conditions in Assumption 11 are adapted from assumptions used to prove consistency of quantile regression forests [21]. Tree construction algorithms that satisfy these properties or variants of them have been referred to in recent random forest literature as "regular", "balanced", or "random-split" [27, 1, 9]

Assumption 12 *$F_E(e|Z = z)$ is Lipschitz continuous with parameter L. That is, for all $z, z' \in \mathcal{X} \times \mathcal{S}$,*

$$\sup_{e \in \mathbb{R}} |F_E(\epsilon|z) - F_E(e|z')| \leq L\|z - z'\|_1$$

All existing results on pointwise consistency of random forests have required an analogous smoothness condition in the distribution of interest, including Biau(2012)[3], meinshausen(2006)[21], and Wager and Athey(2018)[27].

Assumption 13 *$F_E(e|Z = z)$ is strictly monotone in e for all $z \in \mathcal{X} \times \mathcal{S}$.*

We assume that the distribution of prediction errors is strictly monotone so that consistency of quantile estimates follows from consistency of distribution estimates.

Assumption 14 BEHAVIOR OF THE OUT-OF-BAG WEIGHT *Define $\mathcal{M}_i(\delta) := \{\hat{m}_\Gamma(Z_i) - m(Z_i)\}$ be the event for any given $\delta > 0$. We say that δ-stability of the i^{th} unit has been realized if and only if $\mathcal{M}_i(\delta)$ holds. For all $z \in \mathcal{X} \times \mathcal{S}$, there exists $\delta_0 > 0$ such that for any $\delta_0 \in (0, \delta_0)$, $\mathbb{E}[v_i(z)|\mathcal{M}_i(\delta)] = \mathcal{O}(n^{-1})$ and $\mathbb{E}[v_i(z)| - \mathcal{M}_i(\delta)] = \mathcal{O}(n^{-1})$*

Assumption 14 characterizes the stability of the random forest and the underlying population distribution. It states that the expected out-of-bag weight of the i^{th} observation is of order $1/n$ whether δ-stability has been realized for the observation or not. The expected values are taken over all training units and all random parameters governing the sample-splitting and tree-growing mechanisms.

[**Proof of Theorem 2.1**] Fix $x \in [0,1]^p$ and $s \in [0,1]^d$ (Assumption 10). Denote $E = Y - m(z)$ as the true underlying error random variable (Assumption 1). Let E_i^* follow the distribution $F_E(\epsilon|Z_i)$, and E_i follows $F_E(\epsilon|z)$, $i = 1, \cdots, n$. The goal is to prove that $\hat{F}_n(\epsilon|z)$ is a consistent estimator of $F_E(\epsilon|z)$ by showing its convergence in probability, i.e. $|\hat{F}_n(\epsilon|z) - F_E(\epsilon|z)| \xrightarrow{p} 0$ as $n \to \infty$, for any $\epsilon \in \mathbb{R}$. We have

$$\begin{aligned}
\hat{F}_n(\epsilon|z) &= \sum_{i=1}^{n} v_i(z) \int_{-\infty}^{\epsilon} K_h(E_i^* - e)de \\
&= \sum_{i=1}^{n} v_i(z) \int_{-\infty}^{\epsilon} K_h(E_i - e)de + \sum_{i=1}^{n} v_i(z) \int_{-\infty}^{\epsilon} \left(K_h(E_i^* - e) - K_h(E_i - e)\right) de
\end{aligned}$$

$$\begin{aligned}
\left|\hat{F}_n(\epsilon|z) - F_E(\epsilon|z)\right| \leq & \left| \sum_{i=1}^{n} v_i(z) \int_{-\infty}^{\epsilon} K_h(E_i - e)de - F_E(\epsilon|z) \right| \\
& + \left| \sum_{i=1}^{n} v_i(z) \int_{-\infty}^{\epsilon} \left(K_h(E_i^* - e) - K_h(E_i - e)\right) de \right|
\end{aligned} \tag{3}$$

We wish to show that the right side of (3) converges to zero in probability. By following the terminologies in Meinshausen(2006)[21], we will call the first term on the right side as a "Variance Term", and the second term as a "Shift Term", that is, $\left|\hat{F}_n(\epsilon|z) - F_E(\epsilon|z)\right| \leq$ (Variance Term) + (Shift Term). Next, we will verify that each term converges to zero in probability.

Bounding the Variance Term We will use the asymptotics for the marginal density function estimator for spatial data[19]. In order to verify the variance term converges to zero in probability, it suffices to show that

$$\sum_{i=1}^{n} v_i(z) \int_{-\infty}^{\epsilon} K_{h_n}(E_i - e)de - F_E(\epsilon|z) \xrightarrow{d} \Phi\left(\frac{1}{2}\mu_{K,2}\ddot{f}_E(\epsilon|z)h_n^2, \frac{\nu_K f_E(\epsilon|z)}{nh_n}\right), \tag{4}$$

which implies that both the mean and the variance of the limiting distribution are close to zero for sufficiently large sample size.

Since the weights $\{v_i(z)\}$ are built on a randomly chosen subset of bootstrapped samples that are not containing Y_i, conditioning on $Z = z$ yields sufficient independence $(v_i \perp\!\!\!\perp E_i)|Z = z$. Then we can evaluate the expectation of the weighted average in (4) as follows:

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} v_i(z)K_h(E_i - e)\right) &= \mathbb{E}\left[\mathbb{E}\left(\sum_{i=1}^{n} v_i(z)K_h(E_i - e)\middle| Z = z\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left(v_i(z)\middle| Z = z\right)\mathbb{E}\left(K_h(E_i - e)\middle| Z = z\right)\right]
\end{aligned} \tag{5}$$

After separating the one conditional expectation into the out-of-bag weight part and the kernel part as shown in (5), we can apply the similar process illustrated in Zudi Lu *et al.* [19] and the assumption of kernel functions (Assumption 7) to the kernel part:

$$
\begin{aligned}
\mathbb{E}\left(K_h(E_i-e)\middle|Z=z\right) &= h^{-1}\int K((u-e)/h)f_E(u|z)du \\
&= \int K(u)f_E(e+hu|z)du \\
&= \int K(u)\left[f_E(e|z)+\dot{f}_E(e|z)(hu)+\frac{1}{2}\ddot{f}_E(e+\xi hu|z)(hu)^2\right]du \\
&= f_E(e|z)+\frac{1}{2}\ddot{f}_E(e|z)h^2\int u^2K(u)du
\end{aligned}
$$

This completes the equation (5) as follows:

$$
\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^n v_i(z)K_h(E_i-e)\right) &= \mathbb{E}\left[\mathbb{E}\left(\sum_{i=1}^n v_i(z)K_h(E_i-e)\middle|Z=z\right)\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left(v_i(z)\middle|Z=z\right)\mathbb{E}\left(K_h(E_i-e)\middle|Z=z\right)\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left(v_i(z)\middle|Z=z\right)\left(f_E(e|z)+\frac{1}{2}\ddot{f}_E(e|z)h^2\int u^2K(u)du\right)\right] \\
&= \left(f_E(e|z)+\frac{1}{2}\ddot{f}_E(e|z)h^2\int u^2K(u)du\right)\mathbb{E}\sum_{i=1}^n v_i(z) \\
&= f_E(e|z)+\frac{1}{2}\mu_{K,2}\ddot{f}_E(e|z)h^2
\end{aligned}
\tag{6}
$$

Furthermore, since the variance of a summation is equal to the summation of the covariances,

$$
\begin{aligned}
\mathbb{V}ar\left(\sum_{i=1}^n v_i(z)K_h(E_i-e)\right) &= \sum_{i=1}^n \mathbb{V}ar\Big(v_i(z)K_h(E_i-e)\Big) \\
&\quad + \sum_{i\neq j}\mathbb{C}ov\Big(v_i(z)K_h(E_i-e), v_j(z)K_h(E_j-e)\Big)
\end{aligned}
$$

Each summation of the right hand side converges to zero by Lemma A.1 and Lemma A.2.

Bounding the Shift Term We will first show that

$$
\sum_{i=1}^n v_i(z)K_{h_n}(E_i^*-e)-\sum_{i=1}^n v_i(z)f_E(e|Z_i)\xrightarrow{p}0, \tag{7}
$$

as $n \to \infty$. Since $v_i(z)$ and $K_{h_n}(E_i^* - e)$ are not exactly but asymptotically independent conditioning on Z_i (Lemma A.3),

$$\begin{aligned}
&\mathbb{E}\left(K_h(E_i^* - e)\middle| Z_i\right) = \mathbb{E}\left(K_h(\tilde{E}_i^* + o_p(1) - e)\middle| Z_i\right) \\
&= h^{-1}\int K((u + o_p(1) - e)/h) f_{\tilde{E}}(u|z)du = \int K(u) f_{\tilde{E}}(e - o_p(1) + hu|z)du \\
&= \int K(u)\left[f_{\tilde{E}}(\epsilon|z) + \dot{f}_{\tilde{E}}(\epsilon|z)(hu - o_p(1)) + \frac{1}{2}\ddot{f}_{\tilde{E}}(e + \xi hu|z)(hu - o_p(1))^2\right] du \\
&= f_{\tilde{E}}(\epsilon|z) - \dot{f}_{\tilde{E}}(\epsilon|z)o_p(1) + \frac{1}{2}\ddot{f}_{\tilde{E}}(\epsilon|z)h^2(1 + o_p(1))\int u^2 K(u)du \\
&\to f_E(\epsilon|z) + \frac{1}{2}\mu_{K,2}\ddot{f}_E(\epsilon|z)h^2
\end{aligned} \tag{8}$$

Consequently, we find the left side of the equation (7) has expectation zero by the linearity of expectation such that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^n v_i(z)K_{h_n}(E_i^* - e) - \sum_{i=1}^n v_i(z) f_E(e|Z_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\sum_{i=1}^n v_i(z)K_h(\tilde{E}_i^* + o_p(1) - e)\middle| Z_i\right) - \sum_{i=1}^n v_i(z) f_E(e|Z_i)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}\left(v_i(z)\middle| Z_i\right)\mathbb{E}\left(K_h(\tilde{E}_i^* + o_p(1) - e)\middle| Z_i\right) - \sum_{i=1}^n v_i(z) f_E(e|Z_i)\right] \\
&\to \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}\left(v_i(z)\middle| Z_i\right)\left(f_E(\epsilon|Z_i) + \frac{1}{2}\mu_{K,2}\ddot{f}_E(\epsilon|Z_i)h^2\right) - \sum_{i=1}^n v_i(z) f_E(e|Z_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\sum_{i=1}^n v_i(z) f_E(\epsilon|Z_i)\middle| Z_i\right) - \sum_{i=1}^n v_i(z) f_E(e|Z_i)\right] + \frac{1}{2}\mu_{K,2}\ddot{f}_E(\epsilon|Z_i)h^2\left(\mathbb{E}\sum_{i=1}^n v_i(z)\right) \\
&= \frac{1}{2}\mu_{K,2}\ddot{f}_E(\epsilon|Z_i)h^2
\end{aligned} \tag{9}$$

Also, we can consider the left side of the equation (7) has decreasing variance by showing that

$$\begin{aligned}
&\mathbb{V}ar\left[\sum_{i=1}^n v_i(z)K_{h_n}(E_i^* - e) - \sum_{i=1}^n v_i(z) f_E(e|Z_i)\right] \\
&= \sum_{i=1}^n \mathbb{V}ar\left[v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big)\right] \\
&+ \sum_{i\neq j} \mathbb{C}ov\left[v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big), v_j(z)\Big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\Big)\right]
\end{aligned}$$

with each summation of the right side converging to zero by Lemma A.4 and Lemma A.5. By the triangle inequality,

$$
\begin{aligned}
&\Big| \sum_{i=1}^{n} v_i(z) \left(K_h(E_i^* - e) - K_h(E_i - e) \right) - \sum_{i=1}^{n} v_i(z) \left(f_E(e|Z_i) - f_E(\epsilon|z) \right) \Big| \\
&= \Big| \sum_{i=1}^{n} v_i(z) \left(K_h(E_i^* - e) - f_E(e|Z_i) \right) - \sum_{i=1}^{n} v_i(z) \left(K_h(E_i - e) - f_E(\epsilon|z) \right) \Big| \\
&\leq \Big| \sum_{i=1}^{n} v_i(z) \left(K_h(E_i^* - e) - f_E(e|Z_i) \right) \Big| + \Big| \sum_{i=1}^{n} v_i(z) \left(K_h(E_i - e) - f_E(\epsilon|z) \right) \Big|
\end{aligned}
$$

Recall that $\Big| \sum_{i=1}^{n} v_i(z) \left(K_h(E_i - e) - f_E(\epsilon|z) \right) \Big| = o_p(1)$ by equation (4), and $\Big| \sum_{i=1}^{n} v_i(z) \left(K_h(E_i^* - e) - f_E(e|Z_i) \right) \Big| = o_p(1)$ by equation (7). Therefore, we can reduce the task of bounding the shift term by

$$
\sum_{i=1}^{n} v_i(z) \left(K_h(E_i^* - e) - K_h(E_i - e) \right) \xrightarrow{p} \sum_{i=1}^{n} v_i(z) \left(f_E(e|Z_i) - f_E(\epsilon|z) \right),
$$

The Lipschitz continuity of the conditional prediction error distribution (Assumption 12) shows

$$
\Big| \sum_{i=1}^{n} v_i(z) \left(f_E(e|Z_i) - f_E(\epsilon|z) \right) \Big| \leq \sum_{i=1}^{n} v_i(z) \|Z_i - z\|_1 \tag{10}
$$

To complete the proof, we need to verify $\sum_{i=1}^{n} v_i(z) \|Z_i - z\|_1 = o_p(1)$. This is followed by the argument in the Lemma 2 of Meinshausen(2006) [21]. In particular, we want to show that

$$
\lim_{B \to \infty} \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{\#\{Z_i \notin \mathcal{D}_n^*\} \mathbb{1}(X_i \in R_{l_b(z)})}{\sum_{j=1}^{n} \#\{Z_j \notin \mathcal{D}_n^*\} \mathbb{1}(Z_j \in R_{l_b(z)})} \|Z_i - z\|_1 \xrightarrow{p} 0
$$

Then it suffices to show that, for any given $b \in \{1, \cdots, B\}$,

$$
\sum_{i=1}^{n} \frac{\#\{Z_i \notin \mathcal{D}_n^*\} \mathbb{1}(X_i \in R_{l_b(z)})}{\sum_{j=1}^{n} \#\{Z_j \notin \mathcal{D}_n^*\} \mathbb{1}(Z_j \in R_{l_b(z)})} \|Z_i - z\|_1 \xrightarrow{p} 0
$$

Note that we can decompose the rectangular subspace $R_{l_b(z)} \subseteq [0,1]^p$ of leaf $l_b(z)$ of the tree into the intervals $I(x, m, b) \subseteq [0, 1]$ for $m = 1, \cdots, p$: $R_{l_b(z)} = \bigotimes_{m=1}^{p} I(x, m, b)$. Then the arguments in Lemma2 of Meinshausen [21] can assure that $\max_m |I(x, m, b)| = o_p(1)$, which completes our proof.

Lemma A.1 *Under Assumptions 10-14, as $n \to \infty$,*

$$\sum_{i=1}^{n} \mathbb{V}ar\Big(v_i(z)K_{h_n}(E_i - e)\Big) \to 0$$

[Proof of Lemma A.1]

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{V}ar\Big(v_i(z)K_{h_n}(E_i - e)\Big) \\
&\quad = \sum_{i=1}^{n} \Big[\mathbb{V}ar\Big\{\mathbb{E}\Big(v_i(z)K_{h_n}(E_i - e)\Big|\Omega \setminus \{Y_i\}\Big)\Big\} + \mathbb{E}\Big\{\mathbb{V}ar\Big(v_i(z)K_{h_n}(E_i - e)\Big|\Omega \setminus \{Y_i\}\Big)\Big\}\Big] \\
&\quad = \sum_{i=1}^{n} \Big[\mathbb{V}ar\Big\{v_i(z)\mathbb{E}\Big(K_{h_n}(E_i - e)\Big|\Omega \setminus \{Y_i\}\Big)\Big\} + \mathbb{E}\Big\{v_i^2(z)\mathbb{V}ar\Big(K_{h_n}(E_i - e)\Big|\Omega \setminus \{Y_i\}\Big)\Big\}\Big] \\
&\quad = \sum_{i=1}^{n} \Big[(f_E(e|z) + \frac{1}{2}\mu_{K,2}\ddot{f}_E(e|z)h^2)^2\mathbb{V}ar(v_i(z)) + \frac{\nu_K f_E(\epsilon|z)}{nh_n}\mathbb{E}(v_i^2(z))\Big] \\
&\quad = A\sum_{i=1}^{n} \mathbb{V}ar(v_i(z)) + B\sum_{i=1}^{n} \mathbb{E}(v_i^2(z)),
\end{aligned} \tag{11}$$

where $A := (f_E(e|z) + \frac{1}{2}\mu_{K,2}\ddot{f}_E(e|z)h^2)^2$, $B := \frac{\nu_K f_E(\epsilon|z)}{nh_n}$. Let $C = \max\{A, B\}$, and M_n be the maximum possible weight given to any observation, which is decreasing in n (Assumption 11). Thus, equation 10 yields

$$\begin{aligned}
0 \le \sum_{i=1}^{n} \mathbb{V}ar\Big(v_i(z)K_{h_n}(E_i - e)\Big) &= A\sum_{i=1}^{n} \mathbb{V}ar(v_i(z)) + B\sum_{i=1}^{n} \mathbb{E}(v_i^2(z)) \\
&\le (A + B)\sum_{i=1}^{n} \mathbb{E}(v_i^2(z)) \\
&\le 2C\sum_{i=1}^{n} \mathbb{E}(v_i^2(z)) \\
&\le 2C\sum_{i=1}^{n} M_n\mathbb{E}(v_i(z)) \\
&= 2CM_n \xrightarrow{n\to\infty} 0
\end{aligned}$$

Lemma A.2 *Under Assumptions 10-14, as $n \to \infty$,*

$$\sum_{i\neq j} \mathbb{C}ov\Big(v_i(z)K_h(E_i - e), v_j(z)K_h(E_j - e)\Big) \to 0$$

[Proof of Lemma A.2]

$$
\begin{aligned}
&\sum_{i\neq j}\mathbb{C}ov\Big(v_i(z)K_h(E_i-e), v_j(z)K_h(E_j-e)\Big)\\
&\quad=\sum_{i\neq j}\Big[\mathbb{E}\big\{v_i(z)K_h(E_i-e)v_j(z)K_h(E_j-e)\big\}-\mathbb{E}\big\{v_i(z)K_h(E_i-e)\big\}\mathbb{E}\big\{v_j(z)K_h(E_j-e)\big\}\Big]\\
&\quad=\sum_{i\neq j}\mathbb{E}\big\{v_i(z)v_j(z)K_h(E_i-e)K_h(E_j-e)\big\}-\mathbb{E}\big\{\sum_{i=1}^{n}v_i(z)K_h(E_i-e)\big\}\mathbb{E}\big\{\sum_{j=1}^{n}v_j(z)K_h(E_j-e)\big\}\Big]\\
&\quad+\sum_{i=1}^{n}\mathbb{E}\big\{v_i^2(z)K_h^2(E_i-e)\big\}\\
&\quad\to\sum_{i\neq j}\mathbb{E}\big\{v_i(z)v_j(z)K_h(E_i-e)K_h(E_j-e)\big\}-\Big(f_E(e|z)+\frac{1}{2}\mu_{K,2}\ddot{f}_E(e|z)h^2\Big)^2,
\end{aligned}
$$

since, by equation (4),

$$
\mathbb{E}\big\{\sum_{j=1}^{n}v_j(z)K_h(E_j-e)\big\}=\mathbb{E}\big\{\sum_{j=1}^{n}v_j(z)K_h(E_j-e)\big\}=f_E(e|z)+\frac{1}{2}\mu_{K,2}\ddot{f}_E(e|z)h_n^2,
$$

and

$$
\begin{aligned}
\sum_{i=1}^{n}\mathbb{E}\big\{v_i^2(z)K_h^2(E_i-e)\big\}&=\sum_{i=1}^{n}\mathbb{E}\Big[\mathbb{E}\big\{v_i^2(z)K_h^2(E_i-e)\Big|Z=z\big\}\Big]\\
&=\sum_{i=1}^{n}\mathbb{E}\Big[\mathbb{E}\big\{v_i^2(z)\Big|Z=z\big\}\mathbb{E}\big\{K_h^2(E_i-e)\Big|Z=z\big\}\Big]\\
&=\mu_{K,2}\sum_{i=1}^{n}\mathbb{E}\big\{v_i^2(z)\big\}\to 0
\end{aligned}
$$

Let $A:=f_E(e|z)+\frac{1}{2}\mu_{K,2}\ddot{f}_E(e|z)h^2$. Then it suffices to show

$$
\lim_{n\to\infty}\left|\sum_{i\neq j}\mathbb{E}\big\{v_i(z)v_j(z)K_h(E_i-e)K_h(E_j-e)\big\}-A^2\right|=0 \tag{12}
$$

We apply the second Taylor's expansion to the expectation of the kernel products condition on z.

$$
\begin{aligned}
&\mathbb{E}\big\{K_h(E_i-e)K_h(E_j-e)\Big|Z=z\big\}\\
&\quad= h^{-2}\int\int K((u-e)/h)K((v-e)/h)f_E(u|z)f_E(v|z)dudv\\
&\quad= \int\int K(u)K(v)f_E(e+uh|z)f_E(e+vh|z)dudv\\
&\quad= \int\int K(u)K(v)\Big[f_E(e+uh|z)f_E(e+vh|z)+\dot{f}_E(e+uh|z)f_E(e+vh|z)(uh)\\
&\quad+ f_E(e+uh|z)\dot{f}_E(e+vh|z)(vh)+\frac{1}{2}\big\{\ddot{f}_E(e+uh|z)f_E(e+vh|z)(uh)^2\\
&\quad+ 2\dot{f}_E(e+uh|z)\dot{f}_E(e+vh|z)(uvh^2)+f_E(e+uh|z)\ddot{f}_E(e+vh|z)(vh)^2\big\}\Big]dudv\\
&\quad= f_E(e|z)^2+\ddot{f}_E(e|z)f_E(e|z)\mu_{K,2}h^2(1+o_p(1))
\end{aligned}
$$

This proves $\mathbb{E}\big\{K_h(E_i-e)K_h(E_j-e)\Big|Z=z\big\}-A^2=o_p(1)$, for a sufficiently small choice of $h>0$. Finally, we can verify the equation (12) as follows:

$$
\begin{aligned}
&\sum_{i\neq j}\mathbb{E}\big\{v_i(z)v_j(z)K_h(E_i-e)K_h(E_j-e)\big\}-A^2\\
&\quad=\sum_{i\neq j}\mathbb{E}\left[\mathbb{E}\big\{v_i(z)v_j(z)K_h(E_i-e)K_h(E_j-e)\Big|Z=z\big\}\right]-A^2\\
&\quad=\sum_{i\neq j}\mathbb{E}\left[\mathbb{E}\big\{v_i(z)v_j(z)\Big|Z=z\big\}\mathbb{E}\big\{K_h(E_i-e)K_h(E_j-e)\Big|Z=z\big\}\right]-A^2\\
&\quad=\sum_{i\neq j}\mathbb{E}\left[\mathbb{E}\big\{v_i(z)v_j(z)\Big|Z=z\big\}\left\{\mathbb{E}\big\{K_h(E_i-e)K_h(E_j-e)\Big|Z=z\big\}-A^2+A^2\right\}\right]-A^2\\
&\quad=\sum_{i\neq j}\mathbb{E}\left[\mathbb{E}\big\{v_i(z)v_j(z)\Big|Z=z\big\}\left(o_p(1)+A^2\right)\right]-A^2\\
&\quad\to\sum_{i\neq j}A^2\mathbb{E}\big[\mathbb{E}\big\{v_i(z)v_j(z)\Big|Z=z\big\}\big]-A^2\\
&\quad=A^2\Big(\sum_{i\neq j}\mathbb{E}\big\{v_i(z)v_j(z)\big\}-1\Big)=A^2\Big(\sum_{i=1}^{n}\mathbb{E}\big\{v_i(z)(1-v_i(z))\big\}-1\Big)\to 0
\end{aligned}
$$

as $n\to\infty$.

Lemma A.3 *Under Assumptions 1-5, the GLS-style Random Forest error $E_i^*:=Y_i-\hat{m}_{\hat{\Gamma}}(X_i)$ is $\mathbb{L}_2$-consistent to the true underlying Random Forest error $\tilde{E}_i^*:=Y_i-m(X_i)$, that is, $\lim_{n\to\infty}\mathbb{E}\int(E_i^*-\tilde{E}_i^*)dX_i=0$, and E_i^* is asymptotically independent on the out-of-bag weight $v_i(z)$ conditioning on Z_i, for $i=1,\cdots,n$, respectively.*

[Proof of Lemma A.3]

If we consider the RF-GLS predictor $\hat{m}_{\hat{\Gamma}}$ is a $\mathcal{L}_2$-consistent estimator of the true underlying function m under assumptions 1-5 [23], we have

$$
E_i^*-\tilde{E}_i^*=\big(Y_i-\hat{m}_{\hat{\Gamma}}(X_i)\big)-(Y_i-m(X_i))=m(X_i)-\hat{m}_{\hat{\Gamma}}(X_i)
$$

This implies

$$
\lim_{n\to\infty}\mathbb{E}\int(E_i^*-\tilde{E}_i^*)dX_i=\lim_{n\to\infty}\mathbb{E}\int(m(X_i)-\hat{m}_{\hat{\Gamma}}(X_i))dX_i=0
$$

The out-of-bag weights $\{v_i(z)\}$ are built on a randomly chosen subset of bootstrapped samples that are not containing Y_i, so that conditioning on Z_i yields sufficient independence on the true underlying random forest error $\tilde{E}_i^*$, i.e. $(v_i \perp\!\!\!\perp \tilde{E}_i^*)|Z_i$. However, the weight $\{v_i(z)\}$ is not exactly independent on the estimated random forest error E_i^* conditioning on Z_i, i.e. $(v_i \not\perp\!\!\!\perp E_i^*)|Z_i$, since the same observations are used for the out-of-bag weights and for the tree construction of the predictor $\hat{m}_{\hat{\Gamma}}$. Therefore, we suggest an asymptotic independence between $v_i(z)$ and E_i^* by considering the above $\mathcal{L}_2$-consistency that implies $E_i^* \xrightarrow{p} \tilde{E}_i^*$, for $i = 1, \cdots, n$. Once we assume sufficiently large training sample size n, we can replace E_i^* by $(\tilde{E}_i^* + o_p(1))$ for the rest of our proof.

Lemma A.4 *Under Assumptions 10-14, as $n \to \infty$,*

$$\sum_{i=1}^{n} \mathbb{V}ar\Big[v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|z_i)\Big)\Big] \to 0$$

[Proof of Lemma A.4]

We decompose the sum of variances via the law of total variance:

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{V}ar\Big[v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big)\Big] \\
&\quad = \sum_{i=1}^{n} \mathbb{V}ar\left\{\mathbb{E}\Big(v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big)\Big|\Omega \setminus \{Y_i\}\Big)\right\} \\
&\quad + \sum_{i=1}^{n} \mathbb{E}\left\{\mathbb{V}ar\Big(v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big)\Big|\Omega \setminus \{Y_i\}\Big)\right\}
\end{aligned}$$

The first term of the right side converges to zero:

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{V}ar\left\{\mathbb{E}\Big(v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big)\Big|\Omega \setminus \{Y_i\}\Big)\right\} \\
&= \sum_{i=1}^{n} \mathbb{V}ar\left\{v_i(z)\mathbb{E}\Big(\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\Big|\Omega \setminus \{Y_i\}\Big)\right\} \\
&= \sum_{i=1}^{n} \mathbb{V}ar\left\{v_i(z)\mathbb{E}\Big(\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\Big|Z_i\Big)\right\} \\
&= \sum_{i=1}^{n} \mathbb{V}ar\left\{v_i(z) \cdot \frac{1}{2}\mu_{K,2}\ddot{f}_E(\epsilon|z)h^2\right\} \to 0,
\end{aligned}$$

where the last equality follows by (8). The second term of the right side converges to zero:

$$
\begin{aligned}
0 &\leq \sum_{i=1}^{n} \mathbb{E}\left\{ \mathbb{V}ar\Big(v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big)\Big|\Omega \setminus \{Y_i\}\Big)\right\} \\
&= \sum_{i=1}^{n} \mathbb{E}\left\{ v_i^2(z)\mathbb{V}ar\Big(\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\Big|\Omega \setminus \{Y_i\}\Big)\right\} \\
&= \sum_{i=1}^{n} \mathbb{E}\left\{ v_i^2(z)\mathbb{V}ar\Big(\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\Big|Z_i\Big)\right\} \\
&= \sum_{i=1}^{n} \mathbb{E}\left\{ v_i^2(z)\mathbb{V}ar\Big(K_{h_n}(E_i^* - e)\Big|Z_i\Big)\right\} \leq \sum_{i=1}^{n} \mathbb{E}\left\{v_i^2(z)\right\} \\
&\leq M_n \sum_{i=1}^{n} \mathbb{E}\left\{v_i(z)\right\} \xrightarrow{n\to\infty} 0
\end{aligned}
$$

Lemma A.5 *Under Assumptions 10-14, as $n \to \infty$,*

$$
\sum_{i\neq j} \mathbb{C}ov\Big[v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big), v_j(z)\Big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\Big)\Big] \to 0
$$

[Proof of Lemma A.5]

$$
\begin{aligned}
&\sum_{i\neq j} \mathbb{C}ov\Big[v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big), v_j(z)\Big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\Big)\Big] \\
&= \sum_{i\neq j} \mathbb{E}\Big\{ v_i(z)v_j(z)\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\big)\Big\} \\
&- \sum_{i\neq j} \mathbb{E}\Big\{ v_i(z)\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\Big\}\mathbb{E}\Big\{ \big(v_j(z)K_{h_n}(E_j^* - e) - f_E(e|Z_j)\big)\Big\} \\
&= \sum_{i\neq j} \mathbb{E}\Big\{ v_i(z)v_j(z)\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\big)\Big\}
\end{aligned}
$$

where the last equality follows by (9). Note that

$$
\begin{aligned}
&\mathbb{E}\big\{\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\big)\Big|Z_i, Z_j\big\} \\
&\quad = \mathbb{E}\big\{K_{h_n}(E_i^* - e)K_{h_n}(E_j^* - e)\Big|Z_i, Z_j\big\} - \mathbb{E}\big\{K_{h_n}(E_i^* - e)f_E(e|Z_j)\Big|Z_i, Z_j\big\} \\
&\quad - \mathbb{E}\big\{f_E(e|Z_i)K_{h_n}(E_j^* - e)\Big|Z_i, Z_j\big\} + \mathbb{E}\big\{f_E(e|Z_i)f_E(e|Z_j)\Big|Z_i, Z_j\big\} \\
&\quad = \mathbb{E}\big\{K_{h_n}(E_i^* - e)K_{h_n}(E_j^* - e) - f_E(e|Z_i)f_E(e|Z_j)\Big|Z_i, Z_j\big\}
\end{aligned}
$$

This is because the equation (8) implies

$$
\mathbb{E}\big\{K_{h_n}(E_i^* - e)f_E(e|Z_j)\Big|Z_i, Z_j\big\} = f_E(e|Z_i)f_E(e|Z_j)
$$

$$
\mathbb{E}\big\{f_E(e|Z_i)K_{h_n}(E_j^* - e)\Big|Z_i, Z_j\big\} = f_E(e|Z_i)f_E(e|Z_j)
$$

Then, conditioning on (Z_i, Z_j), we apply the second Taylor's expansion to the above kernel products:

$$
\begin{aligned}
&\mathbb{E}\big\{K_{h_n}(E_i^* - e)K_{h_n}(E_j^* - e)\Big|Z_i, Z_j\big\} \\
&= h^{-2}\int\int K((u-e)/h)K((v-e)/h)f_E(u|Z_i)f_E(v|Z_j)dudv \\
&= \int\int K(u)K(v)f_E(e+uh|Z_i)f_E(e+vh|Z_j)dudv \\
&= \int\int K(u)K(v)\Big[f_E(e+uh|Z_i)f_E(e+vh|Z_j) + \dot{f}_E(e+uh|Z_i)f_E(e+vh|Z_j)(uh) \\
&+ f_E(e+uh|Z_i)\dot{f}_E(e+vh|Z_j)(vh) + \frac{1}{2}\big\{\ddot{f}_E(e+uh|Z_i)f_E(e+vh|Z_j)(uh)^2 \\
&+ 2\dot{f}_E(e+uh|Z_i)\dot{f}_E(e+vh|Z_j)(uvh^2) + f_E(e+uh|Z_i)\ddot{f}_E(e+vh|Z_j)(vh)^2\big\}\Big]dudv \\
&= f_E(e|Z_i)f_E(e|Z_j) + \frac{1}{2}\{\ddot{f}_E(e|Z_i)f_E(e|Z_j) + f_E(e|Z_i)\ddot{f}_E(e|Z_j)\}\mu_{K,2}h^2(1+o_p(1))
\end{aligned}
$$

This proves $\mathbb{E}\big\{K_{h_n}(E_i^* - e)K_{h_n}(E_j^* - e) - f_E(e|Z_i)f_E(e|Z_j)\Big|Z_i, Z_j\big\} = o_p(1)$, for a sufficiently small choice of $h > 0$. Therefore, we conclude

$$
\begin{aligned}
&\sum_{i\neq j}\mathbb{C}ov\Big[v_i(z)\Big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\Big), v_j(z)\Big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\Big)\Big] \\
&= \sum_{i\neq j}\mathbb{E}\Big\{v_i(z)v_j(z)\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\big)\Big\} \\
&= \sum_{i\neq j}\mathbb{E}\Big[\mathbb{E}\Big\{v_i(z)v_j(z)\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\big)\Big\}\Big|Z_i, Z_j\Big] \\
&= \sum_{i\neq j}\mathbb{E}\Big[\mathbb{E}\Big\{v_i(z)v_j(z)\Big|Z_i, Z_j\Big\}\mathbb{E}\Big\{\big(K_{h_n}(E_i^* - e) - f_E(e|Z_i)\big)\big(K_{h_n}(E_j^* - e) - f_E(e|Z_j)\big)\Big|Z_i, Z_j\Big\}\Big] \\
&= \sum_{i\neq j}\mathbb{E}\Big[\mathbb{E}\Big\{v_i(z)v_j(z)\Big|Z_i, Z_j\Big\}\mathbb{E}\big\{K_{h_n}(E_i^* - e)K_{h_n}(E_j^* - e) - f_E(e|Z_i)f_E(e|Z_j)\Big|Z_i, Z_j\big\}\Big] \\
&= \sum_{i\neq j}\mathbb{E}\Big[\mathbb{E}\Big\{v_i(z)v_j(z)\Big|Z_i, Z_j\Big\}\cdot o_p(1)\Big] \to 0
\end{aligned}
$$

as $n \to \infty$.

A.2 Supplementary Simulation Results

Table 2: MARGINAL PERFORMANCES UNDER A FEW SPATIAL ERROR Average coverage rates of 90% prediction intervals, widths, and interval score across 100 simulations constructed by Quantile Regression Forests (QRF), split conformal prediction (SC), the unweighted out-of-bag method (OOB), Local Spatial Conformal Prediction (LSCP), the weighted out-of-bag method (OOBW), and the generalized out-of-bag kernel method (OOBGK). Bold quantities represents the case showing the lowest values in interval length or average interval score, respectively, among the candidates.

	LINEAR			STEP			FRIEDMAN		
	CPR	LEN	AIS90	CPR	LEN	AIS90	CPR	LEN	AIS90
QRF	0.94	6.92	7.77	0.95	7.96	8.65	0.92	7.47	8.83
SC	0.92	6.37	7.48	0.90	7.98	9.80	0.91	7.18	9.01
OOB	0.91	5.95	**7.20**	0.91	6.37	**8.00**	0.91	6.48	**8.02**
LSCP	0.91	5.92	**7.20**	0.90	6.33	8.04	0.91	6.51	**8.02**
OOBW	0.90	5.92	7.43	0.89	6.18	8.24	0.90	6.46	8.19
OOBGK	0.87	**5.47**	7.52	0.87	**5.86**	8.19	0.87	**6.04**	8.31
	SINUSOIDAL(HOMO)			SINUSOIDAL(HEAVY)			SINUSOIDAL(HETERO)		
	CPR	LEN	AIS90	CPR	LEN	AIS90	CPR	LEN	AIS90
QRF	0.91	6.48	7.80	0.92	4.48	5.97	0.91	9.09	11.24
SC	0.90	6.47	7.96	0.91	4.52	6.08	0.91	9.21	11.58
OOB	0.90	5.89	**7.28**	0.90	3.71	**5.45**	0.90	8.46	**10.93**
LSCP	0.90	5.89	7.29	0.91	3.69	5.47	0.90	8.49	10.94
OOBW	0.90	5.92	7.47	0.88	3.78	5.71	0.90	8.61	10.93
OOBGK	0.85	**5.32**	7.64	0.87	**3.44**	5.62	0.86	**7.71**	11.28

Figure 3: SENSITIVITY TO A MEASUREMENT ERROR DISTRIBUTION The top left panel represents boxplots of estimated marginal coverage probabilities with 90% confidence level under homoscadestic (HOMO), heavy-tailed (HEAVY), and heteroscadestic (HETERO) measurement error distribution, respectively, as illustrated in section 3. The rest of the panels are scatterplots of estimated versus nominal miscoverage rate ranged from 0.02 to 0.20 assuming HOMO, HEAVY, HETERO measurement error distribution, respectively. For all the panels, blue color represents OOBGK method, green represents OOBW, and red represents LSCP.

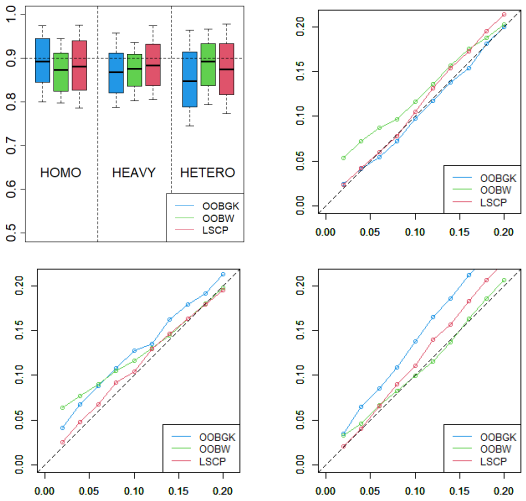


Figure 4: SENSITIVITY TO A TUNING PARAMETERS Based on the simulation settings in section 3, the first row of panels represent the relationship between tuning parameters and the estimated coverage rate on average, and the second row of panels represent the relationship between tuning parameters and the estimated prediction interval width on average. The first column of panels come from the kernel bandwidth(h), the second column come from the number of trees(n_{tree}), the third column come from the number of predictors for node-splitting (m_{try}), and the fourth colmn come from the minimum number of final nodes ($nodesize$)

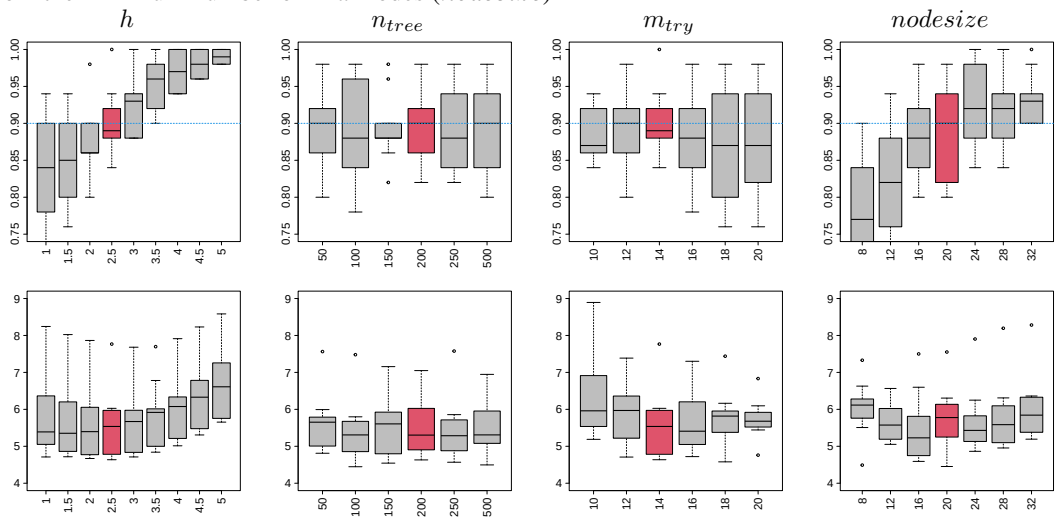


Table 3: SENSITIVITY TO TRUE UNDERLYING COVARIANCE Based on the simulation settings in section 3, we assume the Matérn covariance but three different cases of the true underlying parameters, denoted as Bumpy Matérn ($\nu = 0.1$), Exponential ($\nu = 0.5$), and Smooth Matérn ($\nu = 2.0$). For each cases, we provide average coverage rates of 90% prediction intervals, widths, and interval score across 100 simulations constructed by Quantile Regression Forests (QRF), split conformal prediction (SC), the unweighted out-of-bag method (OOB), Local Spatial Conformal Prediction (LSCP), the weighted out-of-bag method (OOBW), and the generalized out-of-bag kernel method (OOBGK). OOBGK(oracle) represents the proposed method using the true specified spatial covariance. OOBGK(ν) represents the proposed method but using the arbitrary specified parameters ν. Bold quantities represents the case showing the lowest values in interval length or average interval score, respectively, among the candidates.

	Bumpy Matérn ($\nu = 0.1$)			Exponential ($\nu = 0.5$)			Smooth Matérn ($\nu = 2.0$)		
	CPR	LEN	AIS90	CPR	LEN	AIS90	CPR	LEN	AIS90
QRF	0.90	7.09	8.77	0.94	6.60	7.37	0.95	6.00	6.42
SC	0.90	7.34	9.11	0.92	6.16	7.59	0.92	5.47	6.64
OOB	0.88	6.10	7.75	0.90	5.19	6.28	0.92	4.65	5.59
LSCP	0.91	6.06	7.42	0.91	4.98	6.06	0.92	4.55	5.45
OOBW	0.88	6.08	7.79	0.90	5.20	6.39	0.91	4.58	5.45
OOBGK($\nu = 0.1$)	0.91	7.98	9.21	0.90	6.12	7.70	0.93	5.74	7.01
OOBGK($\nu = 2.0$)	0.94	8.07	9.18	0.92	6.10	7.11	0.95	5.65	6.34
OOBGK(oracle)	0.92	6.32	**7.14**	0.91	4.89	5.71	0.94	4.71	5.44
OOBGK	0.86	**5.86**	7.56	0.89	**4.49**	**5.63**	0.90	**4.24**	**5.35**

Table 4: MARGINAL PERFORMANCES UNDER DOMINANT SPATIAL ERROR WITH DIFFERENT LEVEL OF THE NOMINAL MISCOVERAGE RATE For the different set of nominal miscoverage rates $\alpha \in \{0.05, 0.10, 0.20\}$, we provide average coverage rates of 90% prediction intervals, widths, and interval score across 100 simulations constructed by Quantile Regression Forests (QRF), split conformal prediction (SC), the unweighted out-of-bag method (OOB), Local Spatial Conformal Prediction (LSCP), the weighted out-of-bag method (OOBW), and the generalized out-of-bag kernel method (OOBGK). Bold quantities represents the case showing the lowest values in interval length or average interval score, respectively, among the candidates.

α		LINEAR			SINUSOIDAL			STEP			FRIEDMAN		
	Method	CPR	LEN	AIS90	CPR	LEN	AIS90	CPR	LEN	AIS90	CPR	LEN	AIS90
0.2	QRF	0.91	4.59	5.07	0.85	3.97	4.83	0.92	5.52	5.99	0.86	5.02	6.07
	SC	0.85	4.08	4.98	0.82	4.00	5.17	0.82	5.88	7.73	0.80	4.76	6.69
	OOB	0.83	3.50	4.54	0.80	3.35	4.53	0.81	3.87	5.41	0.80	4.02	5.57
	LSCP	0.83	3.24	4.18	0.80	3.05	4.13	0.81	3.60	5.07	0.80	3.80	5.24
	OOBW	0.81	3.42	4.54	0.80	3.28	4.47	0.78	3.63	5.36	0.80	3.98	5.37
	OOBGK	0.81	**3.00**	**4.06**	0.78	**2.89**	**4.01**	0.79	**3.16**	**4.83**	0.77	**3.56**	**5.05**
0.1	QRF	0.96	5.73	6.14	0.93	5.04	5.77	0.97	6.85	7.21	0.93	6.32	7.19
	SC	0.93	4.99	5.73	0.91	4.98	6.05	0.91	6.76	8.20	0.91	5.92	7.32
	OOB	0.91	4.41	5.30	0.90	4.18	5.25	0.90	4.85	6.52	0.90	5.05	6.35
	LSCP	0.91	4.11	4.89	0.90	3.86	4.79	0.90	4.49	6.11	0.90	4.77	5.95
	OOBW	0.90	4.29	5.35	0.89	4.12	5.31	0.87	4.63	6.70	0.90	4.98	6.27
	OOBGK	0.91	**3.86**	**4.73**	0.89	**3.74**	**4.78**	0.88	**4.24**	**6.04**	0.88	**4.64**	**5.93**
0.05	QRF	0.99	7.19	7.39	0.97	6.36	6.93	1.00	9.46	9.55	0.97	8.02	8.84
	SC	0.97	6.23	6.67	0.96	6.03	6.81	0.95	8.93	10.19	0.95	7.61	9.11
	OOB	0.96	5.31	6.02	0.95	5.03	6.00	0.95	6.04	7.82	0.95	6.13	7.46
	LSCP	0.96	4.97	5.60	0.95	4.61	**5.43**	0.95	5.61	7.41	0.95	5.88	7.10
	OOBW	0.95	5.19	6.26	0.94	4.93	6.19	0.93	5.72	8.17	0.94	6.06	7.36
	OOBGK	0.95	**4.60**	**5.45**	0.94	**4.50**	5.49	0.93	**5.10**	**7.35**	0.93	**5.50**	**6.95**

Figure 5: SENSITIVITY TO KERNEL BANDWIDTH SELECTION Based on the simulation settings in section 3, the first row of panels represent the relationship between tuning parameters and the estimated coverage rate on average, and the second row of panels represent the relationship between tuning parameters and the estimated prediction interval width on average. Each column of panels com form the linear, sinusoidal, step, and Friedman mean function, respectively.